SCORE-BASED GREEDY SEARCH FOR STRUCTURE IDENTIFICATION OF PARTIALLY OBSERVED CAUSAL MODELS

Xinshuai Dong[1] **Ignavier Ng** [1] **Haoyue Dai**[1] **Jiaqi Sun**[1] **Xiangchen Song**[1]
Peter Spirtes[1] **Kun Zhang**[1,2]
[1]Carnegie Mellon University
[2]Mohamed bin Zayed University of Artificial Intelligence

ABSTRACT

Identifying the structure of a partially observed causal system is essential to various scientific fields. Recent advances have focused on constraint-based causal discovery to solve this problem, and yet in practice these methods often face challenges related to multiple testing and error propagation. These issues could be mitigated by a score-based method and thus it has raised great attention whether there exists a score-based greedy search method that can handle the partially observed scenario. In this work, we propose the first score-based greedy search method for the identification of structure involving latent variables with identifiability guarantees. Specifically, we propose the Generalized N Factor Model and establish its global consistency: the true structure including latent variables can be identified up to the Markov equivalence class by using score. We then design Latent variable Greedy Equivalence Search (LGES), a greedy search algorithm for this class of models with well-defined operators, which searches very efficiently over the graph space to find the optimal structure. Our experiments on both synthetic and real-life data validate the effectiveness of our method (code will be available at `https://github.com/dongxinshuai/scm-identify`).

1 INTRODUCTION AND RELATED WORK

Causal discovery aims at identifying the causal relations from observational data and it is crucial to many scientific fields (Spirtes et al., 2001; Pearl, 2009). However, traditional methods such as PC (Spirtes et al., 2001), GES (Chickering, 2002), and LiNGAM (Shimizu et al., 2006), rely on the causal sufficiency assumption, i.e., the absence of latent variables, which hardly holds in many real-world scenarios. Therefore, extensive efforts are being made towards structure identification of a partially observed causal model.

To handle this problem, the earliest attempts make use of conditional independence including 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), as well as over-complete ICA-based techniques (Hoyer et al., 2008; Salehkaleybar et al., 2020). Yet, these methods only focus on identifiable relations among observed variables, and the results provide limited information about structure among latent variables.

To this end, recent advance has been on the discovery of the entire structure including latent variables, by introducing additional parametric or graphical assumptions. Typical methods include rank or tetrad constraints based under linearity assumption (Silva et al., 2003; 2006; Silva & Scheines, 2005; Choi et al., 2011; Kummerfeld & Ramsey, 2016; Huang et al., 2022; Dong et al., 2024a), high-order moments (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; Améndola et al., 2023; Wang & Drton, 2023), matrix decomposition (Anandkumar et al., 2013), mixture oracles (Kivva et al., 2021), and multiple domains (Zeng et al., 2021; Sturma et al., 2023). Despite of the asymptotic correctness, however, these methods generally fall into the category of constraint-based methods; they rely heavily on statistical tests to iteratively construct the structure and thus suffer from the problem of

multiple-testing and error propogation (Spirtes, 2010; Colombo et al., 2012), especially with small sample size and large number of variables.

On the other hand, score-based causal discovery may not suffer from these issues and thus is believed to be more practially favorable (Nandy et al., 2018; Ramsey et al., 2017). One of the most classical methods is Greedy Equivalence Search (GES) (Chickering, 2002), and yet it cannot handle the existence of latent variables. Later, various score-based methods that allow the existence of latent variables have been proposed (Shpitser et al., 2012; Triantafillou & Tsamardinos, 2016; Nowzohour et al., 2017; Bhattacharya et al., 2021a; Shahin & Chechik, 2020; Bernstein et al., 2020; Bellot & van der Schaar, 2021; Claassen & Bucur, 2022), and yet they still focus only on relations among observed variables, except the one by Zhang (2004) without identifiability and another with exact search by Ng et al. (2024) (a more detailed discussion between this work and the exact search (Ng et al., 2024) can be found in Section A). As a consequence, the following crucial question naturally arises: Is it possible to develop a score-based greedy search method that can efficiently recover the entire underlying structure involving latent variables with an asymptotic correctness guarantee?

To address such a challenging problem, we are confronted with three fundamental questions: (i) What is the essential relations between structure identifiability and likelihood scores? (ii) What graphical assumptions are needed to uniquely recover the structure by using score? (iii) How can we design an efficient algorithm to search over the graph space to find the optimal structure? We provide our answers to these questions respectively in Sections 3.1, 3.2 and 4 and our contributions can be summarized as follows.

- We characterize how the likelihood score can be used for the structure identification of partially observed linear causal models. Specifically, we show that the structure with the best likelihood score and minimal dimension is algebraically equivalent to the ground truth (in Theorem 1).
- We propose the GNFM (in Def. 2), and accordingly establish the global consistency of score for it - the whole underlying structure can be uniquely recovered up to the bMarkov Equivalence Class (MEC) by using the score (in Theorem 2 and Corollary 1). This graphical condition is rather mild and takes the prevalent one factor model (Silva et al., 2003) as a special case.
- We develop the Latent variable Greedy Equivalence Search (LGES), an efficient causal discovery algorithm for identifying structure involving latent variables. To our best knowledge, this is the first score-based greedy search with identifiability guarantees in the partially observed scenario (in Thm. 3). Our experiments on both synthetic and real-world dataset empirically validate its effectiveness.

2 PRELIMINARIES

2.1 PROBLEM SETTING

We aim to identify the structure of a partially observed linear causal model, defined as follows.

Definition 1 (Partially Observed Linear Causal Models)**.** *Let $\mathcal{G} := (\mathbf{V}_\mathcal{G}, \mathbf{E}_\mathcal{G})$ be a Directed Acyclic Graph (DAG) and variables follow a linear SEM as $\mathbf{V}_\mathcal{G} = F^T\mathbf{V}_\mathcal{G} + \epsilon_{\mathbf{V}_\mathcal{G}}$, where $\mathbf{V}_\mathcal{G} = \mathbf{L}_\mathcal{G} \cup \mathbf{X}_\mathcal{G} = \{\mathsf{V}_i\}_1^{m+n} = \{\mathsf{L}_i\}_1^m \cup \{\mathsf{X}_i\}_1^n$ contains m latent variables and n observed variables, $F = (f_{j,i})$ is the weighted adjacency matrix and $f_{j,i} \neq 0$ if and only if V_j is a parent of V_i in $\mathcal{G}$, and ϵ_{V_i} represents the Gaussian noise term of V_i.*

Our goal is to identify the underlying structure $\mathcal{G}$ over all the variables $\mathbf{L}_\mathcal{G} \cup \mathbf{X}_\mathcal{G}$, given i.i.d. samples of observed variables $\mathbf{X}_\mathcal{G}$ only. Note that the name/order of latent variables can never be identified so we focus on structure identification up to permutation of latent variables. Without loss of generality, we can assume that all variables have zero mean, and thus the observational data can also be summarized as the empirical covariance matrix over observed variables, i.e., $\hat{\Sigma}_{\mathbf{X}_\mathcal{G}}$. We use V and $\mathbf{V}$, to denote a random variable and a set of variables, respectively. We drop the subscript $\mathcal{G}$ in $\mathbf{L}_\mathcal{G}$ and $\mathbf{X}_\mathcal{G}$ when the context is clear. For a matrix M, we define its support set as $\mathrm{supp}(M) \coloneqq \{(i,j) : M_{i,j} \neq 0\}$. $\mathcal{G}_1$ and $\mathcal{G}_2$ belong to the same MEC iff they share the same skeleton and set of v-structures (in Definition 9) over the entire graph including latent variables.

2.2 LIKELIHOOD SCORE

Despite the asymptotic correctness of constraint-based causal discovery approaches, in practice these methods often suffer from the problem of multi-testing and error propagation (Spirtes, 2010; Colombo et al., 2012). In the finite sample case, they rely heavily on statistical tests to iteratively build the

result, while the power of each test might be limited especially when the sample size is small and number of variables is large. On the contrary, score-based causal discovery methods may not suffer from these problems and could be practically more favorable (Nandy et al., 2018; Ramsey et al., 2017), especially when the sample size is small (also empirically validated in Section 5.2). A more detailed discussion about error propagation can be found in Section C.8.

To this end, in this work we aim at structure identification based on the use of likelihood scores in the partially observed scenario. In contrast to the fully observed case, in the presence of latent variables, the formulation of the likelihood is not trivial, and we provide it in what follows.

