Score-Based Causal Discovery of Latent Variable Causal Models

Ignavier Ng [*1] **Xinshuai Dong** [*1] **Haoyue Dai** [1] **Biwei Huang** [2] **Peter Spirtes** [1] **Kun Zhang** [1 3]

Abstract

Identifying latent variables and the causal structure involving them is essential across various scientific fields. While many existing works fall under the category of constraint-based methods (with e.g. conditional independence or rank deficiency tests), they may face empirical challenges such as testing-order dependency, error propagation, and choosing an appropriate significance level. These issues can potentially be mitigated by properly designed score-based methods, such as Greedy Equivalence Search (GES) (Chickering, 2002) in the specific setting without latent variables. Yet, formulating score-based methods with latent variables is highly challenging. In this work, we develop score-based methods that are capable of identifying causal structures containing causally-related latent variables with identifiability guarantees. Specifically, we show that a properly formulated scoring function can achieve score equivalence and consistency for structure learning of latent variable causal models. We further provide a characterization of the degrees of freedom for the marginal over the observed variables under multiple structural assumptions considered in the literature, and accordingly develop both exact and continuous score-based methods. This offers a unified view of several existing constraint-based methods with different structural assumptions. Experimental results validate the effectiveness of the proposed methods.

1. Introduction

At the core of understanding complex systems lies causal discovery, the identification of causal relations from observational data (Spirtes et al., 2001; Pearl, 2009). One common assumption in causal discovery algorithms is the absence of latent confounders, known as *causal sufficiency*, positing that the observed correlations stem either from true causation or can be sufficiently explained by other observed variables. Yet, real-world scenarios often defy this assumption. For instance, in psychological studies, the measured questionnaires are indirect proxies of latent mental factors. In unstructured data like images and texts, the observed pixels and words are confounded by latent semantic variables. Directly applying causal discovery methods without considering these latent variables can lead to false discoveries, as latent variables may introduce spurious correlations among observed ones that cannot be attributed to true causation.

Notable efforts have thus been made to identify the true causal relations in the presence of latent variables. Earliest attempts include Fast Causal Inference (FCI) (Spirtes et al., 2001; Zhang, 2008) and its variants (Colombo et al., 2012; Spirtes et al., 2013; Claassen et al., 2013; Akbari et al., 2021) that exploit conditional independence information. There are two main limitations of FCI: First, the results, presented by partial ancestral graphs (PAG) (Richardson, 1996), tend to be overgeneralized – e.g., whenever two observed variables may be confounded, it indicates so. Second, it focuses solely on causal relations among observed variables and does not provide information about those among latent variables. In short, FCI does not require specific assumptions about the latent structure, at the cost of having a less informative output. In contrast, one may often be interested in identifying the causal relations among latent variables (e.g., the latent mental and semantic variables in the above examples).

Hence, another line of work has been developed to discover the causal structure also among latent variables. For the identifiability conditions, these methods typically introduce additional parametric assumptions to mitigate the large model indeterminacies faced by FCI. This includes rank or tetrad condition-based methods with linearity assumption (Silva et al., 2003; 2006; Silva & Scheines, 2005; Choi et al., 2011; Kummerfeld & Ramsey, 2016; Huang et al., 2022; Dong et al., 2023), high-order moments-based methods (Shimizu et al., 2009; Zhang et al., 2018; Cai et al., 2019; Salehkaleybar et al., 2020; Xie et al., 2020; Adams et al., 2021; Dai et al., 2022; Chen et al., 2022; Améndola et al., 2023; Wang & Drton, 2023), matrix decomposition-based methods (Anandkumar et al., 2013), copula model-based methods (Cui et al., 2018), mixture oracles-based

[*]Equal contribution [1]Carnegie Mellon University [2]University of California, San Diego [3]Mohamed bin Zayed University of Artificial Intelligence.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

methods (Kivva et al., 2021), and multiple domains-based methods (Zeng et al., 2021; Sturma et al., 2023). For the algorithmic procedures, these methods generally fall under the category of constraint-based methods, by matching the statistical properties to possible structural patterns and constructing the whole causal structure iteratively. A typical constraint-based method in the causally sufficient case is PC (Spirtes & Glymour, 1991). Despite the asymptotic consistency, the empirical reliability of constraint-based methods may be limited due to *testing-order dependency* and *error propagation* (Spirtes, 2010; Colombo et al., 2012), especially when the number of variables is large.

To address such empirical issues of constraint-based methods, score-based causal discovery methods have been introduced, and may be more favored in practical applications (Nandy et al., 2018; Ramsey et al., 2017). Unlike the iterative construction of a single causal graph by constraint-based methods, score-based methods assign a score to each potential graph reflecting how well it explains the observed data and generally search over the graph space to find the optimal graph. In the causally sufficient case, one typical score-based method is the Greedy Equivalence Search (GES) (Chickering, 2002). There also exists several score-based methods that can handle latent variables (Shpitser et al., 2012; Triantafillou & Tsamardinos, 2016; Nowzohour et al., 2017; Bhattacharya et al., 2021; Shahin & Chechik, 2020; Bernstein et al., 2020; Bellot & van der Schaar, 2021; Claassen & Bucur, 2022). Similar to FCI, most of them do not discover the causal relations among latent variables, except the method by Zhang (2004) without identifiability guarantee. When latent variables are introduced and relations among them are further allowed in the causal structures, challenges arise in characterizing the degrees of freedom (Geiger et al., 1996; 2001), formulating a scoring function, and structuring the search procedure. We tackle these challenges in this paper, and to the best of our knowledge, this is the first score-based method that identifies causal structures containing causally-related latent variables with identifiability guarantees.

Contributions. We develop score-based methods, called SALAD (which stands for Score-bAsed Latent cAusal Discovery), for causal discovery of latent variable causal models, providing a unified view for several existing constraint-based methods (Silva et al., 2003; Huang et al., 2022). Our contributions can be summarized as follows:

- We develop a formulation of scoring function for identifying linear latent variable causal models. We show (1) that it is score equivalent and (2) that minimizing it yields a structure that is algebraic equivalent to the true structure. The latter implies that both structures have the same equality constraints (on the marginal over the observed variables), including conditional independence and rank deficiency constraints.
- We provide a characterization of the degrees of freedom for the marginal over the observed variables under the structural assumptions considered by Silva et al. (2003); Huang et al. (2022).
- We develop exact score-based methods for estimating the causal structure, and show that they can asymptotically identify the true equivalence class of the whole structure. We also provide continuous score-based methods in some of the settings to improve the computational efficiency.
- We demonstrate that the proposed score-based methods achieve improved performance over existing constraint-based methods for estimating the structures of latent variable causal models, which further validate the effectiveness of score-based methods.

Notations. For a matrix M, we define its support set as $\mathrm{supp}(M) \coloneqq \{(i,j) : M_{i,j} \neq 0\}$. We denote by $M_{\mathbf{S},:}$ the rows in M indexed by set $\mathbf{S}$, and similarly by $M_{:,\mathbf{S}}$ for the columns. For a directed acyclic graph (DAG) $\mathcal{G}$, we denote by $|\mathcal{G}|$ the number of edges in $\mathcal{G}$. Also, let $\mathrm{diag}(\mathbb{R}^m_{>0})$ be the set of $m \times m$ diagonal matrices with positive diagonal entries, $\mathbb{U}^m$ be the set of $m \times m$ strictly upper triangular matrices, and $\mathbb{G}^m$ be the set of graphs with m measured variables that follow Equation (1). For set $\mathbf{S}$, we define its k-partition as a partition of its elements into k non-empty subsets.

2. Latent Variable Causal Models

In this section, we discuss several aspects of latent variable causal models. Specifically, we describe the preliminaries and problem setting in Section 2.1, as well as the formulation of likelihood function in Section 2.2. We provide a discussion of latent variable causal models in Appendix A.1.

2.1. Preliminaries and Problem Setting

We consider a linear latent variable causal model with DAG $\mathcal{G}$, in which the measured variables $X = (X_1, \ldots, X_m)$ and latent (unmeasured) variables $L = (L_1, \ldots, L_n)$ follow the data generating procedure:

$$L = CL + E_L \quad \text{and} \quad X = BL + E_X, \tag{1}$$

where E_X and E_L are jointly independent noise terms that follow Gaussian distributions. The structure of DAG $\mathcal{G}$ is defined by the support of matrices B and C, i.e., $L_j \to L_i$ is an edge in $\mathcal{G}$ if $C_{i,j} \neq 0$ and $L_j \to X_i$ is an edge in $\mathcal{G}$ if $B_{i,j} \neq 0$. For DAG $\mathcal{G}$, we denote by $B_\mathcal{G} \in \{0,1\}^{m \times n}$ the binary adjacency matrix that represent the edges from latent variables L to measured variables X, and by $C_\mathcal{G} \in \{0,1\}^{n \times n}$ the binary adjacency matrix that represent the edges among latent variables L. Without loss of generality, we assume that matrices C and $C_\mathcal{G}$ are strictly upper triangular.

Let Σ_X and Σ_L be the population covariance matrices of measured variables X and latent variables L respectively. Also let Ω_X and Ω_L be the (diagonal) covariance matrices of noise terms E_X and E_L respectively. Σ_L can be written as

$$\Sigma_L = (I - C)^{-1}\Omega_L(I - C)^{-\top}.$$

By $\Sigma_X = B\Sigma_L B^\top + \Omega_X$, we then have

$$\Sigma_X = B(I - C)^{-1}\Omega_L(I - C)^{-\top}B^\top + \Omega_X. \quad (2)$$

We say that a DAG $\mathcal{G}$ can generate a covariance matrix if there exists a parameterization of $\mathcal{G}$ such that Equation (2) holds. Furthermore, since the labeling of latent variables in general cannot be identified, we say that two DAGs are Markov equivalent if they are Markov equivalent after relabeling of latent variables. Given T i.i.d. samples of variables X, denoted as $\mathbf{D}$ with empirical covariance matrix S, the goal is to estimate the structure $\mathcal{G}$ up to certain type of model equivalence (specified in Sections 4 and 5).

2.2. Formulation of Likelihood Function

We first discuss about the indeterminacy of parameter Ω_L via the following lemma, since it affects how we formulate the likelihood. The proof is given in Appendix B.1.

Lemma 1 (Indeterminacy of Ω_L)**.** *For any parameters B, C, Ω_X, Ω_L, and Σ_X that follow Equation (2), there exist parameters $\tilde{B}$ and $\tilde{C}$ with $\mathrm{supp}(B) = \mathrm{supp}(\tilde{B})$ and $\mathrm{supp}(C) = \mathrm{supp}(\tilde{C})$ such that*

$$\Sigma_X = \tilde{B}(I - \tilde{C})^{-1}(I - \tilde{C})^{-\top}\tilde{B}^\top + \Omega_X.$$

In other words, any covariance matrix Σ_X resulting from DAG $\mathcal{G}$ and arbitrary Ω_L can be generated by alternative parameters from the same DAG with $\tilde{\Omega}_L = I$. This implies that the parameter Ω_L cannot be estimated from Σ_X without additional information and further assumption. Furthermore, since the goal is to estimate the structure $\mathcal{G}$, this suggests that one may assume Ω_L to be an identity matrix during estimation without loss of generality. It is worth noting that such indeterminacy of Ω_L has been discussed in various existing works (Squires et al., 2023), which we make precise here, as it is crucial for formulating the likelihood.

We now provide the likelihood formulation for the linear latent variable causal model in Equation (1). As suggested by Lemma 1, we set $\Omega_L = I$ in the likelihood. Given the empirical covariance matrix S obtained from T samples, the negative log-likelihood is given up to additive constant by

$$\begin{aligned}&\mathcal{L}(B, C, \Omega_X; \mathbf{D})\\ &= \frac{T}{2}\operatorname{tr}\left(S\left(B(I - C)^{-1}(I - C)^{-\top}B^\top + \Omega_X\right)^{-1}\right)\\ &\quad + \frac{T}{2}\log\det\left(B(I - C)^{-1}(I - C)^{-\top}B^\top + \Omega_X\right).\end{aligned}$$

3. Score-Based Identification of Latent Variable Causal Models

In this section, we discuss how to learn linear latent variable causal models with scoring function. First, we introduce the notion of distribution sets and equality constraints in Section 3.1. We formulate the scoring function in Section 3.2, and show how it enables structure identification up to algebraic equivalence in Section 3.3. We then discuss about the BIC score in Section 3.4.

3.1. Distribution Sets and Equality Constraints

We describe the notion of distribution set that is a key ingredient of our score-based search procedure. It refers to the set of marginal distributions generated by a specific structure.

Definition 1 (Distribution set)**.** *The distribution set of DAG $\mathcal{G}$, denoted by $\mathcal{M}(\mathcal{G})$, is defined as*

$$\begin{aligned}\mathcal{M}(\mathcal{G}) \coloneqq \{&B(I - C)^{-1}\Omega_L(I - C)^{-\top}B^\top + \Omega_X :\\ &\mathrm{supp}(B) \subseteq \mathrm{supp}(B_\mathcal{G}), \mathrm{supp}(C) \subseteq \mathrm{supp}(C_\mathcal{G}),\\ &\Omega_X \in \mathrm{diag}(\mathbb{R}^m_{>0}), \Omega_L \in \mathrm{diag}(\mathbb{R}^n_{>0})\}.\end{aligned}$$

Specifically, $\mathcal{M}(\mathcal{G})$ is the set of covariances matrices Σ_X that can be generated by DAG $\mathcal{G}$ by varying the parameters in matrices B, C, Ω_X, and Ω_L. Moreover, since $C_\mathcal{G}$ is acyclic by assumption, we have $(I - C)^{-1} = \sum_{k=0}^{n-1} C^k$. It follows that Equation (2) is a polynomial map, and thus the distribution set $\mathcal{M}(\mathcal{G})$ is, by Tarski–Seidenberg theorem (Benedetti & Risler, 1990), a semialgebraic set. Note that a set is said to be *semialgebraic* if it can be equivalently represented by a finite number of polynomial equalities and inequalities (Benedetti & Risler, 1990).

