On the Parameter Identifiability of Partially Observed Linear Causal Models

Xinshuai Dong[1]* **Ignavier Ng**[1]* **Biwei Huang**[2] **Yuewen Sun**[3] **Songyao Jin**[3]
Roberto Legaspi[4] **Peter Spirtes**[1] **Kun Zhang**[1,3]
[1]Carnegie Mellon University [2]University of California San Diego
[3]Mohamed bin Zayed University of Artificial Intelligence [4]KDDI Research

Abstract

Linear causal models are important tools for modeling causal dependencies and yet in practice, only a subset of the variables can be observed. In this paper, we examine the parameter identifiability of these models by investigating whether the edge coefficients can be recovered given the causal structure and partially observed data. Our setting is more general than that of prior research—we allow all variables, including both observed and latent ones, to be flexibly related, and we consider the coefficients of all edges, whereas most existing works focus only on the edges between observed variables. Theoretically, we identify three types of indeterminacy for the parameters in partially observed linear causal models. We then provide graphical conditions that are sufficient for all parameters to be identifiable and show that some of them are provably necessary. Methodologically, we propose a novel likelihood-based parameter estimation method that addresses the variance indeterminacy in a specific way and can asymptotically recover the underlying parameters up to trivial indeterminacy. Empirical studies on both synthetic and real-world datasets validate our identifiability theory and the effectiveness of the proposed method in the finite-sample regime.

1 Introduction and Related Work

Causal models, which serve as a fundamental tool to capture causal relations among random variables, have achieved great success in many fields [49, 39, 40, 44]. A fundamental problem in the field is how and to what extent can we identify the underlying causal model given observational data. When all variables are observed, the problem has been well studied: the underlying structure can be identified up to the Markov equivalence class, e.g., by the PC [49] or GES [13] algorithm; when the structure is given, the causal coefficient (direct causal effect) between two variables can also be identified [8, 39].

However, in real-world systems, the variables of interest may only be partially observed. Thus, considerable efforts have been dedicated to identification of causal models in the presence of latent variables. One line of research focuses on structure learning given partially observed variables. Notable approaches include FCI and its variants [49, 38, 14, 2], as well as ICA-based [23, 43], tetrad-based [48, 28], high-order moments-based [46, 11, 58, 1, 12], and rank constraint-based [48, 24, 18] methods.

In this paper, we focus on the the identification of parameters of a partially observed model. Specifically, given the causal structure of and observational data from a partially observed causal model, we are interested in identifying all the parameters, and thus the underlying causal model can be fully specified. To identify the parameters, a classical way is to project the directed acyclic graph (DAG) with latent variables to an acyclic directed mixed graph (ADMG) or partially ancestral graph [42], without explicitly modeling the latent confounders. Based on ADMG, graphical criteria such as half-trek [20], G-criterion [9], and some further developments [51, 29] have been proposed to establish the

*Equal contribution.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

parameter identifiability. Another way is to leverage do-calculus, proxy variables, and instrumental variables [47, 39, 25] to identify the direct causal effect, which corresponds to the edge coefficient in linear causal models. For a more detailed discussion of related work, please refer to Appendix D.

Despite the effectiveness of current methods for parameter identification, however, they have two main drawbacks: they require all the variables to be connected in specific ways, and only focus on identifying the edge coefficients between observed variables. To this end, in this paper we propose a novel framework that considers a more general setting for parameter identification. To be specific, we allow all variables, including both observed and latent ones, to be flexibly related, and we aim to recover the edge coefficients among all variables, even including those from a latent variable to another latent variable or another observed variable, which previous methods cannot handle. We summarize our contributions as follows.

- To the best of our knowledge, we are the first to consider parameter identifiability of partially observed causal model in the most general scenario—all variables, including both observed and latent ones, are allowed be flexibly related, and edge coefficients between any pair of variables are concerned. In contrast, most existing works consider only the edges between observed variables.
- Theoretically, we identify three types of parameter indeterminacy in partially observed linear causal models. We then provide graphical conditions that are sufficient for all parameters to be identifiable and show that some of them are provably necessary. These necessary conditions also offer insights into scenarios where the parameters are guaranteed to be non-identifiable.
- Methodologically, we propose a novel likelihood-based parameter estimation method, which parameterizes population covariance in specific ways to address variance indeterminacy. Our empirical studies on both synthetic and real-world data validate the effectiveness of our method in the finite-sample regime, even under certain misspecification of the underlying causal model.

2 Preliminaries

2.1 Problem Setting

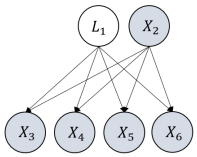
In this work, we focus on partially observed linear causal models, defined as follows.

Definition 1 (Partially Observed Linear Causal Models)**.** *Let $\mathcal{G} := (\mathbf{V}_\mathcal{G}, \mathbf{E}_\mathcal{G})$ be a DAG. Each variable $V_i \in \mathbf{V}_\mathcal{G}$ follows a linear structural equation model $\mathsf{V}_i = \sum_{\mathsf{V}_j \in Pa_\mathcal{G}(\mathsf{V}_i)} f_{j,i}\mathsf{V}_j + \epsilon_{\mathsf{V}_i}$, where $\mathbf{V}_\mathcal{G} := \mathbf{L}_\mathcal{G} \cup \mathbf{X}_\mathcal{G} = \{\mathsf{L}_i\}_{i=1}^m \cup \{\mathsf{X}_i\}_{i=m+1}^{m+n}$ contains m latent variables and n observed variables. $Pa_\mathcal{G}(\mathsf{V}_i)$ denotes the parent set of V_i, $f_{j,i}$ denotes the edge coefficient from V_j to V_i, and ϵ_{V_i} represents the Gaussian noise term of V_i.*

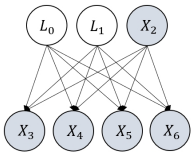
We drop the subscript $\mathcal{G}$ in $\mathbf{L}_\mathcal{G}$ and $\mathbf{X}_\mathcal{G}$ when the context is clear. We use V, $\mathbf{V}$, and $\mathcal{V}$ to denote a random variable, a set of variables, and a set of sets of variables, respectively. In Definition 1, the relations between variables can also be written in the matrix form as $\mathbf{V}_\mathcal{G} = F^T\mathbf{V}_\mathcal{G} + \epsilon_{\mathbf{V}_\mathcal{G}}$, where $F = (f_{j,i})_{i,j\in[m+n]}$ is the weighted adjacency matrix. Here, $f_{j,i} \neq 0$ if and only if V_j is a parent of V_i in $\mathcal{G}$. We also write

$$F = \begin{matrix} \\ \mathbf{L}_\mathcal{G} \\ \mathbf{X}_\mathcal{G} \end{matrix} \begin{matrix} \begin{matrix} \mathbf{L}_\mathcal{G} & \mathbf{X}_\mathcal{G} \end{matrix} \\ \begin{pmatrix} A & B \\ C & D \end{pmatrix} \end{matrix} \quad \text{and} \quad \Omega = \begin{pmatrix} \Omega_{\mathbf{L}_\mathcal{G}} & 0 \\ 0 & \Omega_{\mathbf{X}_\mathcal{G}} \end{pmatrix},$$

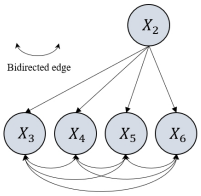
where Ω is the diagonal covariance matrix of $\epsilon_{\mathbf{V}_\mathcal{G}}$. Our objective is to identify F, the causal edge coefficients of the model, given observational data and the causal structure $\mathcal{G}$. Denote by $\Sigma_{\mathbf{L}_\mathcal{G}}$ and $\Sigma_{\mathbf{X}_\mathcal{G}}$ the population covariance matrix of latent variables $\mathbf{L}_\mathcal{G}$ and observed variables $\mathbf{X}_\mathcal{G}$, respectively; their precise formulations are provided in Proposition 1. We also denote by $\sigma_{i,j}$ the (i,j)-th entry of $\Sigma_{\mathbf{X}_\mathcal{G}}$. In this work, we assume that the noise variances $\epsilon_{\mathbf{L}_\mathcal{G}}$ of latent variables have unit variance, i.e., $\Omega_{\mathbf{L}_\mathcal{G}} = I$, which will be justified later in Section 3.1. Note that variables are partially observed and thus we only have access to i.i.d. samples of observed variables $\mathbf{X}_\mathcal{G}$. As variables are jointly Gaussian, the observations can asymptotically be summarized as the population covariance matrix $\Sigma_{\mathbf{X}_\mathcal{G}}$. In other words, we aim to identify F and Ω given $\Sigma_{\mathbf{X}_\mathcal{G}}$ and $\mathcal{G}$. The identification of parameters is important in that, once we identify the parameters, the underlying causal model is fully specified, and thus we can flexibly calculate causal effects, infer interventional distributions, and finally answer counterfactual queries [39]. It is worth noting that, for parameter identification, the structure $\mathcal{G}$ is assumed to be known, which is different from the setting of causal discovery where the goal is to identify $\mathcal{G}$ from data.



(a) Graph $\mathcal{G}_1$. Its parameters can be fully identifiable.



(b) Graph $\mathcal{G}_2$. Its parameters cannot be fully identified.



(c) $\mathcal{G}_1$ $\mathcal{G}_2$ have the same ADMG in the latent projection framework.

Figure 1: Illustrations of the advantage of our framework. Within our framework, it can be shown that $\mathcal{G}_1$'s parameters can be identified (up to sign) while $\mathcal{G}_2$'s cannot. In contrast, the latent projection framework cannot even differentiate $\mathcal{G}_1$ from $\mathcal{G}_2$ as they share the same ADMG (c) after projection. Furthermore, with ADMG, any edge coefficient that involves a latent variable cannot be considered.

2.2 Framework Comparison

Without latent variables, it has been shown all parameters are identifiable [8]. However, the problem becomes very challenging when latent variables exist. There are two lines of research. One focuses on the use of do-calculus, proxy variables, and instrumental variables to identify direct causal effects among observed variables [47, 39, 25] (in linear models the direct causal effect is captured by the edge coefficient). Another line addresses latent confounders by projecting a DAG with latent variables into an ADMG, where the confounding effects of latent variables are simplified and represented by correlation among noise terms [20, 9, 51, 29]. An example is in Figure 1, where (a) is the original graph and (c) is the projected ADMG whose bidirected edges correspond to correlated noise terms.

Compared to the two previous lines of thought, our framework has two advantages. To begin with, we additionally considers the identifiability of coefficients of edges that involve latent variables. For example, in Figure 1, we aim to identify all the coefficients including the one from L_1 to X_3, i.e., $f_{1,3}$. In contrast, the proxy variable framework and the latent projection framework identify only the coefficients among observed variables: the proxy variable framework focuses only on the direct causal effect from one observed variable to another observed variable, while the latent projection framework transforms all latent variables into bidirected edges and thus can never identify the coefficient of the edge that has a latent variable as the head or tail.

Furthermore, the projection framework deals with latent variables in a rather brute-force way: dense latent confounding effects among observed variables may be caused by only a small number of latent variables, but that information is lost during projection. For example, in Figure 1, (a) and (b) share the same ADMG after projection, i.e., (c). However, as we will show later, parameters in (a) can be identified, while in (b) the parameters cannot. If we only consider the ADMG in (c), then we can never capture this nuance and thus cannot identify the coefficients that we might be able to.

3 Identifiability Theory

3.1 Definition of Parameter Identifiability and Indeterminacy

We follow the notion of generic identifiability and define parameter identifiability as follows.

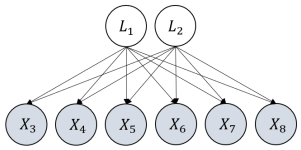
Definition 2 (Identifiability of Parameters of Partially Observed Linear Causal Models)**.** *Let $\theta = (F, \Omega) \in \Theta$. We say that θ is generically identifiable, if the mapping $\phi(\theta) = \Sigma_{\mathbf{X}_\mathcal{G}}$ is injective, for almost all $\theta \in \Theta$ with respect to the Lebesgue measure.*

Definition 2 indicates if parameter θ is identifiable, then there does not exist $\theta' \in \Theta$ that entails the same observations as those of θ. As in the typical literature of parameter identification, we consider generic identifiability to rule out some rare cases where the parameters for that structure is generally identifiable, but with some specific parameterization, the parameters cannot be identified. This is similar to faithfulness in causal discovery [49] and we will provide an example in Example 1. We next introduce three important indeterminacies about parameter identification when latent variables exist.

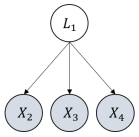
Theorem 1 (Indeterminacy of Scaling of $\Omega_{\mathbf{L}_\mathcal{G}}$)**.** *Consider a model that follows Definition 1 with number of latent variables $m \geq 1$ and $\theta = (A, B, C, D, \Omega_{\mathbf{X}_\mathcal{G}}, \Omega_{\mathbf{L}_\mathcal{G}})$. Let Λ be any invertible diagonal matrix, and $\tilde{\theta} = (\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{\Omega}_{\mathbf{X}_\mathcal{G}}, \tilde{\Omega}_{\mathbf{L}_\mathcal{G}})$, where*

$$\tilde{A} \coloneqq \Lambda^{-1} A \Lambda, \quad \tilde{B} \coloneqq \Lambda^{-1} B, \quad \tilde{C} \coloneqq C\Lambda, \quad \tilde{D} \coloneqq D, \quad \tilde{\Omega}_{\mathbf{L}_\mathcal{G}} \coloneqq \Lambda^2 \Omega_{\mathbf{L}_\mathcal{G}}, \quad \textit{and} \quad \tilde{\Omega}_{\mathbf{X}_\mathcal{G}} \coloneqq \Omega_{\mathbf{X}_\mathcal{G}}.$$

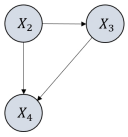
Then, $\tilde{\theta}$ and θ entail the same observations, i.e., $\tilde{\Sigma}_{\mathbf{X}_\mathcal{G}} = \Sigma_{\mathbf{X}_\mathcal{G}}$. Furthermore, we have $\tilde{\Sigma}_{\mathbf{L}_\mathcal{G}} = \Lambda \Sigma_{\mathbf{L}_\mathcal{G}} \Lambda$.



(a) $\mathcal{G}_1$. Its structure is identifiable but its parameters are not identifiable even if the structure is given (due to orthogonal indeterminacy).



(b) $\mathcal{G}_2$. Its structure is not identifiable but params are identifiable.



(c) $\mathcal{G}_3$. $\mathcal{G}_2$'s structure is not identifiable due to the existence of $\mathcal{G}_3$.

Figure 2: Illustrative examples to show that the graphical condition for structure-identifiability and parameter-identifiability could be very different.

A similar theoretical result is provided in [4], and yet our setting is much more general and takes that of [4] as a special case: in our setting, all variables including latent and observed ones can be arbitrarily related while in [4] latent variables cannot be the effect of observed variables.