Proposition 1 (Parameterization of Population Covariance (Dong et al., 2024b)). *Consider the model defined in Def. 1, and let* $F = \begin{pmatrix} F_{\mathbf{LL}} & F_{\mathbf{LX}} \\ F_{\mathbf{XL}} & F_{\mathbf{XX}} \end{pmatrix}$, *and* $\Omega = \begin{pmatrix} \Omega_{\epsilon_{\mathbf{L}}} & 0 \\ 0 & \Omega_{\epsilon_{\mathbf{X}}} \end{pmatrix}$, *where* Ω *is the diagonal covariance matrix of* $\epsilon_{\mathbf{V}_\mathcal{G}}$. *Let* $M = ((I - F_{\mathbf{LL}} - F_{\mathbf{LX}}(I - F_{\mathbf{XX}})^{-1}F_{\mathbf{XL}}))^{-1}$, $N = (((I - F_{\mathbf{LL}})F_{\mathbf{XL}}^{-1}(I - F_{\mathbf{XX}}) - F_{\mathbf{LX}}))^{-1}$, *and* $\Sigma_{\mathbf{L}} = M^T\Omega_{\epsilon_{\mathbf{L}}}M + N^T\Omega_{\epsilon_{\mathbf{X}}}N$. *Then the population covariance of* $\mathbf{X}$ *can be formulated as*

$$\Sigma_{\mathbf{X}} = (I - F_{\mathbf{XX}})^{-T}\Big(F_{\mathbf{LX}}^T\Sigma_{\mathbf{L}}F_{\mathbf{LX}} + \Omega_{\epsilon_{\mathbf{X}}}NF_{\mathbf{LX}} + \Omega_{\epsilon_{\mathbf{X}}} + F_{\mathbf{LX}}^TN^T\Omega_{\epsilon_{\mathbf{X}}}\Big)(I - F_{\mathbf{XX}})^{-1}. \quad (1)$$

By making use of Proposition 1 to parametrize $\Sigma_{\mathbf{X}}$, the maximum log-likelihood of a given structure $\mathcal{G}$ and observation $\hat{\Sigma}_{\mathbf{X}}$ is as follows (tr and $\det$ refers to matrix trace and determinant, respectively).

$$\text{score}_{\text{ML}}(\mathcal{G}, \hat{\Sigma}_{\mathbf{X}}) = \max_{(F,\Omega):\ \text{supp}(F)\subseteq\text{supp}(F_\mathcal{G}),\Omega\in\text{diag}(\mathbb{R}_{>0}^{n+m})} \mathcal{L}, \quad (2)$$

$$\mathcal{L} = -(N/2)(\mathrm{tr}((\Sigma_{\mathbf{X}})^{-1}\hat{\Sigma}_{\mathbf{X}}) + \log\det\Sigma_{\mathbf{X}}). \quad (3)$$

We next show the theoretical foundation of using maximum likelihood score for the structure identification of partially observed linear causal models.

3 Score-Based Identifiability Theory for Partially Observed Causal Models

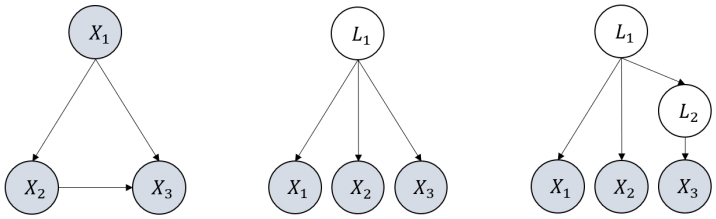
3.1 Algebraic Equivalence by Score and Dimension

Consider a model in Def. 1. Its structure imposes various types of equality (i.e., algebraic) constraints on the covariance matrices (over observed variables), no matter how its parameters (F, Ω) may change. The imposed equality constraints are properties of the observational distribution and contain crucial graphical information about the underlying structure $\mathcal{G}$. Such constraints include conditional independence (i.e., vanishing partial correlation) constraints (Spirtes et al., 2001), rank constraints (i.e., vanishing determinant) (Spirtes et al., 2001; Sullivant et al., 2010), and possibly Verma constraints (Verma & Pearl, 1991). An overview can be found in Drton (2018).

Let $H(\mathcal{G})$ be the set of equality constraints imposed by structure $\mathcal{G}$ on the generated covariance matrices over observed variables (detailed in Definition 6), $\mathbb{G}^n$ be the set of all DAG structures that has n measured variables, and $\mathbb{H}^n := \bigcup_{\mathcal{G}\in\mathbb{G}^n} B(\mathcal{G})$, where $B(\mathcal{G})$ consists of the canonical and minimal set of equality constraints with respect to reduced Gröbner basis (detailed in Definition 7). We say two structures $\mathcal{G}_1$ and $\mathcal{G}_2$ are algebraically equivalent, if they lead to the same equality constraints (on the observational distribution), i.e., $H(\mathcal{G}_1) = H(\mathcal{G}_2)$ (van Ommen & Mooij, 2017). Similar to the classical CI faithfulness assumption in causal discovery (Spirtes et al., 2001), to better relate the constraints to the underlying structure, we assume the generalized faithfulness as follows.

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

In Assumption 1, it suffices to use $\Sigma_{\mathbf{X}}$ to denote the distribution, as $\mathbf{X}$ are jointly gaussian and mean do not contain any information about structure (Ghassami et al., 2020). Note that different types of faithfulness assumptions have been adopted in causal discovery, e.g., CI faithfulness and rank faithfulness (Spirtes et al., 2001; Ghassami et al., 2020; Huang et al., 2022; Dong et al., 2024a) and Assumption 1 is the generalized version of them for linear causal models. Similar to CI faithfulness, generalized faithfulness is justified by that the set of parameters that result in violation has Lebesgue measure zero (Ghassami et al., 2020) and it has been widely adopted in the field (Ng et al., 2020; Bhattacharya et al., 2021b; Sethuraman et al., 2023). On the other hand, without



(a) Ground truth $\mathcal{G}^*$. (b) $\hat{\mathcal{G}}_1$ by Theorem 1. (c) $\hat{\mathcal{G}}_2$ by Theorem 1.

Figure 1: Without further graphical assumption, the algebraic equivalence class is very large and not very informative: suppose the ground truth $\mathcal{G}^*$ in (a), by Theorem 1 we may arrive at either $\hat{\mathcal{G}}_1$ (b) or $\hat{\mathcal{G}}_2$ (c), both are algebraically equivalent to $\mathcal{G}^*$.

faithfulness, graphical information extracted from observations cannot be trusted, which makes structure identification extremely hard, if not impossible.

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 also be viewed as the number of free parameters of the set of distribution it can generate). In the absence of latent variables, the degrees of freedom are nothing but the sum of the number of edges and the number measured variables. However, in the presence of latent variables, it does not necessarily hold (Geiger et al., 1996; 2001). Without any specific graphical assumption, capturing the dimension could be highly non-trivial; e.g., the analysis of the degrees of freedom for sparse factor analysis, where latent variables are independent, already involves complex techniques from algebraic statistics (Drton et al., 2023). We note that the focus of this work is not to characterize the dimension of each structure. In contrast, we only need to know the basic idea of dimension here and later we will show that a greedy search does not necessarily rely on knowing the exact dimension of a graph.

Now we are ready to present the key result of this subsection, achieving algebraic equivalence by making use of the likelihood score, captured in the following theorem.

Theorem 1 (Algebraic Equivalence by Score and Dimension)**.** *Suppose a model follows Definition 1 with $\mathcal{G}^*$ and distribution $\Sigma^*_{\mathbf{X}}$ satisfies the generalized faithfulness assumption. Given observation $\hat{\Sigma}_{\mathbf{X}}$ and let $\mathbb{G}^* = \arg\max_{\mathcal{G}\in\mathbb{G}^n} score_{ML}(\mathcal{G}, \hat{\Sigma}_{\mathbf{X}})$. If $\hat{\mathcal{G}} \in \mathbb{G}^*$ and $\hat{\mathcal{G}} \in \arg\min_{\mathcal{G}\in\mathbb{G}^*} \dim(\mathcal{G})$, 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. *Theorem 1 (partially inspired by Ng et al. (2024)) says that, if $\hat{\mathcal{G}}$ can generate the observation $\hat{\Sigma}_{\mathbf{X}}$, and $\hat{\mathcal{G}}$ has the smallest dimension among those graphs that can generate the observation, then $\hat{\mathcal{G}}$ and $\mathcal{G}^*$ are algebraically equivalent. In other words, we can enumerate all the graphs in the assumed graph space, and utilize Theorem 1 to find a graph $\hat{\mathcal{G}}$ that is algebraically equivalent to the ground truth $\mathcal{G}^*$. In the absence of latent variables, if $\hat{\mathcal{G}}$ and $\mathcal{G}^*$ are algebraically equivalent, $\hat{\mathcal{G}}$ and $\mathcal{G}^*$ belong to the same MEC, and thus we can make use of score to identify the structure up to MEC. In this sense, Theorem 1 takes the theoretical guarantee of score, e.g., in GES (Chickering, 2002) as a special case and generalizes it to the partially observed scenario.*

Algebraic equivalence is a general sense of equivalence: if two graphs are algebraically equivalent, then we cannot differentiate them purely by observational data without any further assumption. The reason lies in that, in the linear gaussian case, all the information from the distribution are just equality and inequality constraints (in Definition 6), and generally inequality constraints cannot be used; we have to assume inequality-constraint-faithfulness to utilize inequality-constraints, but the set of parameters that results in violation of such faithfulness is not of Lebesgue measure zero.

At this point, we have characterized how the likelihood score can be used for structure identification in the partially observed scenario. Yet, there still exist two main challenges. First, in the partially observed scenario, without any graphical assumption, relating $\hat{\mathcal{G}}$ to $\mathcal{G}^*$ can be very challenging, as the algebraic equivalence class is very large. For example, suppose the ground truth graph is $\mathcal{G}^*$ in Figure 1(a). Even though by making use of all the equality constraints, we still just arrive at some elements of the algebraic equivalence class of $\mathcal{G}^*$, e.g., $\hat{\mathcal{G}}_1$ or $\hat{\mathcal{G}}_2$ in Figure 1(b) and (c). In fact, without any further graphical assumption, the cardinality of the class is infinity, as we can always add one more latent variable to the structure without changing any constraints. Therefore, a general recipe

may involve *identifying suitable structural assumptions that allow algebraic equivalence to translate into more fine-grained notions of model equivalence, such as Markov equivalence*, as in Section 3.2.

