RECOVERY OF CAUSAL GRAPH INVOLVING LATENT VARIABLES VIA HOMOLOGOUS SURROGATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal discovery with latent variables is an important and challenging problem. To identify latent variables and infer their causal relations, most existing works rely on the assumption that latent variables have pure children. Considering that this assumption is potentially restrictive in practice and not strictly necessary in theory, by introducing the concept of homologous surrogate, this paper eliminates the need for pure child in the context of causal discovery with latent variables. The homologous surrogate fundamentally differs from the pure child in the sense that the latter is characterized by having strictly restricted parents while the former allows for much more flexible parents. We formulate two assumptions involving homologous surrogates and develop theoretical results under each assumption. Under the weaker assumption, our theoretical results imply that we can determine each variable's ancestors, that is, partially recover the causal graph. The stronger assumption further enables us to determine each variable's parents exactly, that is, fully recover the causal graph. Building on these theoretical results, we derive an algorithm that fully leverages the properties of homologous surrogates for causal graph recovery. Also, we validate its efficacy through experiments. Our work broadens the applicability of causal discovery.

1 INTRODUCTION

Causality is a basic concept in natural and social sciences, playing a pivotal role in explanation, prediction, decision making and control (Zhang et al., 2018). While the gold standard for uncovering causality is conducting randomized experiments, this is usually prohibitively expensive and time-consuming. Consequently, researchers have increasingly turned to causal discovery, which aims to infer causal relations from observational data alone. Traditional causal discovery methods typically assume that all task-relevant variables have been enumerated and measured (Spirtes & Glymour, 1991; Chickering, 2002; Shimizu et al., 2006; Hoyer et al., 2009; Zhang & Hyvärinen, 2009). However, this assumption is not always valid in real-world scenarios, prompting the development of causal discovery with latent variables. These methods can be classified into three categories. The first category assume that latent variables are mutually independent (Hoyer et al., 2008; Salehkaleybar et al., 2020; Maeda & Shimizu, 2020; Yang et al., 2022; Cai et al., 2023). The second category allow the presence of causally-related latent variables but cannot identify latent variables, let alone their causal relations (Spirtes et al., 1995; Claassen et al., 2013; Claassen & Bucur, 2022). The third category not only allows the presence of causally-related latent variables but also can identify latent variables along with their causal relations (Silva et al., 2006; Cai et al., 2019; Xie et al., 2020; Huang et al., 2022; Chen et al., 2023; Jin et al., 2024). Our work belongs to the third category.

Recent works in the third category predominantly rely on the pure children assumption that latent variables have pure children. They not only identify latent variables by locating their pure children but also use their pure children as proxies to infer their causal relations. These works can be further categorized into two groups. Some works (Silva et al., 2006; Shimizu et al., 2009; Kummerfeld & Ramsey, 2016; Cai et al., 2019; Chen et al., 2022; Zeng et al., 2021; Xie et al., 2022; Chen et al., 2023) make the special pure children assumption that each latent variable has multiple pure children. Here, a variable is said a pure child of anther only if the latter is the only parent of the former. Other works (Xie et al., 2020; 2024; Huang et al., 2022; Dong et al., 2024; Jin et al., 2024) make the general pure children assumption that each latent variable belongs to a latent set (comprising one or more latent variables) which has sufficient pure children. Here, a variable is said a pure child of a

latent set only if all parents of the former are within the latter. It should be noted that the weaker general pure children assumption is often accompanied by local unidentifiability: for multiple latent variables within a latent set, even the existence (let alone directions) of the causal relations between them might be undeterminable. In summary, the concept of pure children is characterized by having strictly restricted parents.

Although the pure children assumption is widely employed for the sake of tractability, Adams et al. (2021) argue that this assumption is restrictive in practice and prove that it is not necessary for identifiability of linear non-Gaussian acyclic models with latent variables. In this paper, by introducing the concept of homologous surrogate, we eliminate the need for pure children. The homologous surrogate fundamentally differs from the pure child in that the former allows for much more flexible parents. For instance, if an observed variable O is a pure child of a latent variable L, O must have only one parent L; but if O is a homologous surrogate of L, O is allowed to have other parents besides L, such as other latent parents provided that they are all L's ancestors. Taking Fig. 1 as an example, although O_3 has two parents L_1, L_2, it can still serve as a homologous surrogate of L_2. On the one hand, the existence of other parents is not mandatory, so if O is L's pure child, it is also L's homologous surrogate[1]. On the other hand, even if O is not L's pure child, it might still be L's homologous surrogate. The latter case is quite common in practice. For instance, a company's stock price is caused by both its performance and macroeconomic environment, and macroeconomic environment also impacts its performance, so its stock price is a homologous surrogate but not a pure child of its performance.

We begin with the assumption that each latent variable has at least one (rather than multiple) homologous surrogate. Under this assumption, we develop theoretical results implying that the causal graph can be partially recovered, that is, we can determine whether any variable is an ancestor of any other variable. From these theoretical results, we derive an algorithm that sequentially identifies latent variables by locating their homologous surrogates, progressing from roots to leaves, during which process the causal graph is also partially recovered. We then develop further theoretical results under the above assumption plus an extra assumption that if a latent variable is an ancestor of another, the latter must have two generalized homologous surrogates that are not children of the former, where the generalized homologous surrogate is a variant of the homologous surrogate but is subject to fewer restrictions. These theoretical results imply that the causal graph can be fully recovered, that is, we can determine whether any variable is a parent of any other variable, from which we also derive an algorithm. Building on the partial recovery result, this algorithm first locates latent variables' generalized homologous surrogates, then uses them to infer the causal relations between latent variables. Combining this information with the partial recovery result, the causal graph can be fully recovered. A causal graph that can be fully recovered by our algorithm is shown as Fig. 1, where no latent variable has multiple pure children.

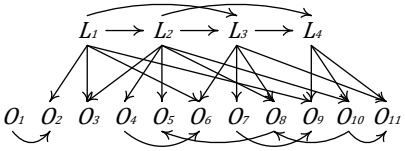


Figure 1: A causal graph that can be fully recovered by our algorithm.

The major innovations of our work are summarized as follows.

- We investigate a new problem setting where latent variables leave footprints in observed variables via homologous surrogates rather than conventional pure children. The homologous surrogate fundamentally differs from the pure child in the sense that the latter is characterized by having strictly restricted parents while the former allows for much more flexible parents.
- We formulate two assumptions involving homologous surrogates and develop novel theoretical results under each assumption. These theoretical results imply that the causal graph can be partially/fully recovered under the weaker/stronger assumption.
- Building on our theoretical results, we derive a systematic and innovative algorithm which fully leverages the properties of homologous surrogates for causal graph recovery. We demonstrate the efficacy of our algorithm through experiments.

In summary, our work broadens the applicability of causal discovery. It may not only inspire further research in this direction but also benefit research in natural and social sciences.

[1] Strictly speaking, to derive "O is L's homologous surrogate" from "O is L's pure child", we need an additional condition that O has no child. This condition is satisfied in most cases because "O is L's pure child" implies "O has no child" in most of the literature. Only in very few works (Dong et al., 2024; Jin et al., 2024), an observed pure child of a latent variable is allowed to have children of its own.

2 PRELIMINARY

In this paper, we focus on the linear non-Gaussian acyclic model (LiNGAM) with latent variables whose causal graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is a directed acyclic graph (DAG). $\mathbf{V} = \mathbf{L} \cup \mathbf{O}$ where $\mathbf{L}$ and $\mathbf{O}$ respectively denote the set of latent and observed variables. Each variable in $\mathbf{L}$ and $\mathbf{O}$ follows

$$L_i = \sum_{L_j \in \mathbf{L}} a_{L_i}^{L_j} L_j + \epsilon_{L_i}, \quad O_i = \sum_{L_j \in \mathbf{L}} a_{O_i}^{L_j} L_j + \sum_{O_j \in \mathbf{O}} a_{O_i}^{O_j} O_j + \epsilon_{O_i}. \tag{1}$$

$a_{V_j}^{V_i}$ denotes the direct causal strength from V_i to V_j, $a_{V_j}^{V_i} \neq 0$ if and only if V_i is a parent of V_j. Given $V \in \mathbf{V}$, we denote its parents, children, neighbors, ancestors, and descendants by $\mathrm{Pa}(V), \mathrm{Ch}(V), \mathrm{Ne}(V), \mathrm{An}(V)$, and $\mathrm{De}(V)$. Particularly, a variable's ancestors/descendants do not include itself. In the following, for any $\mathbf{V}' \subset \mathbf{V}$, we abbreviate $\cup_{V \in \mathbf{V}'} \mathrm{Pa}(V)$ to $\mathrm{Pa}(\mathbf{V}')$. The latent parents and observed parents of V are denoted by $\mathrm{Pa}_{\mathbf{L}}(V)$ and $\mathrm{Pa}_{\mathbf{O}}(V)$ respectively. ϵ_V refers to the exogenous noise of V, all exogenous noises have non-Gaussian distributions and are independent of each other. Without loss of generality, we assume that each exogenous noise has zero mean and the exogenous noise of each latent variable has unit variance.

We can rewrite Eq. (1) in a matrix form as

$$\begin{bmatrix} \mathbf{L} \\ \mathbf{O} \end{bmatrix} = \mathbf{A} \begin{bmatrix} \mathbf{L} \\ \mathbf{O} \end{bmatrix} + \begin{bmatrix} \epsilon_{\mathbf{L}} \\ \epsilon_{\mathbf{O}} \end{bmatrix}, \text{ where } \mathbf{A} := \begin{bmatrix} \mathbf{A}_{\mathbf{L}}^{\mathbf{L}} & \mathbf{0} \\ \mathbf{A}_{\mathbf{O}}^{\mathbf{L}} & \mathbf{A}_{\mathbf{O}}^{\mathbf{O}} \end{bmatrix} \tag{2}$$

is the *adjacency matrix*. For $\mathbf{V}_1, \mathbf{V}_2 \subset \mathbf{V}$, $\mathbf{A}_{\mathbf{V}_2}^{\mathbf{V}_1}$ refers to the adjacent matrix from $\mathbf{V}_1$ to $\mathbf{V}_2$. Since $\mathbf{I} - \mathbf{A}$ is invertible (Shimizu et al., 2006), we can further rewrite Eq. (2) as

$$\begin{bmatrix} \mathbf{L} \\ \mathbf{O} \end{bmatrix} = \mathbf{M} \begin{bmatrix} \epsilon_{\mathbf{L}} \\ \epsilon_{\mathbf{O}} \end{bmatrix}, \text{ where } \mathbf{M} = (\mathbf{I} - \mathbf{A})^{-1} = \begin{bmatrix} (\mathbf{I} - \mathbf{A}_{\mathbf{L}}^{\mathbf{L}})^{-1} & \mathbf{0} \\ (\mathbf{I} - \mathbf{A}_{\mathbf{O}}^{\mathbf{O}})^{-1} \mathbf{A}_{\mathbf{O}}^{\mathbf{L}} (\mathbf{I} - \mathbf{A}_{\mathbf{L}}^{\mathbf{L}})^{-1} & (\mathbf{I} - \mathbf{A}_{\mathbf{O}}^{\mathbf{O}})^{-1} \end{bmatrix} := \begin{bmatrix} \mathbf{M}_{\mathbf{L}}^{\mathbf{L}} & \mathbf{0} \\ \mathbf{M}_{\mathbf{O}}^{\mathbf{L}} & \mathbf{M}_{\mathbf{O}}^{\mathbf{O}} \end{bmatrix} \tag{3}$$

is the *mixing matrix* whose elements are called *mixing coefficients*. By convention, we assume the distribution over $\mathbf{V}$ is Markov and rank-faithful to $\mathcal{G}$, which means that $m_{V_j}^{V_i} \neq 0$ if and only if V_i is an ancestor of V_j or $V_i = V_j$. In the latter case, $m_{V_j}^{V_i} = 1$.

Assumption 1. *(Rank faithfulness) Given a probability distribution p and a DAG $\mathcal{G}$, p is rank-faithful to $\mathcal{G}$ if every rank constraint on a sub-covariance matrix that holds in p is entailed by every linear structural model with respect to $\mathcal{G}$.*

3 PARTIAL RECOVERY

Definition 1. *(homologous surrogate) $O \in \mathbf{O}$ is called a homologous surrogate of $L \in \mathbf{L}$, denoted by $O \in \mathrm{HSu}(L)$, if $O \in \mathrm{Ch}(L)$, $\mathrm{Ch}(O) = \emptyset$, $\mathrm{An}_{\mathbf{L}}(O) = \mathrm{An}(L) \cup \{L\}$ and $\mathrm{An}_{\mathbf{O}}(O) \cap \mathrm{De}_{\mathbf{O}}(L) = \emptyset$.*

Example. In Fig. 1, $\mathrm{HSu}(L_1) = \{O_2\}$, $\mathrm{HSu}(L_2) = \{O_3\}$, $\mathrm{HSu}(L_3) = \{O_6\}$, $\mathrm{HSu}(L_4) = \{O_9\}$.

Remark. Given $O \in \mathbf{O}$ and $L \in \mathbf{L}$, we detail the the connections and differences between "O is L's pure child" and "O is L's homologous surrogate" in the following. There is a consensus among previous works that if an observed variable O is a pure child of a latent variable L, then O must have no other parent except L. On this basis, some studies (Silva et al., 2006; Kummerfeld & Ramsey, 2016; Xie et al., 2023; Li et al., 2024) explicitly require that O has no child while others (Shimizu et al., 2009; Cai et al., 2019; Xie et al., 2020; 2022; Huang et al., 2022; Chen et al., 2023) directly assume that there exists no edge between observed variables. O is allowed to have children of its own only in very few works (Dong et al., 2024; Jin et al., 2024). That is, if O is a pure child of L in most senses, then O is also a homologous surrogate of L, but the reverse is not necessarily true because L's homologous surrogate is allowed to have other parents.

Intuition. Suppose O is L's homologous surrogate. If L is a root variable, then O has no child and only one latent parent L which is a root variable. Otherwise, with both L's ancestors and observed variables whose latent ancestors are a subset of L's ancestors removed, O still has no child and

only one latent parent L which is a root variable. Because of this, homologous surrogates can be located from observed variables.

Assumption 2. $\forall L \in \mathbf{L}$, $\mathrm{HSu}(L) \neq \emptyset$ *and* $|\mathrm{Ch}(L)| \geq 2$.

Example. The causal graph shown as Fig. 1 satisfies Asmp. 2.

Remark. In previous works making the pure children assumption, the number of latent variables must be strictly smaller than their pure children. Instead, we only require one homologous surrogate per latent variable.

Intuition. Assuming that each latent variable has a homologous surrogate, latent variables can be sequentially identified by locating their homologous surrogates, progressing from roots to leaves, during which process the causal graph is also partially recovered. Also, we assume each latent variable has multiple children, otherwise it can be modeled as a noise (Silva et al., 2006).

§ High-level Overview. First, we identify observed root variables (Thm. 1), estimate their effects on others (Cor. 1), and then remove them (Cor. 2). Second, we identify latent root variables (Thm. 2, Props. 1 and 2), estimate their effects on others (Cor. 3), and then remove them (Cor. 4). Repeating these two procedures until all observed variables are removed, we can identify all latent variables and partially recover the causal graph. During this process, all operations on latent variables are implemented through their homologous surrogates.

§ Initialization. We denoted the set of removed variables by $\mathbf{J} \cup \mathbf{K}$ where $\mathbf{J} \subset \mathbf{L}, \mathbf{K} \subset \mathbf{O}$. In addition, for each $O_i \in \mathbf{O} \backslash \mathbf{K}$, there is an auxiliary variable $\tilde{O}_i$ which is a linear combination of O_i and variables in $\mathbf{K}$ where the coefficient of O_i is always 1 while that of each variable in $\mathbf{K}$ is not fixed. Initially, we let $\mathbf{J} = \mathbf{K} = \emptyset$, so $\tilde{O}_i = O_i$ for each $O_i \in \mathbf{O}$, it is trivial that Cond. 1 is valid.

Condition 1. *(1) For each* $V \in \mathbf{V} \backslash (\mathbf{J} \cup \mathbf{K})$, $\mathrm{De}(V) \cap (\mathbf{J} \cup \mathbf{K}) = \emptyset$. *(2) For each* $L \in \mathbf{J}$ *and* $O \in \mathbf{K}$ *where* $\mathrm{Ch}(O) \neq \emptyset$, $m^L_{\tilde{O}_i} = m^O_{\tilde{O}_i} = 0$.