Remark 1 (Implication of Theorem 1)**.** *A key implication of Theorem 1 is that, without further assumption, the edge coefficients involving latent variables, i.e.,* (A, B, C)*, can never be identified, as there always exists a diagonal matrix* Λ *such that* $\tilde{\theta}$ *and* θ *entail the same observations but* $(\tilde{A}, \tilde{B}, \tilde{C}) \neq (A, B, C)$*. Thus, in the rest of this paper, we assume that the noise variances* $\epsilon_{\mathbf{L}_\mathcal{G}}$ *of latent variables have unit variance, i.e.,* $\Omega_{\mathbf{L}_\mathcal{G}} = I$*. Under this assumption, we have* $(\tilde{\Omega}_{\mathbf{L}_\mathcal{G}})_{i,i} = \Lambda_{i,i}^2 (\Omega_{\mathbf{L}_\mathcal{G}})_{i,i} = 1, i \in [m]$*, which implies* $\Lambda_{i,i} = \pm 1$*. As such, this assumption makes parameter identifiability possible. However, even though we fix the scaling of* $\Omega_{\mathbf{L}_\mathcal{G}}$*, there still exists indeterminacy about the sign of parameters, captured by Theorem 2.*

Theorem 2 (Group Sign Indeterminacy)**.** *Consider a model that follows Definition 1 with number of latent variables* $m \geq 1$*,* $\theta = (A, B, C, D, \Omega_{\mathbf{X}_\mathcal{G}}, \Omega_{\mathbf{L}_\mathcal{G}})$*, and* $\Omega_{\mathbf{L}_\mathcal{G}} = I$*. Let* S *be a diagonal sign matrix where the diagonal entries are either* 1 *or* -1*, and* $\tilde{\theta} = (\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{\Omega}_{\mathbf{X}_\mathcal{G}}, \tilde{\Omega}_{\mathbf{L}_\mathcal{G}})$*, where*

$$\tilde{A} \coloneqq SAS, \quad \tilde{B} \coloneqq SB, \quad \tilde{C} \coloneqq CS, \quad \tilde{D} \coloneqq D, \quad \tilde{\Omega}_{\mathbf{L}_\mathcal{G}} \coloneqq \Omega_{\mathbf{L}_\mathcal{G}} = I, \quad \textit{and} \quad \tilde{\Omega}_{\mathbf{X}_\mathcal{G}} \coloneqq \Omega_{\mathbf{X}_\mathcal{G}}.$$

Then, $\tilde{\theta}$ *and* θ *entail the same observations, i.e.,* $\tilde{\Sigma}_{\mathbf{X}_\mathcal{G}} = \Sigma_{\mathbf{X}_\mathcal{G}}$*, and* $(\tilde{\Sigma}_{\mathbf{L}_\mathcal{G}})_{ii} = (\Sigma_{\mathbf{L}_\mathcal{G}})_{ii}, \ \forall i \in [m]$*.*

Remark 2 (Remark on Theorem 2)**.** *The indeterminacy described in Theorem 2 is referred to as group sign indeterminacy for the following reason: According to the theorem, flipping the sign of* $S_{i,i}$ *is equivalent to flipping the signs of all coefficients of edges involving the latent variable* L_i*. This transformation preserves the resulting observations* $\Sigma_{\mathbf{X}_\mathcal{G}}$*. In essence, each group consists of coefficients of edges involving a particular latent variable.*

Example 1 (Example for Group Sign Indeterminacy and Generic Identifiability)**.** *In Figure 2 (b), given the structure and* $\Sigma_{\mathbf{X}_\mathcal{G}}$*, by assuming* $\Omega_{\mathbf{L}_\mathcal{G}} = I$*, the parameters are generally identifiable up to group sign indeterminacy. Specifically, there exist three equality constraints with three free parameters:* $f_{1,2} f_{1,3} = \sigma_{2,3}$*,* $f_{1,2} f_{1,4} = \sigma_{2,4}$*, and* $f_{1,3} f_{1,4} = \sigma_{3,4}$*. The solutions are: (i)* $f_{1,2} = \sqrt{\frac{\sigma_{2,3}\sigma_{2,4}}{\sigma_{3,4}}}$*,* $f_{1,3} = \sigma_{2,3}/f_{1,2}$*,* $f_{1,4} = \sigma_{2,4}/f_{1,2}$ *and (ii)* $f_{1,2} = -\sqrt{\frac{\sigma_{2,3}\sigma_{2,4}}{\sigma_{3,4}}}$*,* $f_{1,3} = -\sigma_{2,3}/f_{1,2}$*,* $f_{1,4} = -\sigma_{2,4}/f_{1,2}$*. The two solutions are different only in terms of group sign. However, if we set* $f_{1,2} = 0$*, then the parameters are not identifiable (as we will encounter division where the divisor is zero). These rare cases of parameters are of zero Lebesgue measure so we rule out these cases for the definition of identifiability, as in Definition 2.*

Intuitively speaking, group sign indeterminacy arises because one may multiply the latent variable L_i by -1 and accordingly flip the signs of all edge coefficients involving L_i. Note that such an indeterminacy is rather minor for the following reason. (i) In practice, we can always anchor the sign of some edges according to our preference or prior knowledge in order to eliminate the group sign indeterminacy. For example, in Figure 4, if we expect that L2 should be understood as Extraversion instead of non-Exterversion, we can add one additional constraint during our parameter estimation such that the edge coefficient from L2 to E1 ("I am the life of party.") will be positive (as we believe E1 should be positively related to Extraversion). (ii) On the other hand, there are some application scenarios that are not influenced by the group sign indeterminacy, such as causal effect estimations between certain variables. We note that, as the indeterminacy of group sign is rather minor, in the following if the parameters are identifiable only up to group sign indeterminacy, we still say that the parameters are identifiable.

Definition 3 (Orthogonal Transformation Indeterminacy)**.** *Consider a model that follows Definition 1 with number of latent variables $m \geq 1$, $\theta = (A, B, C, D, \Omega_{\mathbf{X}_\mathcal{G}}, \Omega_{\mathbf{L}_\mathcal{G}})$, and $\Omega_{\mathbf{L}_\mathcal{G}} = I$. We say that there exists an orthogonal transformation indeterminacy in the identification of parameters if there exists a non-diagonal orthogonal matrix Q such that $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{\Omega}_{\mathbf{X}_\mathcal{G}}, \tilde{\Omega}_{\mathbf{L}_\mathcal{G}})$ and $(A, B, C, D, \Omega_{\mathbf{X}_\mathcal{G}}, \Omega_{\mathbf{L}_\mathcal{G}})$ share the same support and entail the same observations, where*

$$\tilde{A} \coloneqq Q^T A Q, \quad \tilde{B} \coloneqq Q^T B, \quad \tilde{C} \coloneqq CQ, \quad \tilde{D} \coloneqq D, \quad \tilde{\Omega}_{\mathbf{L}_\mathcal{G}} \coloneqq Q^T \Omega_{\mathbf{L}_\mathcal{G}} Q = I, \quad \textit{and} \quad \tilde{\Omega}_{\mathbf{X}_\mathcal{G}} \coloneqq \Omega_{\mathbf{X}_\mathcal{G}}.$$

The orthogonal transformation indeterminacy is the major indeterminacy we consider in the presence of latent variables. Such an indeterminacy also arises in factor analysis [45, 7], which can be viewed as a special case of the data generating procedure considered in Definition 1. Here we only give the definition and will later provide Theorem 4 with an example that captures the scenarios where such indeterminacy exists.

It is worth noting that the graphical condition for structure identifiability and parameter identifiability could be very different. For example, $\mathcal{G}_1$ in Figure 2 (a) is structure-identifiable, and yet the parameters are not identifiable even if the structure is given. In contrast $\mathcal{G}_2$ in Figure 2 (b) is not structure-identifiable, as there exists another structure $\mathcal{G}_3$ in Figure 2 (c) such that $\mathcal{G}_2$ and $\mathcal{G}_3$ can never be differentiated from observational distribution; and yet if $\mathcal{G}_2$ is given, its parameters are identifiable (as in Example 1). Therefore, in this paper, we first consider the cases where the structure can be identified and then study which further conditions are needed for the identifiability of parameters. This will give rise to conditions under which the whole causal model can be fully specified.

3.2 Graphical Condition for Structure Identifiability

To explore the conditions for the whole causal model to be specified, we start with the structure identifiability of partially observed linear causal models. Recent advances have shown that if certain graphical conditions are satisfied [24, 18], even though all variables including latent ones are allowed to be very flexibly related, the causal structure can still be identified. Next, we focus on the conditions by [18], which takes that of [24] as special cases. Roughly speaking, the identifiability of the structure of a partially observed linear causal model is built upon the identifiability of atomic covers, defined as follows (with *effective cardinality* defined as $||\mathcal{V}|| = |(\cup_{\mathbf{V}\in\mathcal{V}}\mathbf{V})|$ and $\text{PCh}_\mathcal{G}$ defined in Appendix B.2).

Definition 4 (Atomic Cover [18])**.** *Let $\mathbf{V} \in \mathbf{V}_\mathcal{G}$ be a set of variables, where l out of $|\mathbf{V}|$ are latent, and the remaining $|\mathbf{V}| - l$ are observed. $\mathbf{V}$ is an atomic cover if $\mathbf{V}$ contains a single observed variable, or if the following conditions hold:*

(i) There exists a set of atomic covers $\mathcal{C}$, with $||\mathcal{C}|| \geq l + 1$, such that $\cup_{\mathbf{C}\in\mathcal{C}}\mathbf{C} \subseteq PCh_\mathcal{G}(\mathbf{V})$.

(ii) There exists a set of covers $\mathcal{N}$, with $||\mathcal{N}|| \geq l + 1$, such that every element in $\cup_{\mathbf{N}\in\mathcal{N}}\mathbf{N}$ is a neighbour of $\mathbf{V}$ and $(\cup_{\mathbf{N}\in\mathcal{N}}\mathbf{N}) \cap (\cup_{\mathbf{C}\in\mathcal{C}}\mathbf{C}) = \emptyset$.

(iii) There is not a partition of $\mathbf{V} = \mathbf{V_1} \cup \mathbf{V_2}$ such that both $\mathbf{V_1}$ and $\mathbf{V_2}$ are atomic covers.

The intuition that we build structure identifiability upon the notion of atomic covers is as follows. When a set of latent variables share the same set of children and neighbors, it is impossible to differentiate these latent variables from each other, and thus we need to consider them together as the minimal identifiable group to build up the identifiability of the whole structure. Such a minimal identifiable group of variables is defined as an atomic cover. Roughly, for a group of variables to be qualified as an atomic cover, it has to have enough children and neighbors. An example is as follows.

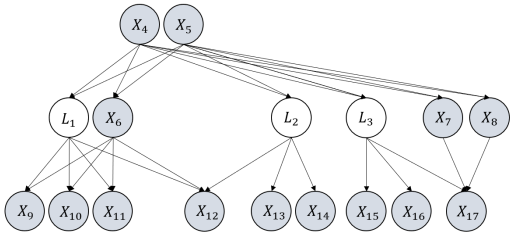


Figure 3: An illustrative graph that satisfies the conditions for structure-identifiability. At the same time, it also satisfies the condition for parameter identifiability - given the structure and $\Sigma_{\mathbf{X}_\mathcal{G}}$, all the parameters are identifiable only up to group sign indeterminacy.

Example 2 (Example of Atomic Cover)**.** *Consider the graph in Fig. 3. $\mathbf{V} = \{\mathsf{L}_1, \mathsf{X}_6\}$ is an atomic cover. This is because there exist $\mathcal{C} = \{\{\mathsf{X}_9\}, \{\mathsf{X}_{10}\}\}$ with $||\mathcal{C}|| \geq l + 1 = 2$ such that (i) in Def. 4 is satisfied. And there exist $\mathcal{N} = \{\{\mathsf{X}_{11}\}, \{\mathsf{X}_{12}\}\}$ (or, $\mathcal{N} = \{\{\mathsf{X}_4\}, \{\mathsf{X}_5\}\}$) with $||\mathcal{N}|| \geq l + 1 = 2$ such that (ii) in Def. 4 is satisfied. We can also find that (iii) in Def. 4 is satisfied. Thus $\{\mathsf{L}_1, \mathsf{X}_6\}$ is an atomic cover. Another example would be in Figure 8, where $\{\mathsf{L}_1, \mathsf{L}_2\}$ is an atomic cover.*

Condition 1 (Basic Conditions for Structure Identifiability [18]). *$\mathcal{G}$ satisfies the basic graphical condition for identifiability, if every latent variable belongs to at least one atomic cover in $\mathcal{G}$ and for each atomic cover with latent variables, any of its children is not adjacent to any of its neighbours.*

Condition 2 (Condition on Colliders [18]). *In $\mathcal{G}$, if (i) there exists sets of variables $\mathbf{V}$, $\mathbf{V_1}$, $\mathbf{V_2}$, and $\mathbf{T}$ such that every variable in $\mathbf{V}$ is a collider of two atomic covers $\mathbf{V_1}$, $\mathbf{V_2}$, and $\mathbf{T}$ is a minimal set of variables that d-separates $\mathbf{V_1}$ from $\mathbf{V_2}$, and (ii) there exists at least one latent variable in $\mathbf{V} \cup \mathbf{V_1} \cup \mathbf{V_2} \cup \mathbf{T}$, then we must have $|\mathbf{V}| + |\mathbf{T}| \geq |\mathbf{V_1}| + |\mathbf{V_2}|$.*

Example 3 (Example that satisfies Conditions 1 and 2). *Consider Figure 3. All latent variables in the graph belong to at least one atomic cover and thus Condition 1 is satisfied. Plus, Condition 2 is also satisfied. This is because the sets of variables $\mathbf{V}$, $\mathbf{V_1}$, $\mathbf{V_2}$, and $\mathbf{T}$ that satisfy (i) and (ii) in Condition 2 are $\mathbf{V} = \{\mathsf{X}_{12}\}$, $\mathbf{V_1} = \{\mathsf{L}_1, \mathsf{X}_6\}$, $\mathbf{V_2} = \{\mathsf{L}_2\}$, and $\mathbf{A} = \{\mathsf{X}_4, \mathsf{X}_5\}$, and we also have $|\mathbf{V}| + |\mathbf{A}| \geq |\mathbf{V_1}| + |\mathbf{V_2}|$. Therefore, the graph in Figure 3 satisfies both Conditions 1 and 2.*

The identifiability theory of structure is as follows. For a graph $\mathcal{G}$, if Condition 1 and Condition 2 are satisfied, then asymptotically the structure is identifiable up to the Markov equivalence class (MEC) of $\mathcal{O}_{\min}(\mathcal{O}_s(\mathcal{G}))$ (definitions of $\mathcal{O}_{\min}(\cdot)$ and $\mathcal{O}_s(\cdot)$ can be found in Appendix B.3). Roughly speaking, the underlying causal structure of $\mathcal{G}$ can be identified except that the directions of some edges cannot be determined. Next, we will show that, given any DAG in the identified equivalence class together with $\Sigma_{\mathbf{X}_\mathcal{G}}$, the parameters of the model are also identifiable, if certain conditions are satisfied.

3.3 Identifiability of Parameters

In this section we show that, given graphical Conditions 1 and 2, the causal coefficients F in Definition 1 are also identifiable, if certain conditions are satisfied.

Theorem 3 (Sufficient Condition for Parameter identifiability (up to group sign)). *Assume that $\mathcal{G}$ satisfies Conditions 1 and 2 and thus the structure can be identified up to the MEC of $\mathcal{O}_{min}(\mathcal{O}_s(\mathcal{G}))$. For any DAG in the equivalence class, the parameters are identifiable, if both the following hold:*

(i) For any atomic cover $\mathbf{V} = \mathbf{X} \cup \mathbf{L}$, $|\mathbf{L}| \leq 1$.

(ii) If an atomic cover $\mathbf{V} = \mathbf{X} \cup \mathbf{L}$ satisfies $|\mathbf{L}| \neq 0$ and $|\mathbf{X}| \geq 1$, then there exists $\hat{\mathbf{X}} \subseteq \mathbf{X}_\mathcal{G}$ such that $\hat{\mathbf{X}}$ d-separates $\mathbf{X}$ and $\mathbf{L}$.

Theorem 3 provides a sufficient condition such that the parameters are identifiable. We note that this sufficient condition has certain extent of necessity and we will detail this point later after we introduced some necessary conditions that can guide us under which scenarios parameters are guaranteed to be not identifiable. Now, for a better understanding of Theorem 3, we provide an example of it as follows.