Second, Theorem 1 only implies a search procedure that requires the exact enumeration of all possible graphs in the assumed graph space. Yet, such an exact search is impractical due to the computational overhead. To be specific, the number of possible graphs grows super-exponentially with the increase of the number of observed variables, even when we rule out a lot of structures with latent variables that cannot be identified (e.g., $\mathsf{X}_1 \to \mathsf{L}_1 \to \mathsf{X}_2$ where we can never know whether L_1 exists or not). More specifically, if we consider the GNFM (which will be detailed in the next section with definition in Definition 2) with 10 observed variables, the number of all possible graphs that satisfy Definition 2 is more than 3×10^8 and this number grows to 2×10^{13} when the number of observed variables increase only by 1. Therefore, it is crucial to design *an efficient way to search over the graph space for the identification of the optimal structure*, as in Section 4.

3.2 Graphical Assumption and Global Consistency

In this section, we propose the generalized N factor model, a graphical condition under which algebraic equivalence can be translated into Markov equivalence, defined as follows.

Definition 2 (Generalized N Factor Model)**.** *DAG $\mathcal{G}$ satisfies the definition of generalized N factor model if observed variables are the effects of latent variables and there exists a partition of all latent variables $\mathbf{L}_\mathcal{G}$ such that for each element in the partition, $\mathbf{L}_p$, (i) there exist at least $|\mathbf{L}_p| * 2$ observed variables $\mathbf{X}_p$ such that for all $\mathsf{X} \in \mathbf{X}_p, Pa_\mathcal{G}(\mathsf{X}) = \mathbf{L}_p$, (ii) if a variable $\mathsf{V} \in \mathbf{V}_\mathcal{G}$ causes or is caused by another variable in $\mathbf{L}_p$, then V also causes or is caused by all the variables in $\mathbf{L}_p$, respectively, and (iii) elements in $\mathbf{L}_p$ are mutually nonadjacent.*

For brevity, in the rest of the paper, we use $\mathbb{G}_{\text{GNFM}}$ to denote the set of all graphs that satisfy the definition of generalized N factor model. For a better understanding of the generalized N factor model, we provide an example to elaborate each requirement in Definition 2.

Fig. 2 shows an illustrative graph that satisfies Def. 2. Specifically, all the observed variables in Figure 2 are leaf nodes and caused by only latent variables. Further, all latent variables can be partitioned into groups $\{\mathsf{L}_1\}, \{\mathsf{L}_2\}, \{\mathsf{L}_3\}, \{\mathsf{L}_4, \mathsf{L}_5\}, \{\mathsf{L}_6, \mathsf{L}_7\}$ and thus the rest requirements (i), (ii), (iii) are also satisfied. For example, for $\mathbf{L}_p = \{\mathsf{L}_6, \mathsf{L}_7\}$, there exist $\mathbf{X}_p = \{\mathsf{X}_{15}, \mathsf{X}_{16}, \mathsf{X}_{17}, \mathsf{X}_{18}\}$ such that their parents are $\{\mathsf{L}_6, \mathsf{L}_7\}$ and $|\mathbf{X}_p| \geq |\mathbf{L}_p| \times 2$. These groups have the properties that the relation within a group can not be identified, and the relation between groups applies to every element in the group. Thus, requirements in Def. 2 are satisfied in Figure 2.

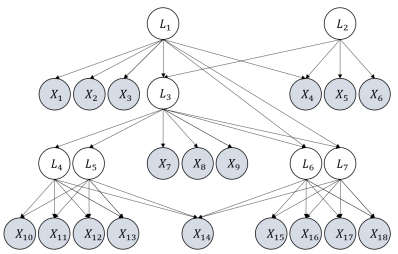


Figure 2: An illustrative example of the graph that satisfies generalized N factor model in Definition 2.

Notably, the graphical condition required for generalized N factor model is rather weak. It takes the prevalent one factor model assumption (Silva et al., 2003) (defined in Definition 8 and illustrated in Section C.2 for a comparison) as a special case, and further allows latent variables to share observed variables as children. In GNFM, we want to make very weak assumption about how latent variables are related - they can be partitioned into groups, and the relation among each group of latent variables can be very flexible. At this premise, the key requirement of $2|\mathbf{L}_p|$ observed children of $\mathbf{L}_p$ in GFNM is minimal. On the other hand, if we make a stronger assumption about how latent variables are related to each others than what has been made in GNFM, the requirement of having at least twice number of observed children can be relaxed.

Here we take $\{\mathsf{L}_3\}$ in Fig. 2 as $\mathbf{L}_p$ to show why we need $2|\mathbf{L}_p|$ observed children of $\mathbf{L}_p$. If we want to determine whether two latent groups $\{\mathsf{L}_1\}$ and $\{\mathsf{L}_4, \mathsf{L}_5\}$ are adjacent, we need to check whether $\{\mathsf{L}_1\}$ and $\{\mathsf{L}_4, \mathsf{L}_5\}$ can be d-separated by $\{\mathsf{L}_3\}$, which can be done by checking whether there exists a pure observed child of $\{\mathsf{L}_1\}$ (e.g., X_1) and another pure observed child of $\{\mathsf{L}_4, \mathsf{L}_5\}$ (e.g., X_{10}) can be d-separated by $\{\mathsf{L}_3\}$. As $\{\mathsf{L}_3\}$ cannot be observed, we have to rely on t-separation (Sullivant et al., 2010) to check whether this d-separation holds, as t-separation allows us to use the observed children of $\{\mathsf{L}_3\}$ as surrogates. This can be translated into checking whether there exists two distinct groups $\mathbf{X}_a$ and $\mathbf{X}_b$ as the observed pure children of $\{\mathsf{L}_3\}$ such that, $|\mathbf{X}_b| = |\mathbf{X}_b| = |\{\mathsf{L}_3\}|$, and $\{\mathsf{X}_1\}$ and $\{\mathsf{X}_{10}\}$ can be t-separated by $(\mathbf{X}_a, \mathbf{X}_b)$. Thus, for each $\mathbf{L}_p$ we need $2|\mathbf{L}_p|$ observed children of it.

The reason why GNFM chooses this specific trade-off point is that we believe it is more practically meaningful. In real-life problems, the way latent variables are related could be complicated and we do not want to rule out the possibility of certain structural patterns in advance. At the same time, if the number of observed children is insufficient, we can still gather more relevant observations or measurements of the underlying system.

Next, we show in Theorem 2 that for GNFM, the notion of algebraic equivalence leads to Markov equivalence, and thus by using score we can identify the structure up to MEC, as in Corollary 1.

Theorem 2 (Identifiability of Generalized N Factor Models by Equality Constraint up to MEC)**.** *For $\mathcal{G}_1, \mathcal{G}_2 \in \mathbb{G}_{GNFM}$, if they are algebraically equivalent, i.e., $H(\mathcal{G}_1) = H(\mathcal{G}_2)$, then $\mathcal{G}_1$ and $\mathcal{G}_2$ belong to the same MEC (same skeleton and v-structures over all variables).*

Corollary 1 (Global Consistency by Score for Generalized N Factor Models)**.** *Suppose a model follows Definition 1 with $\mathcal{G}^* \in \mathbb{G}_{GNFM}$ and distribution $\Sigma^*_{\mathbf{X}}$ satisfies Assumption 1. Given observation $\hat{\Sigma}_{\mathbf{X}}$ and let $\mathbb{G}^* = \arg\max_{\mathcal{G} \in \mathbb{G}_{GNFM}} score_{ML}(\mathcal{G}, \hat{\Sigma}_{\mathbf{X}})$. If $\hat{\mathcal{G}} \in \mathbb{G}^*$ and $\hat{\mathcal{G}} \in \arg\min_{\mathcal{G} \in \mathbb{G}^*} \dim(\mathcal{G})$, then $\hat{\mathcal{G}}$ and $\mathcal{G}^*$ are Markov equivalent in the large sample limit.*

Corollary 1 establishes the global consistency of using likelihood score for structure identification. Yet, it still requires impractical exact enumeration. Furthermore, it is also unclear how to calculate the exact dimension for each graph with latent variables. Fortunately, an efficient greedy search can be designed to identify the optimal structure without capturing the dimension, as shown in what follows.

4 Score-based Greedy Search for Partially Observed Linear Causal Models

We begin with the general design for greedy search in Section 4.1, specify the design in LGES for generalized N factor model in Section 4.2, and establish the asymptotic correctness in Section 4.3.

4.1 General Design for Greedy Search

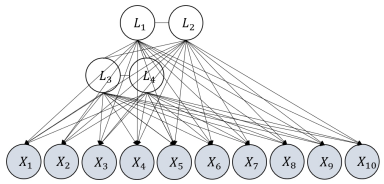
Our overall objective is to find a graph $\hat{\mathcal{G}}$ such that it can generate $\hat{\Sigma}_{\mathbf{X}}$ equally well as the ground truth $\mathcal{G}^*$, while having a dimension that is as small as possible (as in Theorem 1 and Corollary 1). To efficiently search over the graph space, we follow the traditional wisdom GES (Chickering, 2002) and define three key elements for a greedy search.