Structure $\mathcal{G}$ imposes various types of equality (i.e., algebraic) constraints on the covariance matrices, such as conditional independence (i.e., vanishing partial correlation) constraints (Spirtes et al., 2001), rank deficiency (i.e., vanishing determinant) constraints (Spirtes et al., 2001; Sullivant et al., 2010), and possibly Verma constraints (Verma & Pearl, 1991). We refer the readers to Drton (2018) for an overview. Let $H(\mathcal{G})$ be the set of equality constraints imposed by structure $\mathcal{G}$ on the distribution set $\mathcal{M}(\mathcal{G})$, and $\mathbb{H}^m \coloneqq \bigcup_{\mathcal{G}\in\mathbb{G}^m} H(\mathcal{G})$ be the set of possible equality constraints imposed by any structure $\mathcal{G}$ (with m measured variables). Two structures $\mathcal{G}_1$ and $\mathcal{G}_2$ are said to be *algebraic equivalent* if they lead to the same equality constraints, i.e., $H(\mathcal{G}_1) = H(\mathcal{G}_2)$ (van Ommen & Mooij, 2017).[1]

Furthermore, let $\dim(\mathcal{G})$ denote the model dimension or degrees of freedom of DAG $\mathcal{G}$ for the marginal over the observed variables, which can be viewed as the number of free parameters for the distribution set $\mathcal{M}(\mathcal{G})$. In general, the

[1] In the terminology of algebraic geometry, $\mathcal{M}(\mathcal{G}_1)$ and $\mathcal{M}(\mathcal{G}_2)$ share the same vanishing ideal or Zariski closure (Cox et al., 2015).

degrees of freedom are not necessarily equal to the number of parameters (i.e., sum of number of edges and measured variables) in the presence of latent variables (Geiger et al., 1996; 2001). In Sections 4 and 5, we further characterize the degrees of freedom under specific structural assumptions.

3.2. Formulation of Scoring Function

We now provide the formulation of the scoring function for identifying linear latent variable causal models. Specifically, our score-based method involves searching for the structure with the smallest degrees of freedom that can generate the covariance matrix. Given structure $\mathcal{G}$ with samples $\mathbf{D}$ and empirical covariance matrix S, the scoring function is

$$\text{score}_{\text{dim}}(\mathcal{G}, \mathbf{D}) := \begin{cases} \dim(\mathcal{G}) & \text{if } \mathcal{G} \text{ can generate } S, \\ \infty & \text{otherwise.} \end{cases}$$

To determine whether structure $\mathcal{G}$ can generate S, one may minimize the squared errors between S and the covariance matrix parameterized by $\mathcal{G}$, or compare the maximum likelihood w.r.t. $\mathcal{G}$ (e.g., see Equation (3)) to the likelihood of S. Moreover, we show in Section 3.4 that the scoring function satisfies the property of score equivalence.

Similar scoring function has been discussed by Raskutti & Uhler (2014). Roughly speaking, there may exist multiple structures that can generate the same distribution; the scoring function identifies one of them with the smallest degrees of freedom. In Sections 3.3, 4 and 5, we discuss how this scoring function identifies structures up to different types of model equivalence in the large sample limit. Specifically, we show in Section 3.3 that it yields a structure that is algebraic equivalent to the ground truth.

Since the key idea is to identify the structure that generates the population covariance matrix (in the large sample limit) with the smallest degrees of freedom, different types of scoring functions that can achieve so can also be used. In Section 3.4, we further discuss the use of the BIC score.

3.3. Identifying Structures up to Algebraic Equivalence

Having formulated the scoring function in Section 3.2, the question remains as how to leverage it to identify the underlying structure $\mathcal{G}$. To do so, a key ingredient is to establish the correspondence between the covariance matrix and the structure $\mathcal{G}$. As discussed in Section 3.1, the structure $\mathcal{G}$ imposes different types of constraints on the entries of covariance matrices, including equality and inequality constraints. Here, we adopt the following assumption which requires that the equality constraints are imposed by the structure $\mathcal{G}$.

Assumption 1 (Generalized faithfulness (Ghassami et al., 2020))**.** *A distribution Σ_X is said to be generalized faithful to DAG $\mathcal{G}$ if the entries of Σ_X satisfy an equality constraint $\kappa \in \mathbb{H}^m$ only if $\kappa \in H(\mathcal{G})$.*

It is worth noting that different types of faithfulness assumptions have been adopted in causal discovery (Spirtes et al., 2001; Ghassami et al., 2020; Huang et al., 2022) to relate the constraints of the distributions (e.g., conditional independence and rank deficiency constraints) to the underlying structure. This is often motivated by the fact that the set of parameters violating these assumptions has Lebesgue measure zero (see, e.g., Ghassami et al. (2020, Proposition 8)).

We then present the following result that describes the notion of equivalence achieved by minimizing the scoring function. The proof is provided in Appendix B.3, which is partly inspired by the proof of Ghassami et al. (2020, Theorem 3).

Theorem 1 (Algebraic equivalence)**.** *Suppose the true DAG $\mathcal{G}^*$ and the distribution Σ_X satisfy the generalized faithfulness assumption. Let $\hat{\mathcal{G}} \in \arg\min_{\mathcal{G} \in \mathbb{G}^m} \text{score}_{dim}(\mathcal{G}, \mathbf{D})$. Then, $\hat{\mathcal{G}}$ and $\mathcal{G}^*$ are algebraic equivalent, i.e., $H(\hat{\mathcal{G}}) = H(\mathcal{G}^*)$, in the large sample limit.*

Remark 1. *Under generalized faithfulness assumption, Theorem 1 implies that minimizing the scoring function leads to a structure with the same equality constraints (on the marginal over the measured variables) as the true structure.*

In general, relating the estimated structure to the true one, which are algebraically equivalent, can be challenging without any restrictions on the structures. In Sections 4 and 5, we show that, under specific structural assumptions, Theorem 1 helps achieve notions of model equivalence that are more fine-grained than algebraic equivalence (including Markov equivalence in Section 4). Therefore, a general recipe may involve identifying suitable structural assumptions that allow algebraic equivalence to translate into more fine-grained notions of model equivalence. This enables the application of the score based procedure in Theorem 1, given an appropriate characterization of the degrees of freedom. We give a further discussion of generalized faithfulness and algebraic equivalence in Appendices A.2 and A.3, respectively.

3.4. Remark on the BIC Score

The scoring function discussed in Section 3.2 is justified in the large sample limit and may not perform well for finite-sample cases. We consider the BIC score (Schwarz, 1978; Chickering, 2002) that maximizes the likelihood while penalizing the degrees of freedom of structure $\mathcal{G}$:

$$\text{score}_{\text{BIC}}(\mathcal{G}, \mathbf{D}) := \text{score}_{\mathcal{L}}(\mathcal{G}, \mathbf{D}) + \frac{\log T}{2} \dim(\mathcal{G}).$$

where $\text{score}_{\mathcal{L}}(\mathcal{G}, \mathbf{D})$ denotes the maximum likelihood w.r.t. structure $\mathcal{G}$, given by

$$\text{score}_{\mathcal{L}}(\mathcal{G}, \mathbf{D}) := \min_{\substack{(B,C,\Omega_X): \\ \text{supp}(B) \subseteq \text{supp}(B_{\mathcal{G}}), \\ \text{supp}(C) \subseteq \text{supp}(C_{\mathcal{G}}), \\ \Omega_X \in \text{diag}(\mathbb{R}^m_{>0})}} \mathcal{L}(B, C, \Omega_X; \mathbf{D}). \tag{3}$$

Since it may not be straightforward to derive a closed-form solution, various numerical solvers or continuous optimization methods, such as L-BFGS (Byrd et al., 1995) and gradient descent, as well as the expectation-maximization algorithm (Dempster et al., 1977), can be used to compute the maximum likelihood above.

It is worth noting that the BIC score has been widely adopted in score-based causal discovery (Chickering, 2002). Haughton (1988) showed that it is an asymptotic approximation for the log marginal likelihood of *curved* exponential families, which include Gaussian DAG models without latent variables (Geiger et al., 2001; Richardson & Spirtes, 2002). In the presence of latent variables, the models are *stratified* exponential families, and complications arise in using BIC for model selection. Although the typical theoretical justifications of using BIC (Schwarz, 1978; Haughton, 1988) may not apply for identifying latent variable causal models in our setting, we apply it in place of $\text{score}_{\text{dim}}(\mathcal{G}, \mathbf{D})$ in our experiments, since the latter is justified in the large sample limit and may not perform well for finite-sample cases. Surprisingly, using the BIC score leads to a superior empirical performance, specifically under the structural assumptions described in Sections 4 and 5. This suggests that BIC may be a valid scoring criterion in these cases. Therefore, future works involve studying the theoretical justifications of using BIC score under these structural assumptions.

Recall that a scoring function is *score equivalent* if every pair of Markov equivalent structures have the same score (Chickering, 2002). This is a desirable property for score-based procedure as it implies that we can search in the space of Markov equivalence classes (MECs) instead of DAGs. That is, one does not have to compute the score multiple times for the DAGs in the same MEC, which may help improve the runtime. We show that our scoring functions satisfy such a property, with a proof given in Appendix B.2.

Proposition 1 (Score equivalence)**.** *Suppose that DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent. Then, we have* $\text{score}_{dim}(\mathcal{G}_1, \mathbf{D}) = \text{score}_{dim}(\mathcal{G}_2, \mathbf{D})$ *and* $\text{score}_{BIC}(\mathcal{G}_1, \mathbf{D}) = \text{score}_{BIC}(\mathcal{G}_2, \mathbf{D})$.

4. Linear 1-Factor Latent Variable Models

In the previous section, we show that the scoring function can produce a structure algebraic equivalent to the ground truth. We now discuss how such result helps estimate a structure up to Markov equivalence. In this section, we focus on the structural assumption by Silva et al. (2003; 2006).[2]

Assumption 2 (Silva et al. (2003))**.** *Each measured variable has a single latent parent, and each latent variable has at least three measured variables as children.*

[2]In our setting, it may be sufficient to require that each latent variable has at least two measured variables as children.

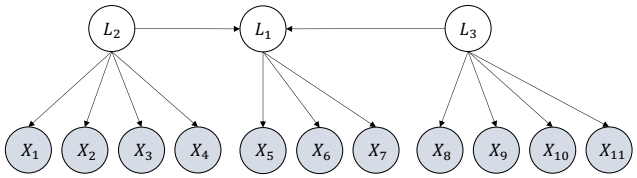


Figure 1: Example of 1-factor latent variable model.

An example illustrating the above assumption is provided in Figure 1. Silva et al. (2003) proposed a search procedure based on statistical tests of tetrad constraints that can identify structures under this assumption. In this section, we develop a score-based method based on this structural assumption. We first characterize the degrees of freedom of the structure in Section 4.1, as required by the scoring function. We then establish the consistency and provide an exact score-based search procedure in Section 4.2. We also develop a continuous search procedure in Section 4.3 that may be more computationally efficient.

4.1. Degrees of Freedom

The scoring function requires a proper specification of the degrees of freedom during the search procedure. For the structural assumption in Assumption 2, the degrees of freedom, as one may expect, equals the number of edges in DAG $\mathcal{G}$ plus the number of measured variables. Here, the number of edges include those among the latent variables and those from the latent variables to the measured ones. The proof follows straightforwardly from parameter identifiability under Assumption 2 (Bollen, 1989), which is provided in Appendix B.4 for completeness.

Proposition 2 (Degrees of freedom)**.** *Suppose that DAG $\mathcal{G}$ satisfies Assumption 2. Then,* $\dim(\mathcal{G}) = |\mathcal{G}| + m$.

To illustrate, the degrees of freedom of the example in Figure 1 are simply equal to 24. The above property holds in many other settings such as the typical setting without latent confounders (Chickering, 2002), as well as those with bow-free acyclic mixed graphs (Brito & Pearl, 2002) and cycles (excluding 2-cycles) (Amendola et al., 2020). Note that such property, while desirable, does not hold in general for structures with latent variables (Geiger et al., 1996). For instance, the degrees of freedom of the structures that we consider in Section 5 are generally not equal to $|\mathcal{G}| + m$.

4.2. Consistency and Exact Score-Based Search

Having characterized the degrees of freedom, we now establish the correctness of score-based approach under Assumption 2 and accordingly develop an exact search procedure. Specifically, under Assumption 2 and the generalized faithfulness assumption, we show that the structure with the optimal score is Markov equivalent to the true structure.

Algorithm 1 Enumerating structures under Assumption 2

Input: Measured variables $X_1, \ldots, X_m$
Output: Set of DAGs $\mathbf{A}$
Initialize $\mathbf{A}$ as an empty set;
for $n = 1$ **to** $\lfloor m/3 \rfloor$ **do**
  foreach latent MEC with n variables **do**
    Generate a latent DAG $C_{\mathcal{G}}$ from the latent MEC;
    foreach ordered n-partition $(\mathbf{P}_j)_{j=1}^n$ of $\{X_i\}_{i=1}^m$ **do**
      if $|\mathbf{P}_j| \geq 3$ for $j = 1, \ldots, n$ **then**
        Construct DAG $\mathcal{G}$ with latent DAG $C_{\mathcal{G}}$ and each latent L_j pointing to the variables in $\mathbf{P}_j$;
        if $\mathcal{G}$ is not Markov equivalent to all DAGs in $\mathbf{A}$ **then**
          $\mathbf{A} \leftarrow \mathbf{A} \cup \{\mathcal{G}\}$;
return set $\mathbf{A}$

Theorem 2 (Correctness)**.** *Suppose that the true DAG $\mathcal{G}^*$ and the distribution Σ_X satisfy the generalized faithfulness assumption, and that $\mathcal{G}^*$ satisfies Assumption 2. Let $\hat{\mathcal{G}}$ be a global minimizer of the following optimization problem:*

$$\begin{aligned} \min_{\mathcal{G} \in \mathbb{G}^m} \quad & \text{score}_{dim}(\mathcal{G}, \mathbf{D}) \\ \text{subject to} \quad & \mathcal{G} \textit{ satisfies Assumption 2}, \end{aligned} \tag{4}$$

where $\dim(\mathcal{G}) = |\mathcal{G}| + m$. Then, $\hat{\mathcal{G}}$ and $\mathcal{G}^*$ are Markov equivalent in the large sample limit.