Example 4 (Example for Theorem 3). *The graph $\mathcal{G}$ in Figure 3 satisfies the conditions for parameter identifiability in Theorem 3. Specifically, condition (i) in Theorem 3, is satisfied as all atomic covers contain no more than one latent variable. Plus, condition (ii) in Theorem 3 is also satisfied, as the atomic cover $\mathbf{V} = \mathbf{X} \cup \mathbf{L} = \{\mathsf{L}_1\} \cup \{\mathsf{X}_6\}$ satisfies $|\mathbf{L}| \neq 0$ and $|\mathbf{X}| \geq 1$ and there exist $\hat{\mathbf{X}} = \{\mathsf{X}_4, \mathsf{X}_5\}$ such that $\hat{\mathbf{X}}$ d-sep $\mathbf{X}$ and $\mathbf{L}$. Therefore, the parameters are identifiable for the graph in Figure 3.*

Next, we discuss under which conditions the parameters are guaranteed to be not identifiable. As discussed in Section 3.1, there are three kinds of indeterminacy. The first one can be solved by assuming unit variance of the noise terms of latent variables while the second one group sign indeterminacy is rather trivial such that we still consider parameters as identifiable even if group sign indeterminacy exists. Therefore, we will focus on the third one, orthogonal transformation indeterminacy, in what follows.

Theorem 4 (Condition for the Existence of Orthogonal Transformation Indeterminacy). *Consider the model in Definition 1. If a set of latent variables $\mathbf{L}$ with $|\mathbf{L}| \geq 2$, have the same parents and children, then there must exist orthogonal transformation indeterminacy regarding the edge coefficients F. In other words, F can at most be identified up to orthogonal transformation indeterminacy.*

Example 5 (Example for Thm. 4). *Consider Fig. 8. The graph satisfies the conditions in Thm. 4 as the parents and children of L_1 and L_2 are exactly the same. Therefore, there must exist orthogonal transformation indeterminacy for the edge coefficients F and thus the parameters are not identifiable.*

The Theorem 4 above indicates that, if there exist two latent variables that share the same parents and children, then the edge parameters can at most be identified up to orthogonal transformation. This directly implies a necessary condition for parameter identifiability as follows.

Corollary 1 (General Necessary Condition for Parameter Identifiability)**.** *For parameters to be identifiable, every pair of latent variables has to have at least one different parent or child.*

Corollary 1 captures a necessary condition in the general cases such that parameters are identifiable. If we further consider the graphs that are also structure identifiable (as we need to identify the structure first to fully specified the causal model), we further have the following Corollary 2 by considering the notion of atomic covers (the proofs of both corollaries can be found in the Appendix).

Corollary 2 (Necessary Condition about Atomic Covers for Parameter Identifiability)**.** *Assume $\mathcal{G}$ satisfies Conditions 1 and 2 and thus the structure can be identified up to the MEC of $\mathcal{O}_{min}(\mathcal{O}_s(\mathcal{G}))$. For any DAG $\mathcal{G}$ in the equivalence class, for $\mathcal{G}$'s parameters to be identifiable, every atomic cover must contain no more than one latent variable.*

Now we are ready to discuss the necessity of the graphical condition proposed in Theorem 3.

Remark 3 (Necessity of Conditions in Theorem 3)**.** *Theorem 3 provides a sufficient condition for parameter identifiability and it has certain extent of necessity with reasons as follows. First, condition (i) in Theorem 3 is provably necessary: by Corollary 2, for parameters to be identifiable, one has to assume (i) in Theorem 3. Second, if condition (ii) in Theorem 3 does not hold, then there are only some rare cases where parameters can be identified.*

We would like to mention that establishing a necessary and sufficient condition is always highly non-trivial in various tasks. For example, for the identification of linear non-Gaussian causal structure with latent variables, researchers initially developed sufficient conditions with three pure children in [46], later relaxed to two in [11, 58], before ultimately achieving both necessary and sufficient conditions in [1]. Similarly, for parameter identification, although the condition we proposed is not a necessary and sufficient one, it has considerable extents of necessity, and could be expected to serve as a stepping stone towards tighter and ultimately the necessary and sufficient condition for the field.

Identifiability theory often focuses on the asymptotic case, i.e., we assume that we know the structure and the population covariance matrix $\Sigma_{\mathbf{X}_\mathcal{G}}$. However, in practice, we only have access to i.i.d. data with finite sample size and thus only have the sample covariance matrix. Therefore, in the next section, we will propose a novel method to estimate the parameters in the finite sample cases.

4 Parameter Estimation Method

4.1 Objective

Our goal is to estimate F in Definition 1, given the causal structure $\mathcal{G}$ and observational data. The key is to parameterize the population covariance $\Sigma_{\mathbf{X}_\mathcal{G}}$ using $\theta = (F, \Omega)$ and then maximize the likelihood of observed sample covariance $\hat{\Sigma}_{\mathbf{X}_\mathcal{G}}$. To make this technically precise, we provide a closed-form expression of $\Sigma_{\mathbf{X}_\mathcal{G}}$ in terms of θ in the following proposition, with a proof given in Appendix A.6.

Proposition 1 (Parameterization of Population Covariance Matrix)**.** *Consider the model defined in Definition 1. Let $M := \big(I - A - B(I-D)^{-1}C\big)^{-1}$ and $N := \big((I-A)C^{-1}(I-D) - B\big)^{-1}$. Then, the population covariance matrices of $\mathbf{L}_\mathcal{G}$ and $\mathbf{X}_\mathcal{G}$ can be formulated as*

$$\Sigma_{\mathbf{L}_\mathcal{G}} = M^T \Omega_{\mathbf{L}_\mathcal{G}} M + N^T \Omega_{\mathbf{X}_\mathcal{G}} N, \tag{1}$$

$$\Sigma_{\mathbf{X}_\mathcal{G}} = (I-D)^{-T}\Big(B^T \Sigma_{\mathbf{L}_\mathcal{G}} B + \Omega_{\mathbf{X}_\mathcal{G}} + \Omega_{\mathbf{X}_\mathcal{G}} NB + B^T N^T \Omega_{\mathbf{X}_\mathcal{G}}\Big)(I-D)^{-1}. \tag{2}$$

The formulations of $\Sigma_{\mathbf{L}_\mathcal{G}}$ and $\Sigma_{\mathbf{X}_\mathcal{G}}$ are rather complicated due to the general scenario we considered, i.e., latent variables can be the cause or the effect of latent and observed variables. That is, the submatrices A, B, C and D defined in the above proposition can all have nonzero entries. In most existing works, at least one of these submatrices are assumed to be zero. For instance, factor analysis assumes that A, C and D are zero, while [32] assumes that A and C are zero. Furthermore, Proposition 1 also provides insight into the indeterminacy involved when identifying the parameters, such as the indeterminacy of variance in Theorem 1 and the orthogonal transformation indeterminacy in Theorem 4.

Similar to the typical factor analysis setting [45, 7, 21], we assume that $\epsilon_{\mathbf{V}_\mathcal{G}}$ are all Gaussian and thus $\mathbf{X}_\mathcal{G}$ are jointly Gaussian. Thus, the negative log-likelihood of observational data can be formulated as

$$\mathcal{L} = (K/2)(\mathrm{tr}((\Sigma_{\mathbf{X}_\mathcal{G}})^{-1}\hat{\Sigma}_{\mathbf{X}_\mathcal{G}}) + \log\det\Sigma_{\mathbf{X}_\mathcal{G}}), \tag{3}$$

where K is the number of i.i.d. observations. With the parameterized negative log-likelihood, we estimate the edge coefficients by minimizing the negative log-likelihood, as

$$\hat{F}, \hat{\Omega} = \arg\min_{F,\Omega} \mathcal{L}, \quad \text{subject to } \Omega_{\mathbf{L}_\mathcal{G}} = I, \tag{4}$$

where the entries of matrix F that do not correspond to an edge in $\mathcal{G}$ are constrained to be zero during the optimization. Note that in Eq. 4 the constraint that the noise terms of latent variables have unit variance is crucial to deal with the variance indeterminacy defined in Theorem 1. In practice, it is also favorable to use another constraint to address the variance indeterminacy, i.e., the constraint that all the latent variables have unit variance. This leads to an alternative objective as

$$\hat{F}, \hat{\Omega} = \arg\min_{F,\Omega} \mathcal{L}, \quad \text{subject to } (\Sigma_{\mathbf{L}_\mathcal{G}})_{ii} = 1,\ i \in [m], \tag{5}$$

where the entries of F that do not correspond to an edge in $\mathcal{G}$ are also constrained to be zero.

Both objectives in Eq. 4 and Eq. 5 can be employed, and yet using the second one gives rise to edge coefficients that are easier to understand. To be concerete, if we normalize all observed variables to have unit variance, then using Eq. 5 would give rise to $\hat{F}$ such that $-1 \leq \hat{F}_{i,j} \leq 1, \forall i, j \in [m]$. An example can be found in Figure 4. However, it may not be straightforward to realize the constraint in Eq. 5. To this end, in the next section we introduce a way to parameterize $\Sigma_{\mathbf{X}_\mathcal{G}}$ using F, such that the required constraint in Eq. 5 can be automatically satisfied. Later in Section 5.2, we also empirically compare the performance of using Eq. 4 with that of using Eq. 5.

4.2 Parameterization Trick of Covariance Matrix

In this section, we introduce how trek rules can be employed to parameterize $\Sigma_{\mathbf{X}_\mathcal{G}}$ while the unit variance constraint on latent variables in Eq. 5 can be elegantly satisfied. We start with the definition of trek. For readers who are less familiar with treks, please refer to Appendix B.1 for examples.

Definition 5 (Treks [50])**.** *In $\mathcal{G}$, a trek from X to Y is an ordered pair of directed paths (P_1, P_2) where P_1 has a sink X, P_2 has a sink Y, and both P_1 and P_2 have the same source Z, i.e., top$(P_1, P_2) = \mathsf{Z}$. A Trek is simple if P_1 and P_2 have no intersection except their common source Z.*

At this point, we are able to parameterize each entry of $\Sigma_{\mathbf{X}_\mathcal{G}}$ using $(F, \{\sigma_{ii}\}_{i=1}^{n+m})$, instead of (F, Ω), by making use of the (simple) trek rule [50], as follows:

$$\sigma_{ij} = \sum_{P_1, P_2 \in \mathcal{S}(\mathsf{V_i}, \mathsf{V_j})} \sigma_{\text{top}(P_1,P_2)} f^{P_1} f^{P_2}, \tag{6}$$

where $\mathcal{S}(\mathsf{V_i}, \mathsf{V_j})$ is the set of all simple treks between $\mathsf{V_i}$ and $\mathsf{V_j}$, and f^P is the path monomial along P defined as $f^P := \Pi_{k \to l \in P} f_{kl}$. By this form of parameterization, we can simply set all entries of $\{\sigma_{ii}\}_{i=1}^{n+m}$ as 1 (which is equivalent to requiring all variables to have unit variance), such that the constraint in Eq. 5 can be automatically satisfied. For a better understanding of how to use the simple trek rule for parameterization, we provide an example as follows.

Example 6 (Example for Parameterization using Simple Trek)**.** *In Figure 7 (a), there are four simple treks between $\mathsf{X_4}$ and $\mathsf{X_5}$, as shown in (b). By the simple trek rule and further assuming that all variables have unit variance, the covariance between $\mathsf{X_4}$ and $\mathsf{X_5}$, $\sigma_{4,5}$, can be formulated as* $f_{1,4}f_{1,5} + f_{3,4}f_{3,5} + f_{2,1}f_{1,4}f_{2,3}f_{3,5} + f_{2,3}f_{3,4}f_{2,1}f_{1,5}$.

5 Experiments

We validate our identifiability theory and parameter estimation method on synthetic and real-life data.

5.1 Setting and Evaluation Metric

We begin with our experimental setting of synthetic data. The causal strength f_{ij} is uniformly sampled from $[-2, 2]$ and the noise terms are Gaussian with variance uniformly from $[1, 5]$. We consider 20 graphs. 10 of them should be parameter-identifiable up to group sign indeterminacy according to our identifiability theory and we refer to them as *GS Case* (examples in Figure 10 in Appendix). Another 10 should be parameter-identifiable up to group sign and orthogonal transformation indeterminacy and we refer to them as *OT Case* (examples in Figure 11 in Appendix). On average each graph contains 15 variables, 3 out of them are latent. We consider three different sample sizes: 2k, 5k, and 10k. We use three random seeds to generate the causal model and report the mean performance as well as the std.

As the optimization in Eq 4 is nonconvex, we will rely on 30 random starts and choose the one with the best likelihood. We report the performance of the proposed method with two different objectives.

Table 1: Experimental result on synthetic data using MSE (mean (std)).

		MSE up to group sign	
Method		Estimator	Estimator-TR
GS Case	2k	0.0023 (0.002)	0.0012 (0.0005)
	5k	0.0014 (0.002)	0.0005 (0.0005)
	10k	0.0012 (0.001)	0.0003 (0.0004)

(a) MSE up to group sign indeterminacy.

		MSE up to orthogonal.	
Method		Estimator	Estimator-TR
OT Case	2k	0.0278 (0.008)	0.0355 (0.015)
	5k	0.0194 (0.002)	0.0352 (0.012)
	10k	0.0182 (0.003)	0.0351 (0.015)

(b) MSE up to orthogonal transformation.

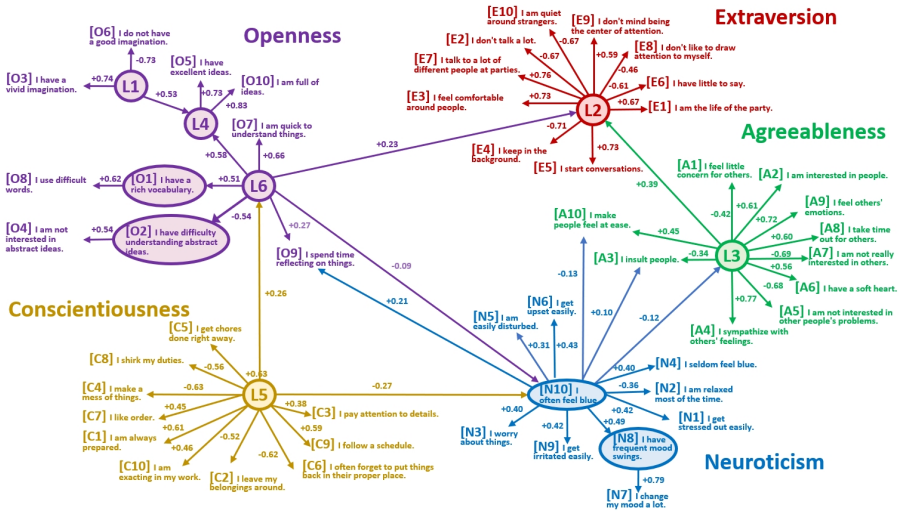


Figure 4: Estimated edge coefficients by the proposed method for Big Five human personality dataset. Variables whose name starts with "L" are latent variables while the others are observed variables.

(i) Parameter Estimator with objective defined in Eq. 4, referred to as Estimator, and **(ii)** Parameter Estimator with objective defined in Eq. 5 and Trek Rule parameterization trick in Eq 6, referred to as Estimator-TR.

It is worth noting that our setting is very general in that we allow latent variables and observed variables to be causally connected in a very flexible way, and we consider the identification of parameters of edges that can involve both observed and latent variables. Therefore, to the best of our knowledge, no current method can achieve the same goal to serve as the baseline (which also shows the novelty of the proposed method). As such, we mainly focus on comparing our estimation result with the ground truth parameters.