- A set of states.
- A representation scheme for the states.
- A set of operators.

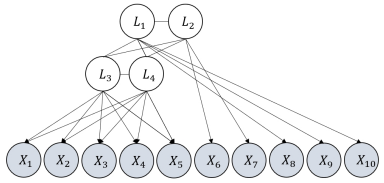
A state represents a solution to the search problem and we use $\mathcal{S}$ to represent state and $\mathbb{S}$ to represent a set of states. The representation scheme defines an efficient way to represent the states. As the structure can only be identified up to MEC, we follow GES such that each state is a MEC, which corresponds to an unique Completed Partially Directed Graph (CPDAG in Definition 10). Thus we also use $\mathcal{S}$ to refer to a CPDAG and $\mathcal{S}(\mathcal{G})$ to transform a DAG $\mathcal{G}$ into a CPDAG. Finally, the set of operators is to transform one state to another state, in order to traverse the whole graph space systematically and efficiently.

For any two graphs belong to the same state (i.e., the same MEC), they share the same dimension and maximum likelihood score (Prop. 1 in (Ng et al., 2024)). Thus, we also define score_{ML} for a CPDAG $\mathcal{S}$, as $\text{score}_{\text{ML}}(\mathcal{S}, \hat{\Sigma}_{\mathbf{X}}) = \text{score}_{\text{ML}}(\mathcal{G}, \hat{\Sigma}_{\mathbf{X}})$, for all $\mathcal{G} \in \text{MEC}(\mathcal{S})$.

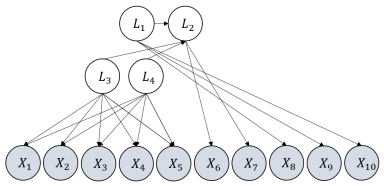
Furthermore, the initial state $\mathcal{S}_{\text{init}}$ is often designed as a state that is a super graph of the ground truth, and $\mathcal{S}_{\text{init}}$ can



(a) $\mathcal{S}_{\text{init}}$ by Def. 3, where latents are mutually fully connected by undirected edges and all latents cause all observed.



(b) $\mathcal{S}_{\text{phase1}}$, the output of Alg. 1, where the number of latent and the edges from latent to observed variables are determined.



(c) $\mathcal{S}_{\text{final}}$, the output of Alg. 2, represents the MEC of the ground truth $\mathcal{G}^*$ (here the MEC only contains $\mathcal{G}^*$).

Figure 3: LGES illustration, where (a) is the initial state of Alg. 1, (b) is the output of Alg. 1, and (c) is the final output.

generate the observation equally well as the ground truth. For example, the phase 1 of GES is to find such an initial state and thus the phase 2 of GES can focus on the delete operation. Next, we introduce the detailed design of our novel method, LGES.

4.2 ALGORITHM: LATENT VARIABLE GREEDY EQUIVALENCE SEARCH (LGES)

In this section we discuss the detailed design for Latent variable Greedy Equivalence Search (LGES) for structure identification of generalized N factor models. We begin with the initial state $\mathcal{S}_{\text{init}}$.

Definition 3 (Initial State for Generalized N Factor Model)**.** *Given* $\mathbf{X}$*,* $\mathcal{S}_{init}(\mathbf{X})$ *outputs a CPDAG such that it contains observed variables* $\mathbf{X}$ *and* $\lfloor \frac{|\mathbf{X}|}{2} \rfloor$ *latent variables, all latent variables are fully connected with undirected edge, and all latent variables cause all observed variables.*

In Definition 3 it implicitly requires that the number of observed variables is at least twice the number of latent ones. We note that this latent-to-observed ratio is a property of the graphical assumption of GNFM in Definition 2, and thus invariant to the design of a method. An example of $\mathcal{S}_{\text{init}}(\mathbf{X})$ can be found in Figure 3, where (c) is the ground truth $\mathcal{G}^*$ and (a) is the initial state $\mathcal{S}_{\text{init}}(\mathbf{X})$ by Definition 3. The reason why we design $\mathcal{S}_{\text{init}}$ as such lies in the properties of $\mathcal{S}_{\text{init}}$ formalized in Lemma 1.

Lemma 1 (Properties of Initial State)**.** *Suppose a model follows Definition 1 with* $\mathcal{G}^* \in \mathbb{G}_{GNFM}$ *and we are given observation* $\hat{\Sigma}_{\mathbf{X}}$*. Then* $\mathcal{S}_{init}$ *is a supergraph of* $\mathcal{S}(\mathcal{G}^*)$ *and* $\mathcal{S}_{init}$ *can generate the observed distribution, i.e.,* $score_{ML}(\mathcal{S}_{init}, \hat{\Sigma}_{\mathbf{X}}) = \max_{\mathcal{G} \in \mathbb{G}_{GNFM}} score_{ML}(\mathcal{G}, \hat{\Sigma}_{\mathbf{X}})$ *in the large sample limit.*

The core spirit of LGES is that we begin with a state that can generate the observation. Each time we delete some edges and see whether the new CPDAG can still generate the observation. If so, we keep it as the new state; otherwise we try a different deletion. As an edge deletion leads to smaller or equal dimension, the process is in essence finding the CPDAG with smallest dimension while keeping the ability of generating the observation. Next we discuss the operators that connect between states.

Definition 4 (Delete Operator $\mathcal{O}_{\mathbf{LX}}$)**.** $\mathcal{O}_{\mathbf{LX}}(\mathcal{S}, \mathbf{L}, \mathbf{X})$ *returns a CPDAG that is the same as* $\mathcal{S}$ *except that all edges from* $\mathbf{L}$ *to* $\mathbf{X}$ *are deleted.*

Definition 5 (Delete Operator $\mathcal{O}_{\mathbf{LL}}$)**.** $\mathcal{O}_{\mathbf{LL}}(\mathcal{S}, \mathbf{L}_1, \mathbf{L}_2, \mathbf{H})$ *returns a CPDAG that is the same as* $\mathcal{S}$ *except that (i) all edges between* $\mathbf{L}_1$ *and* $\mathbf{L}_2$ *are deleted, (ii) for each* $\mathsf{H} \in \mathbf{H}$*, directing the previously undirected edge between* $\mathbf{L}_1$ *and* H *as* $\mathbf{L}_1 \to \mathsf{H}$ *and directing the previously undirected edge between* $\mathbf{L}_2$ *and* H *as* $\mathbf{L}_2 \to \mathsf{H}$*.*

$\mathcal{O}_{\mathbf{LX}}$ is designed to delete edges from latent variables to observed variables while $\mathcal{O}_{\mathbf{LL}}$ (partially inspired by Chickering (2002)) is designed to delete relations among two groups of latent variables $\mathbf{L}_1$ and $\mathbf{L}_2$. Our LGES is built upon these two operators and includes two phases. Phase 1 (summarized in Alg 1) is to recover the structure between latent variables and observed variables. Roughly speaking, each time the algorithm chooses some latent variables and some observed variables from the current state, deletes all the edges between them to get a neighbouring state, and check whether this neighbouring state can generate the observation $\hat{\Sigma}_{\mathbf{X}}$ (up to a certain tolerance level δ). If so, this neighboring state is taken as the new state; otherwise the algorithm looks for a new combination for edge deletion. An example of the output of phase 1 is in Fig 3 (b), where the structure between latent and observed are expected to be the same as the ground truth (also formalized in Lemma 2).

Phase 2 (summarized in Alg 2) of LGES, which is partially inspired by Chickering (2002), aims to identify the structure among latent variables. Roughly, each time the algorithm chooses two groups of latent variables, deletes all the edges between them to get a neighbouring state; if this state can generate the observation (up to δ), then takes it as the new state; otherwise looks for another two groups. δ controls the sparsity level in the finite sample case: with bigger δ, deletions are more likely to be kept (ablation study in Section 5.2). An illustration of the output of phase 2 is in Fig 3 (c), which is expected to be the same as the ground truth (which is formalized in Theorem 3).

As implied by Theorem 1, our objective is to search over the graph space to find a graph $\hat{\mathcal{G}}$ such that (i) $\hat{\mathcal{G}}$ has the best likelihood and (ii) at the premise of ensuring (i), the dimension of $\hat{\mathcal{G}}$ should be as small as possible. Below are the key designs of LGES to ensure this goal.

- We start with a graph that is guaranteed to be a super graph of the ground truth.
- At each step, we try to delete some edges. Only when the likelihood after the deletion is still the best do we keep the deletion. This guarantees that (i) always holds.

Table 1: F1 score and SHD (mean(standard error)) of each compared method (LGES, FOFC, GIN, and RLCD) across different sample sizes. ↑ means the bigger the better while ↓ the smaller the better.