The proof can be found in Appendix B.5, which leverages Theorem 1 that shows how the scoring function produces a structure that is algebraic equivalent to the true structure. Moreover, the theorem above indicates that one could perform exact search for all structures under Assumption 2. A naive approach is to iterate over all possible structures and check if each of them satisfy Assumption 2. This may be computationally infeasible because much of the time may be spent on structures that do not fall within the model class.

The question is then how to efficiently enumerate and perform exact search for these structures. We provide an algorithm to do so in Algorithm 1. Leveraging the score equivalence property in Proposition 1, we consider only structures that are not Markov equivalent to one another, since they are indistinguishable based on Theorem 2 and give rise to the same score. First, we generate the possible structures $C_{\mathcal{G}}$ among the latent variables that are not Markov equivalent to one another. To construct the structure $B_{\mathcal{G}}$ from latent variables to measured variables, we then find all ordered partitions of measured variables and add each subset from the partition to be the children of each latent variable. We compute the score for each structure enumerated by Algorithm 1, and find the structure with the optimal score. Under Theorem 2, such an exact search procedure will output a structure Markov equivalent to the true one.

4.3. Continuous Search

The exact search procedure presented in the previous section requires computing the score for each structure satisfying the structural assumption, which can be computationally intensive when there is a large number of variables. For instance, when using the BIC score, each computation involves solving a continuous optimization problem in Equation (3); the same applies to $\text{score}_{\text{dim}}(\mathcal{G}, \mathbf{D})$. This is inherent to discrete score-based search that assigns a score to each structure. A key question naturally arises: how do we unify the structure search part and likelihood computation into a single continuous optimization problem? Such a unified procedure helps reduce the computational burden of separately computing the score for each structure in a discrete search. Furthermore, this aligns with recent studies in continuous optimization for causal discovery (Zheng et al., 2018; Ng et al., 2020; Vowels et al., 2022).

We first provide a reformulation of Equation (4) with the BIC score that is more amenable to continuous optimization. The key lies in characterizing the penalty term $|\mathcal{G}|$ and the constraint involving Assumption 2. Specifically, we solve the following constrained optimization problem:

$$\begin{aligned} \min_{\substack{M_B \in \{0,1\}^{m \times \bar{n}}, \\ M_C \in \{0,1\}^{\bar{n} \times \bar{n}}, \\ B \in \mathbb{R}^{m \times \bar{n}}, C \in \mathbb{U}^{\bar{n}}, \\ \Omega_X \in \text{diag}(\mathbb{R}^m_{>0})}} \quad & \Big(\frac{1}{T} \mathcal{L}\left(M_B \odot B, M_C \odot C, \Omega_X; \mathbf{D}\right) \\ & \quad + \lambda \|M_B\|_1 + \lambda \|M_C\|_1 \Big) \\ \text{subject to} \quad & \sum_{k=1}^{\bar{n}} (M_B)_{i,k} - 1 = 0, \ i \in [m], \\ & \left(\left(\sum_{k=1}^{m} (M_B)_{k,j} + \sum_{k=1}^{\bar{n}} (M_C)_{k,j} \right) \right. \\ & \quad \left. \left(\sum_{k=1}^{m} (M_B)_{k,j} - 3 \right) \right) \geq 0, \ j \in [\bar{n}], \end{aligned} \tag{5}$$

where $\bar{n} = \lfloor m/3 \rfloor$ is an upper bound of the number of latent variables and $\lambda = \log T / 2T$. In the formulation above, the matrices M_B and M_C can be viewed as the support matrices of B and C; they act as binary masks which indicate which edges are present in the structure. Furthermore, the two constraints in Equation (5) serve as a characterization of Assumption 2 using the support matrices M_B and M_C. Specifically, the first constraint requires each row of M_B to have one nonzero entry (i.e., each measured variable has one a single latent parent). The second constraint requires each column of M_B and M_C to satisfy the following: either the column of M_B has at least three nonzero entries, or the column of M_B and M_C have no nonzero entries (i.e., each latent variable has at least 3 measured variables as children, or no child at all).

We now discuss how to solve Equation (5) using continuous constrained optimization procedure. We first introduce slack

variable $t_i \geq 0$ and convert the inequality constraints into equality constraints. To estimate the binary matrices M_B and M_C, we apply the Gumbel-Softmax technique (Maddison et al., 2017; Jang et al., 2017) that is widely used to sample and approximate samples from a categorical distribution, which has also been adopted in continuous optimization approaches for causal discovery (Ng et al., 2022; Brouillard et al., 2020). Specifiaclly, we apply Gumbel-Sigmoid for each entry of M_C, and Gumbel-Softmax with $\bar{n}$ categories for each row of M_B; the latter also incorporates the first constraint of Equation (5) that requires each row of M_B to have one nonzero entry. The resulting continuous constrained optimization problem can then be solved using standard methods such as augmented Lagrangian method and quadratic penalty method (Bertsekas, 1982; 1999; Nocedal & Wright, 2006). These methods transform the constrained problem into a series of unconstrained problems, each of which can be solved via continuous optimization methods such as gradient descent or L-BFGS (Byrd et al., 1995). In this work, we adopt augmented Lagragian method that is commonly used in causal discovery with continuous optimization (Zheng et al., 2018; Vowels et al., 2022).

5. Linear Latent Hierarchical Structures

The structural assumption in Section 4 requires (i) each measured variable to have only one latent parent and (ii) each latent variable to have measured children. In real-world cases, the structure may be more complex – the measurement model may not be a tree and some latent variables may not have measured children. Thus, we also consider a more general assumption formulated by Huang et al. (2022).

We first explain the notions of pure children, pure descendants, and latent atomic cover, which serve as the fundamental building blocks of the whole structure.

Definition 2 (Pure children (Huang et al., 2022))**.** *Variables* $\mathbf{V}$ *are pure children of variables* $\mathbf{L}$ *in structure* $\mathcal{G}$*, iff* $Pa_{\mathcal{G}}(\mathbf{V}) = \bigcup_{V_i \in \mathbf{V}} Pa_{\mathcal{G}}(V_i) = \mathbf{L}$ *and* $\mathbf{L} \cap \mathbf{V} = \emptyset$*. We denote the pure children of* $\mathbf{L}$ *in* $\mathcal{G}$ *by* $PCh_{\mathcal{G}}(\mathbf{L})$.

Accordingly, pure descendants of a set of variables $\mathbf{L}$, i.e., $\text{PDe}_{\mathcal{G}}(\mathbf{L})$, are defined as all recursive pure children of $\mathbf{L}$.

Definition 3 (Latent atomic cover (Huang et al., 2022))**.** *Let* $\mathbf{L}$ *be a set of latent variables in* $\mathcal{G}$ *with* $|\mathbf{L}| = k$*.* $\mathbf{L}$ *is an atomic cover if the following conditions hold:*

(i) There exists a set of variables $\mathbf{C}$ *such that* $\mathbf{C} \subseteq PCh_{\mathcal{G}}(\mathbf{L})$ *and* $|\mathbf{C}| \geq k + 1$.

(ii) There exists a set of variables $\mathbf{N}$ *such that every element in* $\mathbf{N}$ *is a neighbour of* $\mathbf{L}$*,* $|\mathbf{N}| \geq k + 1$*, and* $\mathbf{N} \cap \mathbf{C} = \emptyset$.

(iii) There does not exist a partition of $\mathbf{L} = \mathbf{L}_1 \cup \mathbf{L}_2$ *such that both* $\mathbf{L}_1$ *and* $\mathbf{L}_2$ *are latent atomic covers.*

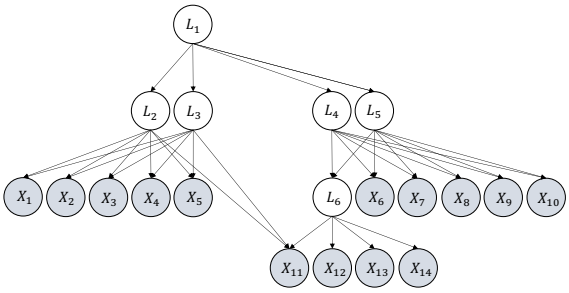


Figure 2: Example of latent hierarchical structure.

Having introduced the required notions, we now provide the structural assumptions considered by Huang et al. (2022).

Assumption 3 (Identifiable linear latent hierarchical graph (Huang et al., 2022))**.** *A graph* $\mathcal{G}$ *is an identifiable linear latent hierarchical graph if (i) every latent variable* L_i *belongs to at least one latent atomic cover and there is no triangle structure in the graph, and (ii) if there exists a set of variables* $\mathbf{V}$ *such that every variable in* $\mathbf{V}$ *is a collider of two latent atomic covers* $\mathbf{V}_1$*,* $\mathbf{V}_2$*, and denote by* $\mathbf{T}$ *the minimal set of variables that d-separates* $\mathbf{V_1}$ *from* $\mathbf{V_2}$*, then we must have* $|\mathbf{V}| + |\mathbf{T}| \geq |\mathbf{V}_1| + |\mathbf{V}_2|$.

The assumption above requires that each latent variable belongs to at least one latent atomic cover, since a latent atomic cover is the minimal identifiable substructure in a graph using rank constraints of covariance over observed variables. Also, the assumption requires certain graphical patterns that are related to the common pure descendants across different latent atomic covers for the identifiability of the whole latent structure. The assumption above may be more general than Assumption 2 as it allows each measured variable to have multiple latent variables as parents, and also allow some latent variables to not have any measured child at all; see Appendix A.4 for more details. An example is given in Figure 2. Under the assumption above, Huang et al. (2022) developed a constraint-based method based on rank deficiency test to estimate the equivalence class of the true structure.

In this section, we develop a score-based method under Assumption 3. We characterize the degrees of freedom in Section 5.1 and provide an exact search method in Section 5.2. We do not provide a continuous search method for this structural assumption, since it cannot be straightforwardly formulated as inequality constraints, similar to Equation (5).

5.1. Degrees of Freedom

As noted in Sections 3.2 and 4, a key ingredient of the score-based method is the specification of the degrees of freedom. It may be natural to expect that the degrees of freedom are equal to the number of edges and measured variables, similar to the structural assumption considered in Section 4 and the

Algorithm 2 Degrees of freedom under Assumption 3

Input: Structure $\mathcal{G}$
Output: Degrees of freedom d
Initialize degrees of freedom $d \leftarrow |\mathcal{G}| + m$;
foreach subset $\mathbf{L}$ of latent variables where $|\mathbf{L}| \geq 2$ **do**
  if variables in $\mathbf{L}$ have the same parents and children **and** no proper superset of $\mathbf{L}$ satisfies the previous condition **then**
    $d \leftarrow d - |\mathbf{L}|(|\mathbf{L}| - 1)/2$;
return degrees of freedom d

standard setting without latent variables (Chickering, 2002). However, this property does not hold for latent hierarchical structures, as illustrated by the following lemma.

Proposition 3. *Suppose that DAG $\mathcal{G}$ follows the linear latent variable causal model in Equation (1). Suppose also that there exist $k \geq 2$ latent variables in $\mathcal{G}$ with the same set of parents and children, where either the number of parents or children is at least k. Then,* $\dim(\mathcal{G}) \leq |\mathcal{G}| + m - k(k-1)/2$.

The proof is given in Appendix B.6. As shown in the proof, under these circumstances, there exists an alternative structure $\tilde{\mathcal{G}}$ obtained by removing $k(k-1)/2$ edges (corresponding to edges involving the parents or children) from $\mathcal{G}$ such that $\tilde{\mathcal{G}}$ leads to the same distribution set as $\mathcal{G}$, i.e., $\mathcal{M}(\tilde{\mathcal{G}}) = \mathcal{M}(\mathcal{G})$. Thus, the degrees of freedom in this scenario are reduced compared to the number of edges and measured variables. This reduction can be intuitively explained by the redundancy of certain edges in such situations. For instance, the degrees of freedom for the structure are 44 instead of 46 (i.e., the sum of number of edges and measured variables), because the variables $\{L_2, L_3\}$ and $\{L_4, L_5\}$ (i.e., in the same atomic covers) have the same parents and children.

Building on the result above, we develop a procedure in Algorithm 2 to calculate the degrees of freedom under Assumption 3. Specifically, the algorithm iterates over all subsets of latent variables and calculate the degrees of freedom that can be reduced. The following proposition shows that the algorithm outputs the upper bound of the degrees of freedom, with a proof provided in Appendix B.7.

Proposition 4 (Degrees of freedom)**.** *Suppose that DAG $\mathcal{G}$ satisfies Assumption 2. Then, Algorithm 2 outputs the upper bound of* $\dim(\mathcal{G})$.

We conjecture, supported by simulations over $10,000$ examples (by computing the rank of Jacobian matrices (Geiger et al., 2001)) and the experiments in Section 6, that the upper bound provided by this algorithm is tight, although a proof seems to involve tools from algebraic statistics and is not straightforward. For instance, Drton et al. (2023) analyzed the degrees of freedom for sparse factor analysis, which is technically complex even with independent latent variables.

Algorithm 3 Enumerating structures under Assumption 3

Input: Measured variables $X_1, \ldots, X_m$
Output: Set of DAGs $\mathbf{A}$
Initialize $\mathbf{A}$ as an empty set;
for $n = 1$ **to** $\bar{n}$ **do**
  for partition $\{\mathbf{C}_i\}_{i=1}^{l}$ of $\{L_i\}_{i=1}^{n}$ as atomic covers **do**
    for DAG $\mathcal{G}_{\mathbf{C}\to\mathbf{C}}$ among $\{\mathbf{C}_i\}_{i=1}^{l}$ **do**
      for DAG $\mathcal{G}_{\mathbf{C}\to X}$ from $\{\mathbf{C}_i\}_{i=1}^{l}$ to $\{X_i\}_{i=1}^{m}$ **do**
        Construct DAG $\mathcal{G}$ by combining $\mathcal{G}_{\mathbf{C}\to\mathbf{C}}$ and $\mathcal{G}_{\mathbf{C}\to X}$;
        if $\mathcal{G}$ satisfies Assumption 3 and is not Markov equivalent to all DAGs in $\mathbf{A}$ **then**
          $\mathbf{A} \leftarrow \mathbf{A} \cup \{\mathcal{G}\}$;
return set $\mathbf{A}$

5.2. Consistency and Exact Score-Based Search

With the algorithm to compute the degrees of freedom, we now develop a score-based method to estimate latent hierarchical structures. We first establish the correctness of the score-based method. Under Assumption 3 and the generalized faithfulness assumption, we prove that the structure with the optimal score is Markov equivalent to the true hierarcahical structure, up to certain rank equivalent graph operators. The proof and definition of the operators together with illustrative examples can be found in Appendix B.8.