We use two MSE-based metrics defined as follows. **MSE up to group sign:** suppose the ground truth parameter is F and our estimation is $\hat{F}$. The MSE up to group sign is defined as $\frac{\||F|-|\hat{F}|\|_2^2}{\|F\|_0}$, where $|\cdot|$ takes element wise absolute value, $\|\cdot\|_2$ denotes the Frobenius norm and $\|\cdot\|_0$ denotes the number of nonzero entries of a matrix. **MSE up to orthogonal transformation:** the MSE up to orthogonal transformation is defined as

$$\min_{Q:Q^TQ=I} \frac{\||A|-|Q^T\hat{A}Q|\|_2^2+\||B|-|Q^T\hat{B}|\|_2^2+\||C|-|\hat{C}Q|\|_2^2+\||D|-|\hat{D}|\|_2^2}{\|F\|_0}, \hat{F}=\begin{pmatrix}\hat{A} & \hat{B}\\ \hat{C} & \hat{D}\end{pmatrix}, \quad (7)$$

where the optimization is solved by Adam [27] and the orthogonal matrix Q can be directly parameterized in PyTorch.

5.2 Synthetic Data Performance

We report the performance using synthetic data in Tab 1a and 1b, where both our Estimator and Estimator-TR achieve very good identification performance. For example, in the GS scenario with 10k samples, our Estimator achieves 0.0.0012 MSE up to group sign and our Estimator-TR achieves 0.0003 MSE up to group sign. The good performance by Estimator and Estimator-TR not only validates our estimation method, but also empirically verifies our identifiability theory.

5.3 Misspecification Behavior

In this section, we show that the proposed estimation method still performs well, even under model misspecification: violation of normality and violation of linearity.

As for violation of normality, we use uniform noise terms for the underlying model, and thus the distribution is not jointly Gaussian anymore. We aim to see to what extent can the proposed method still recover the correct parameters. The result is shown in Table 2 in the Appendix, which shows even when the normality is violated, we can still estimate the parameters pretty well. The reason lies in that our proposed asymptotic identifiability result holds true, even when we do not assume Gaussianity; as we only make use of the second-order statistics of the distribution, the additive noise in Definition 1 can follow any other continuous distribution.

To simulate the violation of linearity, we employ the leaky ReLU function during the generation process, as $\mathsf{V_i} = \text{LRELU}(\sum_{\mathsf{V_j} \in \text{Pa}(\mathsf{V_i})} f_{ji}\mathsf{V_j} + \epsilon_{\mathsf{V_i}})$, $\text{LRELU}(x) = \max(\alpha x, x)$. When α is close to 1, the function is close to a linear one, and when α is close to 0, the model is very nonlinear. The result is shown in Table 3 and we found that our estimation method is quite robust to small violations of linearity. For example, for Estimator-TR in GS case with 10k sample size, if we set $\alpha = 0.8$, we still get a small MSE of 0.001. Even when α decreases to 0.6, the MSE is around 0.005, which is still small. However, when α is decreased to 0.3, the underlying model is considerably nonlinear, and the MSE increases to 0.027.

5.4 Implementation Details, Runtime Analysis, and Scalability

Our code is based on Python3.7 and PyTorch [37]. Data is standardized and the optimization in Eq 4, Eq 5, and Eq. 7 are solved by Adam [27], with a learning rate of 0.02. We conduct all the experiments with single Intel(R) Xeon(R) CPU E5-2470. All experiments can be finished within 2 hours. We note that our method is very computationally efficient. First, the computational cost is almost irrelevant to sample size: we only need to calculate the sample covariance matrix once and cache it for further use during the optimization. Plus, our estimation method can handle a large number of variables. For example, the running time of our method are roughly 10 seconds, 2 minutes, and 10 minutes for 20 variables, 50 variables, and 100 variables respectively. For 300 variables, which is a considerably large number for typical experiments considered in causal discovery papers, the estimation can still be finished within around one hour.

It is also worth noting that model misspecifications do not influence the computation cost of our method. We briefly discuss the efficiency of checking whether conditions in Theorem 3 hold, together with what if conditions do not hold in solving real-life problems in Appendix A.7 and A.8.

5.5 Real-World Data Performance

In this section, we employ a famous psychometric dataset - Big Five dataset `https://openpsychometrics.org/`, to validate our method. It consists of 50 indicators and close to 20,000 data points. There are five dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (O-C-E-A-N). Each is measured with 10 indicators. Data is standardized.We employ the RLCD method [18] to determine the MEC and GIN [58] to decide the remaining directions. Then we employ the proposed Estimator-TR to estimate all the edge coefficients.

The estimated edge coefficients are shown in Figure 4. We found that our estimated coefficients are well aligned with existing psychology studies. For example, according to [16, 17], being successful in exploratory endeavors depends on the stability to pursue them. This is illustrated in our result where $\text{L5} \xrightarrow{+0.26} \text{L6}$ and $\text{L3} \xrightarrow{+0.39} \text{L2}$ indicates that Conscientiousness positively influence openness and Agreeableness positively influences Extraversion. Moreover, it has been shown that people are likely to weigh the outcomes of their actions, thus, their level of Conscientiousness coupled with Neuroticism may prohibit them from engaging in risky behaviors ($\text{L5} \xrightarrow{-0.27} \text{N10} \xrightarrow{-0.12} \text{L3} \xrightarrow{+0.39} \text{L2}$) [54]. Such consistency with current psychometric studies again validates the effectiveness of the proposed method in parameter estimation of real-life systems.

6 Conclusion

In this paper, we characterize indeterminacy of parameter identification and provide conditions for identifiability. Finally, we propose a novel estimation method and validate it by empirical study.

7 Acknowledgement

This material is based upon work supported by NSF Award No. 2229881, AI Institute for Societal Decision Making (AI-SDM), the National Institutes of Health (NIH) under Contract R01HL159805, and grants from Salesforce, Apple Inc., Quris AI, and Florin Court Capital.

References

[1] Jeffrey Adams, Niels Hansen, and Kun Zhang. Identification of partially observed linear causal models: Graphical conditions for the non-gaussian and heterogeneous cases. *Advances in Neural Information Processing Systems*, 34:22822–22833, 2021.

[2] Sina Akbari, Ehsan Mokhtarian, AmirEmad Ghassami, and Negar Kiyavash. Recursive causal structure learning in the presence of latent variables and selection bias. In *Advances in Neural Information Processing Systems*, volume 34, pages 10119–10130, 2021.

[3] Animashree Anandkumar, Daniel Hsu, Adel Javanmard, and Sham Kakade. Learning linear bayesian networks with latent variables. In *International Conference on Machine Learning*, pages 249–257, 2013.

[4] Ankur Ankan, Inge Wortel, Kenneth Bollen, and Johannes Textor. Combining graphical and algebraic approaches for parameter identification in latent variable structural equation models. In *International Conference on Artificial Intelligence and Statistics*, pages 7252–7264. PMLR, 2023.

[5] Rina Foygel Barber, Mathias Drton, Nils Sturma, and Luca Weihs. Half-trek criterion for identifiability of latent variable models. *The Annals of Statistics*, 50(6):3174–3196, 2022.

[6] Elias Bareinboim and Jin Tian. Recovering causal effects from selection bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[7] Paul A Bekker and Jos MF ten Berge. Generic global indentification in factor analysis. *Linear Algebra and its Applications*, 264:255–263, 1997.

[8] Carlos Brito and Judea Pearl. A new identification condition for recursive models with correlated errors. *Structural Equation Modeling*, 9(4):459–474, 2002.

[9] Carlos Brito and Judea Pearl. Generalized instrumental variables. *arXiv preprint arXiv:1301.0560*, 2012.

[10] Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. In *Advances in Neural Information Processing Systems*, 2020.

[11] Ruichu Cai, Feng Xie, Clark Glymour, Zhifeng Hao, and Kun Zhang. Triad constraints for learning causal structure of latent variables. *Advances in neural information processing systems*, 32, 2019.

[12] Zhengming Chen, Feng Xie, Jie Qiao, Zhifeng Hao, Kun Zhang, and Ruichu Cai. Identification of linear latent variable model with arbitrary distribution. In *Proceedings 36th AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[13] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

[14] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

[15] Philipp Dettling, Mathias Drton, and Mladen Kolar. On the lasso for graphical continuous lyapunov models. In *Causal Learning and Reasoning*, pages 514–550. PMLR, 2024.

[16] Colin G. DeYoung. Higher-order factors of the big five in a multi-informant sample. *J Pers Soc Psychol*, 91(6)(6):1138–1151, 2006.

[17] Colin G. DeYoung. Cybernetic big five theory. *J Res Pers*, 56:33–56, 2015.

[18] 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 Representation*, 2024.

[19] Mathias Drton, Rina Foygel, and Seth Sullivant. Global identifiability of linear structural equation models. *The Annals of Statistics*, 39(2):865–886, 2011.

[20] Rina Foygel, Jan Draisma, and Mathias Drton. Half-trek criterion for generic identifiability of linear structural equation models. *The Annals of Statistics*, pages 1682–1713, 2012.

[21] Richard L Gorsuch. *Factor analysis: Classic edition*. Routledge, 2014.

[22] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012.

[23] 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(2):362–378, 2008.

[24] Biwei Huang, Charles Jia Han Low, Feng Xie, Clark Glymour, and Kun Zhang. Latent hierarchical causal structure discovery with rank constraints. *arXiv preprint arXiv:2210.01798*, 2022.

[25] Yimin Huang and Marco Valtorta. Pearl's calculus of intervention is complete. *arXiv preprint arXiv:1206.6831*, 2012.

[26] Yonghan Jung, Jin Tian, and Elias Bareinboim. Learning causal effects via weighted empirical risk minimization. *Advances in neural information processing systems*, 33:12697–12709, 2020.

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[28] Erich Kummerfeld and Joseph Ramsey. Causal clustering for 1-factor measurement models. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1655–1664. ACM, 2016.

[29] Daniel Kumor, Carlos Cinelli, and Elias Bareinboim. Efficient identification in linear structural causal models with auxiliary cutsets. In *International Conference on Machine Learning*, pages 5501–5510. PMLR, 2020.

[30] Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

[31] Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural DAG learning. In *International Conference on Learning Representations*, 2020.

[32] Dennis Leung, Mathias Drton, and Hisayuki Hara. Identifiability of directed gaussian graphical models with one latent source. *Electronic Journal of Statistics*, 10, 05 2015.

[33] Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

[34] Ignavier Ng, Xinshuai Dong, Haoyue Dai, Biwei Huang, Peter Spirtes, and Kun Zhang. Score-based causal discovery of latent variable causal models. In *Forty-first International Conference on Machine Learning*, 2024.

[35] Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. *Advances in Neural Information Processing Systems*, 33:17943–17954, 2020.

[36] R. M. O'BRIEN. Identification of simple measurement models with multiple latent variables and correlated errors. *Sociological Methodology*, 24, 1994.

[37] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[38] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2000.

[39] Judea Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

[40] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[41] O. Reiersøl. On the identifiability of parameters in thurstone's multiple factor analysis. *Psychometrika*, 15(2), 1950.

[42] Thomas Richardson and Peter Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

[43] 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(39):1–24, 2020.

[44] Bernhard Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

[45] Alexander Shapiro. Identifiability of factor analysis: Some results and open problems. *Linear Algebra and its Applications*, 70:1–7, 1985.

[46] Shohei Shimizu, Patrik O Hoyer, and Aapo Hyvärinen. Estimation of linear non-gaussian acyclic models for latent factors. *Neurocomputing*, 72(7-9):2024–2027, 2009.

[47] Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1219. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[48] Ricardo Silva, Richard Scheines, Clark Glymour, Peter Spirtes, and David Maxwell Chickering. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7(2), 2006.

[49] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[50] Seth Sullivant, Kelli Talaska, and Jan Draisma. Trek separation for gaussian graphical models. *arXiv:0812.1938.*, 2010.

[51] Jin Tian. Parameter identification in a class of linear structural equation models. In *Twenty-First International Joint Conference on Artificial Intelligence*, 2009.

[52] Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Aaai/iaai*, pages 567–573, 2002.

[53] Sivan Toledo. Locality of reference in lu decomposition with partial pivoting. *SIAM Journal on Matrix Analysis and Applications*, 18(4):1065–1081, 1997.

[54] Nicholas A. Turiano, Daniel K. Mroczek, Jan Moynihan, and Benjamin P. Chapman. Big 5 personality traits and interleukin-6: evidence for "healthy neuroticism" in a us population sample. *Brain Behav Immun*, pages 83–89, 2013.

[55] Gherardo Varando and Niels Richard Hansen. Graphical continuous lyapunov models. In *Conference on Uncertainty in Artificial Intelligence*, pages 989–998. Pmlr, 2020.

[56] L. L. WEGGE. Local identifiability of the factor analysis and measurement error model parameter. *Journal of Econometrics*, 70, 1996.

[57] B. Williams. Identification of the linear factor model. *Econometric Reviews*, 39(1), 2020.

[58] Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. *Advances in neural information processing systems*, 33:14891–14902, 2020.

[59] Xun Zheng. *Learning DAGs with Continuous Optimization*. PhD thesis, PhD thesis, Carnegie Mellon University, 2020.

A Proofs

A.1 Proof of Theorem 1

Theorem 1 (Indeterminacy of Scaling of $\Omega_{\mathbf{L}_\mathcal{G}}$)**.** *Consider a model that follows Definition 1 with number of latent variables $m \geq 1$ and $\theta = (A, B, C, D, \Omega_{\mathbf{X}_\mathcal{G}}, \Omega_{\mathbf{L}_\mathcal{G}})$. Let Λ be any invertible diagonal matrix, and $\tilde{\theta} = (\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{\Omega}_{\mathbf{X}_\mathcal{G}}, \tilde{\Omega}_{\mathbf{L}_\mathcal{G}})$, where*

$$\tilde{A} \coloneqq \Lambda^{-1}A\Lambda, \quad \tilde{B} \coloneqq \Lambda^{-1}B, \quad \tilde{C} \coloneqq C\Lambda, \quad \tilde{D} \coloneqq D, \quad \tilde{\Omega}_{\mathbf{L}_\mathcal{G}} \coloneqq \Lambda^2\Omega_{\mathbf{L}_\mathcal{G}}, \quad \textit{and} \quad \tilde{\Omega}_{\mathbf{X}_\mathcal{G}} \coloneqq \Omega_{\mathbf{X}_\mathcal{G}}.$$

Then, $\tilde{\theta}$ and θ entail the same observations, i.e., $\tilde{\Sigma}_{\mathbf{X}_\mathcal{G}} = \Sigma_{\mathbf{X}_\mathcal{G}}$. Furthermore, we have $\tilde{\Sigma}_{\mathbf{L}_\mathcal{G}} = \Lambda\Sigma_{\mathbf{L}_\mathcal{G}}\Lambda$.