§ Identifying Observed Root Variables. This can be accomplished based on Thm. 1.

Definition 2. *(Pseudo-residual (Cai et al., 2019)) Given three variables* V_1, V_2, V_3 *s.t.* $\mathrm{Cov}(V_2, V_3) \neq 0$, *the pseudo-residual of* V_1, V_2 *relative to* V_3 *is defined as*

$$\mathrm{R}(V_1, V_2 | V_3) = V_1 - \frac{\mathrm{Cov}(V_1, V_3)}{\mathrm{Cov}(V_2, V_3)} V_2. \tag{4}$$

Intuition. Pseudo-residual is a simple variant of the conventional residual. The former reduces to the latter when $V_2 = V_3$. Before Cai et al. (2019), similar concepts have already been used by earlier works (Drton & Richardson, 2004; Chen et al., 2017).

Theorem 1. *Suppose* $O_i \in \mathbf{O} \backslash \mathbf{K}$, *then* $\mathrm{An}(O_i) \subset (\mathbf{J} \cup \mathbf{K})$ *if and only if* $\forall O_j \in \mathbf{O} \backslash (\mathbf{K} \cup \{O_i\})$, $\mathrm{R}(O_j, O_i | \tilde{O}_i) \perp\!\!\!\perp \tilde{O}_i$.

Intuition. The part before "if and only if" means that all ancestors of O_i are in $\mathbf{J} \cup \mathbf{K}$, that is, O_i is a root variable among $\mathbf{V} \backslash (\mathbf{J} \cup \mathbf{K})$; the part after "if and only if" means that O_i satisfies certain independence constraints. Therefore, this theorem provides a method for identifying observed root variables via statistical analysis.

Example. Suppose the underlying causal graph is shown as Fig. 1. Initially, $\mathbf{J} = \mathbf{K} = \emptyset$. We can identify O_1 as an observed root because $\forall O_j \in \{O_2, ..., O_{11}\}, \mathrm{R}(O_j, O_1 | \tilde{O}_1) \perp\!\!\!\perp \tilde{O}_1$.

§ Estimating the Effects of Observed Root Variables. This can be accomplished based on Cor. 1.

Corollary 1. *Suppose* O_i *satisfies Thm. 1, then* $\forall O_j \in \mathbf{O} \backslash (\mathbf{K} \cup \{O_i\})$, $m^{O_i}_{O_j} = \frac{\mathrm{Cov}(\tilde{O}_i, O_j)}{\mathrm{Cov}(\tilde{O}_i, O_i)}$.

Remark. With the rank-faithfulness assumption, O_j is a descendant of O_i if and only if $m^{O_i}_{O_j} \neq 0$.

§ Removing Observed Root Variables. This can be accomplished based on Cor. 2.

Corollary 2. *Suppose* O_i *satisfies Thm. 1, if we update* $\mathbf{K}$ *to* $\mathbf{K} \cup \{O_i\}$ *and* $\tilde{O}_j$ *to* $\tilde{O}_j - m^{O_i}_{O_j} \tilde{O}_i$ *for each* $O_j \in \mathbf{O} \backslash \mathbf{K}$, *Cond. 1 is still valid.*

Remark. With observed root variables removed, some observed non-root variables before removal might become roots after removal, so we need to repeat the above three steps until there is no observed root variable, that is, no observed variable satisfies Thm. 1.

§ Identifying Latent Root Variables. We identify latent root variables by locating their respective homologous surrogates from observed variables. However, this cannot be achieved in a single step. Instead, we first locate observed variables that might be homologous surrogates (called candidate homologous surrogates) of latent root variables (Thm. 2), then check whether any two of them share a common latent parent (Prop. 1), and finally find true homologous surrogates (Prop. 2).

Theorem 2. *Suppose* $\forall O \in \mathbf{O}\backslash\mathbf{K}, \mathrm{An}(O) \not\subset \mathbf{J} \cup \mathbf{K}$. *Given* $O_i \in \mathbf{O}\backslash\mathbf{K}$, *then* $\mathrm{Ch}(O_i) = \emptyset$, $\mathrm{Pa}_{\mathbf{O}}(O_i)\backslash\mathbf{K} = \emptyset$, $|\mathrm{Pa}_{\mathbf{L}}(O_i)\backslash\mathbf{J}| = 1$, *and* $\mathrm{An}(\mathrm{Pa}_{\mathbf{L}}(O_i)\backslash\mathbf{J}) \subset \mathbf{J}$ *if and only if* $\forall\{O_j, O_k\} \subset \mathbf{O}\backslash(\mathbf{K} \cup \{O_i\})$ *where* $\mathrm{Cov}(\tilde{O}_i, O_j)\mathrm{Cov}(\tilde{O}_i, O_k) \neq 0$, $\mathrm{R}(O_j, O_k|\tilde{O}_i) \perp\!\!\!\perp \tilde{O}_i$.

Intuition. The part before "if and only if" means that O_i has no child and only one latent parent in $\mathbf{J} \cup \mathbf{K}$ which has no ancestor in $\mathbf{J} \cup \mathbf{K}$. This is a necessary but not sufficient condition for O_i being a homologous surrogate of a latent root variable among $\mathbf{J} \cup \mathbf{K}$ (this is further explained in Remark later); the part after "if and only if" means that O_i satisfies certain independence constraints. Therefore, this theorem provides a method for identifying candidate homologous surrogates of latent root variables via statistical analysis.

Example. Suppose the underlying causal graph is shown as Fig. 1. After removing the observed variable O_1 based on Thm. 1, $\mathbf{J} = \emptyset, \mathbf{K} = \{O_1\}$. We can identify O_2 as a candidate homologous surrogate of a latent root because $\forall\{O_j, O_k\} \in \{O_3, ..., O_{11}\}, \mathrm{R}(O_i, O_j|\tilde{O}_2) \perp\!\!\!\perp \tilde{O}_2$.

Remark. Based on Def. 1, it is trivial that the part before "if and only if" is a necessary condition for O_i being a homologous surrogate of a latent root variable among $\mathbf{J} \cup \mathbf{K}$, so we will not omit any homologous surrogate of any latent root variable. Moreover, the part before "if and only if" is not a sufficient condition for O_i being a homologous surrogate of a latent root variable among $\mathbf{J} \cup \mathbf{K}$, an example is shown as Fig. 2, so this theorem can only be used to locate candidate homologous surrogate.

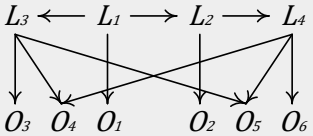


Figure 2: When $\mathbf{J} = \{L_1, L_2, L_3\}$ and $\mathbf{K} = \{O_1, O_2, O_3\}$, O_4 satisfies Thm. 2 but $O_4 \notin \mathrm{HSu}(L_4)$.

By the way, this theorem significantly differs from Thm. 2 in Cai et al. (2019) although they both utilize pseudo-residuals to identify latent variables. With the pure children assumption, the latter provides a sufficient and necessary condition for two observed variables O_i, O_j to be pure children of a same latent (not necessarily root) variable: for any other O_k, $R(O_i, O_j|O_k) \perp\!\!\!\perp O_k$. In contrast, with the homologous surrogates assumption, the former only provides a necessary but not sufficient condition for a single observed variables O_i to be a homologous surrogate of a latent root variable: for any other O_j and O_k, $R(O_j, O_k|\tilde{O}_i) \perp\!\!\!\perp \tilde{O}_i$.

Proposition 1. *Suppose* O_i *and* O_j *satisfy Thm. 2, then* $\mathrm{Pa}_{\mathbf{L}}(O_i)\backslash\mathbf{J} = \mathrm{Pa}_{\mathbf{L}}(O_j)\backslash\mathbf{J}$ *if and only if* $\mathrm{Cov}(\tilde{O}_i, O_j) \neq 0$.

Proposition 2. *Suppose* O_i *satisfies Thm. 2, then* $O_i \in \mathrm{HSu}(\mathrm{Pa}_{\mathbf{L}}(O_i)\backslash\mathbf{J})$ *if and only if* $\forall O_j$ *satisfying Thm. 2 and* $\mathrm{Pa}_{\mathbf{L}}(O_j)\backslash\mathbf{J} = \mathrm{Pa}_{\mathbf{L}}(O_i)\backslash\mathbf{J}$, $\|\mathbf{M}^{\mathbf{J}}_{\{O_i\}}\|_0 \leq \|\mathbf{M}^{\mathbf{J}}_{\{O_j\}}\|_0$.

§ Estimating the Effects of Latent Root Variables. This can be accomplished based on Cor. 3.

Definition 3. *(Cumulant) Given n random variables* $V_1, ..., V_n$, *the k-th order cumulant is defined as a tensor of size* $n \times ... \times n$ *(k times), whole element at position* $(i_1, ..., i_k)$ *is*

$$\mathrm{Cum}(V_{i_1}, ..., V_{i_k}) = \sum_{\pi} (-1)^{|\pi|-1}(|\pi| - 1)! \prod_{B \in \pi} \mathbb{E}\left[\prod_{j \in B} V_j\right], \tag{5}$$

where π *is enumerated over all partitions of* $\{i_1, ..., i_k\}$.

Remark. High-order cumulants have been widely used in the community of signal processing since the last century, especially in the topic of independent component analysis (ICA) (Thi &

Algorithm 1: Partial recovery of the causal graph under Asmp. 2.

Input: $\mathbf{O}$
Output: $\mathrm{An}(V)$ for each $V \in \mathbf{V}$, $\mathbf{M}_{\mathbf{O}}^{\mathbf{L}}$, $\mathbf{M}_{\mathbf{O}}^{\mathbf{O}}$

Initialize $\mathbf{J} = \mathbf{K} = \emptyset$, and $\tilde{\mathbf{O}} = \mathbf{O}$.
while $\mathbf{K} \neq \mathbf{O}$ **do**
  while *there exists* $O \in \mathbf{O} \backslash \mathbf{K}$ *satisfying* Thm. 1 **do**
    Identify all observed root variables $\mathbf{O}'$ based on Thm. 1.
    Estimate $\mathbf{M}_{\mathbf{O}\backslash\mathbf{K}}^{\mathbf{O}'}$ based on Cor. 1 and find $\mathrm{De}(O')$ for each $O' \in \mathbf{O}'$.
    Remove $\mathbf{O}'$ based on Cor. 2.
  end
  Identify all latent root variables $\mathbf{L}'$ based on Thm. 2 plus Props. 1 and 2.
  Estimate $\mathbf{M}_{\mathbf{O}\backslash\mathbf{K}}^{\mathbf{L}'}$ based on Cor. 3 and find $\mathrm{An}(L'), \mathrm{De}_{\mathbf{O}}(L')$ for each $L' \in \mathbf{L}'$.
  Remove $\mathbf{L}'$ based on Cor. 4.
end

Jutten, 1995; Hyvärinen & Oja, 1997; Belkin et al., 2013; Voss et al., 2013; Ge & Zou, 2016), which is closely related to causal discovery (Shimizu et al., 2006; Hoyer et al., 2008).

Corollary 3. *Suppose O_i satisfies Thm. 2 and $O_i \in \mathrm{HSu}(L_i)$, then $\forall O_j \in \mathbf{O}\backslash(\mathbf{K} \cup \{O_i\})$,*

$$m_{O_i}^{L_i} m_{O_j}^{L_i} = \mathrm{Cov}(\tilde{O}_i, O_j), \quad \Big(\frac{m_{O_i}^{L_i}}{m_{O_j}^{L_i}}\Big)^2 = \frac{\mathrm{Cum}(\tilde{O}_i, \tilde{O}_i, \tilde{O}_i, O_j)}{\mathrm{Cum}(\tilde{O}_i, O_j, O_j, O_j)}. \tag{6}$$

Remark. Let $m_{O_i}^{L_i} > 0$ without loss of generality, we can obtain $m_{O_i}^{L_i}$ and $m_{O_j}^{L_i}$. With the rank-faithfulness assumption, O_j is a descendant of L_i if and only if $m_{O_j}^{L_i} \neq 0$, L_h is an ancestor of L_i if and only if $m_{O_i}^{L_h} \neq 0$ since $\mathrm{An}(L_i) = \mathrm{An}_{\mathbf{L}}(O_i)\backslash\{L_i\}$.

§ Removing Latent Root Variables. This can be accomplished based on Cor. 4.

Corollary 4. *Suppose O_i satisfies Thm. 2 and $O_i \in \mathrm{HSu}(L_i)$, if we update $\mathbf{J}$ to $\mathbf{J} \cup \{L_i\}$, $\mathbf{K}$ to $\mathbf{K} \cup \{O_i\}$, and $\tilde{O}_j$ to $\tilde{O}_j - (m_{O_j}^{L_i}/m_{O_i}^{L_i})\tilde{O}_i$ for each $O_j \in \mathbf{O}\backslash\mathbf{K}$, Cond. 1 is still valid.*

§ Summary. The algorithm is summarized in Alg. 1 with $O(|\mathbf{O}|^4)$ complexity, its procedures are shown as Fig. 3 and detailed below.

(1) Initially, $\mathbf{J} = \mathbf{K} = \emptyset, \tilde{O}_i = O_i$.

(2) First iteration shown as Fig. 3(b): Alg. 1 identifies O_1 as an observed root (Thm. 1), calculates $m_{O_i}^{O_1}$ (Cor. 1), and updates $\mathbf{K} := \mathbf{K} \cup \{O_1\}, \tilde{O}_i := \tilde{O}_i - m_{O_i}^{O_1}\tilde{O}_1$ (Cor. 2). It determines $\mathrm{De}(O_1) = \{O_2\}$ as $m_{O_2}^{O_1} \neq 0$. Next, it identifies L_1 as a latent root with $\mathrm{HSu}(L_1) = \{O_2\}$ (Thm. 2 and Props. 1,2), calculates $m_{O_i}^{L_1}$ (Cor. 3), and updates $\mathbf{J} := \mathbf{J} \cup \{L_1\}$, $\mathbf{K} := \mathbf{K} \cup \{O_2\}$, $\tilde{O}_i := \tilde{O}_i - (m_{O_i}^{L_1}/m_{O_2}^{L_1})\tilde{O}_2$ (Cor. 4). It determines $\mathrm{De}_{\mathbf{O}}(L_1) = \{O_2, ..., O_{11}\}$ as $m_{O_2}^{L_1} \neq 0, ..., m_{O_{11}}^{L_1} \neq 0$.

(3) Second iteration shown as Fig. 3(c): Alg. 1 identifies no observed root (Thm. 1). Next, it identifies L_2 as a latent root with $\mathrm{HSu}(L_2) = \{O_3\}$ (Thm. 2 and Props. 1,2), calculates $m_{O_i}^{L_2}$ (Cor. 3), and updates $\mathbf{J} := \mathbf{J} \cup \{L_2\}$, $\mathbf{K} := \mathbf{K} \cup \{O_3\}$, $\tilde{O}_i := \tilde{O}_i - (m_{O_i}^{L_2}/m_{O_3}^{L_2})\tilde{O}_3$ based on Cor. 4. It determines $\mathrm{De}_{\mathbf{O}}(L_2) = \{O_3, ..., O_{11}\}$ as $m_{O_3}^{L_2} \neq 0, ..., m_{O_{11}}^{L_2} \neq 0$ and determines $\mathrm{An}(L_2) = \{L_1\}$ as $O_3 \in \mathrm{HSu}(L_2)$ and $L_1 \in \mathrm{An}(O_3)$.

(4) The following iterations proceed similarly, which are shown as Fig. 3(d,e,f).

Theorem 3. *Suppose the observed variables are generated by a LiNGAM with latent variables satisfying Asmps. 1 and 2, in the limit of infinite data, Alg. 1 identifies latent variables and ancestral relationships correctly.*

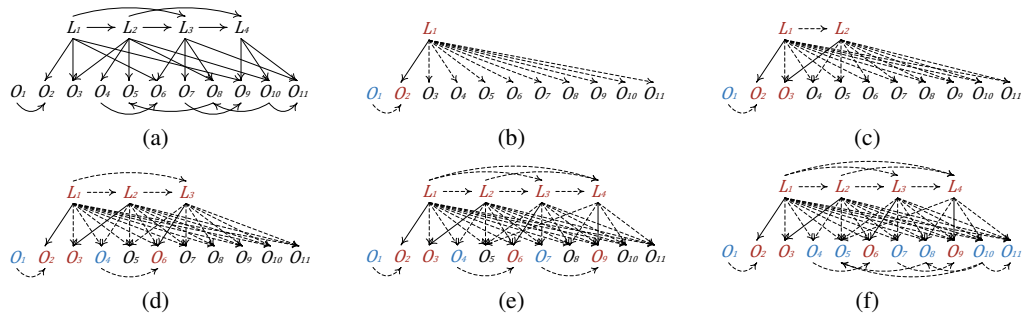


Figure 3: Illustration of Alg. 1. (a) Ground truth. (b) Result of the first iteration. (c) Result of the second iteration. (d) Result of the third iteration. (e) Result of the fourth iteration. (f) Result of the fifth iteration.