Theorem 3 (Correctness)**.** *Suppose that the true DAG $\mathcal{G}^*$ and the distribution Σ_X satisfy the generalized faithfulness assumption, and that $\mathcal{G}^*$ satisfies Assumption 3. Let $\hat{\mathcal{G}}$ be a global minimizer of the following optimization problem:*

$$\begin{aligned} \min_{\mathcal{G}\in\mathbb{G}^m} \quad & \text{score}_{dim}(\mathcal{G}, \mathbf{D}) \\ \text{subject to} \quad & \mathcal{G} \text{ satisfies Assumption 3}, \\ & \mathcal{G} = \mathcal{O}_{min}(\mathcal{O}_{skeleton}(\mathcal{G})). \end{aligned}$$

Then, $\mathcal{O}_{atomic}(\hat{\mathcal{G}})$ and $\mathcal{O}_{atomic}(\mathcal{O}_{min}(\mathcal{O}_{skeleton}(\mathcal{G}^*)))$ are Markov equivalent in the large sample limit.

Based on the theorem above, we develop an exact search procedure for structures under Assumption 3. We introduce a procedure in Algorithm 3 for enumeration of these structures, where $\bar{n}$ is a hyperparameter indicating the maximal number of latent variables. Note that a possible upper bound for $\bar{n}$ is $3m$. Similar to the algorithm developed in Algorithm 1, we enumerate only structures that are not Markov equivalent to one another, leveraging the score equivalence property. The whole procedure of Algorithm 3 is roughly as follows. We maintain a set of DAGs, $\mathbf{A}$. Given the number of observed variables, we first decide the possible number of latent variables, and then enumerate all possible combinations of atomic covers. For each combination of atomic covers, we enumerate all possible DAGs among atomic covers and all possible DAGs from atomic covers to observed

Table 1: F1 scores of skeletons across various structural assumptions and sample sizes. For each setting, the top two methods are in bold. For FOFC, the number within the brackets indicates the number of valid runs (for which an error did not occur).

Model type	Sample size	SALAD	SALAD-CS	HUANG	FOFC	GIN
1-Factor model	100	$\mathbf{0.99 \pm 0.03}$	$\mathbf{0.97 \pm 0.07}$	0.50 ± 0.20	0.90 ± 0.12 (8)	0.35 ± 0.02
	300	$\mathbf{0.99 \pm 0.01}$	$\mathbf{0.99 \pm 0.02}$	0.85 ± 0.15	0.98 ± 0.03 (9)	0.35 ± 0.02
	1000	$\mathbf{0.99 \pm 0.01}$	$\mathbf{0.99 \pm 0.01}$	0.93 ± 0.03	0.98 ± 0.03 (12)	0.35 ± 0.02
	3000	$\mathbf{1 \pm 0}$	$\mathbf{0.99 \pm 0.02}$	0.93 ± 0.03	0.99 ± 0.01 (12)	0.35 ± 0.02
	10000	$\mathbf{1 \pm 0}$	$\mathbf{0.99 \pm 0.02}$	0.93 ± 0.03	0.99 ± 0.01 (11)	0.37 ± 0.07
Hierarchical structure	100	$\mathbf{0.92 \pm 0.06}$	N/A	$\mathbf{0.57 \pm 0.13}$	N/A (0)	0.48 ± 0.05
	300	$\mathbf{0.92 \pm 0.07}$	N/A	$\mathbf{0.66 \pm 0.09}$	N/A (0)	0.48 ± 0.05
	1000	$\mathbf{0.95 \pm 0.04}$	N/A	$\mathbf{0.76 \pm 0.13}$	N/A (0)	0.48 ± 0.05
	3000	$\mathbf{0.97 \pm 0.03}$	N/A	$\mathbf{0.87 \pm 0.11}$	N/A (0)	0.48 ± 0.05
	10000	$\mathbf{0.98 \pm 0.03}$	N/A	$\mathbf{0.88 \pm 0.15}$	N/A (0)	0.47 ± 0.05

variables. Finally, we combine both enumerated DAGs to get a possible graph $\mathcal{G}$; if $\mathcal{G}$ satisfies Assumption 3 and is not Markov equivalent to all structures in $\mathbf{A}$, we add it to $\mathbf{A}$. Once the search space is constructed, the algorithm identifies the structure with the optimal score, with the degrees of freedom for each structure computed using Algorithm 2.

6. Experiments

We conduct experiments to validate our score-based methods, by comparing them to existing methods that support causally-related latent variables, such as FOFC (Kummerfeld & Ramsey, 2016), HUANG (Huang et al., 2022), and GIN (Xie et al., 2020). We do not include the comparison with FCI because, even when working perfectly, it will output complete PAGs over the observed variables for most ground truths considered here, which do not have any information of the orientation and are not informative. Moreover, we denote our exact search method by SALAD and continuous one by SALAD-CS, and adopt the BIC score here.

For the ground truths, we consider the 1-factor models and hierarchical structures provided in Figures 4 and 5 in Appendix C. For each structure, the nonzero elements of matrices B and C are generated uniformly at random from the interval $[-2, -0.5] \cup [0.5, 2.0]$. For GIN, the noise terms E_X and E_L follow $\mathrm{Uniform}[-\alpha, \alpha]$, where α is sampled uniformly from $[\sqrt{6}, \sqrt{15}]$. For the other methods, the noise terms follow Gaussian distributions with variances sampled uniformly from interval $[2, 5]$. We consider sample size $T \in \{100, 300, 1000, 3000, 10000\}$. We evaluate the estimated structures using F1 scores calculated over the skeleton and structural Hamming distance (SHD) over the MEC. We run three random trials for each ground truth, and report the mean and standard devation for each metric. Further details about the metrics and baselines can be found in Appendix C.

The F1 scores of skeletons are reported in Table 1, while the SHDs of MECs are given in Table 2 in the supplementary material. One observes that our methods achieve much better F1 scores and SHDs as compared to the other baselines, especially for small sample sizes. For instance, for 100 samples, our SALAD method achieves average F1 scores of 0.99 and 0.92 for 1-factor model and hierarchical structures, respectively, while the second best baseline achieves F1 scores of 0.90 and 0.57, respectively. Note that although FOFC achieves an F1 score of 0.90 for 1-factor model in this case, four of the runs are not valid (i.e., an error occurred). A possible reason of the improvement is that the existing constraint-based baselines, as discussed in Section 1, may be prone to the issue of error propagation during the estimation procedure, while our score-based method is not susceptible to such an issue. Furthermore, the F1 scores of our method are close to one for both structural assumptions when the sample size is large, which suggest that BIC may be a valid scoring function in our setting and help verify the correctness established in Theorems 2 and 3. The runtime and computational efficiency are discussed in Appendix D.

7. Conclusion and Discussion

In this work, we propose SALAD, a score-based causal discovery method capable of identifying causal relations among latent variables. Achieving score equivalence and consistency, along with degrees of freedom characterization and exact and continuous score-based methods, our work provides a unified view on multiple existing constraint-based methods with latent variables, and further validates the effectiveness of score-based methods. We hope that this work could spur future research on developing score-based methods for latent variable causal models.

Indeed, our exact methods require a relatively long runtime, similar to the exact score-based methods even without latent variables (Singh & Moore, 2005; Yuan & Malone, 2013). Future works include developing greedy approaches similar to GES to make the search procedure more efficient and scalable, and studying the theoretical justifications of using BIC score under the structural assumptions considered.

Acknowledgements

The authors would like to thank the anonymous reviewers for helpful comments and suggestions. The authors would also like to acknowledge the support from NSF Grant 2229881, the National Institutes of Health (NIH) under Contract R01HL159805, and grants from Apple Inc., KDDI Research Inc., Quris AI, and Florin Court Capital.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Adams, J., Hansen, N., and Zhang, K. Identification of partially observed linear causal models: Graphical conditions for the non-gaussian and heterogeneous cases. *Advances in Neural Information Processing Systems*, 34: 22822–22833, 2021.

Akbari, S., Mokhtarian, E., Ghassami, A., and Kiyavash, N. Recursive causal structure learning in the presence of latent variables and selection bias. *Advances in Neural Information Processing Systems*, 34:10119–10130, 2021.

Amendola, C., Dettling, P., Drton, M., Onori, F., and Wu, J. Structure learning for cyclic linear causal models. In *Conference on Uncertainty in Artificial Intelligence*, 2020.

Améndola, C., Drton, M., Grosdos, A., Homs, R., and Robeva, E. Third-order moment varieties of linear non-gaussian graphical models. *Information and Inference: A Journal of the IMA*, 12(3):iaad007, 2023.

Anandkumar, A., Hsu, D., Javanmard, A., and Kakade, S. Learning linear bayesian networks with latent variables. In *International Conference on Machine Learning*, pp. 249–257. PMLR, 2013.

Bellot, A. and van der Schaar, M. Deconfounded score method: Scoring dags with dense unobserved confounding. *arXiv preprint arXiv:2103.15106*, 2021.

Benedetti, R. and Risler, J.-J. *Real algebraic and semi-algebraic sets*. Actualités mathématiques. Hermann, Paris, 1990.

Bernstein, D., Saeed, B., Squires, C., and Uhler, C. Ordering-based causal structure learning in the presence of latent variables. In *International Conference on Artificial Intelligence and Statistics*, pp. 4098–4108. PMLR, 2020.

Bertsekas, D. P. *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, 1982.

Bertsekas, D. P. *Nonlinear Programming*. Athena Scientific, 2nd edition, 1999.

Bhattacharya, R., Nagarajan, T., Malinsky, D., and Shpitser, I. Differentiable causal discovery under unmeasured confounding. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Bollen, K. A. *The General Model, Part I: Latent Variable and Measurement Models Combined*, chapter Eight, pp. 319–394. John Wiley & Sons, Ltd, 1989. ISBN 9781118619179.

Brito, C. and Pearl, J. A new identification condition for recursive models with correlated errors. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(4):459–474, 2002. doi: 10.1207/S15328007SEM0904_1.

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., and Drouin, A. Differentiable causal discovery from interventional data. In *Advances in Neural Information Processing Systems*, 2020.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

Byrne, B. *Structural Equation Modeling With AMOS: Basic Concepts, Applications, and Programming*. Multivariate Applications Series. Taylor & Francis, 2001.

Cai, R., Xie, F., Glymour, C., Hao, Z., and Zhang, K. Triad constraints for learning causal structure of latent variables. *Advances in neural information processing systems*, 32, 2019.

Chen, Z., Xie, F., Qiao, J., Hao, Z., Zhang, K., and Cai, R. Identification of linear latent variable model with arbitrary distribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6350–6357, 2022.

Chickering, D. M. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3 (Nov):507–554, 2002.

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

Claassen, T. and Bucur, I. G. Greedy equivalence search in the presence of latent confounders. In *Conference on Uncertainty in Artificial Intelligence*, 2022.

Claassen, T., Mooij, J., and Heskes, T. Learning sparse causal models is not np-hard. *arXiv preprint arXiv:1309.6824*, 2013.

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

Cox, D. A., Little, J., and O'Shea, D. *Ideals, Varieties, and Algorithms*. Springer, New York, fourth edition, 2015.

Cui, R., Groot, P., Schauer, M., and Heskes, T. Learning the causal structure of copula models with latent variables. 2018.

Dai, H., Spirtes, P., and Zhang, K. Independence testing-based approach to causal discovery under measurement error and linear non-gaussian models. *Advances in Neural Information Processing Systems*, 35:27524–27536, 2022.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39: 1–38, 1977.

Dong, X., Huang, B., Ng, I., Song, X., Zheng, Y., Jin, S., Legaspi, R., Spirtes, P., and Zhang, K. A versatile causal discovery framework to allow causally-related hidden variables. *arXiv preprint arXiv:2312.11001*, 2023.

Drton, M. Algebraic problems in structural equation modeling. In *Advanced Studies in Pure Mathematics*, pp. 35–86. Mathematical Society of Japan, 2018.

Drton, M., Grosdos, A., Portakal, I., and Sturma, N. Algebraic sparse factor analysis. *arXiv preprint arXiv:2312.14762*, 2023.

Forster, M., Raskutti, G., Stern, R., and Weinberger, N. The frugal inference of causal relations. *The British Journal for the Philosophy of Science*, 69, 04 2017.

Geiger, D., Heckerman, D. E., and Meek, C. Asymptotic model selection for directed networks with hidden variables. In *Conference on Uncertainty in Artificial Intelligence*, 1996.

Geiger, D., Heckerman, D., King, H., and Meek, C. Stratified exponential families: Graphical models and model selection. *The Annals of Statistics*, 29(2):505–529, 2001.

Ghassami, A., Yang, A., Kiyavash, N., and Zhang, K. Characterizing distribution equivalence and structure learning for cyclic and acyclic directed graphs. In *International Conference on Machine Learning*, 2020.

Goldberg, L. The development of markers for the big five factor structure. *Psychological Assessment*, 4:26–42, 03 1992.

Haughton, D. M. A. On the choice of a model to fit data from an exponential family. *The Annals of Statistics*, 16 (1):342–355, 1988.

Himi, S. A., Buehner, M., Schwaighofer, M., Klapetek, A., and Hilbert, S. Multitasking behavior and its related constructs: Executive functions, working memory capacity, relational integration, and divided attention. *Cognition*, 189:275–298, 08 2019.