Proof of Theorem 1. Let M and N be matrices defined as in Proposition 1, and similarly for $\tilde{M}$ and $\tilde{N}$. We then have

$$\begin{aligned}
\tilde{M} &= \left(I - \tilde{A} - \tilde{B}(I - \tilde{D})^{-1}\tilde{C}\right)^{-1} \\
&= \left(\Lambda^{-1}\Lambda - \Lambda^{-1}A\Lambda - (\Lambda^{-1}B)(I - D)^{-1}(C\Lambda)\right)^{-1} \\
&= \Lambda^{-1}\left(I - A - B(I - D)^{-1}C\right)^{-1}\Lambda \\
&= \Lambda^{-1}M\Lambda
\end{aligned}$$

and

$$\begin{aligned}
\tilde{N} &= \left((I - \tilde{A})\tilde{C}^{-1}(I - \tilde{D}) - \tilde{B}\right)^{-1} \\
&= \left((\Lambda^{-1}\Lambda - \Lambda^{-1}A\Lambda)(C\Lambda)^{-1}(I - D) - \Lambda^{-1}B\right)^{-1} \\
&= \left((I - A)C^{-1}(I - D) - B\right)^{-1}\Lambda \\
&= N\Lambda.
\end{aligned}$$

By Proposition 1, the latent covariance matrix $\tilde{\Sigma}_{\mathbf{L}}$ after rescaling of the parameters is given by

$$\begin{aligned}
\tilde{\Sigma}_{\mathbf{L}} &= \tilde{M}^T\tilde{\Omega}_{\mathbf{L}}\tilde{M} + \tilde{N}^T\tilde{\Omega}_{\mathbf{X}}\tilde{N} \\
&= (\Lambda^T M^T \Lambda^{-T})(\Lambda\Omega_{\mathbf{L}}\Lambda)(\Lambda^{-1}M\Lambda) + \Lambda^T N^T\Omega_{\mathbf{X}}N\Lambda \\
&= \Lambda(M^T\Omega_{\mathbf{L}}M + N^T\Omega_{\mathbf{X}}N)\Lambda \\
&= \Lambda\Sigma_{\mathbf{L}}\Lambda.
\end{aligned}$$

This implies that the variance of each latent variable L_i is scaled by Λ_{ii}^2. By Proposition 1, the observed covariance matrix $\tilde{\Sigma}_{\mathbf{X}}$ after rescaling of the parameters is given by

$$\begin{aligned}
\tilde{\Sigma}_{\mathbf{X}} &= (I - \tilde{D})^{-T}\left(\tilde{B}^T\tilde{\Sigma}_{\mathbf{L}}B + \tilde{\Omega}_{\mathbf{X}} + \tilde{\Omega}_{\mathbf{X}}\tilde{N}\tilde{B} + \tilde{B}^T\tilde{N}^T\tilde{\Omega}_{\mathbf{X}}\right)(I - \tilde{D})^{-1} \\
&= (I - D)^{-T}\Big((\Lambda^{-1}B)^T(\Lambda\Sigma_{\mathbf{L}}\Lambda)(\Lambda^{-1}B) \\
&\qquad + \Omega_{\mathbf{X}} + \Omega_{\mathbf{X}}(N\Lambda)(\Lambda^{-1}B) + (\Lambda^{-1}B)^T(N\Lambda)^T\Omega_{\mathbf{X}}\Big)(I - D)^{-1} \\
&= (I - D)^{-T}\left(B^T\Sigma_{\mathbf{L}}B + \Omega_{\mathbf{X}} + \Omega_{\mathbf{X}}NB + B^TN^T\Omega_{\mathbf{X}}\right)(I - D)^{-1} \\
&= \Sigma_{\mathbf{X}}.
\end{aligned}$$

□

A.2 Proof of Theorem 2

Theorem 2 (Group Sign Indeterminacy)**.** *Consider a model that follows Definition 1 with number of latent variables $m \geq 1$, $\theta = (A, B, C, D, \Omega_{\mathbf{X}_\mathcal{G}}, \Omega_{\mathbf{L}_\mathcal{G}})$, and $\Omega_{\mathbf{L}_\mathcal{G}} = I$. Let S be a diagonal sign matrix where the diagonal entries are either 1 or -1, and $\tilde{\theta} = (\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{\Omega}_{\mathbf{X}_\mathcal{G}}, \tilde{\Omega}_{\mathbf{L}_\mathcal{G}})$, where*

$$\tilde{A} \coloneqq SAS, \quad \tilde{B} \coloneqq SB, \quad \tilde{C} \coloneqq CS, \quad \tilde{D} \coloneqq D, \quad \tilde{\Omega}_{\mathbf{L}_\mathcal{G}} \coloneqq \Omega_{\mathbf{L}_\mathcal{G}} = I, \quad \textit{and} \quad \tilde{\Omega}_{\mathbf{X}_\mathcal{G}} \coloneqq \Omega_{\mathbf{X}_\mathcal{G}}.$$

Then, $\tilde{\theta}$ and θ entail the same observations, i.e., $\tilde{\Sigma}_{\mathbf{X}_\mathcal{G}} = \Sigma_{\mathbf{X}_\mathcal{G}}$, and $(\tilde{\Sigma}_{\mathbf{L}_\mathcal{G}})_{ii} = (\Sigma_{\mathbf{L}_\mathcal{G}})_{ii}$, $\forall i \in [m]$.

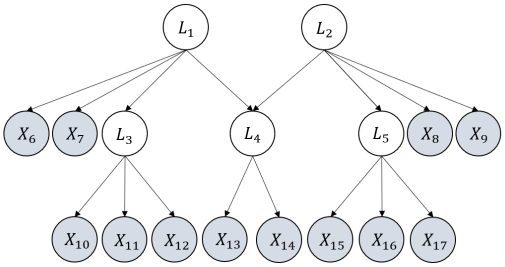


Figure 5: A simple graph that satisfies conditions in Theorem 3, as for each atomic cover with one latent variable, it has no observed variable.

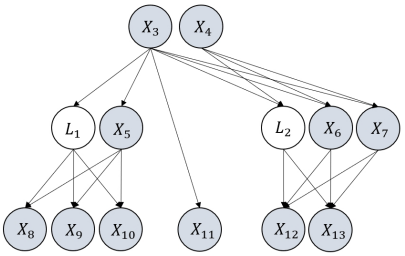


Figure 6: A more complicated graph that satisfies conditions in Theorem 3 and there is an atomic cover that has one latent variable and nonzero observed variables, e.g., $\{\mathsf{L}_1, \mathsf{X}_5\}$ in the graph. The condition (ii) in Theorem 3 is satisfied in that $\mathsf{L}_1, \mathsf{X}_5$ can be d-separated by X_3.

Proof of Theorem 2. Since S is a diagonal sign matrix, we have

$$\tilde{A} \coloneqq S^{-1}AS, \quad \tilde{B} \coloneqq S^{-1}B, \quad \tilde{C} \coloneqq CS, \quad \tilde{D} \coloneqq D, \quad \tilde{\Omega}_{\mathbf{L}_\mathcal{G}} \coloneqq S^2\Omega_{\mathbf{L}_\mathcal{G}}, \quad \text{and} \quad \tilde{\Omega}_{\mathbf{X}_\mathcal{G}} \coloneqq \Omega_{\mathbf{X}_\mathcal{G}}.$$

Note that S is an invertible diagonal matrix. By Theorem 1, we have $\tilde{\Sigma}_{\mathbf{X}_\mathcal{G}} = \Sigma_{\mathbf{X}_\mathcal{G}}$ and $\tilde{\Sigma}_{\mathbf{L}_\mathcal{G}} = S\Sigma_{\mathbf{L}_\mathcal{G}}S$, and thus $(\tilde{\Sigma}_{\mathbf{L}_\mathcal{G}})_{ii} = (\Sigma_{\mathbf{L}_\mathcal{G}})_{ii},\ \forall i \in [m]$. □

A.3 Proof of Theorem 3

The structure identifiability part is that if $\mathcal{G}$ satisfies Condition 1 and Condition 2, the structure of $\mathcal{G}$ can be identified up to the Markov equivalence class of $\mathcal{O}_{\min}(\mathcal{O}_s(\mathcal{G}))$, which is by Theorem 12 in [18].

Next we will focus on the proof of parameter identifiability part, i.e., for any DAG in the equivalence class, if (i) and (ii) in Theorem 3 are satisfied, the parameters are identifiable (up to group sign). Without loss of generality, we assume that all variables have unit variance and zero mean. The reason is that if we can show that the parameters are identfiable (up to group sign) under this assumption, then it is straightforward to show that they are also identifiable under the original assumption where $\Omega_{\mathbf{L}} = I$.

Lemma 1. *Consider a graph $\mathcal{G}$ that satisfies (i) and (ii) in Theorem 3. For an atomic cover $\mathbf{V}$ in $\mathcal{G}$ with one latent variable, $\mathbf{V} = \{\mathsf{L}\} \cup \{\mathsf{X}_\mathsf{i}\}_{i=1}^m$ (m could be zero), if it has an observed pure child C and the coefficients of the edges from $\mathbf{V}$ to C, i.e., $f_{\mathsf{L}\to\mathsf{C}}$, $f_{\mathsf{X}_1\to\mathsf{C}}$, …, $f_{\mathsf{X}_\mathsf{m}\to\mathsf{C}}$, are known, then for any variable A, such that $\mathsf{A} \neq \mathsf{C}$ and A is not a descendant of C, $\sigma_{\mathsf{L},\mathsf{A}}$ can be calculated as $(\sigma_{\mathsf{A},\mathsf{C}} - \Sigma_{i=1}^m \sigma_{\mathsf{X}_\mathsf{i},\mathsf{A}} f_{\mathsf{X}_\mathsf{i}\to\mathsf{C}})/f_{\mathsf{L}\to\mathsf{C}}$.*

Proof of Lemma 1. By the definition of atomic covers, all variables in $\mathbf{V} = \{\mathsf{L}\} \cup \{\mathsf{X}_\mathsf{i}\}_{i=1}^m$ are not adjacent. By trek rule and the fact that C is a pure child of $\mathbf{V}$, all treks from C to A go through $\mathbf{V}$, and thus by the trek rule we have $\sigma_{\mathsf{L},\mathsf{A}} f_{\mathsf{L}\to\mathsf{C}} + \Sigma_{i=1}^m \sigma_{\mathsf{X}_\mathsf{i},\mathsf{A}} f_{\mathsf{X}_\mathsf{i}\to\mathsf{C}} = \sigma_{\mathsf{A},\mathsf{C}}$. □

Remark: This lemma implies that we can find all edge coefficients of the graph in a bottom-up fashion. Roughly speaking, for a latent variable L that belongs to an atomic cover $\mathbf{V}$, once we identify

$f_{\mathsf{L}\to\mathsf{C}}$, $f_{\mathsf{X}_1\to\mathsf{C}}$, ..., $f_{\mathsf{X}_m\to\mathsf{C}}$ where C is an observed pure child of $\mathbf{V}$, we can take L as if it is observed. More specific explanations can be found in the following proof.

Theorem 3 (Sufficient Condition for Parameter identifiability (up to group sign))**.** *Assume that $\mathcal{G}$ satisfies Conditions 1 and 2 and thus the structure can be identified up to the MEC of $\mathcal{O}_{min}(\mathcal{O}_s(\mathcal{G}))$. For any DAG in the equivalence class, the parameters are identifiable, if both the following hold:*

(i) For any atomic cover $\mathbf{V} = \mathbf{X} \cup \mathbf{L}$, $|\mathbf{L}| \leq 1$.

(ii) If an atomic cover $\mathbf{V} = \mathbf{X} \cup \mathbf{L}$ satisfies $|\mathbf{L}| \neq 0$ and $|\mathbf{X}| \geq 1$, then there exists $\hat{\mathbf{X}} \subseteq \mathbf{X}_{\mathcal{G}}$ such that $\hat{\mathbf{X}}$ d-separates $\mathbf{X}$ and $\mathbf{L}$.

Proof of Theorem 3. Consider a graph $\mathcal{G}$ that satisfies (i) and (ii) in Theorem 3. We first show that if all the pure children of an atomic cover $\mathbf{V}$ are observed, then all the edge coefficients from the atomic cover to its children are identifiable (up to group sign). To this end, we categorize the scenarios into four cases and prove them separately with illustrative examples.

(a) $\mathbf{V} = \{\mathsf{X}\}$ contains a single observed variable. The proof for this case is trivial as the edge coefficient from X to its pure child C is simply $\sigma_{\mathsf{X},\mathsf{C}}$.

(b) $\mathbf{V} = \{\mathsf{L}\}$ contains a single latent variable. By Condition 1 there must exist C_1, C_2, and X_N, such that C_1, C_2 are pure children of $\mathbf{V}$ and X_N is an observed variable that has a trek to $\mathbf{V}$. Then we have

$$\sigma_{\mathsf{C}_1,\mathsf{C}_2} = f_{\mathsf{L}\to\mathsf{C}_1} f_{\mathsf{L}\to\mathsf{C}_2}, \quad \sigma_{\mathsf{C}_1,\mathsf{X}_\mathsf{N}} = f_{\mathsf{L}\to\mathsf{C}_1}\sigma_{\mathsf{L},\mathsf{X}_\mathsf{N}}, \quad \text{and} \quad \sigma_{\mathsf{C}_2,\mathsf{X}_\mathsf{N}} = f_{\mathsf{L}\to\mathsf{C}_2}\sigma_{\mathsf{L},\mathsf{X}_\mathsf{N}}.$$

By these three equations, we can solve $f_{\mathsf{L}\to\mathsf{C}_1}$ and $f_{\mathsf{L}\to\mathsf{C}_2}$. If $\mathbf{V}$ has more than two pure children, we can prove the identifiability similarly in a pairwise fashion.

Example. Take the atomic cover $\{\mathsf{L}_4\}$ in Figure 5 as an example. By Condition 1, it must have at least two pure children. To identify $f_{4,13}$ and $f_{4,14}$, we need to borrow an observed variable that has at least one trek from L_4 to it, e.g, X_8. As L must have two other neighbors, there must exist such a variable. Then we have $\sigma_{13,14} = f_{4,13}f_{4,14}$, $\sigma_{13,8} = f_{4,13}\sigma_{4,8}$, and $\sigma_{14,8} = f_{4,14}\sigma_{4,8}$. By these three equations, we can solve $f_{4,13}$ and $f_{4,14}$.

(c) $\mathbf{V} = \{\mathsf{L}\} \cup \{\mathsf{X}_\mathsf{i}\}_{i=1}^m$ ($m \geq 1$) contains a single latent variable and m observed variables, where $\mathbf{V}$ has at least three pure children. We assume that there exist C_1, C_2, and C_3, such that C_1, C_2, and C_3 are pure children of $\mathbf{V}$. By (ii) in Theorem 3, there exist $\hat{\mathbf{X}}$ such that $\hat{\mathbf{X}}$ d-separates L and $\{\mathsf{X}_\mathsf{i}\}_{i=1}^m$. Let $\sigma_{\epsilon_\mathsf{L}} = t$ (the variance of the noise term of L). In this case, we have

$$\sigma_{\mathsf{C}_1,\mathsf{C}_2|\hat{\mathbf{X}}\cup\{\mathsf{X}_\mathsf{i}\}_{i=1}^m} = t f_{\mathsf{L}\to\mathsf{C}_1} f_{\mathsf{L}\to\mathsf{C}_2}, \tag{8}$$

$$\sigma_{\mathsf{C}_1,\mathsf{C}_3|\hat{\mathbf{X}}\cup\{\mathsf{X}_\mathsf{i}\}_{i=1}^m} = t f_{\mathsf{L}\to\mathsf{C}_1} f_{\mathsf{L}\to\mathsf{C}_3}, \tag{9}$$

$$\sigma_{\mathsf{C}_2,\mathsf{C}_3|\hat{\mathbf{X}}\cup\{\mathsf{X}_\mathsf{i}\}_{i=1}^m} = t f_{\mathsf{L}\to\mathsf{C}_2} f_{\mathsf{L}\to\mathsf{C}_3}. \tag{10}$$

By these three equations, we can solve $f_{\mathsf{L}\to\mathsf{C}_1}$,$f_{\mathsf{L}\to\mathsf{C}_2}$, and $f_{\mathsf{L}\to\mathsf{C}_3}$, with the only remaining free parameter t. In other words, we have $f_{\mathsf{L}\to\mathsf{C}_1}(t)$,$f_{\mathsf{L}\to\mathsf{C}_2}(t)$, and $f_{\mathsf{L}\to\mathsf{C}_3}(t)$. Now we consider $\hat{\mathbf{X}} = \{\hat{\mathsf{X}}_\mathsf{j}\}_{j=1}^p$ to solve t. We have additional $pm + p$ free parameters that are $f_{\hat{\mathsf{X}}_\mathsf{j}\to\mathsf{X}_\mathsf{i}}, i \leq m, j \leq p$ and $f_{\hat{\mathsf{X}}_\mathsf{j}\to\mathsf{L}}, j \leq p$, and they can be solved by $\sigma_{\hat{\mathsf{X}}_\mathsf{j},\mathsf{X}_\mathsf{i}}$, $i \leq m, j \leq p$ and $\sigma_{\hat{\mathsf{X}}_\mathsf{j},\mathsf{L}}(t), j \leq p$ (where $\sigma_{\hat{\mathsf{X}}_\mathsf{j},\mathsf{L}}(t)$ is calculated by Lemma A.1). To this point, all the edge coefficients among $\mathbf{V}, \hat{\mathbf{X}}$, and C_1,C_2,C_3 are solved, up to the only free parameter t. Finally, we have that L has unit variance and thus we have $t + \sum_{j=1}^p f^2_{\hat{\mathsf{X}}_\mathsf{j}\to\mathsf{L}}(t) = 1$. By this additional equation, t can also be solved and thus all desired edge coefficients are identified. If $\mathbf{V}$ has more than three pure children, we just choose all the combinations of any three and thus we can identify all the edge coefficients from $\mathbf{V}$ to all its pure children.