4 Full Identification

Definition 4. *(Generalized homologous surrogate) $O \in \mathbf{O}$ is called a generalized homologous surrogate of $L \in \mathbf{L}$, denoted by $O \in \mathrm{GHSu}(L)$, if $O \in \mathrm{Ch}(L)$ and $\mathrm{Pa}_{\mathbf{L}}(O) \subset \mathrm{An}(L) \cup \{L\}$.*

Example. In Fig. 1, $\mathrm{GHSu}(L_1) = \{O_2\}, \mathrm{GHSu}(L_2) = \{O_3, O_4, O_5\}, \mathrm{GHSu}(L_3) = \{O_6, O_7, O_8\}, \mathrm{GHSu}(L_4) = \{O_9, O_{10}, O_{11}\}$.

Remark. Trivially, if O is L's homologous surrogate, it must be L's generalized homologous surrogate, but the reverse is not necessarily true. This is why it is called "generalized" homologous surrogate.

Intuition. With causal relations between observed variables removed, for every latent variable, the ancestors of its each generalized homologous surrogate are exactly its ancestors plus itself. Because of this, generalized homologous surrogates can be located from observed variables.

Assumption 3. *Asmp. 2 holds and $\forall \{L_i, L_j\} \subset \mathbf{L}$ where $L_i \in \mathrm{An}(L_j)$, $\exists \{O_{j_1}, O_{j_2}\} \subset \mathrm{GHSu}(L_j)$ s.t $O_{j_1} \notin \mathrm{Ch}(L_i)$ and $O_{j_2} \notin \mathrm{Ch}(L_i)$.*

Example. The causal graph shown as Fig. 1 satisfies Asmp. 3.

Remark. An extremely special case where this assumption holds is that each latent variable has multiple observed pure children.

Intuition. Since Asmp. 2 holds, $\mathbf{M}_{\mathbf{O}}^{\mathbf{L}}$ and $\mathbf{M}_{\mathbf{O}}^{\mathbf{O}}$ can be estimated by Alg. 1. For every two latent variables L_i, L_j where L_i is an ancestor of L_j, $m_{L_j}^{L_i}$ can be estimated through L_j's two generalized homologous surrogates that are not L_i's children, so we can also estimate $\mathbf{M}_{\mathbf{L}}^{\mathbf{L}}$. $\mathbf{A}$ can be readily recovered from $\mathbf{M}$ composed of $\mathbf{M}_{\mathbf{L}}^{\mathbf{L}}, \mathbf{0}, \mathbf{M}_{\mathbf{O}}^{\mathbf{L}}, \mathbf{M}_{\mathbf{O}}^{\mathbf{O}}$, that is, the causal graph can be fully recovered.

§ High-level Overview. Building on the partial recovery result, we first remove causal relations between observed variables. Second, with these relations removed, we locate generalized homologous surrogates of each latent variable (Lem. 1). Third, through these surrogates, we estimate $\mathbf{M}_{\mathbf{L}}^{\mathbf{L}}$ progressively (Thm. 4). Finally, given $\mathbf{M}$ composed of $\mathbf{M}_{\mathbf{L}}^{\mathbf{L}}, \mathbf{0}, \mathbf{M}_{\mathbf{O}}^{\mathbf{L}}, \mathbf{M}_{\mathbf{O}}^{\mathbf{O}}$, we recover $\mathbf{A}$.

§ Removing Causal Relations Between Observed Variables. $\mathbf{M}_{\mathbf{O}}^{\mathbf{O}}$ is estimated by Alg. 1, from which $\mathbf{A}_{\mathbf{O}}^{\mathbf{O}}$ can be recovered by $\mathbf{A}_{\mathbf{O}}^{\mathbf{O}} = \mathbf{I} - (\mathbf{M}_{\mathbf{O}}^{\mathbf{O}})^{-1}$ following Eq (3). Given $\mathbf{A}_{\mathbf{O}}^{\mathbf{O}}$, we can remove causal relations between observed variables. Specifically, for each $O_i \in \mathbf{O}$, we let

$$O_i^* = O_i - \sum_{O_j \in \mathrm{Pa}(O_i)} a_{O_i}^{O_j} O_j. \tag{7}$$

Note that $\mathbf{M}_{\mathbf{O}}^{\mathbf{L}}$ is also estimated by Alg. 1, for each $L \in \mathbf{L}$,

$$m_{O_i^*}^{L} = m_{O_i}^{L} - \sum_{O_j \in \mathrm{Pa}(O_i)} a_{O_i}^{O_j} m_{O_j}^{L}. \tag{8}$$

Algorithm 2: Full recovery of the causal graph under Asmp. 3.

Input: $\mathbf{O}$, $\mathbf{L}$, $\mathbf{M}_{\mathbf{O}}^{\mathbf{L}}$, $\mathbf{M}_{\mathbf{O}}^{\mathbf{O}}$ returned by Alg. 1.
Output: $\mathbf{A}$

Remove causal relations between $\mathbf{O}$.
Locate generalized homologous surrogates of each latent variable based on Lem. 1.
$i := 1$
while *there exists* $L \in \mathbf{L}$ *s.t.* $\mathrm{De}^i(L) \neq \emptyset$ **do**
  Estimate $m_{L'}^{L}$ and $a_{O'}^{L}$ for each $L \in \mathbf{L}, L' \in \mathrm{De}^i(L), O' \in \mathrm{GHSu}(L')$ based on Thm. 4.
  $i := i + 1$.
end
Recover $\mathbf{A}$ from $\mathbf{M}$.

§ Locating Generalized Homologous Surrogates. This can be accomplished based on Lem. 1.

Lemma 1. *$\forall L_i \in \mathbf{L}$ and $O_i \in \mathbf{O}$, $O_i \in \mathrm{GHSu}(L_i)$ if and only if $m_{O_i^*}^{L_i} \neq 0$ and $\forall O_j \in \mathbf{O}$ where $m_{O_j^*}^{L_i} \neq 0$, $\|\mathbf{M}_{\{O_i^*\}}^{\mathbf{L}}\|_0 \leq \|\mathbf{M}_{\{O_j^*\}}^{\mathbf{L}}\|_0$. Besides, there is $a_{O_i}^{L_i} = m_{O_i^*}^{L_i}$.*

§ Estimating $\mathbf{M}_{\mathbf{L}}^{\mathbf{L}}$. For ease of exposition, we introduce the concept of n-hop descendants (Def. 5). $\mathbf{M}_{\mathbf{L}}^{\mathbf{L}}$ is estimated in a progressive manner. Specifically, for each latent variable, we first estimate the mixing coefficient from itself to its every 1-hop descendant (Thm. 4(1)) and then the direct causal strength from itself to its every 1-hop descendant' each generalized homologous surrogate (Thm. 4(2)). On this basis, we investigate each latent variable's 2-hop descendants. Repeating this process, $\mathbf{M}_{\mathbf{L}}^{\mathbf{L}}$ can be estimated finally.

Definition 5. *(n-hop descendant) Given $\{L_i, L_j\} \subset \mathbf{L}$, we call L_j is an n-hop descendant of L_i, denoted by $L_j \in \mathrm{De}^n(L_i)$, if $L_j \in \mathrm{De}(L_i)$ and the longest directed path from L_i to L_j has length n.*

Intuition. Given any two latent variables, we can determine whether one is an ancestor of the other based on the partial identification result. Therefore, we can find n-hop descendants for each latent variable and each n.

Theorem 4. *Suppose $\{L_i, L_j\} \subset \mathbf{L}$, $L_j \in \mathrm{De}^n(L_i)$. $\forall O_j \in \mathrm{GHSu}(L_j)$, let*

$$\mu_{O_j^*}^{L_i} = m_{O_j^*}^{L_i} - \sum_{L_k \in \mathrm{De}(L_i) \cap \mathrm{An}(L_j)} m_{L_k}^{L_i} a_{O_j}^{L_k}. \tag{9}$$

(a) There exists $\{O_{j_1}, O_{j_2}\} \subset \mathrm{GHSu}(L_j)$ s.t. $\mu_{O_{j_1}^*}^{L_i} / a_{O_{j_1}}^{L_j} = \mu_{O_{j_2}^*}^{L_i} / a_{O_{j_2}}^{L_j}$ and $m_{L_j}^{L_i} = \mu_{O_{j_1}^*}^{L_i} / a_{O_{j_1}}^{L_j}$.

(b) $a_{O_j}^{L_i} = \mu_{O_j^*}^{L_i} - m_{L_j}^{L_i} a_{O_j}^{L_j}$.

Intuition. Given $L_j \in \mathrm{De}^n(L_i)$ and $O_j \in \mathrm{GHSu}(L_j)$, if we have already investigated $\mathrm{De}^1(L), \ldots, \mathrm{De}^{n-1}(L)$ for each $L \in \mathbf{L}$, then $m_{L_k}^{L_i}$ and $a_{O_j}^{L_k}$ in RHS of Eq. (9) are known. Moreover, $m_{O_j^*}^{L_i}$ in RHS of Eq. (9) can be derived from Eq. (8), so $\mu_{O_j^*}^{L_i}$ in LHS of Eq. (9) is known. In fact, $\mu_{O_j^*}^{L_i}$ is exactly $a_{O_j}^{L_i} + m_{L_j}^{L_i} a_{O_j}^{L_j}$ (see Eq. (50) in App. C.3). With $\mu_{O_j^*}^{L_i}$ known, (a) provides a method to estimate $m_{L_j}^{L_i}$ through L_j's some two generalized homologous surrogate O_{j_1} and O_{j_2}, where O_{j_1} and O_{j_2} are exactly those variables that are both not L_i's children (see proof in App. C.3), and (b) provides a method to estimate $a_{O_j}^{L_i}$ for each $O_j \in \mathrm{GHSu}(L_j)$ using the just estimated $m_{L_j}^{L_i}$.

Example. In Fig. 1, $L_3 \in \mathrm{De}^2(L_1)$ and $\mathrm{GHSu}(L_3) = \{O_6, O_7, O_8\}$. According to Eq. (9), $\mu_{O_6^*}^{L_1} = m_{O_6^*}^{L_1} - m_{L_2}^{L_1} a_{O_6}^{L_2}$, $\mu_{O_7^*}^{L_1} = m_{O_7^*}^{L_1} - m_{L_2}^{L_1} a_{O_7}^{L_2}$, and $\mu_{O_8^*}^{L_1} = m_{O_8^*}^{L_1} - m_{L_2}^{L_1} a_{O_8}^{L_2}$. Based on (a), since $\mu_{O_7^*}^{L_1} / a_{O_7}^{L_3} = \mu_{O_8^*}^{L_1} / a_{O_8}^{L_3}$, we have $m_{L_3}^{L_1} = \mu_{O_7^*}^{L_1} / a_{O_7}^{L_3}$. Based on (b), we have $a_{O_6}^{L_1} = \mu_{O_6^*}^{L_1} - m_{L_3}^{L_1} a_{O_6}^{L_3} \neq 0$, $a_{O_7}^{L_1} = \mu_{O_7^*}^{L_1} - m_{L_3}^{L_1} a_{O_7}^{L_3} = 0$, $a_{O_8}^{L_1} = \mu_{O_8^*}^{L_1} - m_{L_3}^{L_1} a_{O_8}^{L_3} = 0$.

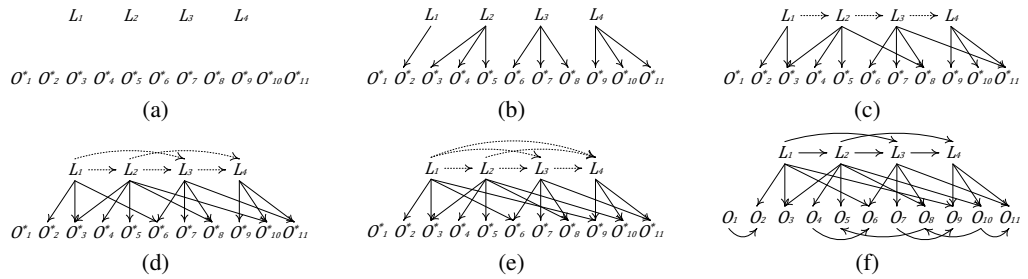


Figure 4: Illustration of Alg. 2, where solid arrow refers to parental relationship while dashed arrow refer to ancestral relationship. (a) Removing causal relations between observed variables. (b) Identifying generalized homologous surrogates of each $L \in \mathbf{L}$. (c) Investigation to $\mathrm{De}^1(L)$ for each $L \in \mathbf{L}$. (d) Investigation to $\mathrm{De}^2(L)$ for each $L \in \mathbf{L}$. (e) Investigation to $\mathrm{De}^3(L)$ for each $L \in \mathbf{L}$. (f) Recovering $\mathbf{A}$ from $\mathbf{M}$.

§ Recovering A. $\mathbf{M}$ is composed of $\mathbf{M}_{\mathbf{L}}^{\mathbf{L}}, \mathbf{0}, \mathbf{M}_{\mathbf{O}}^{\mathbf{L}}, \mathbf{M}_{\mathbf{O}}^{\mathbf{O}}$ where $\mathbf{M}_{\mathbf{L}}^{\mathbf{L}}$ is just estimated and $\mathbf{M}_{\mathbf{O}}^{\mathbf{L}}, \mathbf{M}_{\mathbf{O}}^{\mathbf{O}}$ are estimated by Alg. 1, so we can readily recover $\mathbf{A}$ from $\mathbf{M}$ by $\mathbf{M} = (\mathbf{I} - \mathbf{A})^{-1}$ following Eq (3).

§ Summary. The algorithm is summarized in Alg. 2 with $O(|\mathbf{O}|^2|\mathbf{L}|^2)$ complexity. With the result returned by Alg. 1 as the input, its procedures are shown as Fig. 4 and detailed below.

(1) Fig. 4(a): Alg. 2 removes causal relations between observed variables (Eq. (7)).

(2) Fig. 4(b): Alg. 2 locates generalized homologous surrogates (Lem. 1). Specifically, it determines $\mathrm{GHSu}(L_1) = \{O_2\}$, $\mathrm{GHSu}(L_2) = \{O_3, O_4, O_5\}$, $\mathrm{GHSu}(L_3) = \{O_6, O_7, O_8\}$, $\mathrm{GHSu}(L_4) = \{O_9, O_{10}, O_{11}\}$. Also, it obtains $\mathbf{A}_{\{O_2\}}^{\{L_1\}}, \mathbf{A}_{\{O_3,O_4,O_5\}}^{\{L_2\}}, \mathbf{A}_{\{O_6,O_7,O_8\}}^{\{L_3\}}, \mathbf{A}_{\{O_9,O_{10},O_{11}\}}^{\{L_4\}}$.

(3) Fig. 4(c): Alg. 2 investigates $\mathrm{De}^1(L)$ for each $L \in \mathbf{L}$. Specifically, $\mathrm{De}^1(L_1) = \{L_2\}$, $\mathrm{De}^1(L_2) = \{L_3\}$, $\mathrm{De}^1(L_3) = \{L_4\}$. It first estimates $m_{L_2}^{L_1}$, $m_{L_3}^{L_2}$, $m_{L_4}^{L_3}$ and then $\mathbf{A}_{\{O_3,O_4,O_5\}}^{\{L_1\}}$, $\mathbf{A}_{\{O_6,O_7,O_8\}}^{\{L_2\}}$, $\mathbf{A}_{\{O_9,O_{10},O_{11}\}}^{\{L_3\}}$ (Thm. 4) where only $a_{O_3}^{L_1}, a_{O_8}^{L_2}, a_{O_{11}}^{L_3}$ are non-zero.