Huang, B., Low, C., Xie, F., Glymour, C., and Zhang, K. Latent hierarchical causal structure discovery with rank constraints. In *Advances in Neural Information Processing Systems*, 2022.

Hyvärinen, A., Khemakhem, I., and Monti, R. Identifiability of latent-variable and structural-equation models: from linear to nonlinear. *Annals of the Institute of Statistical Mathematics*, 2023.

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017.

Kingma, D. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2014.

Kivva, B., Rajendran, G., Ravikumar, P., and Aragam, B. Learning latent causal graphs via mixture oracles. *Advances in Neural Information Processing Systems*, 34: 18087–18101, 2021.

Koller, D. and Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA, 2009.

Kummerfeld, E. and Ramsey, J. Causal clustering for 1-factor measurement models. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1655–1664, 2016.

Leung, D., Drton, M., and Hara, H. Identifiability of directed Gaussian graphical models with one latent source. *Electronic Journal of Statistics*, 10, 05 2015.

Maddison, C. J., Mnih, A., and Teh, Y. W. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.

Nandy, P., Hauser, A., and Maathuis, M. H. High-dimensional consistency in score-based and hybrid structure learning. *The Annals of Statistics*, 46(6A):3151–3183, 2018.

Ng, I., Ghassami, A., and Zhang, K. On the role of sparsity and DAG constraints for learning linear DAGs. In *Advances in Neural Information Processing Systems*, 2020.

Ng, I., Zhu, S., Fang, Z., Li, H., Chen, Z., and Wang, J. Masked gradient-based causal structure learning. In *SIAM International Conference on Data Mining*, 2022.

Ng, I., Huang, B., and Zhang, K. Structure learning with continuous optimization: A sober look and beyond. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, 2024.

Nocedal, J. and Wright, S. J. *Numerical optimization*. Springer series in operations research and financial engineering. Springer, 2nd edition, 2006.

Nowzohour, C., Maathuis, M. H., Evans, R. J., and Bühlmann, P. Distributional equivalence and structure learning for bow-free acyclic path diagrams. 2017.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

Pearl, J. *Causality*. Cambridge university press, 2009.

Ramsey, J., Glymour, M., Sanchez-Romero, R., and Glymour, C. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics*, 3(2):121–129, 2017.

Raskutti, G. and Uhler, C. Learning directed acyclic graphs based on sparsest permutations. *arXiv preprint arXiv:1307.0366v3*, 2014.

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

Richardson, T. S. *Models of feedback: interpretation and discovery*. PhD thesis, Carnegie-Mellon University, 1996.

Salehkaleybar, S., Ghassami, A., Kiyavash, N., and Zhang, K. Learning linear non-gaussian causal models in the presence of latent variables. *The Journal of Machine Learning Research*, 21(1):1436–1459, 2020.

Scheines, R., Spirtes, P., Glymour, C., Meek, C., and Richardson, T. The TETRAD project: Constraint based aids to causal model specification. *Multivariate Behavioral Research*, 33:65–117, 1998.

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

Schwarz, G. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

Shahin, R. and Chechik, M. Automatic and efficient variability-aware lifting of functional programs. *Proceedings of the ACM on Programming Languages*, 4 (OOPSLA):1–27, 2020.

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

Shpitser, I., Richardson, T. S., Robins, J. M., and Evans, R. Parameter and structure learning in nested Markov models. *arXiv preprint arXiv:1207.5058*, 2012.

Silva, R. and Scheines, R. Generalized measurement models. Technical report, Carnegie-Mellon Univ Pittsburgh PA School of Computer Science, 2005.

Silva, R., Scheines, R., Glymour, C., and Spirtes, P. Learning measurement models for unobserved variables. In *Conference on Uncertainty in Artificial Intelligence*, 2003.

Silva, R., Scheines, R., Glymour, C., and Spirtes, P. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7(8):191–246, 2006. URL http://jmlr.org/papers/v7/silva06a.html.

Singh, A. P. and Moore, A. W. Finding optimal Bayesian networks by dynamic programming. Technical report, Carnegie Mellon University, 2005.

Spirtes, P. Introduction to causal inference. *Journal of Machine Learning Research*, 11(5), 2010.

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

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2001.

Spirtes, P. L., Meek, C., and Richardson, T. S. Causal inference in the presence of latent variables and selection bias. *arXiv preprint arXiv:1302.4983*, 2013.

Squires, C., Seigal, A., Bhate, S., and Uhler, C. Linear causal disentanglement via interventions. In *International Conference on Machine Learning*, 2023.

Sturma, N., Squires, C., Drton, M., and Uhler, C. Unpaired multi-domain causal representation learning. *arXiv preprint arXiv:2302.00993*, 2023.

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

Triantafillou, S. and Tsamardinos, I. Score-based vs constraint-based causal learning in the presence of confounders. In *Cfa@ uai*, pp. 59–67, 2016.

van Ommen, T. and Mooij, J. M. Algebraic equivalence of linear structural equation models. In *Conference on Uncertainty in Artificial Intelligence*, 2017.

Verma, T. and Pearl, J. Equivalence and synthesis of causal models. In *Conference on Uncertainty in Artificial Intelligence*, 1991.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C., Polat, İ., Feng, Y., Moore, E. W., Vand erPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and Contributors, S. . . SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272, 2020.

Vowels, M. J., Camgoz, N. C., and Bowden, R. D'ya like DAGs? a survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4), nov 2022. ISSN 0360-0300.

Wang, Y. S. and Drton, M. Causal discovery with unobserved confounding and non-Gaussian data. *Journal of Machine Learning Research*, 24(271):1–61, 2023.

Xie, F., Cai, R., Huang, B., Glymour, C., Hao, Z., and Zhang, K. Generalized independent noise condition for estimating latent variable causal graphs. *Advances in neural information processing systems*, 33:14891–14902, 2020.

Yuan, C. and Malone, B. Learning optimal Bayesian networks: A shortest path perspective. *Journal of Artificial Intelligence Research*, 48(1):23–65, 2013.

Zeng, Y., Shimizu, S., Cai, R., Xie, F., Yamamoto, M., and Hao, Z. Causal discovery with multi-domain LiNGAM for latent factors. In *Causal Analysis Workshop Series*, pp. 1–4. PMLR, 2021.

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

Zhang, K., Gong, M., Ramsey, J., Batmanghelich, K., Spirtes, P., and Glymour, C. Causal discovery with linear non-Gaussian models under measurement error: Structural identifiability results. In *Conference on Uncertainty in Artificial Intelligence*, 2018.

Zhang, K., Xie, S., Ng, I., and Zheng, Y. Causal representation learning from multiple distributions: A general setting. *arXiv preprint arXiv:2402.05052*, 2024.

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

Zheng, X., Aragam, B., Ravikumar, P., and Xing, E. P. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, 2018.

Supplementary Material

A. Further Discussions

We provide supplementary discussions below as complements to various sections in the main paper.

A.1. Latent Variable Causal Models

In real-world scenarios, one may often encounter the latent variable causal model in Equation (1), where the measured variables do not influence each other and are effects of latent variables. Thus, there have been many works that aim to estimate this type of linear latent variable causal models (Silva et al., 2003; 2006; Silva & Scheines, 2005; Zhang, 2004; Choi et al., 2011; Kummerfeld & Ramsey, 2016; Cai et al., 2019; Xie et al., 2020; Dai et al., 2022; Huang et al., 2022; Chen et al., 2022).

To provide some examples, in psychometrics, multiple questions are often used as indirect proxies for each latent personality dimension (e.g., openness, extraversion, self-esteem) (Goldberg, 1992; Byrne, 2001; Himi et al., 2019), forming a latent variable causal model. When analyzing fMRI data, a large number of voxels are measured, which do not necessarily have clear semantic meanings. A hierarchical structure can then be used to model functionally meaningful brain regions at different levels (Huang et al., 2022). In representation learning, recent works typically assumed that measured variables (e.g., image pixels) are effects of latent variables and that there are no direct causal influences among the measured variables (Schölkopf et al., 2021; Hyvärinen et al., 2023; Zhang et al., 2024).

A.2. Generalized Faithfulness Assumption

We discuss the necessity of the generalized faithfulness assumption adopted in our results. One of the advantages of score-based causal discovery is that it typically relies on the sparsest Markov representation (SMR) assumption (or unique frugality assumption) (Forster et al., 2017; Raskutti & Uhler, 2014), which is strictly weaker than the faithfulness assumption (Spirtes et al., 2001) in the setting without latent variables. In our setting with latent variables, it is possible to modify Theorems 1, 2, and 3 to replace generalized faithfulness with a formulation similar to the SMR assumption. However, doing so may not be informative because (1) SMR may not be strictly weaker than the faithfulness assumption in our setting, and (2) simply assuming SMR in our setting may not provide insights into what structural assumptions the true structure should obey. Thus, we adopt the generalized faithfulness assumption and various structural assumptions to make the results more informative.

A.3. Algebraic Equivalence

We provide a further discussion of algebraic equivalence (van Ommen & Mooij, 2017) as a complement to Section 3.3. First note that two algebraic equivalent structures are not necessarily Markov equivalent. The reason is that, without any restriction on the structures, one may construct different structures that entail the same equality constraints. For instance, consider the structure in Figure 5(a), denoted as $\mathcal{G}_1$, and another structure $\mathcal{G}_2$ that is identical to $\mathcal{G}_1$, except that the edges $L_1 \to X_1$, $L_2 \to X_1$, and $L_1 \to X_2$ are removed in $\mathcal{G}_2$. One can show $\mathcal{G}_1$ and $\mathcal{G}_2$ are algebraic equivalent, but clearly they are not Markov equivalent.

Nonetheless, algebraic equivalence may be a reasonable way for estimating linear latent variable causal models, because equality constraints (of the covariance matrices) are some of the major footprints in the data that one could leverage (without considering higher-order statistics) to identify the underlying structures. This can be done by relating these constraints to the structures via the generalized faithfulness assumption.

A.4. Structural Assumptions

We discuss the similarities and differences between the structural assumptions considered in our work. First, both Assumptions 2 and 3 require that the observed variables are leaf nodes, and that there are no direct causal influences among observed variables. The key differences between them are as follows. (1) Assumption 2 requires that each latent variable has at least three measured variables as its children, while Assumption 3 allows latent variables to form a hierarchical structure - some latent variables may only have latent variables as their children. (2) Assumption 2 requires each observed variable to be caused by a single latent variable, while Assumption 3 allows an observed variable to be caused by a group of latent variables. Since Assumption 3 does not require each latent variable to have measured variables as children, the structure among latent variables cannot be arbitrary, and thus there is a tradeoff between Assumptions 2 and 3.

B. Proofs

B.1. Proof of Lemma 1

Lemma 1 (Indeterminacy of Ω_L)**.** *For any parameters B, C, Ω_X, Ω_L, and Σ_X that follow Equation (2), there exist parameters $\tilde{B}$ and $\tilde{C}$ with $\mathrm{supp}(B) = \mathrm{supp}(\tilde{B})$ and $\mathrm{supp}(C) = \mathrm{supp}(\tilde{C})$ such that*

$$\Sigma_X = \tilde{B}(I - \tilde{C})^{-1}(I - \tilde{C})^{-\top}\tilde{B}^\top + \Omega_X.$$

Proof. Let $\tilde{B} \coloneqq B\Omega_L^{\frac{1}{2}}$ and $\tilde{C} \coloneqq \Omega_L^{-\frac{1}{2}} C \Omega_L^{\frac{1}{2}}$. We have $\mathrm{supp}(B) = \mathrm{supp}(\tilde{B})$, $\mathrm{supp}(C) = \mathrm{supp}(\tilde{C})$, and

$$\begin{aligned}
\Sigma_X &= B(I - C)^{-1}\Omega_L(I - C)^{-\top}B^\top + \Omega_X \\
&= B\Omega_L^{\frac{1}{2}}\Omega_L^{-\frac{1}{2}}(I - C)^{-1}\Omega_L^{\frac{1}{2}}\Omega_L^{\frac{1}{2}}(I - C)^{-\top}\Omega_L^{-\frac{1}{2}}\Omega_L^{\frac{1}{2}}B^\top + \Omega_X \\
&= (B\Omega_L^{\frac{1}{2}})(I - \Omega_L^{-\frac{1}{2}}C\Omega_L^{\frac{1}{2}})^{-1}(I - \Omega_L^{-\frac{1}{2}}C\Omega_L^{\frac{1}{2}})^{-\top}(B\Omega_L^{\frac{1}{2}})^\top + \Omega_X \\
&= \tilde{B}(I - \tilde{C})^{-1}(I - \tilde{C})^{-\top}\tilde{B}^\top + \Omega_X
\end{aligned}$$

□

B.2. Proof of Proposition 1

The following proof is partly inspired by that of the score equivalence property in the setting without latent variables (Koller & Friedman, 2009).

Proposition 1 (Score equivalence)**.** *Suppose that DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent. Then, we have $\mathrm{score}_{dim}(\mathcal{G}_1, \mathbf{D}) = \mathrm{score}_{dim}(\mathcal{G}_2, \mathbf{D})$ and $\mathrm{score}_{BIC}(\mathcal{G}_1, \mathbf{D}) = \mathrm{score}_{BIC}(\mathcal{G}_2, \mathbf{D})$.*

Proof. Because structures $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent, they can generate the same set of covariance matrices over variables X and L. Thus, for any parameters B, C, Ω_X of $\mathcal{G}_1$ with $\Omega_L = I$, there exists parameters $B', C', \Omega'_X, \Omega'_L$ of $\mathcal{G}_2$ that can generate the same covariance matrix over X and L, which imply that $B', C', \Omega'_X, \Omega'_L$ can generate the same covariance matrix over X. By Lemma 1, there exists parameters of $\mathcal{G}_2$, denoted as $\tilde{B}, \tilde{C}, \tilde{\Omega}_X = \Omega'_X$ and $\tilde{\Omega}_L = I$, that can generate the covariance matrix. Note that the likelihood function depends only on the covariance matrix, which indicates

$$\mathrm{score}_{\mathcal{L}}(\mathcal{G}_1, \mathbf{D}) = \mathcal{L}(\hat{B}, \hat{C}, \hat{\Omega}_X; \mathbf{D}) = \mathcal{L}(\tilde{B}, \tilde{C}, \tilde{\Omega}_X; \mathbf{D}) \geq \mathrm{score}_{\mathcal{L}}(\mathcal{G}_2, \mathbf{D}),$$

where $\hat{B}, \hat{C}, \hat{\Omega}_X$ are the solutions of the optimization problem in Equation (3) for $\mathcal{G} = \mathcal{G}_1$, and, as described above, $\tilde{B}, \tilde{C}, \tilde{\Omega}_X$ are the corresponding parameters of structure $\mathcal{G}_2$.