	F1 score for skeleton ↑				SHD for MEC ↓			
Sample size	**LGES**	FOFC	GIN	RLCD	**LGES**	FOFC	GIN	RLCD
100	**0.60** (0.01)	0.55 (0.02)	0.27 (0.01)	0.33 (0.01)	**20.87** (0.72)	20.8 (0.34)	26.76 (0.36)	35.84 (3.01)
200	**0.72** (0.02)	0.56 (0.02)	0.36 (0.01)	0.49 (0.02)	**14.03** (0.85)	18.0 (0.29)	24.14 (0.45)	28.58 (2.59)
500	**0.79** (0.02)	0.58 (0.02)	0.42 (0.01)	0.69 (0.02)	**10.02** (0.69)	16.5 (0.31)	22.46 (0.39)	14.68 (0.96)
1000	**0.82** (0.02)	0.61 (0.03)	0.47 (0.01)	0.76 (0.02)	**8.80** (0.70)	16.1 (0.29)	20.88 (0.54)	11.24 (0.86)

Table 2: Under model misspecification, F1 score and SHD (mean(standard error)) of each compared method across different sample sizes. ↑ means the bigger the better while ↓ the smaller the better.

Non-gaussian	F1 score for skeleton ↑				SHD for MEC ↓			
Sample size	**LGES**	FOFC	GIN	RLCD	**LGES**	FOFC	GIN	RLCD
100	**0.57** (0.03)	0.40 (0.03)	0.27 (0.01)	0.35 (0.01)	**21.70** (1.67)	22.72 (0.34)	26.76 (0.36)	39.31 (7.37)
200	**0.74** (0.04)	0.42 (0.03)	0.36 (0.01)	0.46 (0.03)	**12.67** (1.97)	18.10 (0.50)	24.14 (0.45)	35.12 (6.52)
500	**0.78** (0.02)	0.40 (0.03)	0.42 (0.01)	0.68 (0.01)	**10.72** (0.93)	17.66 (0.49)	22.46 (0.39)	14.81 (0.63)
1000	**0.79** (0.02)	0.39 (0.03)	0.47 (0.01)	0.75 (0.02)	**10.12** (0.82)	17.82 (0.44)	20.88 (0.54)	12.45 (0.77)
Non-linear	F1 score for skeleton ↑				SHD for MEC ↓			
Sample size	**LGES**	FOFC	GIN	RLCD	**LGES**	FOFC	GIN	RLCD
100	**0.58** (0.02)	0.35 (0.03)	0.32 (0.01)	0.36 (0.02)	**19.21** (1.29)	21.48 (0.44)	29.28 (0.11)	37.01 (4.77)
200	**0.65** (0.01)	0.33 (0.03)	0.38 (0.01)	0.45(0.02)	**18.22** (1.11)	21.80 (0.41)	27.28 (0.28)	31.73 (4.33)
500	**0.68** (0.02)	0.26 (0.04)	0.50 (0.02)	0.60 (0.02)	**16.6** (1.24)	19.98 (0.46)	23.38 (0.64)	19.89 (1.08)
1000	**0.71** (0.02)	0.23 (0.04)	0.52 (0.02)	0.65 (0.01)	**15.0** (0.95)	20.10 (0.54)	23.32 (0.73)	17.56 (0.77)

- An edge deletion operation will either keep the dimension the same or decrease the dimension. Thus, throughout our search process, the dimension will decrease monotonically.
- The design of deletion operators ensure at each step, the current graph is always a super-graph of the ground truth, through out the whole process. This is because if the post-deletion graph is not a super-graph of the ground truth, additional equality constraints will be introduced such that the post-deletion graph cannot reach the best likelihood and thus this deletion will not be kept.
- By the end of the process, LGES will arrive at the ground truth. This can be proved by contradiction in a rough sense as follows (detailed in proof). Suppose the final state $\mathcal{G}'$ is not optimal. As it must be a super graph of $\mathcal{G}^*$, there must exist a sequence of deletion to transform $\mathcal{G}'$ into $\mathcal{G}^*$. Suppose the first deletion in the sequence leads to $\mathcal{G}''$. As $\mathcal{G}''$ is also a super graph of $\mathcal{G}^*$, the score of $\mathcal{G}''$ is also the best, and thus the algorithm would not terminate at $\mathcal{G}'$, yielding a contradiction.

4.3 Asymptotic Correctness of LGES

Here we establish the asymptotic correctness of LGES. Specifically, if the ground truth satisfies the generalized N factor model, LGES can asymptotically produce the correct Markov equivalence class over both observed and latent variables, as in Lemma 2 and Theorem 3 (all proofs in Appendix).

Lemma 2 (Correctness of Phase 1 of LGES)**.** *Suppose a model follows Definition 1 with $\mathcal{G}^* \in \mathbb{G}_{GNFM}$ and we are given observation $\hat{\Sigma}_{\mathbf{X}}$. In the large sample limit the output $\mathcal{S}_{phase1}$ of Algorithm 1 is a CPDAG such that the number of latent variables in $\mathcal{S}_{phase1}$ is the same as that of $\mathcal{G}^*$ and the edges from $\mathbf{X}$ to $\mathbf{L}$ in $\mathcal{S}_{phase1}$ is the same as that of $\mathcal{G}^*$, up to permutation of latent variables.*

Theorem 3 (Correctness of LGES)**.** *Suppose a model follows Definition 1 with $\mathcal{G}^* \in \mathbb{G}_{GNFM}$ and we are given observation $\hat{\Sigma}_{\mathbf{X}}$. In the large sample limit the output $\mathcal{S}_{final}$ of Algorithms 1 and 2 is a CPDAG that represent the MEC of $\mathcal{G}^*$, up to permutation of latent variables.*

5 Experiments

5.1 Synthetic Setting and Evaluation Metric

In our synthetic experiments, we validate LGES by comparing it with multiple latent variable causal discovery methods, including FOFC (Kummerfeld & Ramsey, 2016), GIN (Xie et al., 2020), and RLCD (Dong et al., 2024a). The ground truth model follows Def. 1 where each edge coefficient f_{ji} is randomly sampled uniformly from $[-5, 5]$ and the variance for each noise term uniformly from $[0.1, 1]$. As GIN assumes non-gaussianity, we use uniform distribution for the noise when testing GIN. We consider 20 randomly generated structures that satisfy Def. 2 as the ground truth structure (examples in Fig. 8 in Appendix). On average, each ground truth graph contains 20 variables and 5 of them are latent. For the output of each method, we first transform the output into a CPDAG and then compare it with the ground truth CPDAG by calculating F1 score over the skeleton and SHD over the MEC as evaluation metrics, and consider four different sample sizes: 100, 200, 500,

Table 3: Model fitness scores on real-life datasets to validate LGES (↑ the bigger the better).

Scenarios		Model Fitness Scores		
Dataset	Structures	RMSEA ↓	CFI ↑	TLI↑
Big Five	**Figure 5 by LGES**	**0.054**	**0.874**	**0.855**
	By Goldberg (1993)	0.072	0.767	0.746
T Burnout	**Figure 6 by LGES**	**0.067**	**0.876**	**0.865**
	By Byrne (1994)	0.096	0.753	0.727
	By Byrne (2010)	0.072	0.861	0.847
M-tasking	**Figure 4 by LGES**	**0.068**	**0.977**	**0.965**
	By Himi et al. (2019b)	0.087	0.962	0.943

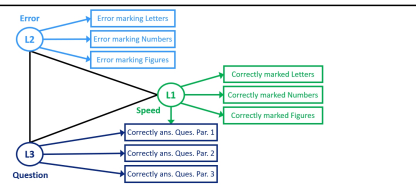


Figure 4: Causal structure (CPDAG) recovered by LGES on Multi-tasking behavior dataset.

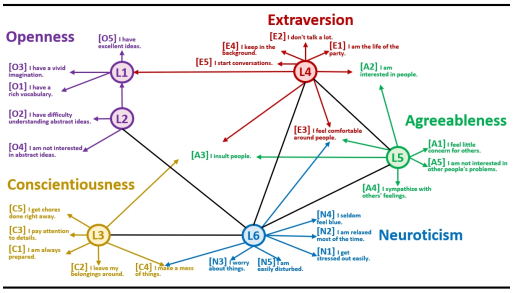


Figure 5: Causal structure (CPDAG) recovered by LGES on Big Five personality dataset.

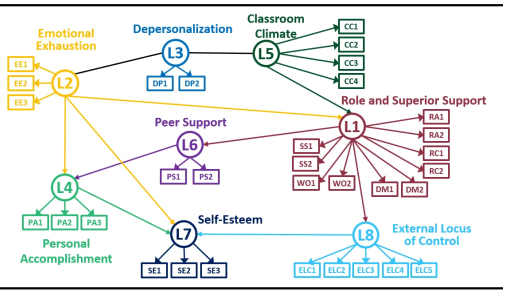


Figure 6: Causal structure (CPDAG) recovered by LGES on Teacher burnout dataset.

1000. We adopt 10 random seeds to generate the ground truth causal model and observational data, and report the mean and standard error.

5.2 Performance on Synthetic Data and Ablation Study on δ

The F1 scores of skeletons and SHDs of MECs are reported in Table 1. As shown, the proposed LGES achieves the best F1 score (bigger better) and SHD (smaller better) performance compared to all baselines. For example, with sample size 1k, LGES achieves a F1 score of 0.82 and a SHD of 8.8, where the runner-up achieves 0.76 and 11.24 respectively. Another key observation is that our score-based method can still work well with a very small sample size, while constraint-based methods may not. E.g., with only 100 datapoints, LGES still achieves a F1 score of 0.60, while constraint-based method GIN (CI-test-based) and RLCD (rank-test-based) only achieves 0.27 and 0.33 respectively. The reason why constraint-based methods do not work well with a small sample size may lie in that with small sample size the power of the test is limited and and the null distribution may be very different from the asymptotic case, both of which aggravate the issue of error propagation. On the contrary, our score-based method may not suffer from these issues.