(4) Fig. 4(d): Alg. 2 investigates $\mathrm{De}^2(L)$ for each $L \in \mathbf{L}$. Specifically, $\mathrm{De}^2(L_1) = \{L_3\}$, $\mathrm{De}^2(L_2) = \{L_4\}$. It first estimates $m_{L_3}^{L_1}$, $m_{L_4}^{L_2}$ and then $\mathbf{A}_{\{O_6,O_7,O_8\}}^{\{L_1\}}, \mathbf{A}_{\{O_9,O_{10},O_{11}\}}^{\{L_2\}}$ (Thm. 4) where only $a_{O_6}^{L_1}, a_{O_{10}}^{L_2}$ are non-zero.

(5) Fig. 4(d): Alg. 2 investigates $\mathrm{De}^3(L)$ for each $L \in \mathbf{L}$. Specifically, $\mathrm{De}^2(L_1) = \{L_4\}$. It first estimates $m_{L_4}^{L_1}$ and then $\mathbf{A}_{\{O_9,O_{10},O_{11}\}}^{\{L_1\}}$ (Thm. 4) where only $a_{O_9}^{L_1}$ is non-zero.

(6) Fig. 4(f): Alg. 2 recovers $\mathbf{A}$ from $\mathbf{M}$ (Eq. (3)).

Theorem 5. *Suppose the observed variables are generated by a LiNGAM with latent variables satisfying Asmps. 1 and 3, in the limit of infinite data, Algs. 1 and 2 together identifies latent variables and parental relationships correctly.*

5 Experiment

We first use four causal graphs shown as Fig. 5 to generate synthetic data. For each causal graph, we draw 10 sample sets of size 2k, 5k, 10k respectively. Each direct causal strength is sampled from a uniform distribution over $[-2.0, -0.5] \cup [0.5, 2.0]$ and each exogenous noise is generated from exponential distribution. We compare our methods with GIN (Xie et al., 2020), LaHME (Xie et al., 2022), and PO-LiNGAM (Jin et al., 2024), all of which focus on causal graph recovery of LiNGAMs with latent variables. We use 3 metrics to evaluate their performances: (1) *Error in Latent Variables*, the absolute difference between the estimated number of latent variables and the ground-truth one; (2) *Correct-Ordering Rate*, the number of correctly estimated causal orderings divided by that of ground-truth causal orderings; (3) *F1-Score* of causal edges. Results are summarized in Tab. 1, where we also report the running time. In particular, we set the size of the largest atomic unit in GIN and PO-LiNGAM to 1 for a fair comparison.

In case 1, each latent variable has at least two observed pure children, GIN demonstrates optimal performance while our algorithm also reaches comparable performance. In other cases, the pure

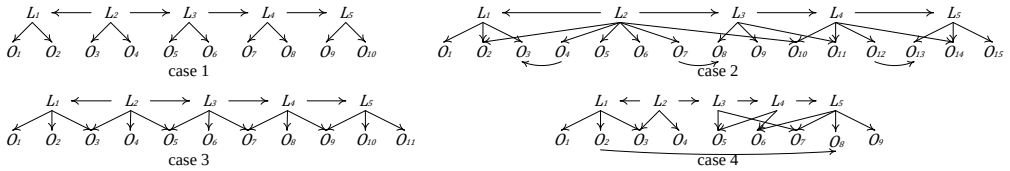


Figure 5: Causal graphs used to generate synthetic data.

Table 1: Comparison on synthetic data. ↑ means higher is better while ↓ means lower is better.

		Error in Latent Variables ↓			Correct-Ordering Rate ↑			F1-Score ↑			Running Time(s) ↓		
		2k	5k	10k	2k	5k	10k	2k	5k	10k	2k	5k	10k
Case 1	GIN	**0.0±0.0**	0.2±0.4	**0.0±0.0**	**1.00±0.00**	0.95±0.11	**1.00±0.00**	**1.00±0.00**	0.97±0.07	**1.00±0.00**	**1.56±0.13**	**1.74±0.14**	**2.07±0.18**
	LaHME	1.1±0.3	1.3±0.6	1.1±0.3	0.87±0.18	0.78±0.21	0.84±0.15	0.79±0.04	0.76±0.06	0.76±0.07	2.01±0.14	2.32±0.25	2.96±0.23
	PO-LiNGAM	0.8±0.6	0.5±0.5	0.1±0.3	0.87±0.09	0.91±0.10	0.98±0.05	0.77±0.16	0.84±0.14	0.98±0.06	121.62±27.26	116.31±27.64	117.39±16.61
	Ours	0.4±0.5	**0.1±0.3**	**0.0±0.0**	0.94±0.07	**0.99±0.03**	**1.00±0.00**	0.92±0.10	**0.98±0.05**	0.99±0.02	2.64±0.27	3.05±0.27	4.01±0.08
Case 2	GIN	3.8±0.4	3.9±0.3	4.1±0.5	0.07±0.04	0.06±0.03	0.05±0.04	0.18±0.07	0.15±0.05	0.13±0.08	**3.13±0.31**	**3.47±0.33**	**4.13±0.44**
	LaHME	1.6±1.0	1.7±0.9	2.1±1.0	0.43±0.09	0.45±0.04	0.33±0.12	0.43±0.06	0.41±0.06	0.39±0.07	36.60±14.43	86.45±71.08	116.15±96.61
	PO-LiNGAM	3.9±1.3	4.3±1.2	3.9±1.4	0.20±0.27	0.12±0.25	0.12±0.20	0.15±0.18	0.12±0.15	0.17±0.23	2214.19±779.92	2073.97±678.85	2482.57±704.83
	Ours	**1.2±0.6**	**0.7±0.8**	**0.2±0.4**	**0.82±0.09**	**0.91±0.10**	**0.97±0.05**	**0.77±0.11**	**0.88±0.13**	**0.94±0.07**	6.27±0.54	7.72±0.33	9.69±0.72
Case 3	GIN	3.0±0.4	3.0±0.0	3.0±0.0	0.12±0.05	0.11±0.00	0.11±0.00	0.33±0.06	0.33±0.00	0.33±0.00	**1.64±0.12**	**1.79±0.11**	**2.22±0.11**
	LaHME	2.0±1.0	2.5±1.0	2.7±0.9	0.37±0.24	0.24±0.17	0.19±0.09	0.58±0.15	0.48±0.10	0.46±0.07	6.36±1.18	8.64±1.13	10.26±1.19
	PO-LiNGAM	1.5±0.7	**0.8±0.9**	0.6±0.9	0.53±0.19	0.56±0.17	0.60±0.17	0.48±0.13	0.52±0.08	0.54±0.07	295.43±79.55	289.59±40.06	369.12±31.68
	Ours	**1.4±0.8**	**0.8±0.4**	**0.5±0.5**	**0.71±0.15**	**0.84±0.07**	**0.91±0.08**	**0.63±0.17**	**0.76±0.06**	**0.81±0.07**	2.58±0.34	3.23±0.28	4.36±0.38
Case 4	GIN	4.8±0.4	4.9±0.3	5.0±0.0	0.01±0.02	0.01±0.02	0.00±0.00	0.04±0.08	0.02±0.06	0.00±0.00	**0.79±0.07**	**0.87±0.08**	**1.05±0.08**
	LaHME	4.0±0.0	4.2±0.4	4.1±0.03	0.11±0.00	0.09±0.05	0.10±0.03	0.30±0.00	0.24±0.12	0.27±0.09	11.88±1.48	13.61±1.67	17.51±0.95
	PO-LiNGAM	4.9±0.3	4.8±0.4	4.9±0.3	0.01±0.03	0.02±0.04	0.01±0.02	0.03±0.01	0.04±0.08	0.02±0.06	63.39±16.79	69.07±17.81	89.61±26.58
	Ours	**2.1±1.1**	**2.1±1.1**	**3.0±0.0**	**0.44±0.21**	**0.41±0.18**	**0.25±0.01**	**0.53±0.16**	**0.53±0.09**	**0.50±0.02**	1.68±0.28	1.85±0.11	2.30±0.08

children assumption is not valid, so previous methods cannot handle these cases properly. In case 2, Asmp. 3 is valid, so our algorithm significantly outperforms others. In case 3, although Asmp. 3 is invalid, Asmp. 2 holds, so our algorithm still reaches the best performance, especially a high correct-ordering rate. In case 4, the violation of Asmp. 3 leads to a remarkable degradation in the performance of our algorithm, although it still exhibits a remarkable advantage over others since it can still identify L_1 and L_2. Moreover, in cases 2, 3, and 4, our algorithm is far more efficient than both LaHME and PO-LiNGAM. This is because LaHME has factorial complexity w.r.t. the number of variables in the worse case while PO-LiNGAM has exponential time complexity in the worst case. In contrast, our algorithm has only polynomial time complexity.

Although our algorithm eliminates the need for pure children, we acknowledge that it cannot yield satisfactory results when the sample size is small. For instance, in Case 2 where Asmp. 3 holds, our algorithm can achieve 0 error in latent variables, 1 correct-ordering rate, and 1 F1-score theoretically, but it performs poorly in practice when the sample size is 5k. This can be attributed to two main factors. First, our algorithm estimates the mixing coefficients from latent to observed variables through high-order cumulants. Compared to covariances, high-order cumulants are more sensitive to extreme values and outliers, especially when the sample size is small. Second, our algorithm operates in a progressive manner, of which each step builds upon the previous one, so errors are propagated and amplified during this process.

Besides synthetic data, we also evaluate our algorithm on a real-world dataset Holzinger and Swineford 1939 (Rosseel, 2012). Our algorithm correctly identifies the textual factor while merges the visual factor and the speed factor into a single factor. This can be attributed to the fact that both the visual factor and speed factor depends on innate abilities, while the textual factor highly depends on learning experience. More details are provided in App. D.

6 Conclusion

In this paper, we investigate a new problem setting where latent variables leave footprints in observed variables via their homologous surrogates rather than conventional pure children. We formulate two assumptions involving homologous surrogates and develop a series of novel theoretical results under each assumption, implying that the causal graph can be partially/fully recovered under the weaker/stronger assumption. Also, building on our theoretical results, we derive an algorithm that fully utilizes the properties homologous surrogates for causal graph recovery. Our work broadens the applicability of causal discovery and may benefit research in natural and social sciences.

Limitations. First, our algorithm cannot handle the latent hierarchical structure where some latent variables have no observed children. Second, this work does not accommodate non-stationary (Liu & Kuang, 2023) and cyclic (Sethuraman et al., 2023) causal relations. We will endeavor to overcome these limitations in the future.

REFERENCES

Jeffrey Adams, Niels Hansen, and Kun Zhang. Identification of partially observed linear causal models: Graphical conditions for the non-gaussian and heterogeneous cases. In *Conference on Neural Information Processing Systems*, pp. 22822–22833, 2021.

Kartik Ahuja, Jason S Hartford, and Yoshua Bengio. Weakly supervised representation learning with sparse perturbations. In *Conference on Neural Information Processing Systems*, pp. 15516–15528, 2022.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning*, pp. 372–407, 2023.

Mikhail Belkin, Luis Rademacher, and James Voss. Blind signal separation in the presence of gaussian noise. In *Conference on Learning Theory*, pp. 270–287, 2013.

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. In *Conference on Neural Information Processing Systems*, pp. 38319–38331, 2022.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. In *Conference on Neural Information Processing Systems*, pp. 45419–45462, 2023.

Ruichu Cai, Feng Xie, Clark Glymour, Zhifeng Hao, and Kun Zhang. Triad constraints for learning causal structure of latent variables. In *Conference on Neural Information Processing Systems*, pp. 12883–12892, 2019.

Ruichu Cai, Zhiyi Huang, Wei Chen, Zhifeng Hao, and Kun Zhang. Causal discovery with latent confounders based on higher-order cumulants. In *International Conference on Machine Learning*, pp. 3380–3407, 2023.

Bryant Chen, Daniel Kumor, and Elias Bareinboim. Identification and model testing in linear structural equation models using auxiliary variables. In *International Conference on Machine Learning*, pp. 757–766, 2017.

Zhengming Chen, Feng Xie, Jie Qiao, Zhifeng Hao, Kun Zhang, and Ruichu Cai. Identification of linear latent variable model with arbitrary distribution. In *AAAI Conference on Artificial Intelligence*, volume 36, pp. 6350–6357, 2022.

Zhengming Chen, Feng Xie, Jie Qiao, Zhifeng Hao, and Ruichu Cai. Some general identification results for linear latent hierarchical causal structure. In *International Joint Conferences on Artificial Intelligence*, 2023.

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

Tom Claassen and Ioan G Bucur. Greedy equivalence search in the presence of latent confounders. In *Conference on Uncertainty in Artificial Intelligence*, pp. 443–452, 2022.

Tom Claassen, Joris M Mooij, and Tom Heskes. Learning sparse causal models is not np-hard. In *Conference on Uncertainty in Artificial Intelligence*, pp. 172–181, 2013.

Diego Colombo, Marloes H Maathuis, et al. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15:3741–3782, 2014.

Xinshuai Dong, Biwei Huang, Ignavier Ng, Xiangchen Song, Yujia Zheng, Songyao Jin, Roberto Legaspi, Peter Spirtes, and Kun Zhang. A versatile causal discovery framework to allow causally-related hidden variables. In *International Conference on Learning Representations*, 2024.

Mathias Drton and Thomas S Richardson. Iterative conditional fitting for gaussian ancestral graph models. In *Conference on Uncertainty in Artificial Intelligence*, pp. 130–137, 2004.

Rong Ge and James Zou. Rich component analysis. In *International Conference on Machine Learning*, pp. 1502–1510, 2016.

Patrik O Hoyer, Shohei Shimizu, Antti J Kerminen, and Markus Palviainen. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49:362–378, 2008.

Patrik O Hoyer, Dominik Janzing, Joris Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Conference on Neural Information Processing Systems*, pp. 689–696, 2009.

Biwei Huang, Charles Jia Han Low, Feng Xie, Clark Glymour, and Kun Zhang. Latent hierarchical causal structure discovery with rank constraints. In *Conference on Neural Information Processing Systems*, pp. 5549–5561, 2022.

Aapo Hyvärinen and Erkki Oja. A fast fixed-point algorithm for independent component analysis. *Neural computation*, pp. 1483–1492, 1997.

Yibo Jiang and Bryon Aragam. Learning nonparametric latent causal graphs with unknown interventions. In *Conference on Neural Information Processing Systems*, pp. 60468–60513, 2023.

Songyao Jin, Feng Xie, Guangyi Chen, Biwei Huang, Zhengming Chen, Xinshuai Dong, and Kun Zhang. Structural estimation of partially observed linear non-gaussian acyclic model: A practical approach with identifiability. In *International Conference on Learning Representations*, 2024.

Abram Meerovich Kagan, Yurij Vladimirovich Linnik, and Calyampudi Radhakrishna Rao. *Characterization problems in mathematical statistics*. 1973.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Learning latent causal graphs via mixture oracles. In *Conference on Neural Information Processing Systems*, pp. 18087–18101, 2021.

Lingjing Kong, Biwei Huang, Feng Xie, Eric Xing, Yuejie Chi, and Kun Zhang. Identification of nonlinear latent hierarchical models. In *Conference on Neural Information Processing Systems*, 2023.

Erich Kummerfeld and Joseph Ramsey. Causal clustering for 1-factor measurement models. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1655–1664, 2016.

Thuc Duy Le, Tao Hoang, Jiuyong Li, Lin Liu, Huawen Liu, and Shu Hu. A fast pc algorithm for high dimensional causal discovery with multi-core pcs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16:1483–1495, 2016.

Xiu-Chuan Li, Kun Zhang, and Tongliang Liu. Causal structure recovery with latent variables under milder distributional and graphical assumptions. In *International Conference on Learning Representations*, 2024.

Chenxi Liu and Kun Kuang. Causal structure learning for latent intervened non-stationary data. In *International Conference on Machine Learning*, pp. 21756–21777, 2023.

Takashi Nicholas Maeda and Shohei Shimizu. Rcd: Repetitive causal discovery of linear non-gaussian acyclic models with latent confounders. In *International Conference on Artificial Intelligence and Statistics*, pp. 735–745, 2020.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17:1103–1204, 2016.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

Yves Rosseel. lavaan: An r package for structural equation modeling. *Journal of statistical software*, 48:1–36, 2012.

Saber Salehkaleybar, AmirEmad Ghassami, Negar Kiyavash, and Kun Zhang. Learning linear non-gaussian causal models in the presence of latent variables. *Journal of Machine Learning Research*, 21:1436–1459, 2020.

Muralikrishnna G Sethuraman, Romain Lopez, Rahul Mohan, Faramarz Fekri, Tommaso Biancalani, and Jan-Christian Hütter. Nodags-flow: Nonlinear cyclic causal structure learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 6371–6387, 2023.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, pp. 2003–2030, 2006.