Similarly, the same reasoning implies $\mathrm{score}_{\mathcal{L}}(\mathcal{G}_2, \mathbf{D}) \geq \mathrm{score}_{\mathcal{L}}(\mathcal{G}_1, \mathbf{D})$. Combining both cases, we have $\mathrm{score}_{\mathcal{L}}(\mathcal{G}_1, \mathbf{D}) = \mathrm{score}_{\mathcal{L}}(\mathcal{G}_2, \mathbf{D})$. Furthermore, since $\mathcal{G}_1$ and $\mathcal{G}_2$ can generate the same set of covariance matrices over variables X and L, they can generate the same set of covariance matrices over variables X. This implies $\dim(\mathcal{G}_1) = \dim(\mathcal{G}_2)$. Therefore, we have $\mathrm{score}_{\mathrm{BIC}}(\mathcal{G}_1, \mathbf{D}) = \mathrm{score}_{\mathrm{BIC}}(\mathcal{G}_2, \mathbf{D})$ and $\mathrm{score}_{\mathrm{dim}}(\mathcal{G}_1, \mathbf{D}) = \mathrm{score}_{\mathrm{dim}}(\mathcal{G}_2, \mathbf{D})$. □

B.3. Proof of Theorem 1

The overall proof strategy below is partly inspired by the proof of Ghassami et al. (2020, Theorem 3).

Theorem 1 (Algebraic equivalence)**.** *Suppose the true DAG $\mathcal{G}^*$ and the distribution Σ_X satisfy the generalized faithfulness assumption. Let $\hat{\mathcal{G}} \in \arg\min_{\mathcal{G} \in \mathbb{G}^m} \mathrm{score}_{dim}(\mathcal{G}, \mathbf{D})$. Then, $\hat{\mathcal{G}}$ and $\mathcal{G}^*$ are algebraic equivalent, i.e., $H(\hat{\mathcal{G}}) = H(\mathcal{G}^*)$, in the large sample limit.*

Proof. Since the search space contains the true DAG $\mathcal{G}^*$ that can generate Σ_X in the large sample limit, the estimated DAG $\hat{\mathcal{G}}$ can also generate Σ_X, because otherwise its score will be infinity and will not be a solution of the optimization problem. Therefore, Σ_X belongs to the distribution set of $\hat{\mathcal{G}}$, i.e., $\Sigma_X \in \mathcal{M}(\hat{\mathcal{G}})$, which implies that Σ_X contains all the equality and inequality constraints of $\hat{\mathcal{G}}$. Under the generalized faithfulness assumption, we have

$$H(\hat{\mathcal{G}}) \subseteq H(\mathcal{G}^*). \tag{6}$$

Now suppose by contradiction that $H(\hat{\mathcal{G}}) \subsetneq H(\mathcal{G}^*)$. This implies $\dim(\hat{\mathcal{G}}) > \dim(\mathcal{G}^*)$, which is a contradiction because the objective function implies $\dim(\hat{\mathcal{G}}) \leq \dim(\mathcal{G}^*)$. Thus, we obtain

$$H(\hat{\mathcal{G}}) \not\subseteq H(\mathcal{G}^*). \tag{7}$$

By Equations (6) and (7), we have $H(\hat{\mathcal{G}}) = H(\mathcal{G}^*)$. □

B.4. Proof of Proposition 2

We first state the following lemma adapted from Leung et al. (2015) that relates the parameter identifiability from a given structure to the underlying degrees of freedom.

Lemma 2 (Leung et al. (2015)). *Suppose $f : \mathbf{S} \to \mathbb{R}^d$ is a polynomial map defined on an open set $\mathbf{S} \subseteq \mathbb{R}^p$. The following statements are equivalent:*

(i) f is generically finite-to-one.

(ii) The Jacobian matrix of f is generically of full column rank.

We now provide the proof of the following proposition.

Proposition 2 (Degrees of freedom). *Suppose that DAG $\mathcal{G}$ satisfies Assumption 2. Then, $\dim(\mathcal{G}) = |\mathcal{G}| + m$.*

Proof. By Corollary 1, it suffices to consider the case where $\Omega_L = I$. Since the structure $\mathcal{G}$ satisfies Assumption 2, by Bollen (1989), the corresponding parameters B, C and Ω_X of $\mathcal{G}$ are identifiable from Σ_X up to certain indeterminacy. Specifically, B is identifiable up to column permutations and sign changes, C is identifiable up to equal row and column permutations, and Ω_X is identifiable. Therefore, the map from B, C, and Ω_X to Σ_X is finite-to-one.

The map from B, C, and Ω_X (with $\Omega_L = I$) to Σ_X is a polynomial map. By Lemma 2, the Jacobian matrix of such map has full column rank. Note that the degrees of freedom (or dimension) of a polynomial map are equal to the maximal rank of the corresponding Jacobian matrix (Geiger et al., 2001, Theorem 10). Therefore, the degrees of freedom are equal to the number of parameters in B, C, and Ω_X, i.e., $\dim(\mathcal{G}) = \|B_\mathcal{G}\|_0 + \|C_\mathcal{G}\|_0 + m = |\mathcal{G}| + m$. □

B.5. Proof of Theorem 2

Theorem 2 (Correctness). *Suppose that the true DAG $\mathcal{G}^*$ and the distribution Σ_X satisfy the generalized faithfulness assumption, and that $\mathcal{G}^*$ satisfies Assumption 2. Let $\hat{\mathcal{G}}$ be a global minimizer of the following optimization problem:*

$$\begin{aligned} \min_{\mathcal{G} \in \mathbb{G}^m} \quad & \text{score}_{dim}(\mathcal{G}, \mathbf{D}) \\ \text{subject to} \quad & \mathcal{G} \text{ satisfies Assumption 2}, \end{aligned} \tag{4}$$

where $\dim(\mathcal{G}) = |\mathcal{G}| + m$. Then, $\hat{\mathcal{G}}$ and $\mathcal{G}^*$ are Markov equivalent in the large sample limit.

Proof. Since the search space contains the true DAG $\mathcal{G}^*$ that can generate Σ_X in the large sample limit, the estimated DAG $\hat{\mathcal{G}}$ can also generate Σ_X, because otherwise its score will be infinity and will not be a solution of the optimization problem. Because $\mathcal{G}^*$ and Σ_X satisfy the generalized faithfulness assumption, we have $H(\hat{\mathcal{G}}) = H(\mathcal{G}^*)$ in the large sample limit by Proposition 2 and restricting the set of structures to those satisfying Assumption 2 in Theorem 1. This indicates that $\hat{\mathcal{G}}$ and Σ_X also satisfy the generalized faithfulness assumption.

Since $\hat{\mathcal{G}}$ and $\mathcal{G}^*$ satisfy Assumption 2 and both can faithfully generate the covariance matrix Σ_X, we have:

- By Silva et al. (2003, Corollary 1), the measurement models of $\hat{\mathcal{G}}$ and $\mathcal{G}^*$ are identical (up to relabeling of latent variables). In other words, the columns of $B_{\hat{\mathcal{G}}}$ are a permutation of the columns of $B_{\mathcal{G}^*}$.
- With a correct measurement model, by leveraging the transitivity of Markov equivalence, it follows straightforwardly from Silva et al. (2006, Theorems 20) that the structural models (i.e., the subgraphs over all and only the latent variables) of $\hat{\mathcal{G}}$ and $\mathcal{G}^*$ are Markov equivalent (up to relabeling of latent variables).

Combining the reasoning for both measurement model and structural model, $\hat{\mathcal{G}}$ and $\mathcal{G}^*$ are Markov equivalent (up to relabeling of latent variables). □

B.6. Proof of Proposition 3

For structure $\mathcal{G}$, we define the following distribution set with the constraint $\Omega_L = I$:

$$\mathcal{M}(\mathcal{G}; \Omega_L = I) \coloneqq \{B(I-C)^{-1}(I-C)^{-\top}B^\top + \Omega_X : \mathrm{supp}(B) \subseteq \mathrm{supp}(B_\mathcal{G}), \mathrm{supp}(C) \subseteq \mathrm{supp}(C_\mathcal{G}), \Omega_X \in \mathrm{diag}(\mathbb{R}^m_{>0})\}.$$

Note that the dimension of the domain and the image space are upper bounds for the dimension of $\mathcal{M}(\mathcal{G}; \Omega_L = I)$. By Lemma 1, it is straightforward to obtain the following corollary.

Corollary 1. *For any structure $\mathcal{G}$ satisfying Equation (1), we have*

$$\mathcal{M}(\mathcal{G}) = \mathcal{M}(\mathcal{G}; \Omega_L = I) \quad \textit{and} \quad \dim(\mathcal{G}) \leq \min\left(|\mathcal{G}| + m, \frac{1}{2}m(m+1)\right).$$

Therefore, it suffices to analyze the degrees of freedom for $\mathcal{M}(\mathcal{G}; \Omega_L = I)$ instead of $\mathcal{M}(\mathcal{G})$.

We first provide the following lemma which shows that an appropriate orthogonal transformation of B and C can generate the same covariance matrix.

Lemma 3 (Orthogonal transformation)**.** *Consider any set of parameters B, C, Ω_X, and Σ_X that satisfy*

$$\Sigma_X = B(I-C)^{-1}(I-C)^{-\top}B^\top + \Omega_X.$$

For any orthogonal matrix Q, i.e., $QQ^\top = I$. the parameters $\tilde{B} = BQ$ and $\tilde{C} = Q^\top C Q$ also satisfy

$$\Sigma_X = \tilde{B}(I-\tilde{C})^{-1}(I-\tilde{C})^{-\top}\tilde{B}^\top + \Omega_X.$$

Proof. The proof follows from straightforward algebraic manipulations:

$$\begin{aligned}
\Sigma_X &= B(I-C)^{-1}(I-C)^{-\top}B^\top + \Omega_X \\
&= BQQ^\top(I-C)^{-1}QQ^\top(I-C)^{-\top}QQ^\top B^\top + \Omega_X \\
&= (BQ)(I-Q^\top CQ)^{-1}(I-Q^\top CQ)^{-\top}(BQ)^\top + \Omega_X \\
&= \tilde{B}(I-\tilde{C})^{-1}(I-\tilde{C})^{-\top}\tilde{B}^\top + \Omega_X.
\end{aligned}$$

□

The following result shows that, in specific cases, some of the edges can be removed from the structure while still leading to the same distribution set.

Lemma 4. *Suppose that DAG $\mathcal{G}$ follows the linear latent variable causal model in Equation (1). Suppose also that there exist $k \geq 2$ latent variables $\mathbf{L}$ in $\mathcal{G}$ with the same set of parents and children, where the number of children (parents) is at least k. Then, there exists a structure, denoted by $\tilde{\mathcal{G}}$, such that: (1) $\tilde{\mathcal{G}}$ is identical to $\mathcal{G}$, except that $k(k-1)/2$ edges among those latent variables $\mathbf{L}$ and their children (parents) are removed in $\tilde{\mathcal{G}}$, and (2) $\mathcal{M}(\tilde{\mathcal{G}}) = \mathcal{M}(\mathcal{G})$.*

Proof. Consider any set of parameters B, C, Ω_X, and Σ_X that satisfy

$$\Sigma_X = B(I-C)^{-1}(I-C)^{-\top}B^\top + \Omega_X. \tag{8}$$

We first consider the case where the number of children is at least k. Denote by $\mathbf{S}$ the set of indices of the latent variables in $\mathbf{L}$. Since the latent variables $\mathbf{L}$ have the same set of children, the rows that correspond to the nonzero entries in each column of $B_{:,\mathbf{S}}$ and $C_{:,\mathbf{S}}$ are the same, which we denote by $\mathbf{R}_1$ and $\mathbf{R}_2$ for $B_{:,\mathbf{S}}$ and $C_{:,\mathbf{S}}$, respectively. Let $D = (B_{\mathbf{R}_1,\mathbf{S}}, C_{\mathbf{R}_2,\mathbf{S}})$ be a matrix by concatenating the rows of $B_{\mathbf{R}_1,\mathbf{S}}$ and $C_{\mathbf{R}_2,\mathbf{S}}$; that is, D is a matrix of dimension $(|\mathbf{R}_1| + |\mathbf{R}_2|) \times k$. Applying orthogonal transformation as in the QR-decomposition, D can be written as $D = \tilde{D}Q$, where $\tilde{D}$ is a lower-triangular matrix and Q is an orthogonal matrix. We rewrite the equation as $\tilde{D} = DQ^{-1}$ where Q^{-1} is also an orthogonal matrix.