Example. Take the atomic cover $\{\mathsf{L}_1, \mathsf{X}_5\}$ in Figure 6 as an example. Let $\sigma_{\epsilon_{\mathsf{L}_1}} = t$. We have $\sigma_{\mathsf{X}_8,\mathsf{X}_9|\{\mathsf{X}_3,\mathsf{X}_5\}} = t f_{\mathsf{L}_1\to\mathsf{X}_8} f_{\mathsf{L}_1\to\mathsf{X}_9}$, $\sigma_{\mathsf{X}_8,\mathsf{X}_{10}|\{\mathsf{X}_3,\mathsf{X}_5\}} = t f_{\mathsf{L}_1\to\mathsf{X}_8} f_{\mathsf{L}_1\to\mathsf{X}_{10}}$, and $\sigma_{\mathsf{X}_9,\mathsf{X}_{10}|\{\mathsf{X}_3,\mathsf{X}_5\}} = t f_{\mathsf{L}_1\to\mathsf{X}_9} f_{\mathsf{L}_1\to\mathsf{X}_{10}}$. By these three equations, we can solve $f_{\mathsf{L}_1\to\mathsf{X}_8}(t)$,$f_{\mathsf{L}_1\to\mathsf{X}_9}(t)$, and $f_{\mathsf{L}_1\to\mathsf{X}_{10}}(t)$, up to t. Then we introduce X_3 and solve $f_{\mathsf{X}_3\to\mathsf{X}_5}$ and $f_{\mathsf{X}_3\to\mathsf{L}_1}$ by $\sigma_{\mathsf{X}_3,\mathsf{X}_5}$ and $\sigma_{\mathsf{X}_3,\mathsf{L}_1}(t)$. Finally, we have $t + f^2_{\mathsf{X}_3\to\mathsf{L}_1}(t) = 1$ and thus all desired edge coefficients are identified.

(d) $\mathbf{V} = \{\mathsf{L}\} \cup \{\mathsf{X}_\mathsf{i}\}_{i=1}^m$ ($m \geq 1$) contains a single latent variable and m observed variables, where $\mathbf{V}$ has two pure children. By Condition 1, there must exist C_1, C_2 as the pure children of $\mathbf{V}$. If there exists one additional pure child, then it is the same as (c). By (ii) in Theorem 3, there exist $\hat{\mathbf{X}} = \{\hat{\mathsf{X}}_j\}_{j=1}^p$ such that $\hat{\mathbf{X}}$ d-separates L and $\{\mathsf{X}_\mathsf{i}\}_{i=1}^m$. Plus, if there only exist these two pure

children, there must exist $\mathsf{X_N}$ such that $\mathsf{X_N}$ is a parent of $\mathbf{V}$, and thus $\mathsf{X_N} \in \hat{\mathbf{X}}$. Without loss of generality, we take $\mathsf{X_N} = \hat{\mathsf{X}}_1$.

We assume that $\sigma_{\epsilon_{\mathsf{L}}} = t$, and thus we have

$$\sigma_{\mathsf{C}_1,\mathsf{C}_2|\hat{\mathbf{X}}\cup\{\mathsf{X_i}\}_{i=1}^m} = t f_{\mathsf{L}\to\mathsf{C}_1} f_{\mathsf{L}\to\mathsf{C}_2}, \tag{11}$$

$$\sigma_{\mathsf{C}_1,\mathsf{X_i}|\hat{\mathbf{X}}} = \sigma_{\mathsf{X_i}|\hat{\mathbf{X}}} f_{\mathsf{X_i}\to\mathsf{C}_1}, i \le m, \tag{12}$$

$$\sigma_{\mathsf{C}_2,\mathsf{X_i}|\hat{\mathbf{X}}} = \sigma_{\mathsf{X_i}|\hat{\mathbf{X}}} f_{\mathsf{X_i}\to\mathsf{C}_2}, i \le m. \tag{13}$$

We further have $\sigma_{\mathsf{C}_1,\mathsf{X_N}|(\hat{\mathbf{X}}\backslash\mathsf{X_N})} = f_{\mathsf{X_N}\to\mathsf{L}} f_{\mathsf{L}\to\mathsf{C}_1} + \sum_{i=1}^m f_{\mathsf{X_N}\to\mathsf{X_i}} f_{\mathsf{X_i}\to\mathsf{C}_1}$, $\sigma_{\mathsf{C}_2,\mathsf{X_N}|(\hat{\mathbf{X}}\backslash\mathsf{X_N})} = f_{\mathsf{X_N}\to\mathsf{L}} f_{\mathsf{L}\to\mathsf{C}_2} + \sum_{i=1}^m f_{\mathsf{X_N}\to\mathsf{X_i}} f_{\mathsf{X_i}\to\mathsf{C}_2}$, and $\sigma_{\mathsf{X_N},\mathsf{X_i}|(\hat{\mathbf{X}}\backslash\mathsf{X_N})} = f_{\mathsf{X_N}\to\mathsf{X_i}}, i \le m$. We now have $3m+3$ edge coefficients and $3m+3$ equations, and by these equations, we can solve these edge coefficients up to t (i.e., all these edge coefficients are functions of t).

Finally we solve t by considering variables in $(\hat{\mathbf{X}}\backslash\mathsf{X_N}) = \{\hat{\mathsf{X}}_j\}_{j=2}^p$. Specifically, we have $f_{\hat{\mathsf{X}}_j\to\mathsf{X_i}} = \sigma_{\hat{\mathsf{X}}_j,\mathsf{X_i}|(\hat{\mathbf{X}}\backslash\hat{\mathsf{X}}_j)}$, where $i \le m, 2 \le j \le p$. We also have $f_{\hat{\mathsf{X}}_j\to\mathsf{L}} = \sigma_{\hat{\mathsf{X}}_j,\mathsf{L}|(\hat{\mathbf{X}}\backslash\hat{\mathsf{X}}_j)}(t)$, where $2 \le j \le p$. By the unit variance of L, we further have $t + \sum_{j=1}^p f^2_{\hat{\mathsf{X}}_j\to\mathsf{L}}(t) = 1$, by which t can be solved. Therefore, all edge coefficients among $\mathbf{V}, \hat{\mathbf{X}}, \mathsf{C}_1, \mathsf{C}_2$ can be identified.

Example. Take the atomic cover $\{\mathsf{L}_2, \mathsf{X}_6, \mathsf{X}_7\}$ in Figure 6 as an example. We assume that $\sigma_{\epsilon_{\mathsf{L}_2}} = t$, and thus we have $\sigma_{\mathsf{X}_{12},\mathsf{X}_{13}|\mathsf{X}_3,\mathsf{X}_4,\mathsf{X}_6,\mathsf{X}_7} = t f_{\mathsf{L}_2\to\mathsf{X}_{12}} f_{\mathsf{L}_2\to\mathsf{X}_{13}}$, and $\sigma_{\mathsf{X}_{12},\mathsf{X_i}|\mathsf{X}_3,\mathsf{X}_4} = \sigma_{\mathsf{X_i}|(\mathsf{X}_3,\mathsf{X}_4)} f_{\mathsf{X_i}\to\mathsf{X}_{12}}, 6 \le i \le 7$ and $\sigma_{\mathsf{X}_{13},\mathsf{X_i}|\mathsf{X}_3,\mathsf{X}_4} = \sigma_{\mathsf{X_i}|(\mathsf{X}_3,\mathsf{X}_4)} f_{\mathsf{X_i}\to\mathsf{X}_{13}}, 6 \le i \le 7$. We further have $\sigma_{\mathsf{X}_{12},\mathsf{X}_3|\mathsf{X}_4} = f_{\mathsf{X}_3\to\mathsf{L}_2} f_{\mathsf{L}_2\to\mathsf{X}_{12}} + \sum_{i=6}^7 f_{\mathsf{X}_3\to\mathsf{X_i}} f_{\mathsf{X_i}\to\mathsf{X}_{12}}$, $\sigma_{\mathsf{X}_{13},\mathsf{X}_3|\mathsf{X}_4} = f_{\mathsf{X}_3\to\mathsf{L}_2} f_{\mathsf{L}_2\to\mathsf{X}_{13}} + \sum_{i=6}^7 f_{\mathsf{X}_3\to\mathsf{X_i}} f_{\mathsf{X_i}\to\mathsf{X}_{13}}$, and $\sigma_{\mathsf{X}_3,\mathsf{X_i}|\mathsf{X}_4} = f_{\mathsf{X}_3\to\mathsf{X_i}}, 6 \le i \le 7$. We now have 9 edge coefficients and 9 equations, and thus by these equations, we can solve these edge coefficients up to t. Next we solve t by considering X_4. Specifically, we have $f_{\mathsf{X}_4\to\mathsf{X_i}} = \sigma_{\mathsf{X}_4,\mathsf{X_i}|\mathsf{X}_3}$, where $6 \le i \le 7$. We also have $f_{\mathsf{X}_4\to\mathsf{L}_2} = \sigma_{\mathsf{X}_4,\mathsf{L}_2|\mathsf{X}_3}(t)$. By these 3 equations we can solve $f_{\mathsf{X}_4\to\mathsf{L}_2}$, $f_{\mathsf{X}_4\to\mathsf{X}_6}$, and $f_{\mathsf{X}_4\to\mathsf{X}_7}$, up to t. Till now all edge coefficients among $\mathsf{X}_3, \mathsf{X}_4, \mathsf{L}_2, \mathsf{X}_6, \mathsf{X}_7, \mathsf{X}_{12}, \mathsf{X}_{13}$ can be taken as a function of t. Finally by the unit variance of L, we further have $t + f^2_{\mathsf{X}_3\to\mathsf{L}_2}(t) + f^2_{\mathsf{X}_4\to\mathsf{L}_2}(t) = 1$, by which t can be solved. Therefore, all edge coefficients among $\mathsf{X}_3, \mathsf{X}_4, \mathsf{L}_2, \mathsf{X}_6, \mathsf{X}_7, \mathsf{X}_{12}, \mathsf{X}_{13}$ can be identified.

Taking (a), (b), (c), (d) into consideration, for a graph that satisfies the conditions in Theorem 3, for an atomic cover $\mathbf{V}$ in the graph, if all pure children of it are observed, then all the edge coefficients from $\mathbf{V}$ to its pure children can be identified.

Now, we will prove by induction to show that, for a graph that satisfies the conditions in Theorem 3, for any atomic cover $\mathbf{V}$ in the graph, all the edge coefficients from $\mathbf{V}$ to its children can be identified, and thus all the edge coefficients of the graph can be identified (the set of all edge coefficients in the graph is the union of the set of edge coefficients from each $\mathbf{V}$ to each $\mathbf{V}$'s children).

To this end, we first index all the atomic covers by the inverse causal ordering, such that leaf nodes have smaller indexes. Then we have a sequence of atomic covers $\mathbf{V_i}, i = 1, ..., C$ in the graph, where C is the number of atomic covers in the graph.

(i) We show for $\mathbf{V_i}, i = 1$, all the edge coefficients from $\mathbf{V_1}$ to its children can be identified. This is proved by considering (a) (b) (c) (d), as $\mathbf{V_1}$'s children must be all observed; otherwise it cannot be indexed as 1.

(ii) We show that, for $i > 1$, if for all $\mathbf{V_j}, 1 \le j < i$, all the edge coefficients from $\mathbf{V_j}$ to $\mathbf{V_j}$'s children has been identified, then all the edge coefficients from $\mathbf{V_i}$ to $\mathbf{V_i}$'s children can also be identified. This can be proved by combining (a) (b) (c) (d) with Lemma 1. If $\mathbf{V_i}$ has children that are latent, then the latent children must belong to an atomic cover with a smaller index. Therefore, as all the edge coefficients from $\mathbf{V_j}$ to $\mathbf{V_j}$'s children have been identified, by the use of Lemma 1, the latent children of $\mathbf{V_i}$ can be taken as if they are observed. Therefore, all the edge coefficients from $\mathbf{V_i}$ to $\mathbf{V_i}$'s children can also be identified.

Taking (i) and (ii) together, all the edge coefficients of the graph can be identified.

Example. An example would be in Figure 6. As illustrated in the example in (d), edge coefficients from $\{\mathsf{L}_2, \mathsf{X}_6, \mathsf{X}_7\}$ to $\{\mathsf{X}_{12}, \mathsf{X}_{13}\}$ can be identified. By Lemma 1, L_2 can be taken as if it is observed. The reason lies in that other than X_{12} and X_{13}, there is no descendant of $\{\mathsf{L}_2, \mathsf{X}_6, \mathsf{X}_7\}$; therefore, from all other variables' perspective, e.g., X_{11}, $\sigma_{\mathsf{X}_{11},\mathsf{L}_2}$ can be calculated as

$(\sigma_{\mathsf{X}_{11},\mathsf{X}_{12}} - \Sigma_{i=6}^{7}\sigma_{\mathsf{X}_i,\mathsf{X}_{11}} f_{\mathsf{X}_i\to\mathsf{X}_{12}})/f_{\mathsf{L}_2\to\mathsf{X}_{12}}$. In other words, $\sigma_{\mathsf{X}_{11},\mathsf{L}_2}$ can be calculated using observational data, so L_2 can be taken as observed from X_{11}'s perspective. Plus, by the example in (c), we know that all edge coefficients from $\{\mathsf{L}_1, \mathsf{X}_5\}$ to $\{\mathsf{X}_8, \mathsf{X}_9, \mathsf{X}_{10}\}$ can be identified and thus similarly by Lemma 1, L_1 can be taken as if it can be observed. Finally, we consider edge coefficients among $\mathsf{L}_1, \mathsf{X}_5, \mathsf{X}_3, \mathsf{X}_4, \mathsf{L}_2, \mathsf{X}_6, \mathsf{X}_7$. In this situation, L_1 and L_2 can be taken as if they can be observed and the other variables are themself observed, therefore the edge coefficients among $\mathsf{L}_1, \mathsf{X}_5, \mathsf{X}_3, \mathsf{X}_4, \mathsf{L}_2, \mathsf{X}_6, \mathsf{X}_7$ can be easily identified. □

A.4 Proof of Theorem 4

Lemma 2. *Let $\Sigma_{\mathbf{X}}$ be the observed covariance matrix entailed by $A, B, C, D, \Omega_{\mathbf{X}}, \Omega_{\mathbf{L}}$. Let Q be an orthogonal matrix, and*

$$\tilde{A} \coloneqq Q^T A Q, \quad \tilde{B} \coloneqq Q^T B, \quad \tilde{C} \coloneqq CQ, \quad \tilde{D} \coloneqq D, \quad \tilde{\Omega}_{\mathbf{L}} \coloneqq Q^T \Omega_{\mathbf{L}} Q, \quad \textit{and} \quad \tilde{\Omega}_{\mathbf{X}} \coloneqq \Omega_{\mathbf{X}}.$$

Then, the matrices $\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{\Omega}_{\mathbf{X}}$, and $\tilde{\Omega}_{\mathbf{L}}$ can also entail the covariance matrix $\Sigma_{\mathbf{X}}$.