Shohei Shimizu, Patrik O Hoyer, and Aapo Hyvärinen. Estimation of linear non-gaussian acyclic models for latent factors. *Neurocomputing*, pp. 2024–2027, 2009.

Ricardo Silva, Richard Scheines, Clark Glymour, and Peter Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7:191–246, 2006.

Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9:62–72, 1991.

Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Conference on Uncertainty in Artificial Intelligence*, pp. 499–506, 1995.

Chandler Squires, Anna Seigal, Salil S Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *International Conference on Machine Learning*, pp. 32540–32560, 2023.

Hoang-Lan Nguyen Thi and Christian Jutten. Blind source separation for convolutive mixtures. *Signal processing*, 45:209–229, 1995.

James Voss, Luis Rademacher, and Mikhail Belkin. Fast algorithms for gaussian noise invariant independent component analysis. In *Conference on Neural Information Processing Systems*, pp. 2544–2552, 2013.

Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. In *Conference on Neural Information Processing Systems*, pp. 14891–14902, 2020.

Feng Xie, Biwei Huang, Zhengming Chen, Yangbo He, Zhi Geng, and Kun Zhang. Identification of linear non-gaussian latent hierarchical structure. In *International Conference on Machine Learning*, pp. 24370–24387, 2022.

Feng Xie, Yan Zeng, Zhengming Chen, Yangbo He, Zhi Geng, and Kun Zhang. Causal discovery of 1-factor measurement models in linear latent variable models with arbitrary noise distributions. *Neurocomputing*, 526:48–61, 2023.

Feng Xie, Biwei Huang, Zhengming Chen, Ruichu Cai, Clark Glymour, Zhi Geng, and Kun Zhang. Generalized independent noise condition for estimating causal structure with latent variables. *Journal of Machine Learning Research*, 25:1–61, 2024.

Yuqin Yang, AmirEmad Ghassami, Mohamed Nafea, Negar Kiyavash, Kun Zhang, and Ilya Shpitser. Causal discovery in linear latent variable models subject to measurement error. In *Conference on Neural Information Processing Systems*, pp. 874–886, 2022.

Yan Zeng, Shohei Shimizu, Ruichu Cai, Feng Xie, Michio Yamamoto, and Zhifeng Hao. Causal discovery with multi-domain lingam for latent factors. In *International Joint Conferences on Artificial Intelligence*, pp. 2097–2103, 2021.

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. In *Conference on Neural Information Processing Systems*, pp. 50254–50292, 2023.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Conference on Uncertainty in Artificial Intelligence*, pp. 647–655, 2009.

Kun Zhang, Bernhard Schölkopf, Peter Spirtes, and Clark Glymour. Learning causality and causality-related learning: some recent progress. *National science review*, pp. 26–29, 2018.

A Related Works

Most traditional causal discovery methods (Spirtes & Glymour, 1991; Colombo et al., 2014; Le et al., 2016; Chickering, 2002; Shimizu et al., 2006; Hoyer et al., 2009; Zhang & Hyvärinen, 2009; Peters et al., 2014; Mooij et al., 2016) assume the absence of latent variables. Since latent variables are ubiquitous in real-world scenarios, extensive research has been devoted to causal discovery with latent variables. Our work is most related to those that not only allow causally-related latent variables but also can identify latent variables along with their causal relations. Most of these works make the pure children assumption. More specifically, some works (Silva et al., 2006; Shimizu et al., 2009; Kummerfeld & Ramsey, 2016; Cai et al., 2019; Chen et al., 2022; Zeng et al., 2021; Xie et al., 2022; Chen et al., 2023) make the special pure children assumption that each latent variable has multiple pure children, where a variable is said a pure child of anther only if the latter is the only parent of the former. Others (Xie et al., 2020; 2024; Huang et al., 2022; Dong et al., 2024; Jin et al., 2024) make the general pure children assumption that each latent variable belongs to a latent set (comprising one or more latent variables) which has sufficient pure children, where a variable is said a pure child of a latent set only if all parents of the former are within the latter. Although the general pure children assumption is weaker than the special one, it comes at the cost of local unidentifiability. Specifically, if a latent set contains multiple latent variables, none of which has its own pure children, even the existence of causal relations between these latent variables cannot be determined, let alone their directions. By introducing the concept of homologous surrogates, our work eliminates the need for pure children, in stark contrast to the above studies.

Although the pure children assumption has been widely adopted by previous works, Adams et al. (2021) argue that it is restrictive in practice and also not necessary for identifiability of linear latent non-Gaussian models in theory. They develop a causal discovery algorithm under the assumption which is exactly sufficient and necessary for identifiability. Unfortunately, this algorithm is unpractical as acknowledged by themselves. First, it estimates $\mathbf{M}_{\mathbf{O}}^{\mathbf{V}}$ via overcomplete independent components analysis, which requires the number of latent variables as prior knowledge and is computationally intractable. Second, to recover $\mathbf{A}$ from $\mathbf{M}_{\mathbf{O}}^{\mathbf{V}}$, it needs to test which submatrices' singular values are exact zeros, which is quite sensitive to noises. Recently, Li et al. (2024) suggest that a pseudo-pure pair, composed of two adjacent observed variables that both become pure children of a same latent variable if the edge between them is removed, can be transformed into two pure children under certain assumption. Based on this finding, they propose a practical causal discovery algorithm free from the pure children assumption. Clearly, our work diverges significantly from existing works these two studies, offering a novel perspective on causal discovery with latent variables.

While the works discussed above all focus on the linear case, several studies have ventured into nonlinear problems, but most assume access to counterfactual data (Brehmer et al., 2022; Ahuja et al., 2022) or interventional data (Ahuja et al., 2023; Jiang & Aragam, 2023; Buchholz et al., 2023; Zhang et al., 2023). Notably, without structural restrictions such as our homologous surrogates assumption, even linear causal models are unidentifiable without comprehensive interventional data obtained by intervening on each latent variable individually (Squires et al., 2023). To the best of our knowledge, only Kivva et al. (2021) and Kong et al. (2023) can handle non-linear problems with latent variables using solely observational data, but they both make strong assumptions, e.g., all latent variables are discrete and the mapping from all exogenous noises to observed variables are invertible. We leave further research on nonlinear problems to our future work.

B Notations

We summarize notations in Tab. 2

Table 2: Summary of notations.

Notation	Description	First appeared
$\mathcal{G}$	Causal graph	Sec. 2
$\mathbf{L}$	Set of latent variables	Sec. 2
$\mathbf{O}$	Set of observed variables	Sec. 2
$\mathbf{V}$	$\mathbf{L} \cup \mathbf{O}$	Sec. 2
L_i	A latent variable	Sec. 2
O_i	An observed variable	Sec. 2
V_i	A latent or observed variable	Sec. 2
ϵ_{V_i}	Exogenous noise of V_i	Sec. 2
$\mathrm{Pa}(V)$	Parents of V	Sec. 2
$\mathrm{Pa}(\mathbf{V}')$	$\bigcup_{V \in \mathbf{V}'} \mathrm{Pa}(V)$	Sec. 2
$\mathrm{Pa}_{\mathbf{L}}(V)$	Latent parents of V	Sec. 2
$\mathrm{Pa}_{\mathbf{O}}(V)$	Observed parents of V	Sec. 2
$\mathrm{Ch}(V)$	Children of V	Sec. 2
$\mathrm{Ne}(V)$	Neighbors of V	Sec. 2
$\mathrm{An}(V)$	Ancestors of V	Sec. 2
$\mathrm{De}(V)$	Descendants of V	Sec. 2
$a_{V_j}^{V_i}$	direct causal strength from V_i to V_j	Sec. 2
$m_{V_j}^{V_i}$	mixing coefficient from V_i to V_j	Sec. 2
$\mathbf{A}_{\mathbf{V}_2}^{\mathbf{V}_1}$	adjacency matrix from $\mathbf{V}_1$ to $\mathbf{V}_2$	Sec. 2
$\mathbf{M}_{\mathbf{V}_2}^{\mathbf{V}_1}$	mixing matrix from $\mathbf{V}_1$ to $\mathbf{V}_2$	Sec. 2
$\mathrm{HSu}(V)$	Homologous surrogates of V	Def. 1 in Sec. 3
$\mathbf{J}$	Set of removed latent variables	Sec. 3
$\mathbf{K}$	Set of removed observed variables	Sec. 3
$\tilde{O}_i$	Auxiliary variable of O_i satisfying Cond. 1(2)	Sec. 3
$\mathrm{R}(V_1, V_2 \mid V_3)$	Pseudo-residual of V_1, V_2 relative to V_3	Def. 2 in Sec. 3
$\mathrm{GHSu}(V)$	Generalized homologous surrogates of V	Def. 4 in Sec. 4
O_i^*	O_i with its all observed parents removed	Eq (7) in Sec. 4
$\mathrm{De}^n(V)$	n-hop descendants of V	Def. 5 in Sec. 4
$\mathbf{K}_1$	$\{O \in \mathbf{K} \mid \mathrm{Ch}(O) \neq \emptyset\}$	Eq. (13) in App. C.2
$\mathbf{K}_2$	$\{O \in \mathbf{K} \mid \mathrm{Ch}(O) = \emptyset\}$	Eq. (14) in App. C.2
$\mathcal{G}^*$	$\mathcal{G}$ with all causal relations between observed variables removed	App. C.3
$\mathbf{O}^*$	$\{O^* \mid O \in \mathbf{O}\}$	App. C.3

C PROOF

C.1 IMPORTANT LEMMAS

Darmois-Skitovitch (D-S) Theorem. (Kagan et al., 1973) Suppose two random variables V_1 and V_2 are both linear combinations of independent random variables $\{n_i\}_i$:

$$V_1 = \sum_i \alpha_i n_i, \quad V_2 = \sum_i \beta_i n_i. \tag{10}$$

Then, if $V_1 \perp\!\!\!\perp V_2$, each n_i for which $\alpha_i \beta_i \neq 0$ follows Gaussian distribution. That is, if there exists a non-Gaussian n_j s.t. $\alpha_j \beta_j \neq 0$, $V_1 \not\perp\!\!\!\perp V_2$.

Lemma 2. *If three variables V_1, V_2, V_3 can be expressed as*

$$V_1 = \gamma_1 e + e_1, \quad V_2 = \gamma_2 e + e_2, \quad V_3 = \gamma_3 e + e_3, \tag{11}$$

where $e \perp\!\!\!\perp \{e_1, e_2, e_3\}$, $e_3 \perp\!\!\!\perp \{e_1, e_2\}$, and $\gamma_1\gamma_2\gamma_3 \neq 0$, then $\mathrm{R}(V_1, V_2|V_3) \perp\!\!\!\perp V_3$.

Proof.

$$\mathrm{R}(V_1, V_2|V_3) = (\gamma_1 e + e_1) - \frac{\mathrm{Cov}(\gamma_1 e + e_1, \gamma_3 e + e_3)}{\mathrm{Cov}(\gamma_2 e + e_2, \gamma_3 e + e_3)}(\gamma_2 e + e_2) = e_1 - \frac{\gamma_1}{\gamma_2} e_2 \perp\!\!\!\perp V_3. \tag{12}$$

□

Lemma 3. *Given three variables V_1, V_2, V_3 where $\mathrm{Cov}(V_1, V_3)\mathrm{Cov}(V_2, V_3) \neq 0$, if $\exists V \in \mathbf{V}$ s.t. only one of $m^V_{V_1}$ and $m^V_{V_2}$ is non-zero and $m^V_{V_3}$ is non-zero, then $\mathrm{R}(V_1, V_2|V_3) \not\perp\!\!\!\perp V_3$.*

Proof. Since $\mathrm{Cov}(V_1, V_3)\mathrm{Cov}(V_2, V_3) \neq 0$ and one of $m^V_{V_1}$ and $m^V_{V_2}$ is non-zero, $\mathrm{R}(V_1, V_2|V_3)$ contains ϵ_V. Because $m^V_{V_3} \neq 0$, $\mathrm{R}(V_1, V_2|V_3) \not\perp\!\!\!\perp V_3$ based on D-S Theorem. □

C.2 PROOF OF THEORETICAL RESULTS IN SEC. 3

Definition 1. (homologous surrogate) $O \in \mathbf{O}$ is called a homologous surrogate of $L \in \mathbf{L}$, denoted by $O \in \mathrm{HSu}(L)$, if $O \in \mathrm{Ch}(L)$, $\mathrm{Ch}(O) = \emptyset$, $\mathrm{An}_{\mathbf{L}}(O) = \mathrm{An}(L) \cup \{L\}$ and $\mathrm{An}_{\mathbf{O}}(O) \cap \mathrm{De}_{\mathbf{O}}(L) = \emptyset$.

Assumption 2. $\forall L \in \mathbf{L}$, $\mathrm{HSu}(L) \neq \emptyset$ and $|\mathrm{Ch}(L)| \geq 2$.

Remark. We can easily derive from Asmp. 2 that $\forall L \in \mathbf{L}$, $|\mathrm{De}_{\mathbf{O}}(L)| \geq 2$.

Condition 1. (1) For each $V \in \mathbf{V} \backslash (\mathbf{J} \cup \mathbf{K})$, $\mathrm{De}(V) \cap (\mathbf{J} \cup \mathbf{K}) = \emptyset$. (2) For each $L \in \mathbf{J}$ and $O \in \mathbf{K}$ where $\mathrm{Ch}(O) \neq \emptyset$, $m^L_{\tilde{O}_i} = m^O_{\tilde{O}_i} = 0$.

Let $\mathbf{K}_1 = \{O \in \mathbf{K} | \mathrm{Ch}(O) \neq \emptyset\}$ and $\mathbf{K}_2 = \{O \in \mathbf{K} | \mathrm{Ch}(O) = \emptyset\}$. Based on Cond. 1, each $O_i \in \mathbf{O} \backslash \mathbf{K}$ and corresponding $\tilde{O}_i$ (which, as mentioned in the main text, is a linear combination of O_i and variables in $\mathbf{K}$ where the coefficient of O_i is always 1 while that of each variable in $\mathbf{K}$ is not fixed) can be expressed as

$$O_i = \sum_{L_i \in \mathbf{J}} m^{L_i}_{O_i} \epsilon_{L_i} + \sum_{L_j \in \mathbf{L} \backslash \mathbf{J}} m^{L_j}_{O_i} \epsilon_{L_j} + \sum_{O_{j_1} \in \mathbf{K}_1} m^{O_{j_1}}_{O_i} \epsilon_{O_{j_1}} + \sum_{O_k \in \mathbf{O} \backslash \mathbf{K}} m^{O_k}_{O_i} \epsilon_{O_k}, \tag{13}$$

$$\tilde{O}_i = \sum_{L_j \in \mathbf{L} \backslash \mathbf{J}} m^{L_j}_{O_i} \epsilon_{L_j} + \sum_{O_{j_2} \in \mathbf{K}_2} \lambda_{ij_2} \epsilon_{O_{j_2}} + \sum_{O_k \in \mathbf{O} \backslash \mathbf{K}} m^{O_k}_{O_i} \epsilon_{O_k}. \tag{14}$$

Lemma 4. *If $\exists V_i \in \mathbf{V} \backslash (\mathbf{J} \cup \mathbf{K})$ and $\{O_i, O_j\} \subset \mathbf{O} \backslash \mathbf{K}$ s.t. $m^{V_i}_{O_i} m^{V_i}_{O_j} \neq 0$, then $\mathrm{Cov}(O_i, \tilde{O}_j) \neq 0$.*

Proof. Based on Eqs. (13) and (14),

$$\mathrm{Cov}(O_i, \tilde{O}_j) = \sum_{V \in \mathbf{V} \backslash (\mathbf{J} \cup \mathbf{K})} m^V_{O_i} m^V_{O_j} \mathrm{Var}(\epsilon_V). \tag{15}$$

Suppose each nonzero element of $\mathbf{M}$ is positive, then $\mathrm{Cov}(O_i, \tilde{O}_j) \neq 0$. According to the rank-faithfulness assumption, there is always $\mathrm{Cov}(O_i, \tilde{O}_j) \neq 0$. □

Theorem 1. Suppose $O_i \in \mathbf{O}\backslash\mathbf{K}$, then $\mathrm{An}(O_i) \subset (\mathbf{J} \cup \mathbf{K})$ if and only if $\forall O_j \in \mathbf{O}\backslash(\mathbf{K} \cup \{O_i\})$, $\mathrm{R}(O_j, O_i|\tilde{O}_i) \perp\!\!\!\perp \tilde{O}_i$.