Consider the reversed mapping of the indices $\mathbf{A}_1 \coloneqq \{1, 2, \ldots, |\mathbf{R}_1|\}$ and $\mathbf{A}_2 \coloneqq \{|\mathbf{R}_1|+1, |\mathbf{R}_1|+2, \ldots, |\mathbf{R}_1|+|\mathbf{R}_2|\}$. We now construct an $n \times n$ orthogonal matrix U as follows: (1) $U_{\mathbf{S},\mathbf{S}} = Q^{-1}$, (2) the other non-diagonal entries are zero, and (3) the other diagonal entries are one. Let $\tilde{B} = BU$ and $\tilde{C} = U^\top C U$. Clearly, the entries in $\tilde{B}$ are the same as B, except that $B_{\mathbf{R}_1,\mathbf{S}}$ is replaced with $(DQ^{-1})_{\mathbf{A}_1,:} = \tilde{D}_{\mathbf{A}_1,:}$. Similarly, the entries in $\tilde{C}$ are the same as C, except that (1) $C_{\mathbf{R}_2,\mathbf{S}}$ is

replaced with $(DQ^{-1})_{\mathbf{A}_2,:} = \tilde{D}_{\mathbf{A}_2,:}$, and (2) $C_{\mathbf{S},:}$ is replaced with $U^\top_{\mathbf{S},\mathbf{S}} C_{\mathbf{S},:}$. Since the latent variables $\mathbf{L}$ have the same set of parents, we have $\text{supp}(U^\top_{\mathbf{S},\mathbf{S}} C_{\mathbf{S},:}) \subseteq \text{supp}(C_{\mathbf{S},:})$. This implies that $\tilde{B}, \tilde{C}$, and Ω_X are parameterization of, e.g., structure $\tilde{\mathcal{G}}$, where $\tilde{\mathcal{G}}$ has the same edges as $\mathcal{G}$, except that $k(k-1)/2$ of the edges from $\mathcal{G}$ are removed in $\tilde{\mathcal{G}}$ (which correspond to the "non-lower-triangular" entries from D that become zero in $\tilde{D}$ after QR-decomposition). Clearly, $\tilde{\mathcal{G}}$ is identical to $\mathcal{G}$, except that $k(k-1)/2$ edges among those latent variables $\mathbf{L}$ and their children are removed in $\tilde{\mathcal{G}}$. Furthermore, by Lemma 3, the parameters $\tilde{B}$, $\tilde{C}$, and Ω_X can generate the same covariance matrix Σ_X.

Since we are able to construct the same structure $\tilde{\mathcal{G}}$ using the above procedure for every parameterization B, C, and Ω_X of $\mathcal{G}$ in Equation (8), we have $\mathcal{M}(\tilde{\mathcal{G}}; \Omega_L = I) = \mathcal{M}(\mathcal{G}; \Omega_L = I)$, which, by Corollary 1, implies $\mathcal{M}(\tilde{\mathcal{G}}) = \mathcal{M}(\mathcal{G})$. The same reasoning also applies when the number of parents is at least k, i.e., such a structure $\tilde{\mathcal{G}}$ can also be constructed. □

We now provide the proof of the following proposition.

Proposition 3. *Suppose that DAG $\mathcal{G}$ follows the linear latent variable causal model in Equation (1). Suppose also that there exist $k \geq 2$ latent variables in $\mathcal{G}$ with the same set of parents and children, where either the number of parents or children is at least k. Then,* $\dim(\mathcal{G}) \leq |\mathcal{G}| + m - k(k-1)/2$.

Proof. By Lemma 4, there exists a structure, denoted by $\tilde{\mathcal{G}}$, such that: (1) $\tilde{\mathcal{G}}$ is identical to $\mathcal{G}$, except that $k(k-1)/2$ edges are removed in $\tilde{\mathcal{G}}$, and (2) $\mathcal{M}(\tilde{\mathcal{G}}) = \mathcal{M}(\mathcal{G})$. By Corollary 1, this implies

$$\dim(\mathcal{G}) = \dim(\tilde{\mathcal{G}}) \leq |\tilde{\mathcal{G}}| + m = |\mathcal{G}| + m - \frac{1}{2}k(k-1).$$

□

B.7. Proof of Proposition 4

Proposition 4 (Degrees of freedom)**.** *Suppose that DAG $\mathcal{G}$ satisfies Assumption 2. Then, Algorithm 2 outputs the upper bound of* $\dim(\mathcal{G})$.

Proof. Let $\mathbf{L}_1, \mathbf{L}_2, \ldots, \mathbf{L}_p$ be the pairwise disjoint sets of latent variables in structure $\mathcal{G}$ such that (1) each $\mathbf{L}_i$ has at least two latent variables, (2) the variables of each $\mathbf{L}_i$ have the same set of parents and children in $\mathcal{G}$, (3) no proper superset of each $\mathbf{L}_i$ has the same set of parents and children $\mathcal{G}$. Since the structure $\mathcal{G}$ is a DAG, we assume without loss of generality that $\mathbf{L}_1, \mathbf{L}_2, \ldots, \mathbf{L}_p$ are sorted based on the reversed causal ordering in $\mathcal{G}$. That is, variables $\mathbf{L}_{i_1}$ cannot be the ancestors of variables $\mathbf{L}_{i_2}$ in structure $\mathcal{G}$ for $i_1 < i_2$.

Although Algorithm 2 does not impose any specific order on the sets of latent variables, we suppose that the algorithm computes the dimension based on $\mathbf{L}_1, \mathbf{L}_2, \ldots, \mathbf{L}_p$ (sorted according to reversed causal ordering), which does not affect the computed degrees of freedom. Denote by d_j the output of the algorithm in the j-th iteration where $j \in [p]$. The final output is then d_p. It suffices to show $\dim(\mathcal{G}) \leq d_j$ in each iteration.

We provide a proof by induction. Specifically, we show that, for the j-th iteration, there exists a structure $\tilde{\mathcal{G}}_j$ such that:

1. $\tilde{\mathcal{G}}_j$ is identical to $\mathcal{G}$, except that $|\mathcal{G}| + m - d_j$ edges among the variables $\bigcup_{i=1}^{j} \mathbf{L}_i$ and their children are removed in $\tilde{\mathcal{G}}_j$.
2. $\mathcal{M}(\tilde{\mathcal{G}}_j) = \mathcal{M}(\mathcal{G})$.

By Corollary 1, this implies the desired outcome

$$\dim(\mathcal{G}) = \dim(\tilde{\mathcal{G}}_j) \leq |\tilde{\mathcal{G}}_j| + m = |\mathcal{G}| + m - (|\mathcal{G}| + m - d_j) = d_j.$$

For induction, we first consider the base case $j = 1$. By assumption, the variables $\mathbf{L}_1$ have the same set of parents and children in $\mathcal{G}$, where, under Assumption 3, the number of children is at least $|\mathbf{L}_1|$. By Lemma 4, there exists a structure, denoted by $\tilde{\mathcal{G}}_1$, such that: (1) $\tilde{\mathcal{G}}_1$ is identical to $\mathcal{G}$, except that $|\mathbf{L}_1|(|\mathbf{L}_1| - 1)/2 = |\mathcal{G}| + m - d_1$ edges among the variables $\mathbf{L}_1$ and their children are removed in $\tilde{\mathcal{G}}_1$, and (2) $\mathcal{M}(\tilde{\mathcal{G}}_1) = \mathcal{M}(\mathcal{G})$. Therefore, the base case is done.

Suppose that the statements hold for $j = t$, i.e., there exists a structure $\tilde{\mathcal{G}}_t$ such that: (1) $\tilde{\mathcal{G}}_t$ is identical to $\mathcal{G}$, except that $|\mathcal{G}| + m - d_t$ edges among the variables $\bigcup_{i=1}^{t} \mathbf{L}_i$ and their children are removed in $\tilde{\mathcal{G}}_t$, and (2) $\mathcal{M}(\tilde{\mathcal{G}}_t) = \mathcal{M}(\mathcal{G})$.

Now consider $j = t+1$. Note that only the edges among $\bigcup_{i=1}^{t} \mathbf{L}_i$ and their children are removed in $\tilde{\mathcal{G}}_t$ (as compared to $\mathcal{G}$); by assumption, the variables $\bigcup_{i=1}^{t} \mathbf{L}_i$ are not ancestors of the variables $\mathbf{L}_{t+1}$. Therefore, the incoming and outgoing edges of variables $\mathbf{L}_{t+1}$ in $\tilde{\mathcal{G}}_t$ are the same as those in $\mathcal{G}$. This implies that $\mathbf{L}_{t+1}$ have the same set of parents and children in $\tilde{\mathcal{G}}_t$, because they have the same set of parents and children in $\mathcal{G}$. Furthermore, under Assumption 3, the number of children is at least $|\mathbf{L}_{t+1}|$. By Lemma 4, there exists a structure, denoted by $\tilde{\mathcal{G}}_{t+1}$, such that: (1) $\tilde{\mathcal{G}}_{t+1}$ is identical to $\tilde{\mathcal{G}}_t$, except that $|\mathbf{L}_{t+1}|(|\mathbf{L}_{t+1}|-1)/2$ edges among the variables $\mathbf{L}_{t+1}$ and their children are removed in $\tilde{\mathcal{G}}_{t+1}$, and (2) $\mathcal{M}(\tilde{\mathcal{G}}_{t+1}) = \mathcal{M}(\tilde{\mathcal{G}}_t)$.

By the induction hypothesis, we have $\mathcal{M}(\tilde{\mathcal{G}}_{t+1}) = \mathcal{M}(\tilde{\mathcal{G}}_t) = \mathcal{M}(\mathcal{G})$. Also, it is clear that $\tilde{\mathcal{G}}_{t+1}$ is identical to $\mathcal{G}$, except that

$$|\mathcal{G}| + m - d_t + \frac{1}{2}|\mathbf{L}_{t+1}|(|\mathbf{L}_{t+1}|-1) = |\mathcal{G}| + m - d_{t+1}$$

edges among the variables $\bigcup_{i=1}^{t+1} \mathbf{L}_i$ and their children are removed in $\tilde{\mathcal{G}}_{t+1}$. Therefore, the induction step is done. □

B.8. Definition of Graph Operators and Proof of Theorem 3

We provide the definition of structure operations $\mathcal{O}_{\text{atomic}}$, $\mathcal{O}_{\text{min}}$, and $\mathcal{O}_{\text{skeleton}}$ below, with an example in Figure 3.

Definition 4 (Minimal-graph operator (Huang et al., 2022; Dong et al., 2023))**.** *For every two atomic covers* $\mathbf{L}$ *and* $\mathbf{P}$ *in structure* $\mathcal{G}$*, we 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 5 (Skeleton operator (Huang et al., 2022; Dong et al., 2023))**.** *Given an atomic cover* $\mathbf{L}$ *in structure* $\mathcal{G}$*. Consider* $\mathcal{S}$ *as the set of atomic covers such that for all* $\mathbf{S} \in \mathcal{S}$*, we have* $\mathbf{S} \subseteq \mathbf{L}$*. Let* $\mathbf{C} = PCh_{\mathcal{G}}(\mathbf{L}) \setminus \cup_{\mathbf{S}\in\mathcal{S}} PCh_{\mathcal{G}}(\mathbf{S})$*. We add edges from elements in* $\mathbf{L}$ *to elements in* $\mathbf{C}$*, and we denote such an operator as skeleton operator* $\mathcal{O}_{skeleton}(\mathcal{G})$*.*

Definition 6 (Intra atomic operator)**.** *For every atomic cover* $\mathbf{L}$ *in structure* $\mathcal{G}$*, if* $|\mathbf{L}| \geq 2$*, then we add edges between elements in* $\mathbf{L}$ *such that* $\mathbf{L}$ *form a fully connected DAG. We denote such an operator as intra atomic operator* $\mathcal{O}_{atomic}(\mathcal{G})$*.*

Example 1 (Example for graph operations)**.** *Let the graph in Figure 3(a) be* $\mathcal{G}$*. By the skeleton operator, we add edges from* L_2 *and* L_3 *to* X_7*, and we arrive at* $\mathcal{O}_{skeleton}(\mathcal{G})$*, which is shown in Figure 3(b). By the minimal graph operator, we delete* L_5 *and directly link* L_1 *to* X_8, X_9, X_{10}*, and arrive at* $\mathcal{O}_{min}(\mathcal{O}_{skeleton}(\mathcal{G}))$*, which is shown in Figure 3(c). Finally, by the intra atomic operator, we add edges among* L_2, L_3, L_4 *such that they are fully connected, and arrive at* $\mathcal{O}_{atomic}(\mathcal{O}_{min}(\mathcal{O}_{skeleton}(\mathcal{G})))$*, which is shown in Figure 3(d).*

Remark 2 (Necessity of graph operators)**.** *These three graph operators do not change the rank constraints (among measured variables), and thus by using rank constraints for causal discovery as in Huang et al. (2022); Dong et al. (2023), we can at most identify the structure up to these graph operators.*

We now provide the proof of the following result.

Theorem 3 (Correctness)**.** *Suppose that the true DAG* $\mathcal{G}^*$ *and the distribution* Σ_X *satisfy the generalized faithfulness assumption, and that* $\mathcal{G}^*$ *satisfies Assumption 3. Let* $\hat{\mathcal{G}}$ *be a global minimizer of the following optimization problem:*

$$\begin{aligned} \min_{\mathcal{G}\in\mathbb{G}^m} \quad & \text{score}_{dim}(\mathcal{G}, \mathbf{D}) \\ \text{subject to} \quad & \mathcal{G} \text{ satisfies Assumption 3}, \\ & \mathcal{G} = \mathcal{O}_{min}(\mathcal{O}_{skeleton}(\mathcal{G})). \end{aligned}$$

Then, $\mathcal{O}_{atomic}(\hat{\mathcal{G}})$ *and* $\mathcal{O}_{atomic}(\mathcal{O}_{min}(\mathcal{O}_{skeleton}(\mathcal{G}^*)))$ *are Markov equivalent in the large sample limit.*

Proof. Since the search space contains the DAG $\mathcal{O}_{\text{min}}(\mathcal{O}_{\text{skeleton}}(\mathcal{G}^*))$ that can generate Σ_X in the large sample limit, the estimated DAG $\hat{\mathcal{G}}$ can also generate Σ_X, because otherwise its score will be infinity and will not be a solution of the optimization problem. Because $\mathcal{G}^*$ and Σ_X satisfy the generalized faithfulness assumption, we have $H(\hat{\mathcal{G}}) = H(\mathcal{G}^*)$ in the large sample limit by restricting the set of structures to those satisfying Assumption 3 and $\mathcal{G} = \mathcal{O}_{\text{min}}(\mathcal{O}_{\text{skeleton}}(\mathcal{G}))$ in Theorem 1. This indicates that $\hat{\mathcal{G}}$ and Σ_X also satisfy the generalized faithfulness assumption.