Proof of Lemma 2. Let M and N be matrices defined as in Proposition 1, and similarly for $\tilde{M}$ and $\tilde{N}$. We then have

$$\begin{aligned}
\tilde{M} &= \left(I - \tilde{A} - \tilde{B}(I-\tilde{D})^{-1}\tilde{C}\right)^{-1} \\
&= \left(Q^TQ - Q^TAQ - (Q^TB)(I-D)^{-1}(CQ)\right)^{-1} \\
&= Q^{-1}\left(I - A - B(I-D)^{-1}C\right)^{-1} Q^{-T} \\
&= Q^T M Q
\end{aligned}$$

and

$$\begin{aligned}
\tilde{N} &= \left((I-\tilde{A})\tilde{C}^{-1}(I-\tilde{D}) - \tilde{B}\right)^{-1} \\
&= \left((Q^TQ - Q^TAQ)(CQ)^{-1}(I-D) - Q^TB\right)^{-1} \\
&= \left((I-A)C^{-1}(I-D) - B\right)^{-1} Q^{-T} \\
&= NQ.
\end{aligned}$$

By Proposition 1, the latent covariance matrix $\tilde{\Sigma}_{\mathbf{L}}$ is given by

$$\begin{aligned}
\tilde{\Sigma}_{\mathbf{L}} &= \tilde{M}^T\tilde{\Omega}_{\mathbf{L}}\tilde{M} + \tilde{N}^T\tilde{\Omega}_{\mathbf{X}}\tilde{N} \\
&= (Q^TM^TQ^{-T})(Q^T\Omega_{\mathbf{L}}Q)(Q^{-1}MQ) + Q^TN^T\Omega_{\mathbf{X}}NQ \\
&= Q^T(M^T\Omega_{\mathbf{L}}M + N^T\Omega_{\mathbf{X}}N)Q \\
&= Q^T\Sigma_{\mathbf{L}}Q.
\end{aligned}$$

By Proposition 1, the observed covariance matrix $\tilde{\Sigma}_{\mathbf{X}}$ is given by

$$\begin{aligned}
\tilde{\Sigma}_{\mathbf{X}} &= (I-\tilde{D})^{-T}\left(\tilde{B}^T\tilde{\Sigma}_{\mathbf{L}}B + \tilde{\Omega}_{\mathbf{X}} + \tilde{\Omega}_{\mathbf{X}}\tilde{N}\tilde{B} + \tilde{B}^T\tilde{N}^T\tilde{\Omega}_{\mathbf{X}}\right)(I-\tilde{D})^{-1} \\
&= (I-D)^{-T}\left((Q^TB)^T(Q^T\Sigma_{\mathbf{L}}Q)(Q^TB) + \Omega_{\mathbf{X}} + \Omega_{\mathbf{X}}(NQ)(Q^TB) + (Q^TB)^T(NQ)^T\Omega_{\mathbf{X}}\right)(I-D)^{-1} \\
&= (I-D)^{-T}\left(B^T\Sigma_{\mathbf{L}}B + \Omega_{\mathbf{X}} + \Omega_{\mathbf{X}}NB + B^TN^T\Omega_{\mathbf{X}}\right)(I-D)^{-1} \\
&= \Sigma_{\mathbf{X}}.
\end{aligned}$$

This indicates that the matrices $\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{\Omega}_{\mathbf{X}}$, and $\tilde{\Omega}_{\mathbf{L}}$ can also entail the covariance matrix $\Sigma_{\mathbf{X}}$. □

Using Lemma 2, we can prove Theorem 4.

Theorem 4 (Condition for the Existence of Orthogonal Transformation Indeterminacy)**.** *Consider the model in Definition 1. If a set of latent variables $\mathbf{L}$ with $|\mathbf{L}| \geq 2$, have the same parents and children, then there must exist orthogonal transformation indeterminacy regarding the edge coefficients F. In other words, F can at most be identified up to orthogonal transformation indeterminacy.*

Proof of Theorem 4. Let $\mathbf{S}_1$, $\mathbf{S}_2$, $\mathbf{S}_3$, $\mathbf{S}_4$, and $\mathbf{S}_5$ be the indices of $\mathbf{L}$, their latent parents, their latent children, their measured parents, and their measured children in $\mathcal{G}$, respectively. Let Q be a $|\mathbf{L}| \times |\mathbf{L}|$ orthogonal matrix. For matrices A, B, C, and D from matrix F, suppose that we replace $A_{\mathbf{S}_2,\mathbf{S}_1}$, $A_{\mathbf{S}_1,\mathbf{S}_3}$, $C_{\mathbf{S}_4,\mathbf{S}_1}$, and $B_{\mathbf{S}_1,\mathbf{S}_5}$ with $A_{\mathbf{S}_2,\mathbf{S}_1}Q$, $Q^T A_{\mathbf{S}_1,\mathbf{S}_3}$, $C_{\mathbf{S}_4,\mathbf{S}_1}Q$, and $Q^T B_{\mathbf{S}_1,\mathbf{S}_5}$, respectively. Then, we will show that the entailed covariance matrix $\Sigma_{\mathbf{X}}$ is unchanged.

Let U be an $(m+n) \times (m+n)$ orthogonal matrix such that: (i) $U_{\mathbf{S}_1,\mathbf{S}_1} = Q$, (ii) the remaining diagonal entries are ones, and (iii) the remaining non-diagonal entries are zeros. Let

$$\tilde{A} \coloneqq U^T A U, \quad \tilde{B} \coloneqq U^T B, \quad \tilde{C} \coloneqq CU, \quad \tilde{D} \coloneqq D, \quad \tilde{\Omega}_{\mathbf{L}} \coloneqq U^T \Omega_{\mathbf{L}} U = I, \quad \text{and} \quad \tilde{\Omega}_{\mathbf{X}} \coloneqq \Omega_{\mathbf{X}}.$$

By Lemma 2, the matrices above can entail the same covariance matrix $\Sigma_{\mathbf{X}}$.

By construction of U, left multiplication of U^T on B only affects $B_{\mathbf{S}_1,*}$; specifically, it is equivalent to replacing $B_{\mathbf{S}_1,*}$ with $Q^T B_{\mathbf{S}_1,*}$. Furthermore, only the columns of $\mathbf{S}_5$ in $B_{\mathbf{S}_1,*}$ will be affected, because those columns correspond to the measured children of $\mathbf{L}$. Therefore, all entries of $\tilde{B}$ are the same as B, except that $B_{\mathbf{S}_1,\mathbf{S}_5}$ is replaced with $Q^T B_{\mathbf{S}_1,\mathbf{S}_5}$. Similar reasoning shows that all entries of $\tilde{C}$ are the same as C, except that $C_{\mathbf{S}_4,\mathbf{S}_1}$ is replaced with $C_{\mathbf{S}_4,\mathbf{S}_1}Q$.

Now consider $U^T A U$. By the reasoning above, left multiplication of U^T on A only is equivalent to replacing $A_{\mathbf{S}_1,\mathbf{S}_3}$ with $Q^T A_{\mathbf{S}_1,\mathbf{S}_3}$. Further right multiplication of U on $U^T A$ is equivalent to replacing $(U^T A)_{\mathbf{S}_2,\mathbf{S}_1}$ with $(U^T A)_{\mathbf{S}_2,\mathbf{S}_1}Q$. Since $\mathbf{S}_1$, $\mathbf{S}_2$, and $\mathbf{S}_3$ are mutually disjoint, all entries of $\tilde{A} = U^T A U$ are the same as A, except that $A_{\mathbf{S}_2,\mathbf{S}_1}$ and $A_{\mathbf{S}_1,\mathbf{S}_3}$ are replaced with $A_{\mathbf{S}_2,\mathbf{S}_1}Q$ and $Q^T A_{\mathbf{S}_1,\mathbf{S}_3}$, respectively.

Hence, for matrices A, B, C, and D, suppose we replace $A_{\mathbf{S}_2,\mathbf{S}_1}$, $A_{\mathbf{S}_1,\mathbf{S}_3}$, $C_{\mathbf{S}_4,\mathbf{S}_1}$, and $B_{\mathbf{S}_1,\mathbf{S}_5}$ with $A_{\mathbf{S}_2,\mathbf{S}_1}Q$, $Q^T A_{\mathbf{S}_1,\mathbf{S}_3}$, $C_{\mathbf{S}_4,\mathbf{S}_1}Q$, and $Q^T B_{\mathbf{S}_1,\mathbf{S}_5}$, respectively. By the reasoning above, this is equivalent to replacing A, B, C, and D with $\tilde{A}$, $\tilde{B}$, $\tilde{C}$, and $\tilde{D}$, respectively, which share the same support and entail the same covariance matrix $\Sigma_{\mathbf{X}}$. □

A.5 Proof of Corollary 1 and Corollary 2

Corollary 1 (General Necessary Condition for Parameter Identifiability)**.** *For parameters to be identifiable, every pair of latent variables has to have at least one different parent or child.*

Proof of Corollary 1. Proof by contradiction. If it is not the case that every pair of latent variables has to have at least one different parent or child, then there exist $\mathbf{L}$ such that $|\mathbf{L}| \geq 2$ and $\mathbf{L}$ share the same parents and children. Therefore by Theorem 4 there must exist orthogonal transformation indeterminacy regarding F, and thus the parameters are not identifiable. □

Corollary 2 (Necessary Condition about Atomic Covers for Parameter Identifiability)**.** *Assume $\mathcal{G}$ satisfies Conditions 1 and 2 and thus the structure can be identified up to the MEC of $\mathcal{O}_{min}(\mathcal{O}_s(\mathcal{G}))$. For any DAG $\mathcal{G}$ in the equivalence class, for $\mathcal{G}$'s parameters to be identifiable, every atomic cover must contain no more than one latent variable.*

Proof of Corollary 2. Proof by contradiction. If for a DAG in the equivalence class, there is an atomic cover that has more than one latent variable, then according to the definition of the concerned equivalence class, the latent variables in that atomic cover share the same parents and children. Then by Theorem 4 there must exist orthogonal transformation indeterminacy regarding F, and thus the parameters are not identifiable. □

A.6 Proof of Proposition 1

Proposition 1 (Parameterization of Population Covariance Matrix)**.** *Consider the model defined in Definition 1. Let $M \coloneqq \left(I - A - B(I-D)^{-1}C\right)^{-1}$ and $N \coloneqq \left((I-A)C^{-1}(I-D) - B\right)^{-1}$. Then, the population covariance matrices of $\mathbf{L}_{\mathcal{G}}$ and $\mathbf{X}_{\mathcal{G}}$ can be formulated as*

$$\Sigma_{\mathbf{L}_{\mathcal{G}}} = M^T \Omega_{\mathbf{L}_{\mathcal{G}}} M + N^T \Omega_{\mathbf{X}_{\mathcal{G}}} N, \tag{1}$$

$$\Sigma_{\mathbf{X}_{\mathcal{G}}} = (I-D)^{-T}\left(B^T \Sigma_{\mathbf{L}_{\mathcal{G}}} B + \Omega_{\mathbf{X}_{\mathcal{G}}} + \Omega_{\mathbf{X}_{\mathcal{G}}} N B + B^T N^T \Omega_{\mathbf{X}_{\mathcal{G}}}\right)(I-D)^{-1}. \tag{2}$$

Proof of Proposition 1. Since matrices A and D are invertible, using the formula of 2×2 block matrix inversion [22, Chapter 0.7], we obtain

$$(I-F)^{-1} = \begin{pmatrix} M & -MB(I-D)^{-1} \\ -(I-D)^{-1}CM & (I-D)^{-1} + (I-D)^{-1}CMB(I-D)^{-1} \end{pmatrix},$$

which implies

$$(I-F)^{-T} = \begin{pmatrix} M^T & -M^TC^T(I-D)^{-T} \\ -(I-D)^{-T}B^TM^T & (I-D)^{-T} + (I-D)^{-T}B^TM^TC^T(I-D)^{-T} \end{pmatrix}$$

and

$$(I-F)^{-T}\Omega = \begin{pmatrix} M^T\Omega_{\mathbf{L}} & -M^TC^T(I-D)^{-T}\Omega_{\mathbf{X}} \\ -(I-D)^{-T}B^TM^T\Omega_{\mathbf{L}} & (I-D)^{-T}\Omega_{\mathbf{X}} + (I-D)^{-T}B^TM^TC^T(I-D)^{-T}\Omega_{\mathbf{X}} \end{pmatrix}.$$

We then have

$$\begin{aligned} \Sigma_L &= M^T\Omega_{\mathbf{L}}M + M^TC^T(I-D)^{-T}\Omega_{\mathbf{X}}(I-D)^{-1}CM \\ &= M^T\Omega_{\mathbf{L}}M + N^T\Omega_{\mathbf{X}}N \end{aligned}$$

and

$$\begin{aligned} \Sigma_X &= (I-D)^{-T}B^TM^T\Omega_{\mathbf{L}}MB(I-D)^{-1} + (I-D)^{-T}\Omega_{\mathbf{X}}(I-D)^{-1} \\ &\quad + (I-D)^{-1}\Omega_{\mathbf{X}}(I-D)^{-1}CMB(I-D)^{-1} + (I-D)^{-T}B^TM^TC^T(I-D)^{-T}\Omega_{\mathbf{X}}(I-D)^{-1} \\ &\quad + (I-D)^{-T}B^TM^TC^T(I-D)^{-T}\Omega_{\mathbf{X}}(I-D)^{-1}CMB(I-D)^{-1} \\ &= (I-D)^{-T}\Big(\Omega_{\mathbf{X}} + B^TM^T\Omega_{\mathbf{L}}MB + \Omega_{\mathbf{X}}(I-D)^{-1}CMB + B^TM^TC^T(I-D)^{-T}\Omega_{\mathbf{X}} \\ &\qquad + B^TM^TC^T(I-D)^{-T}\Omega_{\mathbf{X}}(I-D)^{-1}CMB\Big)(I-D)^{-1} \\ &= (I-D)^{-T}\Big(\Omega_{\mathbf{X}} + B^T\Sigma_{\mathbf{L}}B + \Omega_{\mathbf{X}}NB + B^TN^T\Omega_{\mathbf{X}}\Big)(I-D)^{-1}. \end{aligned}$$

□

Note that matrices $I-D$ and $I-F$ are invertible because structure $\mathcal{G}$ is acyclic. This implies $\det(I-F) \neq 0$ and $\det(I-D) \neq 0$. Define

$$U = \begin{pmatrix} I & 0 \\ -(I-D)^{-1}C & I \end{pmatrix},$$

which implies

$$(I-F)U = \begin{pmatrix} M & B \\ 0 & I-D \end{pmatrix}$$

and thus

$$\det((I-F)U) = \det(M)\det(I-D).$$

Since $\det(U) = 1$ and $\det(I-F) \neq 0$, we have

$$\det((I-F)U) = \det(I-F)\det(U) \neq 0.$$

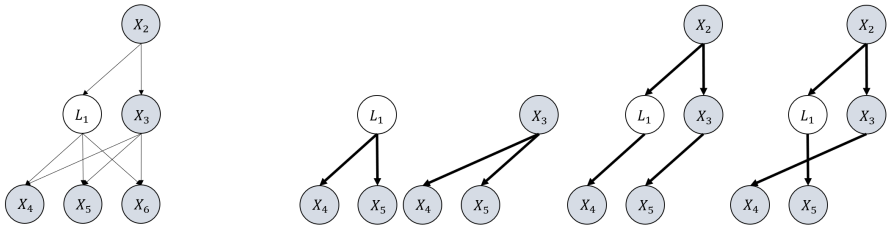
By the statement above and $\det(I-D) \neq 0$, we have

$$\det(M) = \frac{\det((I-F)U)}{\det(I-D)} \neq 0,$$

which implies that M is invertible. Similar reasoning can be used to show that N is invertible.