Proof sketch. If $\mathrm{An}(O_i) \subset (\mathbf{J} \cup \mathbf{K})$, we can prove independence easily for $O_j \notin \mathrm{De}(O_i)$ (in this case, $\mathrm{R}(O_j, O_i|\tilde{O}_i) = O_j \perp\!\!\!\perp \tilde{O}_i$) and based on property (1) in Int. of Def. 2 for $O_j \in \mathrm{De}(O_i)$ (ϵ_{O_i} serves as e in property (1)). Otherwise, there exists an $O_j \in \mathbf{O}\backslash(\mathbf{K} \cup \{O_i\})$ s.t. $\mathrm{Cov}(\tilde{O}_i, O_j) \neq 0$ and $m_{O_j}^{O_i} = 0$, so dependence can be proven by property (2) in Int. of Def. 2.

Proof. "Only if": Based on Eqs. (13) and (14),

$$O_j = \sum_{L_i \in \mathbf{J}} m_{O_j}^{L_i} \epsilon_{L_i} + \sum_{L_j \in \mathbf{L}\backslash\mathbf{J}} m_{O_j}^{L_j} \epsilon_{L_j} + \sum_{O_{k_1} \in \mathbf{K}_1} m_{O_j}^{O_{k_1}} \epsilon_{O_{k_1}} + \sum_{O_l \in \mathbf{O}\backslash\mathbf{K}} m_{O_j}^{O_l} \epsilon_{O_l}, \tag{16}$$

$$O_i = \sum_{L_i \in \mathbf{J}} m_{O_i}^{L_i} \epsilon_{L_i} + \sum_{O_{k_1} \in \mathbf{K}_1} m_{O_i}^{O_{k_1}} \epsilon_{O_{k_1}} + \epsilon_{O_i}, \tag{17}$$

$$\tilde{O}_i = \sum_{O_{k_2} \in \mathbf{K}_2} \lambda_{ik_2} \epsilon_{O_{k_2}} + \epsilon_{O_i}. \tag{18}$$

If $m_{O_j}^{O_i} = 0$, then $\mathrm{Cov}(\tilde{O}_i, O_j) = 0$, $\mathrm{R}(O_j, O_i|\tilde{O}_i) = O_j \perp\!\!\!\perp \tilde{O}_i$. Otherwise, based on Lem. 2, there is also $\mathrm{R}(O_j, O_i|\tilde{O}_i) \perp\!\!\!\perp \tilde{O}_i$, where ϵ_{O_i} serves as e in Eq. (11).

"If": We prove this part by contradiction. Suppose $\mathrm{An}(O_i) \not\subset (\mathbf{J} \cup \mathbf{K})$. There are two possible cases as follows.

1. Suppose $\mathrm{An}_{\mathbf{O}}(O_i) \not\subset \mathbf{K}$, let $O_j \in (\mathbf{O}\backslash\mathbf{K}) \cap \mathrm{An}_{\mathbf{O}}(O_i)$. An illustrative example is shown as Fig. 6(a). As $m_{O_i}^{O_j} m_{O_j}^{O_j} \neq 0$, $\mathrm{Cov}(O_j, \tilde{O}_i) \neq 0$ based on Lem. 4. Also, it is trivial that $\mathrm{Cov}(O_i, \tilde{O}_i) \neq 0$. Since $m_{O_i}^{O_i} m_{\tilde{O}_i}^{O_i} \neq 0$ and $m_{O_j}^{O_i} = 0$, we can derive $\mathrm{R}(O_j, O_i|\tilde{O}_i) \not\perp\!\!\!\perp \tilde{O}_i$ based on Lem. 3.

2. Suppose $\mathrm{An}_{\mathbf{O}}(O_i) \subset \mathbf{K}$, then $\mathrm{An}_{\mathbf{L}}(O_i) \not\subset \mathbf{J}$. Let $L_i \in \mathrm{An}_{\mathbf{L}}(O_i) \cap (\mathbf{L}\backslash\mathbf{J})$, there are two possible sub-cases as follows.

   (a) Suppose $\mathrm{De}_{\mathbf{O}}(O_i) = \emptyset$, let $O_j \in \mathrm{De}_{\mathbf{O}}(L_i)\backslash\{O_i\}$. An illustrative example is shown as Fig. 6(b). As $m_{O_i}^{L_i} m_{O_j}^{L_i} \neq 0$, $\mathrm{Cov}(O_j, \tilde{O}_i) \neq 0$ based on Lem. 4. Also, it is trivial that $\mathrm{Cov}(O_i, \tilde{O}_i) \neq 0$. Since $m_{O_i}^{O_i} m_{\tilde{O}_i}^{O_i} \neq 0$ and $m_{O_j}^{O_i} = 0$, we can derive $\mathrm{R}(O_j, O_i|\tilde{O}_i) \not\perp\!\!\!\perp \tilde{O}_i$ based on Lem. 3.

   (b) Suppose $\mathrm{De}_{\mathbf{O}}(O_i) \neq \emptyset$, let $O_j \in \mathrm{HSu}(L_i)$, it is trivial that $O_j \notin \mathrm{De}(O_i) \cup \{O_i\}$. An illustrative example is shown as Fig. 6(c). As $m_{O_i}^{L_i} m_{O_j}^{L_i} \neq 0$, $\mathrm{Cov}(O_j, \tilde{O}_i) \neq 0$ based on Lem. 4. Also, it is trivial that $\mathrm{Cov}(O_i, \tilde{O}_i) \neq 0$. Since $m_{O_i}^{O_i} m_{\tilde{O}_i}^{O_i} \neq 0$ and $m_{O_j}^{O_i} = 0$, we can derive $\mathrm{R}(O_j, O_i|\tilde{O}_i) \not\perp\!\!\!\perp \tilde{O}_i$ based on Lem. 3.

This finishes the proof. □

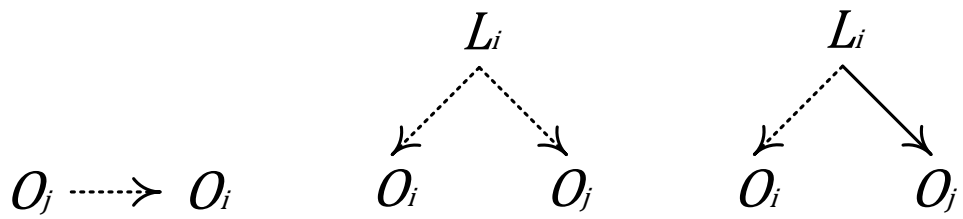
Corollary 1. Suppose O_i satisfies Thm. 1, then $\forall O_j \in \mathbf{O}\backslash(\mathbf{K} \cup \{O_i\})$, $m_{O_j}^{O_i} = \frac{\mathrm{Cov}(\tilde{O}_i, O_j)}{\mathrm{Cov}(\tilde{O}_i, O_i)}$.

Proof. Based on Eqs. (13) and (14),

$$\tilde{O}_i = \sum_{O_{k_2} \in \mathbf{K}_2} \lambda_{ik_2} \epsilon_{O_{k_2}} + \epsilon_{O_i}. \tag{19}$$

$$O_j = \sum_{L_i \in \mathbf{J}} m_{O_j}^{L_i} \epsilon_{L_i} + \sum_{L_j \in \mathbf{L}\backslash\mathbf{J}} m_{O_j}^{L_j} \epsilon_{L_j} + \sum_{O_{k_1} \in \mathbf{K}_1} m_{O_j}^{O_{k_1}} \epsilon_{O_{k_1}} + \sum_{O_l \in \mathbf{O}\backslash\mathbf{K}} m_{O_j}^{O_l} \epsilon_{O_l}, \tag{20}$$

$$O_i = \sum_{L_i \in \mathbf{J}} m_{O_i}^{L_i} \epsilon_{L_i} + \sum_{O_{k_1} \in \mathbf{K}_1} m_{O_i}^{O_{k_1}} \epsilon_{O_{k_1}} + \epsilon_{O_i}, \tag{21}$$



(a) Case 1 (b) Case 2(a) (c) Case 2(b)

Figure 6: Illustration of "If" part in proof of Thm. 1. A dotted arrow from V_1 to V_2 means that $V_2 \in \mathrm{De}(V_1)$.

Therefore,

$$\frac{\mathrm{Cov}(\tilde{O}_i, O_j)}{\mathrm{Cov}(\tilde{O}_i, O_i)} = \frac{m_{O_j}^{O_i}\mathrm{Var}(\epsilon_{O_i})}{\mathrm{Var}(\epsilon_{O_i})} = m_{O_j}^{O_i}. \tag{22}$$

□

Corollary 2. Suppose O_i satisfies Thm. 1, if we update $\mathbf{K}$ to $\mathbf{K} \cup \{O_i\}$ and $\tilde{O}_j$ to $\tilde{O}_j - m_{O_j}^{O_i}\tilde{O}_i$ for each $O_j \in \mathbf{O}\backslash\mathbf{K}$, Cond. 1 is still valid.

Proof. Based on Thm. 1, it is trivial that Cond. 1(1) is valid.

Based on Eq. (14), before removal

$$\tilde{O}_i = \sum_{O_k \in \mathbf{K}_2} \lambda_{ik}\epsilon_{O_k} + \epsilon_{O_i}, \tag{23}$$

$$\tilde{O}_j = \sum_{L_i \in \mathbf{L}\backslash\mathbf{J}} m_{O_j}^{L_i}\epsilon_{L_i} + \sum_{O_k \in \mathbf{K}_2} \lambda_{jk}\epsilon_{O_k} + \sum_{O_l \in \mathbf{O}\backslash\mathbf{K}} m_{O_j}^{O_l}\epsilon_{O_l}, \tag{24}$$

Let $\lambda'_{jk} = \lambda_{jk} - m_{O_j}^{O_i}\lambda_{ik}$, then

$$\tilde{O}_j - m_{O_j}^{O_i}\tilde{O}_i = \sum_{L_i \in \mathbf{L}\backslash\mathbf{J}} m_{O_j}^{L_i}\epsilon_{L_i} + \sum_{O_k \in \mathbf{K}_2} \lambda'_{jk}\epsilon_{O_k} + \sum_{O_l \in \mathbf{O}\backslash(\mathbf{K}\cup\{O_i\})} m_{O_j}^{O_l}\epsilon_{O_l}, \tag{25}$$

so Cond. 1(2) is also valid. □

Lemma 5. *If* $\forall O \in \mathbf{O}\backslash\mathbf{K}, \mathrm{An}(O) \not\subset \mathbf{J} \cup \mathbf{K}$, *then* $\forall O \in \mathbf{O}\backslash\mathbf{K}, \mathrm{An}_{\mathbf{L}}(O) \not\subset \mathbf{J}$.

Proof. We prove it by contradiction. Suppose $\exists O_i \in \mathbf{O}\backslash\mathbf{K}$ s.t. $\mathrm{An}_{\mathbf{L}}(O_i) \subset \mathbf{J}$, then since $\mathrm{An}(O) \not\subset \mathbf{J}\cup\mathbf{K}$, $\mathrm{An}_{\mathbf{O}}(O_i) \not\subset \mathbf{K}$. Let $O_j \in \mathrm{An}_{\mathbf{O}}(O_i)\backslash\mathbf{K}$ s.t. $\mathrm{An}_{\mathbf{O}}(O_j) \cap (\mathrm{An}_{\mathbf{O}}(O_i)\backslash\mathbf{K}) = \emptyset$, that is, $\mathrm{An}_{\mathbf{O}}(O_j) \subset \mathbf{K}$. In addition, $\mathrm{An}_{\mathbf{L}}(O_j) \subset \mathrm{An}_{\mathbf{L}}(O_i) \subset \mathbf{J}$, so $\mathrm{An}(O_j) \subset \mathbf{J} \cup \mathbf{K}$, which leads to contradiction. □

Theorem 2. Suppose $\forall O \in \mathbf{O}\backslash\mathbf{K}, \mathrm{An}(O) \not\subset \mathbf{J} \cup \mathbf{K}$. Given $O_i \in \mathbf{O}\backslash\mathbf{K}$, then $\mathrm{Ch}(O_i) = \emptyset$, $\mathrm{Pa}_{\mathbf{O}}(O_i)\backslash\mathbf{K} = \emptyset$, $|\mathrm{Pa}_{\mathbf{L}}(O_i)\backslash\mathbf{J}| = 1$, and $\mathrm{An}(\mathrm{Pa}_{\mathbf{L}}(O_i)\backslash\mathbf{J}) \subset \mathbf{J}$ if and only if $\forall\{O_j, O_k\} \subset \mathbf{O}\backslash(\mathbf{K} \cup \{O_i\})$ where $\mathrm{Cov}(\tilde{O}_i, O_j)\mathrm{Cov}(\tilde{O}_i, O_k) \neq 0$, $\mathrm{R}(O_j, O_k|\tilde{O}_i) \perp\!\!\!\perp \tilde{O}_i$.

Proof sketch. If O_i satisfies the graphical condition, we can prove independence based on property (1) in Int. of Def. 2 (let $\mathrm{Pa}_{\mathbf{L}}(O_i)\backslash\mathbf{J} = \{L\}$, ϵ_L serves as e in property (1)). Otherwise, there exists $V \in \mathbf{V}\backslash(\mathbf{J} \cup \mathbf{K})$ and $\{O_j, O_k\} \subset \mathbf{O}\backslash(\mathbf{K} \cup \{O_i\})$ s.t. $\mathrm{Cov}(\tilde{O}_i, O_j)\mathrm{Cov}(\tilde{O}_i, O_k) \neq 0$, $m_{\tilde{O}_i}^{V} m_{O_j}^{V} \neq 0$, and $m_{O_k}^{V} = 0$, so dependence can be proven by property (2) in Int. of Def. 2.

Proof. "Only if": Let $\mathrm{Pa}_{\mathbf{L}}(O_i)\backslash\mathbf{J} = \{L_i\}$, note that $\{O_j, O_k\} \cap \mathrm{De}(O_i) = \emptyset$, based on Eqs. (13) and (14),

$$\tilde{O}_i = m_{O_i}^{L_i}\epsilon_{L_i} + \sum_{O_{l_2}\in\mathbf{K}_2} \lambda_{il_2}\epsilon_{O_{l_2}} + \epsilon_{O_i}. \tag{26}$$

$$O_j = \sum_{L_j\in\mathbf{J}} m_{O_j}^{L_j}\epsilon_{L_j} + \sum_{L_k\in\mathbf{L}\backslash\mathbf{J}} m_{O_j}^{L_k}\epsilon_{L_k} + \sum_{O_{l_1}\in\mathbf{K}_1} m_{O_j}^{O_{l_1}}\epsilon_{O_{l_1}} + \sum_{O_m\in\mathbf{O}\backslash(\mathbf{K}\cup\{O_i\})} m_{O_j}^{O_m}\epsilon_{O_m}, \tag{27}$$

$$O_k = \sum_{L_j\in\mathbf{J}} m_{O_k}^{L_j}\epsilon_{L_j} + \sum_{L_k\in\mathbf{L}\backslash\mathbf{J}} m_{O_k}^{L_k}\epsilon_{L_k} + \sum_{O_{l_1}\in\mathbf{K}_1} m_{O_k}^{O_{l_1}}\epsilon_{O_{l_1}} + \sum_{O_m\in\mathbf{O}\backslash(\mathbf{K}\cup\{O_i\})} m_{O_k}^{O_m}\epsilon_{O_m}. \tag{28}$$

Since $\mathrm{Cov}(\tilde{O}_i, O_j)\mathrm{Cov}(\tilde{O}_i, O_k) = (m_{O_i}^{L_i})^2 m_{O_j}^{L_i} m_{O_k}^{L_i}(\mathrm{Var}(\epsilon_{L_i}))^2 \neq 0$, there is $m_{O_j}^{L_i} m_{O_k}^{L_i} \neq 0$. Based on Lem. 2, $\mathrm{R}(O_j, O_k|\tilde{O}_i) \perp\!\!\!\perp \tilde{O}_i$, where ϵ_{L_i} serves as e in Eq. (11).

"If": We prove this part by contradiction.