Let $\mathcal{G}'$ be the structure estimated by Algorithm 1 in Huang et al. (2022) based on the covariance matrix Σ_X. Since $\hat{\mathcal{G}}$ satisfies Assumption 3 and can faithfully generate Σ_X, Huang et al. (2022, Theorem 10) and Dong et al. (2023, Theorem 13)

imply that $\mathcal{O}_{\text{atomic}}(\mathcal{G}')$ and $\mathcal{O}_{\text{atomic}}(\mathcal{O}_{\text{min}}(\mathcal{O}_{\text{skeleton}}(\hat{\mathcal{G}})))$ are Markov equivalent. With similar reasoning, we can show that $\mathcal{O}_{\text{atomic}}(\mathcal{G}')$ and $\mathcal{O}_{\text{atomic}}(\mathcal{O}_{\text{min}}(\mathcal{O}_{\text{skeleton}}(\mathcal{G}^*)))$ are Markov equivalent. Therefore, by the transitivity of Markov equivalence, $\mathcal{O}_{\text{atomic}}(\mathcal{O}_{\text{min}}(\mathcal{O}_{\text{skeleton}}(\hat{\mathcal{G}})))$ and $\mathcal{O}_{\text{atomic}}(\mathcal{O}_{\text{min}}(\mathcal{O}_{\text{skeleton}}(\mathcal{G}^*)))$ are Markov equivalent.

Recall that $\hat{\mathcal{G}} = \mathcal{O}_{\text{min}}(\mathcal{O}_{\text{skeleton}}(\hat{\mathcal{G}}))$. This implies that $\mathcal{O}_{\text{atomic}}(\hat{\mathcal{G}})$ and $\mathcal{O}_{\text{atomic}}(\mathcal{O}_{\text{min}}(\mathcal{O}_{\text{skeleton}}(\mathcal{G}^*)))$ are Markov equivalent. □

C. Supplementary Experiment Details

Implementation details. To improve the efficiency of Algorithm 2, we iterate over the latent atomic covers to identify latent variables with the same set of parents and children. For our exact search method, we use L-BFGS (Byrd et al., 1995) implemented through `SciPy` (Virtanen et al., 2020) and `PyTorch` (Paszke et al., 2019) packages (with the default hyperparameters) to solve the optimization problem in Equation (3) when computing BIC. The experiments for the exact search method are conducted on 16 CPUs in parallel. For the continuous search method, i.e., SALAD-CS, we use augmented Lagrangian method (Bertsekas, 1982; 1999; Nocedal & Wright, 2006) to solve the continuous constrained optimization problem, in which each subproblem is solved using the Adam optimizer (Kingma & Ba, 2014) with 3000 iterations. Furthermore, Equation (5) involves a nonconvex optimization problem; similar to continuous optimization methods for causal discovery (Ng et al., 2024), this procedure may yield suboptimal local solutions. Thus, we run the SALAD-CS method from 10 random initializations, and select the final solution with the best score.

For HUANG, GIN, and RCD, we use the publicly available implementations with default hyperparameters. For FOFC, we use the implementation through the `py-causal` package (Scheines et al., 1998) with Wishart test and significance level of 0.001. Note that we also experimented with significance level of 0.01, 0.05 and 0.1, for which many of the runs are invalid (because an error occurred).

Metrics. Since the goal is to recover the structure up to Markov equivalence, we compute the SHDs over the MECs. Note that the labeling of the latent variables is not important; therefore, we calculate the SHDs of the estimated MECs over all possible permutations of the latent variables, and select the smallest SHD. Similarly, we also compute the F1 scores of the estimated skeletons over all permutations of latent variables, and select the highest F1 score.

For FOFC, an error occurred in some of the experimental runs. Therefore, we additionally report the number of valid runs (for which an error did not occur).

D. Runtime and Computational Efficiency

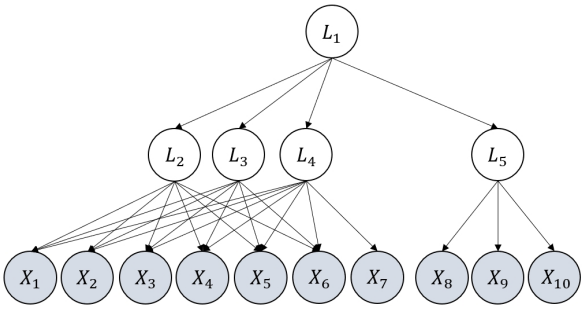
In this section, we report the runtime for different methods considered. For the 1-factor models, our SALAD method has a runtime of 8.77 ± 0.73 and 44.88 ± 8.02 minutes for 10 and 11 measured variables, respectively, while for hierarchical structures, it takes 16.11 ± 2.01 minutes. For the SALAD-CS method, each optimization run takes 14.17 ± 0.69 and 14.73 ± 1.91 minutes for 10 and 11 measured variables, respectively. For the baselines, GIN, HUANG, and FOFC generally finish within one minute. For the 1-factor models, RCD requires 17.23 ± 34.84 and 13.06 ± 24.33 minutes for 10 and 11 measured variables, respectively, while for hierarchical structures, it has a runtime of 7.76 ± 15.33 minutes.

Our methods have a comparable runtime as RCD, but achieve better performance. Although the runtime of our methods exceeds that of GIN, HUANG, and FOFC, the improvement in the causal discovery performance is significant. As discussed in Appendix C, our experiments are conducted on CPUs. It is worth noting that the runtime may be further decreased by (i) conducting experiments with GPU acceleration (specifically when using gradient-based optimization to solve Equations (3) and (5)), or (ii) performing more score computations of different structures (specifically for exact search) concurrently on different CPUs.

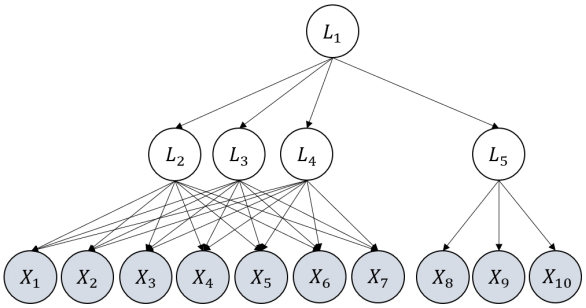
Indeed, the relatively long runtime of our methods may be unsurprising because, even without latent variables, exact score-based methods (Singh & Moore, 2005; Yuan & Malone, 2013) are known to require a long runtime. The search procedure developed in our work serves as a proof of concept, tailored for scenarios involving a relatively small number of variables. Nonetheless, the empirical performance validates the effectiveness of score-based methods for estimating latent variable causal models. Future works include developing greedy approaches similar to GES to make the search procedure more efficient and scalable.

Table 2: SHDs of MECs across various structural assumptions and sample sizes. For each setting, the top two methods are in bold. For FOFC, the number within the brackets indicates the number of valid runs (for which an error did not occur).

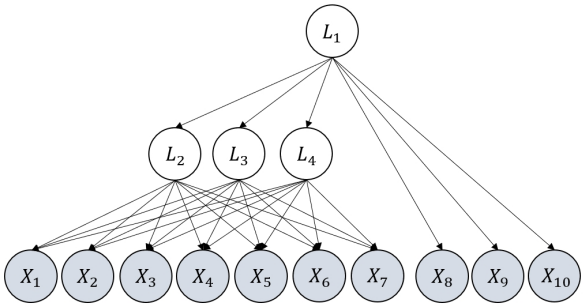
Model type	Sample size	SALAD	SALAD-CS	HUANG	FOFC	GIN
1-Factor model	100	$\mathbf{0.33 \pm 0.65}$	$\mathbf{0.75 \pm 1.76}$	11.50 ± 4.58	3.13 ± 2.64 (8)	15.25 ± 1.36
	300	$\mathbf{0.08 \pm 0.29}$	$\mathbf{0.17 \pm 0.39}$	3.50 ± 3.45	1.67 ± 1.11 (9)	15.25 ± 1.36
	1000	$\mathbf{0.33 \pm 0.89}$	$\mathbf{0.08 \pm 0.29}$	1.67 ± 0.78	1.67 ± 1.07 (12)	15.25 ± 1.36
	3000	$\mathbf{0 \pm 0}$	$\mathbf{0.17 \pm 0.39}$	1.67 ± 0.78	1.25 ± 1.14 (12)	15.17 ± 1.34
	10000	$\mathbf{0 \pm 0}$	$\mathbf{0.17 \pm 0.39}$	1.67 ± 0.78	1.18 ± 1.17 (11)	14.75 ± 2.01
Hierarchical structure	100	$\mathbf{4.50 \pm 3.32}$	N/A	$\mathbf{14.25 \pm 2.73}$	N/A (0)	18.50 ± 3.73
	300	$\mathbf{3.67 \pm 3.06}$	N/A	$\mathbf{11.42 \pm 3.20}$	N/A (0)	18.50 ± 3.73
	1000	$\mathbf{2.75 \pm 2.18}$	N/A	$\mathbf{8.00 \pm 3.64}$	N/A (0)	18.50 ± 3.73
	3000	$\mathbf{1.92 \pm 1.98}$	N/A	$\mathbf{4.92 \pm 4.08}$	N/A (0)	18.50 ± 3.73
	10000	$\mathbf{1.42 \pm 1.56}$	N/A	$\mathbf{4.17 \pm 4.91}$	N/A (0)	18.42 ± 3.75



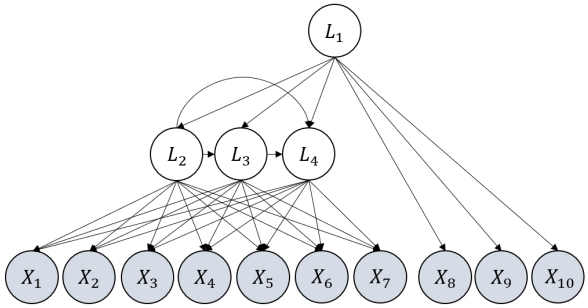
(a) Graph $\mathcal{G}$.



(b) $\mathcal{O}_{\text{skeleton}}(\mathcal{G})$.

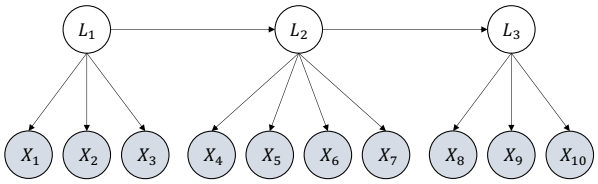


(c) $\mathcal{O}_{\text{min}}(\mathcal{O}_{\text{skeleton}}(\mathcal{G}))$.

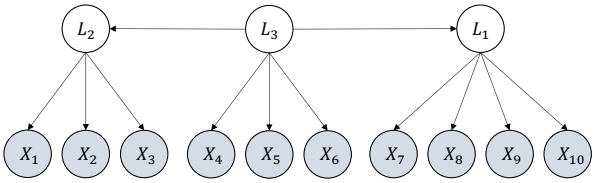


(d) $\mathcal{O}_{\text{atomic}}(\mathcal{O}_{\text{min}}(\mathcal{O}_{\text{skeleton}}(\mathcal{G})))$.

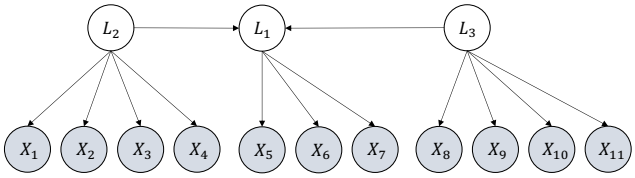
Figure 3: Example to illustrate graph operators $\mathcal{O}_{\text{atomic}}$, $\mathcal{O}_{\text{min}}$, and $\mathcal{O}_{\text{skeleton}}$.



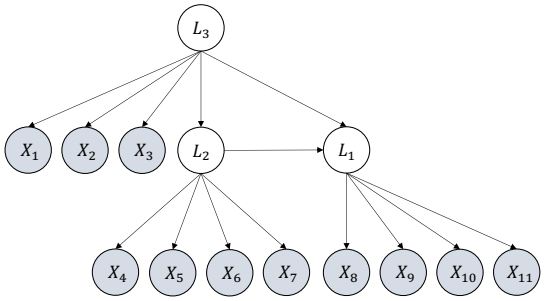
(a) Example 1.



(b) Example 2.

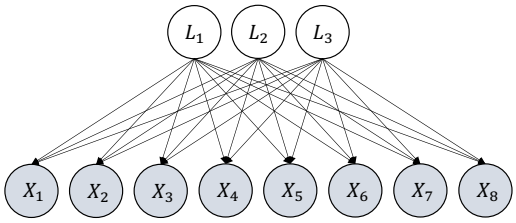


(c) Example 3.

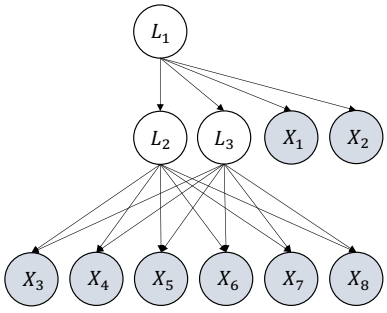


(d) Example 4.

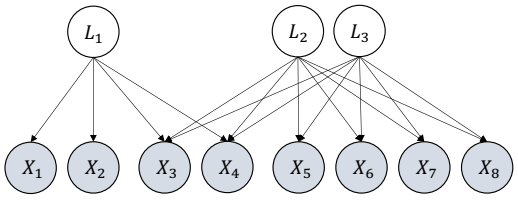
Figure 4: Ground truths for 1-factor latent variable models.



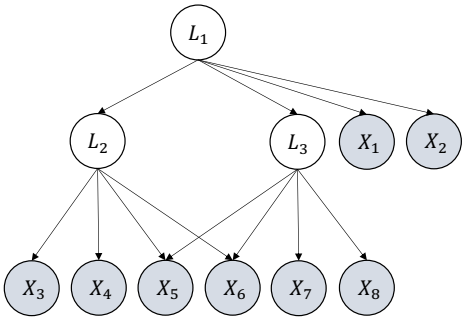
(a) Example 1.



(b) Example 2.



(c) Example 3.



(d) Example 4.

Figure 5: Ground truths for latent hierarchical structures.