Similar to the hyper-parameter λ in GES, in practice we can tune the value of δ to control the sparsity level of the result (the bigger δ the sparser). In our experiments, δ is set as $\delta = 0.25 \times \frac{\log(N)}{N}$, where N is the sample size. This design follows the spirit of BIC score in GES such that $\delta \to 0$ when $N \to \infty$. The ablation study to analyze the sensitivity to δ can be found in Table 5, where LGES is not very sensitive to small change of δ. A more detailed discussion about δ can be found in Section C.5.

5.3 Misspecification Behavior

We investigate the performance of LGES under model misspecification: violation of normality and linearity. For the setting of violation of normality, we use uniform noise terms for the underlying model, and report the result in Table 2. Specifically, when the normality is violated, LGES still performs the best compared to baselines and the result is almost the same as that of the gaussian case. For example, with 1k sample size, the F1 score of LGES under non-gaussianity is 0.79 while the counter part under the standard setting is 0.82. This is not very unexpected: the structure identifiability by score is built upon the hard constraints imposed by structure on the observational covariance matrix, which only relies on the linearity of the underlying causal model and does not rely on gaussianity. As for violation of linearity, we employ leaky ReLU (Xu, 2015) to simulate piecewise linear function, as $\mathsf{V}_\mathsf{i} = \mathrm{LR}(\sum_{\mathsf{V}_\mathsf{j} \in \mathrm{Pa}(\mathsf{V}_\mathsf{i})} f_{ji}\mathsf{V}_\mathsf{j} + \epsilon_{\mathsf{V}_\mathsf{i}})$, $\mathrm{LR}(x) = \max(\alpha x, x)$, where $\alpha = 0.8$. The result is in Table 2, which shows that LGES works reasonably well under certain extent of non-linearity, and still surpasses all baselines. For example, with sample size 1k, LGES still achieves 0.71 F1 score even under nonlinearity, while the runner-up achieves 0.65.

5.4 REAL-WORLD PERFORMANCE

We consider three real-life datasets. Big Five personality dataset (`openpsychometrics.org`), which consists of 50 questions with 19,719 datapoints. There are five dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (O-C-E-A-N), and each dimension with 10 questions (e.g., O1 is the first question for Openness). We use the first 5 questions for each dimension and in total 25 variables. The structure by LGES is shown in Figure 5. Interestingly, without any prior knowledge, the structure recovered by LGES is very aligned with psychology study. To be specific, each item in our result is indeed caused by the supposed dimension (latent variable). Further, we found that some items are caused not only by one latent variable. For example, E3 (I feel comfortable around people) is caused by L4 (Extraversion), L5 (Agreeableness), and L6 (Neuroticism), which may shed new light on the existing factor-analysis-based psychometric study.

Teacher burnout dataset (Byrne, 2001). The term burnout refers to the inability to perform effectively in one's job due to job-related stress. The dataset includes 32 observed variables with 599 datapoints. Multi-tasking behavior dataset (Himi et al., 2019a). To compare with the model proposed by Himi et al. (2019a), we use 9 variables of it with all its 202 datapoints. The structures produced by LGES for teacher burnout data and multi-tasking data are shown in Figures 4 and 6 respectively. Finally, we use three prevalent goodness of fit statistics to validate the structure produced by LGES: RMSEA (Steiger, 1980; 1990), CFI (Bentler, 1990), and TLI (Tucker & Lewis, 1973; Bentler & Bonett, 1980) (detailed in Section C.4). We used these fit indices to compare the structure produced by our method with the well-known hypothesized structures in existing psychological studies. The result is reported in Table 3, where the structures by LGES achieve the best scores compared to all the structures proposed in psychological research, shows that the structures by LGES explain the observational data better than existing study, and validates the proposed method in real-life scenarios.

5.5 COMPARISON WITH THE EXACT SEARCH

In this section we compare the proposed LGES to the score-based exact search method (Ng et al., 2024). The exact search has to enumerate all the possible candidate graphs and thus the complexity grows super-exponentially with the increase of the number of observed variables $|\mathbf{X}_{\mathcal{G}}|$ (which is why we do not include the exact search in our main result in Table 1). In contrast to exact search, LGES search over the candidate graph space in a properly designed way such that each time we only need to decide greedily while the asymptotic correctness can still be guaranteed.

The result is shown in Table 4. Specifically, to handle a single dataset with 8 observed variables, the exact search requires more than 100 hours while LGES only requires 17 seconds. When it comes to 9 observed variables, the exact search requires more than 1000 hours (by estimation) while LGES only requires 19 seconds. As for the causal discovery performance measured by F1 and SHD, the exact search only performs slightly better than LGES. For example, for 7 observed variables, the F1 score of the exact search is 0.89 while the F1 of LGES is 0.88; yet, the the exact search requires nearly 400 times more time (1.5 hours v.s. 14 seconds). This empirical result demonstrates the significant computation efficiency improvement of LGES against the exact search, with only slight degradation in the performance.

5.6 IMPLEMENTATION DETAILS, RUNTIME ANALYSIS, AND EXTENDABILITY

The readers are referred to Section C.5 and Section C.6 for implementation details and runtime analysis, respectively. As for the discussion about the extendability of the proposed method to non-Gaussian or non-linear scenarios, please refer to Section C.7.

6 CONCLUSION

In this paper, we first characterize how likelihood score and minimal dimension are related to the structure identifiability of partially observed linear causal models. Then we propose Generalized N Factor Model under which we prove the global consistency of using score for structure identification. Finally we propose LGES, an asymptotically correct score-based greedy search method to efficiently search over the graph space and identify the causal structure.

ACKNOWLEDGMENTS

We would also like to acknowledge the support from NSF Award No. 2229881, AI Institute for Societal Decision Making (AI-SDM), the National Institutes of Health (NIH) under Contract R01HL159805, and grants from Quris AI, Florin Court Capital, MBZUAI-WIS Joint Program, and the Al Deira Causal Education project.

REFERENCES

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

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

Carlos Améndola, Mathias Drton, Alexandros Grosdos, Roser Homs, and Elina Robeva. Third-order moment varieties of linear non-gaussian graphical models. *Information and Inference: A Journal of the IMA*, 12(3):iaad007, 2023.

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

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. DAGMA: Learning DAGs via M-matrices and a log-determinant acyclicity characterization. In *Advances in Neural Information Processing Systems*, 2022.

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

Peter M Bentler. Comparative fit indexes in structural models. *Psychological bulletin*, 107(2):238, 1990.

Peter M Bentler and Douglas G Bonett. Significance tests and goodness of fit in the analysis of covariance structures. *Psychological bulletin*, 88(3):588, 1980.

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

Rohit Bhattacharya, Tushar Nagarajan, Daniel Malinsky, and Ilya Shpitser. Differentiable causal discovery under unmeasured confounding. In *International Conference on Artificial Intelligence and Statistics*, 2021a.

Rohit Bhattacharya, Tushar Nagarajan, Daniel Malinsky, and Ilya Shpitser. Differentiable causal discovery under unmeasured confounding. In *International Conference on Artificial Intelligence and Statistics*, pp. 2314–2322. PMLR, 2021b.

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

Barbara M. Byrne. Burnout: testing for the validity, replication, and invariance of causal structure across elementary, intermediate, and secondary teachers. *Am. Educ. Res. J.*, 31(3):645–673, 1994. doi: 10.3102/00028312031003645.

Barbara M. Byrne. *Structural Equation Modeling with Amos: Basic Concepts, Applications, and Programming*. Multivariate Application Series. Toutledge, Taylor and Francis Group, New York London, 2nd edition, 2010. ISBN 978-0-8058-6372-7.

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

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

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

Myung Jin Choi, Vincent YF Tan, Animashree Anandkumar, and Alan S Willsky. Learning latent tree graphical models. *Journal of Machine Learning Research*, 12:1771–1812, 2011.

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

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

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

James Cussens. Bayesian network learning with cutting planes. *arXiv preprint arXiv:1202.3713*, 2012.

Haoyue Dai, Peter Spirtes, and Kun Zhang. 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.

Chang Deng, Kevin Bello, Bryon Aragam, and Pradeep Kumar Ravikumar. Optimizing NOTEARS objectives via topological swaps. In *International Conference on Machine Learning*, 2023.

Yanming Di. t-separation and d-separation for directed acyclic graphs. *preprint*, 2009.

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

Xinshuai Dong, Ignavier Ng, Biwei Huang, Yuewen Sun, Songyao Jin, Roberto Legaspi, Peter Spirtes, and Kun Zhang. On the parameter identifiability of partially observed linear causal models. 2024b.

Xinshuai Dong, Ignavier Ng, Boyang Sun, Haoyue Dai, Guang-Yuan Hao, Shunxing Fan, Peter Spirtes, Yumou Qiu, and Kun Zhang. Permutation-based rank test in the presence of discretization and application in causal discovery with mixed data. In *ICML*, 2025.