A.7 Computational Cost of Checking Whether the Conditions in Theorem 3 Hold

Here we want to investigate, given a structure, can we efficiently check whether the proposed sufficient conditions hold? To this end, we generate random graphs and each graph has 100 variables. According to our empirical result, such a check can be done very efficiently. Specifically, on average, given a structure with 100 variables, it only takes our Python code around 3 seconds to check whether the conditions hold.



(a) Graph $\mathcal{G}$ to show treks. (b) The four simple treks between X_4 and X_5 in (a).

Figure 7: Illustrative figure to show how to parameterize $\Sigma_{\mathbf{X}_\mathcal{G}}$ by the use of simple trek rule.

A.8 In Practice, What If the Conditions Do not Hold?

Our condition is useful in solving real-life problems. For example, in the psychometric study, we can properly design the questions with domain knowledge following the condition in Theorem 3 such that each single latent variable has enough observed variables as pure children and thus it can be ensured that all parameters are identifiable (as illustrated in our real-life data result in Figure 4).

On the other hand, even though sometimes the questionnaires and data were designed not so well such that the conditions are not satisfied for the identification of parameters, our Theorem 3 is still useful. In this case, we can still make use of our conditions to check the given structure, and find some local sub-structures where our conditions are satisfied. Consequently, it can be ensured that all the parameters of some sub-structures are identifiable, and we can employ our estimation method to find all the edge coefficients of these sub-structures.

B Additional Definitions, Graphs, Results, and Examples

B.1 Example of Treks

Example 7 (Example of Treks)**.** *In Figure 7 (a), there are four treks between* X_4 *and* X_5*: (i)* $\mathsf{X}_4 \leftarrow \mathsf{L}_1 \rightarrow \mathsf{X}_5$*, (ii)* $\mathsf{X}_4 \leftarrow \mathsf{X}_3 \rightarrow \mathsf{X}_5$*, (iii)* $\mathsf{X}_4 \leftarrow \mathsf{L}_1 \leftarrow \mathsf{X}_2 \rightarrow \mathsf{X}_3 \rightarrow \mathsf{X}_5$*, and (iv)* $\mathsf{X}_4 \leftarrow \mathsf{X}_3 \leftarrow \mathsf{X}_2 \rightarrow \mathsf{L}_1 \rightarrow \mathsf{X}_5$*, illustrated in Figure 7 (b).*

B.2 Definition of Pure Children

Definition 6 (Pure Children [18])**.** $\mathbf{Y}$ *are pure children of variables* $\mathbf{X}$ *in graph* $\mathcal{G}$*, iff* $Pa_\mathcal{G}(\mathbf{Y}) = \cup_{\mathsf{Y}_i \in \mathbf{Y}} Pa_\mathcal{G}(\mathsf{Y}_i) = \mathbf{X}$ *and* $\mathbf{X} \cap \mathbf{Y} = \emptyset$*. We denote the pure children of* $\mathbf{X}$ *in* $\mathcal{G}$ *by* $PCh_\mathcal{G}(\mathbf{X})$*.*

B.3 Definition of Rank-invariant Graph Operator

The definitions are as follows with examples.

Definition 7 (Minimal-Graph Operator [24, 18])**.** *Given two atomic covers* $\mathbf{L}, \mathbf{P}$ *in* $\mathcal{G}$*, we can merge* $\mathbf{L}$ *to* $\mathbf{P}$ *if the following conditions hold: (i)* $\mathbf{L}$ *is the pure children of* $\mathbf{P}$*, (ii) all elements of* $\mathbf{L}$ *and* $\mathbf{P}$ *are latent and* $|\mathbf{L}| = |\mathbf{P}|$*, and (iii) the pure children of* $\mathbf{L}$ *form a single atomic cover, or the siblings of* $\mathbf{L}$ *form a single atomic cover. We denote such an operator as minimal-graph operator* $\mathcal{O}_{min}(\mathcal{G})$*.*

Definition 8 (Skeleton Operator [24, 18])**.** *Given an atomic cover* $\mathbf{V}$ *in a graph* $\mathcal{G}$*, for all* $\mathsf{V} \in \mathbf{V}$*,* V *is latent, and all* $\mathsf{C} \in PCh_\mathcal{G}(\mathbf{V})$*, such that* V *and* C *are not adjacent in* $\mathcal{G}$*, we can draw an edge from* V *to* C*. We denote such an operator as skeleton operator* $\mathcal{O}_s(\mathcal{G})$*.*

Suppose a graph $\mathcal{G}$ of a latent linear model in Figure 9(a) is $\mathcal{G}$. After applying $\mathcal{O}_{\min}(\mathcal{O}_s(\mathcal{G}))$, we have the graph in Figure 9(b). Specifically, the $\mathcal{O}_s$ operator adds an edge from L_2 to X_5 and the $\mathcal{O}_{\min}$ operator delete L_4 and add an edge from L_1 directly to X_6 and X_7. For $\mathcal{G}$, such two operators will not change the rank in the infinite sample case.

B.4 Graphs for Synthetic Data Experiments

Please refer to Figures 10 and 11.

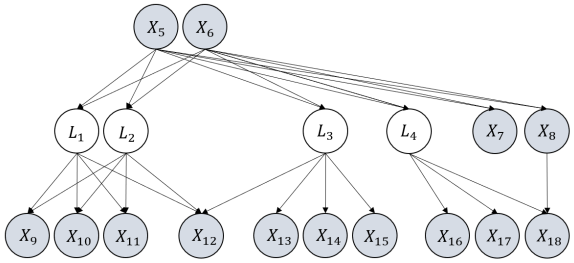
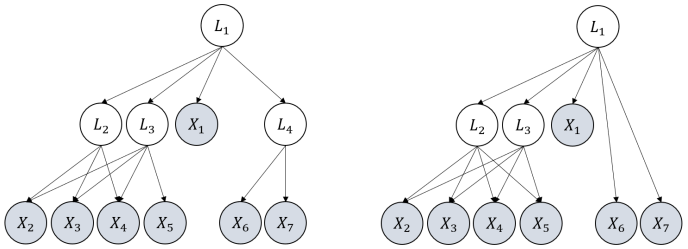


Figure 8: An illustrative graph to show orthogonal transformation indeterminacy. An atomic cover of it, $\{\mathsf{L}_1, \mathsf{L}_2\}$, has more than one latent variable, and thus there exists orthogonal transformation indeterminacy regarding coefficients of edges that involve $\{\mathsf{L}_1, \mathsf{L}_2\}$.



(a) The original graph $\mathcal{G}$. (b) After operators $\mathcal{O}_{\text{min}}$ and $\mathcal{O}_s$.

Figure 9: Example to show graph operators $\mathcal{O}_{\text{min}}$ and $\mathcal{O}_s$ [18].

B.5 Additional Result under Model Misspecification

Please refer to Tables 2and 3.

C Other Discussions

Our optimization problem in Eq. 4 is solved by gradient descent using PyTorch. Our current implementation is based on CPU but it can be further accelerated by using GPU. A very related discussion can also be found in [35].

The optimization problem in Eq.3 is solved by gradient descent, which involves evaluating the LogDet and matrix inverse (for the gradient) terms (which is similar to continuous causal discovery methods based on Gaussian likelihood [35]). According to [53], the computational complexity is $O(td^3)$, where d is the number of variables and t is the number of iterations of gradient descent respectively. Note that the computational cost is largely independent of the sample size as we only need to calculate the sample covariance once and save it for further use.

It is possible to perform inference on the learned parameters in our framework. To be specific, as we use maximum likelihood estimation for the parameters, some standard techniques can be readily used. For example, bootstrap can be employed to provide standard errors on linear coefficients and Chi-square test can also be done to examine the fitness of the model.

D Extended Related Work

One main line of research in latent variable estimation centers on factor-analysis-based methods. Representative studies include [41, 45, 36, 56, 7, 57]. Various other techniques have also been employed for latent structure and parameter identification, including over-complete ICA-based techniques [23, 43, 1] that leverage non-Gaussianity and matrix decomposition-based approaches [3]. However, these approaches typically consider latent variables with observed children, without considering parameter identification in latent hierarchical structures. A more related work is [4], but it considers a much simpler structure.

Table 2: Performance under violation of normality using uniform noise terms in MSE (mean (std)).

Metric		MSE up to group sign	
Method		Estimator	Estimator-TR
GS Case	2k	0.0017	0.0005
	5k	0.0018	0.0004
	10k	0.0018	0.0003

Table 3: Performance under violation of linearity using leaky relu in MSE (mean (std)).

Metric		MSE up to group sign	
Method		Estimator	Estimator-TR
GS Case with 10k sample size	$\alpha = 0.8$ (close to linear)	0.004	0.001
	$\alpha = 0.6$ (quite nonlinear)	0.013	0.005
	$\alpha = 0.3$ (very nonlinear)	0.046	0.027

Another direction would be to project the graph to an ADMG and the latent confounding effects are encoded by correlated noise terms. Following this idea, graphical criteria such as half-trek [20, 5], G-criterion [9], and some further developments [51, 29, 42, 8, 19] has been proposed. Furthermore, another line of works involve studies on causal effect estimation in the presence of latent confounders [52, 6, 30, 33, 26], which often rely on instrumental variables or proxy variables for identification. Notice that in this task, the parameters may not be identified [30], although the causal effect from the treatment variable to the outcome variable can be identified.

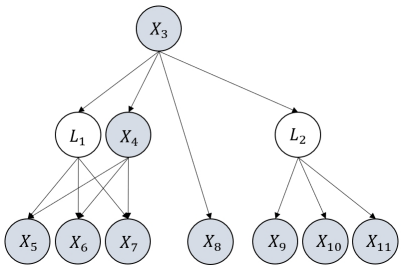
Furthermore, several existing works also solve an optimization problem that involves parameterization of maximum likelihood, such as those in continuous optimization for causal discovery [35, 59, 31, 10, 34] and parameter estimation of Lyapunov models [55, 15]. Differently, our formulation involving likelihood parameterization aims to estimate parameters of partially observed linear causal models.

E Limitations

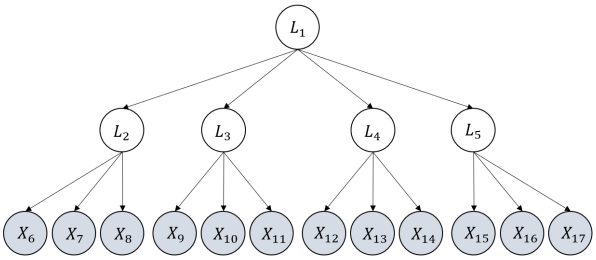
One limitation of this work is that our theoretical results are based on the assumption of linear gaussian causal models. When data is not linear gaussian, we have also conducted experiments to see the performance of our method. It turns out that our method still performs well in the presence of certain extents of violation of normality and linearity. However, theoretical analysis under violation of linearity and normality would be interesting and the focus of future work.

F Broader Impacts

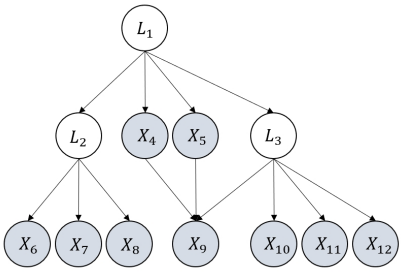
The goal of this paper is to advance the field of machine learning. We do not see any potential negative societal impacts of the work.



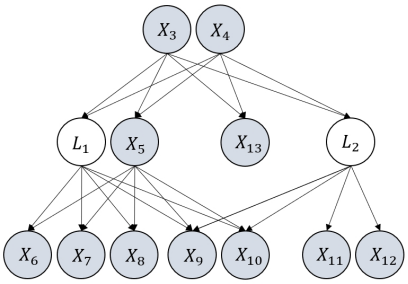
(a) Illustrative graph $\mathcal{G}_1$.



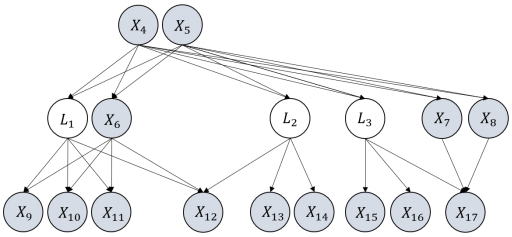
(b) Illustrative graph $\mathcal{G}_2$



(c) Illustrative graph $\mathcal{G}_3$.

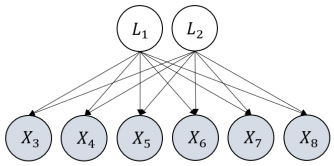


(d) Illustrative graph $\mathcal{G}_4$

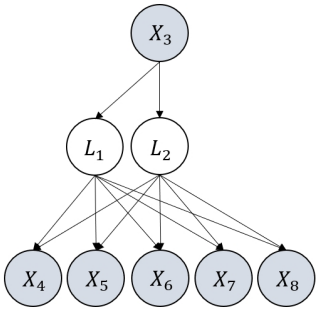


(e) Illustrative graph $\mathcal{G}_5$

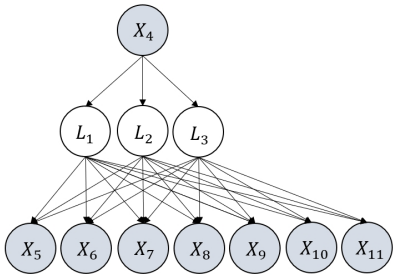
Figure 10: Examples of graphs in the GS case. The parameters of them are identifiable up to group sign indeterminacy.



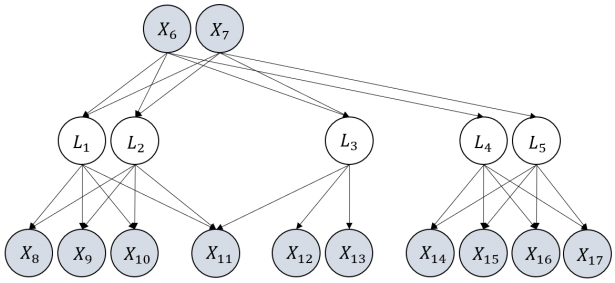
(a) Illustrative graph $\mathcal{G}_6$.



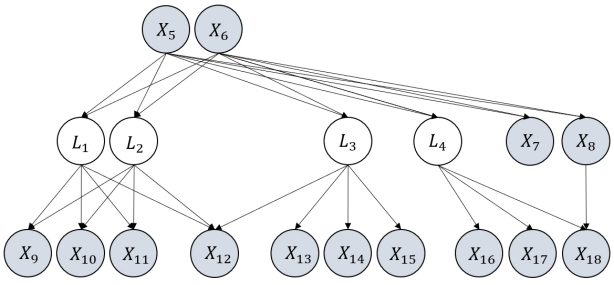
(b) Illustrative graph $\mathcal{G}_7$



(c) Illustrative graph $\mathcal{G}_8$.



(d) Illustrative graph $\mathcal{G}_9$



(e) Illustrative graph $\mathcal{G}_{10}$

Figure 11: Examples of graphs in the OT case. Parameters of them are Identifiable up to group sign and orthogonal transformation indeterminacy.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: the claims stated in the abstract and introduction are the same as what are stated in the theory and method part.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A discussion about the limitations can be found in Appendix E.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All assumptions together with necessary definitions are provided in the main text and all complete proofs are provided in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The implementation details and the experimental settings are all provided in Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code and the data will be publicly available.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See the experiment section for the setting and details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The synthetic experiments were supported by error bars to show statistical significance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Appendix F.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: It is not related to this work. No such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the assets are properly credited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Not involved.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Not involved.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.