1. Suppose $\mathrm{Ch}(O_i) \neq \emptyset$, let $O_j \in \mathrm{Ch}(O_i)$. Based on Lem. 5, $\mathrm{An}_{\mathbf{L}}(O_i) \not\subset \mathbf{J}$ and let $L_i \in \mathrm{An}_{\mathbf{L}}(O_i)\backslash\mathbf{J}$. Besides, let $O_k \in \mathrm{HSu}(L_i)$, it is trivial that $O_k \notin \{O_i, O_j\}$ and $O_i \notin \mathrm{An}(O_k)$. An illustrative example is shown as Fig. 7(a). As $m_{O_i}^{L_i} m_{O_j}^{L_i} m_{O_k}^{L_i} \neq 0$, $\mathrm{Cov}(O_j, \tilde{O}_i)\mathrm{Cov}(O_k, \tilde{O}_i) \neq 0$ based on Lem. 4. Since $m_{\tilde{O}_i}^{O_i} m_{O_j}^{O_i} \neq 0$ and $m_{O_k}^{O_i} = 0$, we can derive $\mathrm{R}(O_j, O_k|\tilde{O}_i) \not\perp\!\!\!\perp \tilde{O}_i$ based on Lem. 3.

2. Suppose $\mathrm{Ch}(O_i) = \emptyset$ and $\mathrm{Pa}_{\mathbf{O}}(O_i)\backslash\mathbf{K} \neq \emptyset$, let $O_j \in \mathrm{Pa}_{\mathbf{O}}(O_i)\backslash\mathbf{K}$. Based on Lem. 5, $\mathrm{An}_{\mathbf{L}}(O_j) \not\subset \mathbf{J}$ and let $L_i \in \mathrm{An}_{\mathbf{L}}(O_j)\backslash\mathbf{J}$. Besides, let $O_k \in \mathrm{HSu}(L_i)$, it is trivial that $O_k \notin \{O_i, O_j\}$ and $O_j \notin \mathrm{An}(O_k)$. An illustrative example is shown as Fig. 7(b). As $m_{O_i}^{L_i} m_{O_j}^{L_i} m_{O_k}^{L_i} \neq 0$, $\mathrm{Cov}(O_j, \tilde{O}_i)\mathrm{Cov}(O_k, \tilde{O}_i) \neq 0$ based on Lem. 4. Since $m_{\tilde{O}_i}^{O_j} m_{O_j}^{O_j} \neq 0$ and $m_{O_k}^{O_j} = 0$, we can derive $\mathrm{R}(O_j, O_k|\tilde{O}_i) \not\perp\!\!\!\perp \tilde{O}_i$ based on Lem. 3.

3. Suppose $\mathrm{Ch}(O_i) = \mathrm{Pa}_{\mathbf{O}}(O_i)\backslash\mathbf{K} = \emptyset$ and $|\mathrm{Pa}_{\mathbf{L}}(O_i)\backslash\mathbf{J}| \geq 2$, let $\{L_i, L_j\} \subset \mathrm{Pa}_{\mathbf{L}}(O_i)\backslash\mathbf{J}$. Without loss of generality, let $L_j \notin \mathrm{An}(L_i)$. Let $O_j \in \mathrm{De}(L_j)\backslash\{O_i\}$ and $O_k \in \mathrm{HSu}(L_i)$. It is trivial that $O_k \notin \{O_i, O_j\}$ and $L_j \notin \mathrm{An}(O_k)$. An illustrative example is shown as Fig. 7(c). As $m_{O_j}^{L_j} m_{O_i}^{L_j} \neq 0$ and $m_{O_k}^{L_i} m_{O_i}^{L_i} \neq 0$, $\mathrm{Cov}(O_j, \tilde{O}_i)\mathrm{Cov}(O_k, \tilde{O}_i) \neq 0$ based on Lem. 4. Since $m_{\tilde{O}_i}^{L_j} m_{O_j}^{L_j} \neq 0$ and $m_{O_k}^{L_j} = 0$, we can derive $\mathrm{R}(O_j, O_k|\tilde{O}_i) \not\perp\!\!\!\perp \tilde{O}_i$ based on Lem. 3.

4. Suppose $\mathrm{Ch}(O_i) = \mathrm{Pa}_{\mathbf{O}}(O_i)\backslash\mathbf{K} = \emptyset$, $|\mathrm{Pa}_{\mathbf{L}}(O_i)\backslash\mathbf{J}| = 1$, and $\mathrm{An}(\mathrm{Pa}_{\mathbf{L}}(O_i)\backslash\mathbf{J}) \not\subset \mathbf{J}$. Let $\mathrm{Pa}_{\mathbf{L}}(O_i)\backslash\mathbf{J} = \{L_i\}$, $L_j \in \mathrm{An}(L_i)\backslash\mathbf{J}$, $O_j \in \mathrm{De}(L_i)\backslash\{O_i\}$, and $O_k \in \mathrm{HSu}(L_j)$. It is trivial that $O_k \notin \{O_i, O_j\}$ and $L_i \notin \mathrm{An}(O_k)$. An illustrative example is shown as Fig. 7(d). As $m_{O_i}^{L_j} m_{O_j}^{L_j} m_{O_k}^{L_j} \neq 0$, $\mathrm{Cov}(O_j, \tilde{O}_i)\mathrm{Cov}(O_k, \tilde{O}_i) \neq 0$ based on Lem. 4. Since $m_{\tilde{O}_i}^{L_i} m_{O_j}^{L_i} \neq 0$ and $m_{O_k}^{L_i} = 0$, we can derive $\mathrm{R}(O_j, O_k|\tilde{O}_i) \not\perp\!\!\!\perp \tilde{O}_i$ based on Lem. 3.

This finishes the proof. □

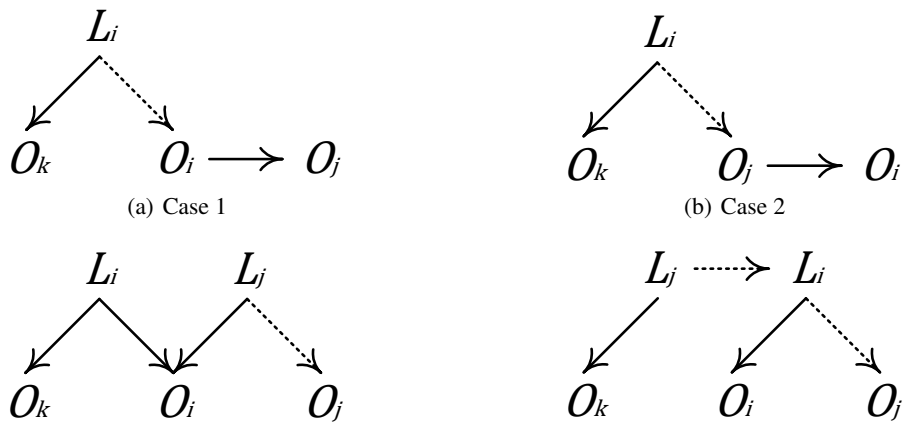
Proposition 1. Suppose O_i and O_j satisfy Thm. 2, then $\mathrm{Pa}_{\mathbf{L}}(O_i)\backslash\mathbf{J} = \mathrm{Pa}_{\mathbf{L}}(O_j)\backslash\mathbf{J}$ if and only if $\mathrm{Cov}(\tilde{O}_i, O_j) \neq 0$.

Proof. Let $\mathrm{Pa}_{\mathbf{L}}(O_i)\backslash\mathbf{J} = \{L_i\}$ and $\mathrm{Pa}_{\mathbf{L}}(O_j)\backslash\mathbf{J} = \{L_j\}$. Based on Eqs. (13) and (14),

$$\tilde{O}_i = m_{O_i}^{L_i}\epsilon_{L_i} + \sum_{O_{l_2}\in\mathbf{K}_2} \lambda_{il_2}\epsilon_{O_{l_2}} + \epsilon_{O_i}. \tag{29}$$

$$O_j = \sum_{L_k\in\mathbf{J}} m_{O_j}^{L_k}\epsilon_{L_k} + m_{O_j}^{L_j}\epsilon_{L_j} + \sum_{O_{l_1}\in\mathbf{K}_1} m_{O_j}^{O_{l_1}}\epsilon_{O_{l_1}} + \epsilon_{O_j}. \tag{30}$$

Obviously, $L_i = L_j$ if and only if $\mathrm{Cov}(\tilde{O}_i, O_j) = m_{O_i}^{L_i} m_{O_j}^{L_j}\mathrm{Cov}(\epsilon_{L_i}, \epsilon_{L_j}) \neq 0$. □



(a) Case 1 (b) Case 2 (c) Case 3 (d) Case 4

Figure 7: Illustration of "If" part in proof of Thm. 2. A dotted arrow from V_1 to V_2 means that $V_2 \in \mathrm{De}(V_1)$.

Proposition 2. Suppose O_i satisfies Thm. 2, then $O_i \in \mathrm{HSu}(\mathrm{Pa}_{\mathbf{L}}(O_i)\backslash\mathbf{J})$ if and only if $\forall O_j$ satisfying Thm. 2 and $\mathrm{Pa}_{\mathbf{L}}(O_j)\backslash\mathbf{J} = \mathrm{Pa}_{\mathbf{L}}(O_i)\backslash\mathbf{J}$, $\|\mathbf{M}^{\mathbf{J}}_{\{O_i\}}\|_0 \leq \|\mathbf{M}^{\mathbf{J}}_{\{O_j\}}\|_0$.

Proof. Let $\mathrm{Pa}_{\mathbf{L}}(O_i)\backslash\mathbf{J} = \{L_i\}$.

"Only if". For each O_j satisfying $O_j \in \mathrm{De}(L_i)$, there is $\mathrm{An}(O_j) \cap \mathbf{J} \supset \mathrm{An}(L_i) \cap \mathbf{J}$. Since $O_i \in \mathrm{HSu}(L_i)$, $\mathrm{An}(L_i) \cap \mathbf{J} = \mathrm{An}(O_i) \cap \mathbf{J}$. Therefore, $\|\mathbf{M}^{\mathbf{J}}_{\{O_i\}}\|_0 \leq \|\mathbf{M}^{\mathbf{J}}_{\{O_j\}}\|_0$.

"If". We prove this part by contradiction. Suppose $O_i \notin \mathrm{HSu}(L_i)$. Let $O_j \in \mathrm{HSu}(L_i)$, then $\|\mathbf{M}^{\mathbf{J}}_{\{O_i\}}\|_0 \leq \|\mathbf{M}^{\mathbf{J}}_{\{O_j\}}\|_0$ and $\|\mathbf{M}^{\mathbf{J}}_{\{O_j\}}\|_0 = |\mathrm{An}(L_i) \cap \mathbf{J}|$. Since $O_i \in \mathrm{De}(L_i)$, $\|\mathbf{M}^{\mathbf{J}}_{\{O_i\}}\|_0 \geq |\mathrm{An}(L_i) \cap \mathbf{J}|$. Therefore, $\|\mathbf{M}^{\mathbf{J}}_{\{O_i\}}\|_0 = \|\mathbf{M}^{\mathbf{J}}_{\{O_j\}}\|_0$, that is, $\mathrm{An}(O_i) \cap \mathbf{J} = \mathrm{An}(O_j) \cap \mathbf{J} = \mathrm{An}(L_i)$, note that O_i satisfies Thm. 2, so $O_i \in \mathrm{HSu}(L_i)$, this leads to contradiction. □

Definition 3. (Cumulant) Given n random variables $V_1, ..., V_n$, the k-th order cumulant is defined as a tensor of size $n \times ... \times n$ (k times), whole element at position $(i_1, ..., i_k)$ is

$$\mathrm{Cum}(V_{i_1}, ..., V_{i_k}) = \sum_{\pi} (-1)^{|\pi|-1} (|\pi| - 1)! \prod_{B \in \pi} \mathbb{E}\left[\prod_{j \in B} V_j\right], \tag{31}$$

where π is enumerated over all partitions of $\{i_1, ..., i_k\}$.

Corollary 3. Suppose O_i satisfies Thm. 2 and $O_i \in \mathrm{HSu}(L_i)$, then $\forall O_j \in \mathbf{O}\backslash(\mathbf{K} \cup \{O_i\})$,

$$m^{L_i}_{O_i} m^{L_i}_{O_j} = \mathrm{Cov}(\tilde{O}_i, O_j), \quad \left(\frac{m^{L_i}_{O_i}}{m^{L_i}_{O_j}}\right)^2 = \frac{\mathrm{Cum}(\tilde{O}_i, \tilde{O}_i, \tilde{O}_i, O_j)}{\mathrm{Cum}(\tilde{O}_i, O_j, O_j, O_j)}. \tag{32}$$

Proof. Note that $O_j \notin \mathrm{De}(O_i)$, based on Eqs. (13) and (14),

$$\tilde{O}_i = m^{L_i}_{O_i} \epsilon_{L_i} + \sum_{O_{l_2} \in \mathbf{K}_2} \lambda_{il_2} \epsilon_{O_{l_2}} + \epsilon_{O_i}. \tag{33}$$

$$O_j = \sum_{L_j \in \mathbf{J}} m^{L_j}_{O_j} \epsilon_{L_j} + \sum_{L_k \in \mathbf{L}\backslash\mathbf{J}} m^{L_k}_{O_j} \epsilon_{L_k} + \sum_{O_{l_1} \in \mathbf{K}_1} m^{O_{l_1}}_{O_j} \epsilon_{O_{l_1}} + \sum_{O_m \in \mathbf{O}\backslash(\mathbf{K} \cup \{O_i\})} m^{O_m}_{O_j} \epsilon_{O_m}. \tag{34}$$

Clearly,

$$\mathrm{Cov}(\tilde{O}_i, O_j) = m^{L_i}_{O_i} m^{L_i}_{O_j} \mathrm{Var}(\epsilon_{L_i}) = m^{L_i}_{O_i} m^{L_i}_{O_j}. \tag{35}$$

The second equation holds because we assume the exogenous noise of each latent variable has unit variance.

Because we assume each exogenous noise has zero mean, $\mathbb{E}[\tilde{O}_i] = \mathbb{E}[O_j] = 0$. There is

$$\mathrm{Cum}(\tilde{O}_i, \tilde{O}_i, \tilde{O}_i, O_j) = \mathbb{E}[\tilde{O}_i^3 O_j] - 3\mathbb{E}[\tilde{O}_i O_j]\mathbb{E}[\tilde{O}_i^2]. \tag{36}$$

Let $\tilde{e}_i = \tilde{O}_i - m_{O_i}^{L_i}\epsilon_{L_i}$ and $e_j = O_j - m_{O_j}^{L_i}\epsilon_{L_i}$. It follows that $\epsilon_{L_i}, \tilde{e}_i, e_j$ are independent of each other. Therefore,

$$\begin{aligned}
&\mathrm{Cum}(\tilde{O}_i, \tilde{O}_i, \tilde{O}_i, O_j)\\
=&\underbrace{\mathbb{E}[(m_{O_i}^{L_i}\epsilon_{L_i} + \tilde{e}_i)^3(m_{O_j}^{L_i}\epsilon_{L_i} + e_j)]}_{\mathbb{E}[\tilde{O}_i^3 O_j]} - 3\underbrace{\mathbb{E}[(m_{O_i}^{L_i}\epsilon_{L_i} + \tilde{e}_i)(m_{O_j}^{L_i}\epsilon_{L_i} + e_j)]}_{\mathbb{E}[\tilde{O}_i O_j]}\underbrace{\mathbb{E}[(m_{O_i}^{L_i}\epsilon_{L_i} + \tilde{e}_i)^2]}_{\mathbb{E}[\tilde{O}_i^2]}\\
=&\underbrace{(m_{O_i}^{L_i})^3 m_{O_j}^{L_i}\mathbb{E}[\epsilon_{L_i}^4] + 3m_{O_i}^{L_i}m_{O_j}^{L_i}\mathbb{E}[\epsilon_{L_i}^2\tilde{e}_i^2]}_{\mathbb{E}[(m_{O_i}^{L_i}\epsilon_{L_i}+\tilde{e}_i)^3(m_{O_j}^{L_i}\epsilon_{L_i}+e_j)]} - 3\underbrace{\left((m_{O_i}^{L_i})^3 m_{O_j}^{L_i}(\mathbb{E}[\epsilon_{L_i}^2])^2 + m_{O_i}^{L_i}m_{O_j}^{L_j}\mathbb{E}[\epsilon_{L_i}^2]\mathbb{E}[\tilde{e}_i^2]\right)}_{\mathbb{E}[(m_{O_i}^{L_i}\epsilon_{L_i}+\tilde{e}_i)(m_{O_j}^{L_i}\epsilon_{L_i}+e_j)]\mathbb{E}[(m_{O_i}^{L_i}\epsilon_{L_i}+\tilde{e}_i)^2]}\\
=&(m_{O_i}^{L_i})^3 m_{O_j}^{L_i}(\mathbb{E}[\epsilon_{L_i}^4] - 3(\mathbb{E}[\epsilon_{L_i}^2])^2).
\end{aligned} \tag{37}$$

The second equation holds because for any two mutually independent random variables V_1, V_2 with $\mathbb{E}[V_1] = \mathbb{E}[V_2] = 0$,

$$\mathbb{E}[V_1 V_2] = \mathrm{Cov}(V_1, V_2) + \mathbb{E}[V_1]\mathbb{E}[V_2] = 0, \tag{38}$$

$$\mathbb{E}[V_1^3 V_2] = \mathrm{Cov}(V_1^3, V_2) + \mathbb{E}[V_1^3]\mathbb{E}[V_2] = 0, \tag{39}$$

where V_1, V_2 can refer to any two of $\tilde{e}_i, e_j, \epsilon_{L_i}$.