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

Mathias Drton, Alexandros Grosdos, Irem Portakal, and Nils Sturma. Algebraic sparse factor analysis. *arXiv preprint arXiv:2312.14762*, 2023.

Gonçalo Rui Alves Faria, Andre Martins, and Mario A. T. Figueiredo. Differentiable causal discovery under latent interventions. In *Conference on Causal Learning and Reasoning*, 2022.

Joseph Felsenstein. Inferring phylogenies. In *Inferring phylogenies*, pp. 664–664. 2004.

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

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

AmirEmad Ghassami, Alan Yang, Negar Kiyavash, and Kun Zhang. Characterizing distribution equivalence and structure learning for cyclic and acyclic directed graphs. In *International Conference on Machine Learning*, 2020.

L. R. Goldberg. The structure of phenotypic personality traits. *Am. Psychol.*, 48(1):26–34, 1993. doi: 0003-066X. URL https://psycnet.apa.org/doiLanding?doi=10.1037%2F0003-066X.48.1.26.

Samsad Afrin Himi, Markus Buehner, Matthias Schwaighofer, Anna Klapetek, and Sven Hilbert. Multitasking behavior and its related constructs: Executive functions, working memory capacity, relational integration, and divided attention. *Cognition*, 189:275–298, 08 2019a.

Samsad Afrin Himi, Markus Bühner, Matthias Schwaighofer, Anna Klapetek, and Sven Hilbert. Multitasking behavior and its related constructs: Executive functions, working memory capacity, relational integration, and divided attentio. *Cognition*, 189:275–298, 2019b. ISSN 0010-0277. doi: 10.1016/j.cognition.2019.04.010. URL https://www.sciencedirect.com/science/article/pii/S0010027719300939.

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

Biwei Huang, Charles Low, Feng Xie, Clark Glymour, and Kun Zhang. Latent hierarchical causal structure discovery with rank constraints. In *Advances in Neural Information Processing Systems*, 2022.

John P Huelsenbeck, Fredrik Ronquist, Rasmus Nielsen, and Jonathan P Bollback. Bayesian inference of phylogeny and its impact on evolutionary biology. *science*, 294(5550):2310–2314, 2001.

Markus Kalisch and Peter Bühlman. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.

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

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Learning latent causal graphs via mixture oracles. *Advances in Neural Information Processing Systems*, 34:18087–18101, 2021.

Mikko Koivisto and Kismat Sood. Exact bayesian structure discovery in bayesian networks. *The Journal of Machine Learning Research*, 5:549–573, 2004.

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

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

Hao-Chih Lee, Matteo Danieletto, Riccardo Miotto, Sarah Cherng, and Joel T. Dudley. Scaling structural learning with NO-BEARS to infer causal transcriptome networks. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 25:391–402, 2019.

Phillip Lippe, Taco Cohen, and Efstratios Gavves. Efficient neural causal discovery without acyclicity constraints. In *International Conference on Learning Representations*, 2022.

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

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

Ignavier Ng, Sébastien Lachapelle, Nan Rosemary Ke, Simon Lacoste-Julien, and Kun Zhang. On the convergence of continuous constrained optimization for structure learning. In *International Conference on Artificial Intelligence and Statistics*, 2022a.

Ignavier Ng, Shengyu Zhu, Zhuangyan Fang, Haoyang Li, Zhitang Chen, and Jun Wang. Masked gradient-based causal structure learning. In *SIAM International Conference on Data Mining*, 2022b.

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

Christopher Nowzohour, Marloes H Maathuis, Robin J Evans, and Peter Bühlmann. Distributional equivalence and structure learning for bow-free acyclic path diagrams. 2017.

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

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

Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. 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.

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

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

Thomas S Richardson, Robin J Evans, James M Robins, and Ilya Shpitser. Nested markov properties for acyclic directed mixed graphs. *The Annals of Statistics*, 51(1):334–361, 2023.

Raanan Y Rohekar, Shami Nisimov, Yaniv Gurwicz, and Gal Novik. Iterative causal discovery in the possible presence of latent confounders and selection bias. *Advances in Neural Information Processing Systems*, 34:2454–2465, 2021.

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

Mauro Scanagatta, Cassio P de Campos, Giorgio Corani, and Marco Zaffalon. Learning bayesian networks with thousands of variables. *Advances in neural information processing systems*, 28, 2015.

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

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

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

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

Ilya Shpitser, Thomas S. Richardson, James M. Robins, and Robin Evans. Parameter and structure learning in nested Markov models. *arXiv preprint arXiv:1207.5058*, 2012.

Ilya Shpitser, Robin J Evans, Thomas S Richardson, and James M Robins. Introduction to nested markov models. *Behaviormetrika*, 41:3–39, 2014.

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

Ricardo Silva, Richard Scheines, Clark Glymour, and Peter Spirtes. Learning measurement models for unobserved variables. In *Conference on Uncertainty in Artificial Intelligence*, 2003.

Ricardo Silva, Richard Scheines, Clark Glymour, and Peter Spirtes. 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.

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

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

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

James H Steiger. Statistically based tests for the number of common factors. In *Paper presented at the Annual Meeting of the Psychometric Society, Iowa Cyty, 1980*, 1980.

James H Steiger. Structural model evaluation and modification: An interval estimation approach. *Multivariate behavioral research*, 25(2):173–180, 1990.

Nils Sturma, Chandler Squires, Mathias Drton, and Caroline Uhler. Unpaired multi-domain causal representation learning. *arXiv preprint arXiv:2302.00993*, 2023.

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

Boyang Sun, Yu Yao, Xinshuai Dong, Zongfang Liu, Tongliang Liu, Yumou Qiu, and Kun Zhang. A sample efficient conditional independence test in the presence of discretization. In *ICML*, 2025.

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

Ledyard R Tucker and Charles Lewis. A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38(1):1–10, 1973.

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

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

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

Dennis Wei, Tian Gao, and Yue Yu. DAGs with no fears: A closer look at continuous optimization for learning Bayesian networks. In *Advances in Neural Information Processing Systems*, 2020.

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

Bing Xu. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In *International Conference on Machine Learning*, 2019.

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

Yan Zeng, Shohei Shimizu, Ruichu Cai, Feng Xie, Michio Yamamoto, and Zhifeng Hao. Causal discovery with multi-domain LiNGAM for latent factors. In *Causal Analysis Workshop Series*, pp. 1–4. PMLR, 2021.

Jiji Zhang. 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.

Kun Zhang, Mingming Gong, Joseph Ramsey, K. Batmanghelich, Peter Spirtes, and Clark Glymour. Causal discovery with linear non-Gaussian models under measurement error: Structural identifiability results. In *Conference on Uncertainty in Artificial Intelligence*, 2018.

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

Zhen Zhang, Ignavier Ng, Dong Gong, Yuhang Liu, Ehsan M Abbasnejad, Mingming Gong, Kun Zhang, and Javen Qinfeng Shi. Truncated matrix power iteration for differentiable DAG learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Zhen Zhang, Ignavier Ng, Dong Gong, Yuhang Liu, Mingming Gong, Biwei Huang, Kun Zhang, Anton van den Hengel, and Javen Qinfeng Shi. Analytic DAG constraints for differentiable DAG learning. In *International Conference on Learning Representations*, 2025.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, 2018.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Learning sparse nonparametric DAGs. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Algorithm 1 Phase 1 of LGES

Input: Empirical covariance $\hat{\Sigma}_{\mathbf{X}}$ over $\mathbf{X}$, tolerance δ
Output: CPDAG $\mathcal{S}_{\text{phase1}}$
Let current state $\mathcal{S} = \mathcal{S}_{\text{init}}$ by Def. 3 with latent variables $\mathbf{L}$, let $k = 1$, $\mathbf{X}_t = \mathbf{X}$, $\mathbf{L}_d = \emptyset$, and P be an empty list;
while $k \leq \frac{|\mathbf{X}_t|}{2}$ **do**
  for each size k subset $\mathbf{L}_i \subseteq \mathbf{L}_d$ (in parallel) **do**
    for each $\mathsf{X}_j \in \mathbf{X}_t$ (in parallel) **do**
      $\mathcal{S}_{ij} = \mathcal{O}_{\mathbf{LX}}(\mathcal{S}, \mathbf{L} \setminus \mathbf{L}_i, \{\mathsf{X}_j\})$;
      Calculate $\text{score}_{\text{ML}}(\mathcal{S}_{ij}, \hat{\Sigma}_{\mathbf{X}})$;
  Let $\text{score}_{\text{ML}}(\mathcal{S}, \hat{\Sigma}_{\mathbf{X}}) = s$;
  for each i, j s.t., $|\text{score}_{\text{ML}}(\mathcal{S}_{ij}, \hat{\Sigma}_{\mathbf{X}}) - s| \leq \delta$ **do**
    $\mathcal{S} = \mathcal{O}_{\mathbf{LX}}(\mathcal{S}, \mathbf{L} \setminus \mathbf{L}_i, \{\mathsf{X}_j\})$ and $\mathbf{X}_t = \mathbf{X}_t \setminus \{\mathsf{X}_j\}$
  Let $\mathbf{L}'$ be any size k subset of $\mathbf{L} \setminus \mathbf{L}_d$;
  for each size $k+1$ subset $\mathbf{X}_j \subseteq \mathbf{X}_t$ (in parallel) **do**
    $\mathcal{S}_j = \mathcal{O}_{\mathbf{LX}}(\mathcal{S}, \mathbf{L} \setminus \mathbf{L}', \mathbf{X}_j)$
    Calculate $\text{score}_{\text{ML}}(\mathcal{S}_j, \hat{\Sigma}_{\mathbf{X}})$;
  Maintain a disjoint set $\mathcal{X}$ for elements in $\mathbf{X}_j, j \in \{j : |\text{score}_{\text{ML}}(\mathcal{S}_j, \hat{\Sigma}_{\mathbf{X}}) - \text{score}_{\text{ML}}(\mathcal{S}, \hat{\Sigma}_{\mathbf{X}})| \leq \delta\}$;
  for each $\mathbf{X}' \in \mathcal{X}$ **do**
    Let $\mathbf{L}'$ be any size k subset of $\mathbf{L} \setminus \mathbf{L}_d$;
    $\mathcal{S} = \mathcal{O}_{\mathbf{LX}}(\mathcal{S}, \mathbf{L} \setminus \mathbf{L}', \mathbf{X}')$;
    $\mathbf{L}_d = \mathbf{L}_d \cup \mathbf{L}'$, $\mathbf{X}_t = \mathbf{X}_t \setminus \mathbf{X}'$, append $\mathbf{L}'$ to P;
  k=k+1;
Remove $\mathbf{L} \setminus \mathbf{L}_d$ from $\mathcal{S}$ and let $\mathcal{S}_{\text{phase1}} = \mathcal{S}$;
return $\mathcal{S}_{\text{phase1}}$, P (which records those latent variables that really exist for the use of Phase 2)