The third equation holds because for any two mutually independent random variables V_1, V_2 with $\mathbb{E}[V_1] = \mathbb{E}[V_2] = 0$,

$$\mathbb{E}[V_1^2 V_2^2] - \mathbb{E}[V_1^2]\mathbb{E}[V_2^2] = \mathrm{Cov}(V_1^2, V_2^2) = 0, \tag{40}$$

where V_1, V_2 refer to $\epsilon_{L_i}, \tilde{e}_i$ respectively.

Similarly,

$$\mathrm{Cum}(\tilde{O}_i, O_j, O_j, O_j) = m_{O_i}^{L_i}(m_{O_j}^{L_i})^3(\mathbb{E}[\epsilon_{L_i}^4] - 3(\mathbb{E}[\epsilon_{L_i}^2])^2). \tag{41}$$

Therefore,

$$\left(\frac{m_{O_i}^{L_i}}{m_{O_j}^{L_i}}\right)^2 = \frac{\mathrm{Cum}(\tilde{O}_i, \tilde{O}_i, \tilde{O}_i, O_j)}{\mathrm{Cum}(\tilde{O}_i, O_j, O_j, O_j)}. \tag{42}$$

□

Corollary 4. Suppose O_i satisfies Thm. 2 and $O_i \in \mathrm{HSu}(L_i)$, if we update $\mathbf{J}$ to $\mathbf{J} \cup \{L_i\}$, $\mathbf{K}$ to $\mathbf{K} \cup \{O_i\}$, and $\tilde{O}_j$ to $\tilde{O}_j - (m_{O_j}^{L_i}/m_{O_i}^{L_i})\tilde{O}_i$ for each $O_j \in \mathbf{O}\backslash\mathbf{K}$, Cond. 1 is still valid.

Proof. Based on Thm. 2, it is trivial that Cond. 1(1) is valid.

Based on Eq. (14), before removal

$$\tilde{O}_i = m_{O_i}^{L_i}\epsilon_{L_i} + \sum_{O_k \in \mathbf{K}_2}\lambda_{ik}\epsilon_{O_k} + \epsilon_{O_i}. \tag{43}$$

$$\tilde{O}_j = \sum_{L_j \in \mathbf{L}\backslash\mathbf{J}} m_{O_j}^{L_j}\epsilon_{L_j} + \sum_{O_k \in \mathbf{K}_2}\lambda_{jk}\epsilon_{O_k} + \sum_{O_l \in \mathbf{O}\backslash(\mathbf{K}\cup\{O_i\})} m_{O_j}^{O_l}\epsilon_{O_l}, \tag{44}$$

Let $\lambda'_{jk} = \lambda_{jk} - \frac{m_{O_j}^{L_i}}{m_{O_i}^{L_i}}\lambda_{ik}$ for each $O_k \in \mathbf{K}_2$ and $\lambda'_{ji} = -\frac{m_{O_j}^{L_i}}{m_{O_i}^{L_i}}$, then

$$\tilde{O}_j - \frac{m_{O_j}^{L_i}}{m_{O_i}^{L_i}}\tilde{O}_i = \sum_{L_j \in \mathbf{L}\backslash(\mathbf{J}\cup\{L_i\})} m_{O_j}^{L_j}\epsilon_{L_j} + \sum_{O_k \in \mathbf{K}_2\cup\{O_i\}}\lambda'_{jk}\epsilon_{O_k} + \sum_{O_l \in \mathbf{O}\backslash(\mathbf{K}\cup\{O_i\})} m_{O_j}^{O_l}\epsilon_{O_l}, \tag{45}$$

so Cond. 1(2) is also valid. □

Theorem 3. Suppose the observed variables are generated by a LiNGAM with latent variables satisfying Asmps. 1 and 2, in the limit of infinite data, Alg. 1 identifies latent variables and ancestral relationships correctly.

Proof. Based on Thm. 1, Cors. 1 and 2, Alg. 1 correctly estimates $\mathbf{M}_{\mathbf{O}}^{\mathbf{O}}$. Based on Thm. 2, Props. 1 and 2, Cors. 3 and 4, Alg. 1 correctly identifies latent variables by locating their respective homologous surrogates and estimates $\mathbf{M}_{\mathbf{O}}^{\mathbf{L}}$. Furthermore, given any two latent variable L_i, L_j, Alg. 1 correctly determines whether L_i is an ancestor of L_j by checking whether L_i is an ancestor of L_j's homologous surrogates. □

C.3 Proof of Theoretical Results in Sec. 4.

Definition 4. (Generalized homologous surrogate) $O \in \mathbf{O}$ is called a generalized homologous surrogate of $L \in \mathbf{L}$, denoted by $O \in \mathrm{GHSu}(L)$, if $O \in \mathrm{Ch}(L)$ and $\mathrm{Pa}_{\mathbf{L}}(O) \subset \mathrm{An}(L) \cup \{L\}$.

Assumption 3. Asmp. 2 holds and $\forall \{L_i, L_j\} \subset \mathbf{L}$ where $L_i \in \mathrm{An}(L_j)$, $\exists \{O_{j_1}, O_{j_2}\} \subset \mathrm{GHSu}(L_j)$ s.t $O_{j_1} \notin \mathrm{Ch}(L_i)$ and $O_{j_2} \notin \mathrm{Ch}(L_i)$.

As mentioned in the main text, we can remove causal relations between observed variables given $\mathbf{A}_{\mathbf{O}}^{\mathbf{O}}$. Specifically, for each $O_i \in \mathbf{O}$, we let

$$O_i^* = O_i - \sum_{O_j \in \mathrm{Pa}(O_i)} a_{O_i}^{O_j} O_j. \tag{46}$$

We denote by $\mathcal{G}^*$ the causal graph among $\mathbf{L} \cup \mathbf{O}^*$ where $\mathbf{O}^* = \{O_i^*\}_i$. Also, we use $\mathrm{Pa}^*(V)$ to denote V's parents in $\mathcal{G}^*$.

Lemma 6. *Given $L \in \mathbf{L}$ and $O \in \mathbf{O}$, $O \in \mathrm{GHSu}(L)$ if and only if* $\mathrm{An}^*(O^*) = \mathrm{An}^*(L) \cup \{L\}$.

Proof. This can be readily derived from the definition of generalized homologous surrogates. □

Lemma 1. $\forall L_i \in \mathbf{L}$ and $O_i \in \mathbf{O}$, $O_i \in \mathrm{GHSu}(L_i)$ if and only if $m_{O_i^*}^{L_i} \neq 0$ and $\forall O_j \in \mathbf{O}$ where $m_{O_j^*}^{L_i} \neq 0$, $\|\mathbf{M}_{\{O_i^*\}}^{\mathbf{L}}\|_0 \leq \|\mathbf{M}_{\{O_j^*\}}^{\mathbf{L}}\|_0$. Besides, there is $a_{O_i}^{L_i} = m_{O_i^*}^{L_i}$.

Proof. "Only if". For each O_j^* where $m_{O_j^*}^{L_i} \neq 0$, there is $\mathrm{An}^*(O_j^*) \supset \mathrm{An}^*(L_i) \cup \{L_i\}$. Based on Lem. 6, $\mathrm{An}^*(O_i^*) = \mathrm{An}^*(L_i) \cup \{L_i\}$, so $\|\mathbf{M}_{\{O_i^*\}}^{\mathbf{L}}\|_0 \leq \|\mathbf{M}_{\{O_j^*\}}^{\mathbf{L}}\|_0$.

"If". We prove this part by contradiction. Suppose $O_i \notin \mathrm{GHSu}(L_i)$. Let $O_j \in \mathrm{GHSu}(L_i)$, then $\|\mathbf{M}_{\{O_i^*\}}^{\mathbf{L}}\|_0 \leq \|\mathbf{M}_{\{O_j^*\}}^{\mathbf{L}}\|_0$ and $\|\mathbf{M}_{\{O_j^*\}}^{\mathbf{L}}\|_0 = |\mathrm{An}^*(L_i) \cup \{L_i\}|$ based on Lem. 6. Since $m_{O_i^*}^{L_i} \neq 0$, $O_i^* \in \mathrm{De}^*(L_i)$, that is, $\|\mathbf{M}_{\{O_i^*\}}^{\mathbf{L}}\|_0 \geq |\mathrm{An}^*(L_i) \cup \{L_i\}|$. Therefore, $\|\mathbf{M}_{\{O_i^*\}}^{\mathbf{L}}\|_0 = \|\mathbf{M}_{\{O_j^*\}}^{\mathbf{L}}\|_0$, that is, $\mathrm{An}^*(O_i^*) = \mathrm{An}^*(L_i) \cup \{L_i\}$, based on Lem. 6, this leads to contradiction.

Finally, based on Lem. 6, it is trivial that if $O_i \in \mathrm{GHSu}(L_i)$, $a_{O_i}^{L_i} = m_{O_i^*}^{L_i}$ because there is only one directed path from L_i to O_i^* in $\mathcal{G}^*$, which is exactly $L_i \to O_i^*$. □

Theorem 4. Suppose $\{L_i, L_j\} \subset \mathbf{L}$, $L_j \in \mathrm{De}^n(L_i)$. $\forall O_j \in \mathrm{GHSu}(L_j)$, let

$$\mu_{O_j^*}^{L_i} = m_{O_j^*}^{L_i} - \sum_{L_k \in \mathrm{De}(L_i) \cap \mathrm{An}(L_j)} m_{L_k}^{L_i} a_{O_j}^{L_k}. \tag{47}$$

(a) There exists $\{O_{j_1}, O_{j_2}\} \subset \mathrm{GHSu}(L_j)$ s.t. $\mu_{O_{j_1}^*}^{L_i} / a_{O_{j_1}}^{L_j} = \mu_{O_{j_2}^*}^{L_i} / a_{O_{j_2}}^{L_j}$ and $m_{L_j}^{L_i} = \mu_{O_{j_1}^*}^{L_i} / a_{O_{j_1}}^{L_j}$.

(b) $a_{O_j}^{L_i} = \mu_{O_j^*}^{L_i} - m_{L_j}^{L_i} a_{O_j}^{L_j}$.

Proof sketch. We can derive that $\mu_{O_j^*}^{L_i} = a_{O_j}^{L_i} + m_{L_j}^{L_i} a_{O_j}^{L_j}$, so (b) holds naturally. For (a), based on the rank-faithfulness, $\mu_{O_{j_1}^*}^{L_i} / a_{O_{j_1}}^{L_j} = \mu_{O_{j_2}^*}^{L_i} / a_{O_{j_2}}^{L_j}$ if and only if $a_{O_{j_1}}^{L_i} = a_{O_{j_2}}^{L_i} = 0$, that is, $O_{j_1} \notin \mathrm{Ch}(L_j)$ and $O_{j_2} \notin \mathrm{Ch}(L_j)$. In this case, $\mu_{O_{j_1}^*}^{L_i} / a_{O_{j_1}}^{L_j} = m_{L_j}^{L_i}$ trivially.

Proof. For any $L \in \mathbf{L}$ and $O^* \in \mathbf{O}^*$, there is

$$m_{O^*}^{L} = a_{O}^{L} + \sum_{L_i \in \mathrm{De}^*(L) \cap \mathrm{Pa}^*(O^*)} m_{L_i}^{L} a_{O}^{L_i}. \tag{48}$$

Given $O_j \in \mathrm{GHSu}(L_j)$, based on Lem. 6, $\mathrm{Pa}^*(O_j^*) \subset \mathrm{An}^*(O_j^*) = \mathrm{An}^*(L_j) \cup \{L_j\}$, so $\big(\mathrm{De}^*(L_i) \cap \mathrm{Pa}^*(O_j^*)\big) \backslash \{L_j\} \subset \mathrm{De}^*(L_i) \cap \mathrm{An}^*(L_j) = \mathrm{De}(L_i) \cap \mathrm{An}(L_j)$. We can rewrite Eq. (48) as

$$m_{O_j^*}^{L_i} = a_{O_j}^{L_i} + \sum_{L_k \in \mathrm{De}(L_i) \cap \mathrm{An}(L_j)} m_{L_k}^{L_i} a_{O_j}^{L_k} + m_{L_j}^{L_i} a_{O_j}^{L_j} \tag{49}$$

because $m_{L_k}^{L_i} a_{O_j}^{L_k} = 0$ if $L_k \notin \mathrm{De}^*(L_i) \cap \mathrm{Pa}^*(O_j^*)$. Therefore,

$$\mu_{O_j^*}^{L_i} = m_{O_j^*}^{L_i} - \sum_{L_k \in \mathrm{De}(L_i) \cap \mathrm{An}(L_j)} m_{L_k}^{L_i} a_{O_j}^{L_k} = a_{O_j}^{L_i} + m_{L_j}^{L_i} a_{O_j}^{L_j}. \tag{50}$$

Let $\{O_{j_1}, O_{j_2}\} \subset \mathrm{GHSu}(L_j)$. On the one hand, if $O_{j_1} \notin \mathrm{Ch}(L_i)$ and $O_{j_2} \notin \mathrm{Ch}(L_i)$, then $a_{O_{j_1}}^{L_i} = a_{O_{j_2}}^{L_i} = 0$, so $\mu_{O_{j_1}^*}^{L_i} / a_{O_{j_1}}^{L_j} = \mu_{O_{j_2}^*}^{L_i} / a_{O_{j_2}}^{L_j} = m_{L_j}^{L_i}$. On the other hand, if $O_{j_1} \in \mathrm{Ch}(L_i)$ or $O_{j_2} \in \mathrm{Ch}(L_i)$, based on the rank-faithfulness assumption, $\mu_{O_{j_1}^*}^{L_i} / a_{O_{j_1}}^{L_j} \neq \mu_{O_{j_2}^*}^{L_i} / a_{O_{j_2}}^{L_j}$. Therefore, (a) holds.

Besides, based on Eq. (50), it is trivial that (b) holds. □

Theorem 5. Suppose the observed variables are generated by a LiNGAM with latent variables satisfying Asmps. 1 and 3, in the limit of infinite data, Algs. 1 and 2 together identifies latent variables and parental relationships correctly.

Proof. Based on Thm. 3, Alg. 1 correctly identifies latent variables and estimates $\mathbf{M}_{\mathbf{L}}^{\mathbf{O}}$ and $\mathbf{M}_{\mathbf{O}}^{\mathbf{O}}$. Based on Thm. 4, Alg. 2 correctly estimates $\mathbf{M}_{\mathbf{L}}^{\mathbf{L}}$. Therefore, $\mathbf{M}$ is estimated correctly, from which $\mathbf{A}$ can be derived based on Eq. (3). □

D Experiment on Real-World Data

The Holzinger and Swineford 1939 dataset consists of mental ability test scores of seventh- and eighth-grade children from two different schools (Pasteur and Grant-White). There are 9 variables, which can be categorized into three dimensions: Visual (O_1, O_2, O_3), Textual (O_4, O_5, O_6), and Speeded (O_6, O_7, O_8). The result returned by our algorithm is shown as Fig. 8. Our algorithm correctly identifies the textual factor while merges the visual factor and the speed factor into a single factor. This can be attributed to the fact that both the visual factor and speed factor depends on innate abilities, while the textual factor highly depends on learning experience.

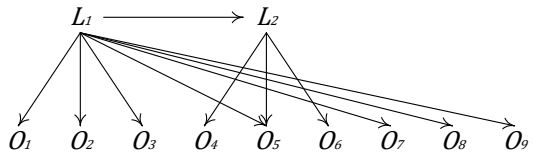


Figure 8: The output of our algorithm on the Holzinger and Swineford dataset.