Algorithm 2 Phase 2 of LGES

Input: Empirical covariance $\hat{\Sigma}_{\mathbf{X}}$ over $\mathbf{X}$, tolerance δ, the output $\mathcal{S}_{\text{phase1}}$ and P of Algorithm 1
Output: CPDAG $\mathcal{S}_{\text{final}}$
Let $\mathcal{S} = \mathcal{S}_{\text{phase1}}$;
while True **do**
  for $\mathbf{L}_i \in P$ (in parallel) **do**
    for $\mathbf{L}_j \in P$ (in parallel) **do**
      if $\mathbf{L}_i - \mathbf{L}_j$ or $\mathbf{L}_i \to \mathbf{L}_j$ **then**
        Let $\mathbf{H} = \{\mathsf{L} : \mathsf{L} - \mathbf{L}_j$ and $(\mathsf{L} - \mathbf{L}_i$ or $\mathsf{L} \to \mathbf{L}_i$ or $\mathsf{L} \leftarrow \mathbf{L}_i)\}$;
        for each subset $\mathbf{H}'_k \subseteq \mathbf{H}$ (in parallel) **do**
          $\mathcal{S}_{ijk} = \mathcal{O}_{\mathbf{LL}}(S, \mathbf{L}_i, \mathbf{L}_j, \mathbf{H}'_k)$;
          Calculate $\text{score}_{\text{ML}}(\mathcal{S}_{ijk}, \hat{\Sigma}_{\mathbf{X}})$;
  Let $i^*, j^*, k^* = \arg\max \text{score}_{\text{ML}}(\mathcal{S}_{ijk}, \hat{\Sigma}_{\mathbf{X}})$;
  if $|\text{score}_{\text{ML}}(\mathcal{S}_{i^*j^*k^*}, \hat{\Sigma}_{\mathbf{X}}) - \text{score}_{\text{ML}}(\mathcal{S}, \hat{\Sigma}_{\mathbf{X}})| \leq \delta$ **then** $\mathcal{S} = \mathcal{S}_{i^*j^*k^*}$; **else** Break;
for $\mathbf{L}_i \in P$ **do**
  Delete edges among $\mathbf{L}_i$ in $\mathcal{S}$;
$\mathcal{S}_{\text{final}} = \mathcal{S}$;
return $\mathcal{S}_{\text{final}}$

A Detailed Discussion about Related Work

In this section, we provide a more comprehensive review of related work based on those discussed in Section 1, expanding on both latent variable causal discovery and score-based causal discovery.

Latent variable causal discovery: The earliest attempts for handling latent variables in causal discovery were the Fast Causal Inference (FCI) algorithm (Spirtes et al., 2001; Richardson & Spirtes, 2002; Zhang, 2008). More recently, LCD (Rohekar et al., 2021), an iterative causal discovery method,

was proposed for structure identification in the possible presence of latent confounders and selection bias. However, this line of work aims at identifying the Maximal Ancestral Graph. In other words, the objective is more of "deconfounding" for the identification of causal relations among observed variables, and does not provide direct insight into the causal structure among latent variables. Moreover, FCI is already proved to be maximally informative under nonparametric conditional independence constraints. Therefore, to go beyond the limitations of conditional independence constraints, new statistical tools have been developed, often relying on additional parametric assumptions.

These tools have been discussed in Section 1. Among them, the most commonly used one and also the earliest developed one might be rank constraints (Sullivant et al., 2010), which generalize the classical Tetrad representation theorem (Spirtes et al., 2001) and basic conditional independence constraints. Based on these new tools, many algorithms have also been developed (Silva et al., 2003; Huang et al., 2022; Dong et al., 2024a), with tests that handle discretizations (Dong et al., 2025; Sun et al., 2025). However, despite their theoretical advancements, most existing methods remain within the constraint-based paradigm, heavily relying on statistical tests that suffer from multiple-testing and error propagation issues. This is exactly our motivation on developing score-based algorithms for latent variable causal discovery.

Score-Based causal discovery: Score-based methods offer an alternative to constraint-based approaches, mitigating some of the issues related to error propagation. These methods search for an optimal structure by maximizing a scoring function, such as the Bayesian Information Criterion (BIC) or likelihood-based scores. Based on their search strategies, they can be broadly categorized as follows:

- Exact Search Methods employ exhaustive graph traversal techniques or exploit minimal pruning strategies, such as permutation search (Raskutti & Uhler, 2014), dynamic programming (Koivisto & Sood, 2004), or integer linear programming (Cussens, 2012), to identify the globally optimal structure. These methods require minimal assumptions about the underlying graph structure or parametric model. However, their computational complexity grows super exponentially with the number of variables, rendering them impractical for large-scale causal discovery.
- A*-based and heuristic search methods integrate heuristic functions into the search process to guide exploration through the graph space (Yuan & Malone, 2013; Scanagatta et al., 2015). These methods strike a balance between computational efficiency and search completeness by prioritizing graph structures with high potential scores while avoiding exhaustive enumeration. Although more scalable than exact search, the quality of the learned structure relies heavily on the effectiveness of the heuristic function.
- Greedy search methods, such as the widely used Greedy Equivalence Search (GES) (Chickering, 2002), formulate the search problem in terms of graphical operators that iteratively modify the structure by adding, deleting, or reversing edges. These methods are computationally efficient and well-suited for large-scale problems. As illustrated in (Nandy et al., 2018), greedy search methods can often be interpreted as progressively refining the graph structure based on conditional independence or other graphical constraints, offering an intuitive connection between constraint-based and score-based paradigms.
- Differentiable approaches, such as the seminal NOTEARS method (Zheng et al., 2018), recast the structure learning problem as a continuous optimization task. Subsequent works have incorporated nonlinear functional forms (Yu et al., 2019; Lachapelle et al., 2020; Zheng et al., 2020; Ng et al., 2022b), interventional data (Brouillard et al., 2020; Faria et al., 2022; Lippe et al., 2022), alternative optimization techniques (Ng et al., 2020; 2022a; Bello et al., 2022; Deng et al., 2023), and improved formulations of the acyclicity constraint (Yu et al., 2019; Lee et al., 2019; Wei et al., 2020; Bello et al., 2022; Zhang et al., 2022; 2025). These methods benefit from direct compatibility with well-established numerical solvers and GPU acceleration, enabling them to efficiently handle large-scale problems.

Lastly, let us note that in the intersection of latent variable causal discovery and score-based algorithms, the only existing approach, to our knowledge, is that of Ng et al. (2024). As for the specific search procedure, their method follows an inefficient exact search paradigm. In contrast, our work is the first to introduce greedy score-based search for latent causal discovery, offering a more practical and scalable solution to real-world problems while maintaining identifiability guarantees. We note that

Table 4: Comparison of LGES with exact search across different graph sizes with 1000 sample size. - means the result is unavailable due to that the complexity of exact search grows super-exponentially.

	F1 score for skeleton ↑		SHD for MEC ↓		Time to find a single graph ↓	
Graph Size	LGES	Exact Search	LGES	Exact Search	LGES	Exact Search
$\lvert\mathbf{X}_{\mathcal{G}}\rvert = 5$	**0.96** (0.06)	**0.96** (0.03)	0.55 (0.43)	**0.37** (0.05)	**9 seconds**	15 seconds
$\lvert\mathbf{X}_{\mathcal{G}}\rvert = 6$	**0.91** (0.02)	**0.91** (0.02)	1.24 (0.75)	**0.98** (0.54)	**11 seconds**	8 minutes
$\lvert\mathbf{X}_{\mathcal{G}}\rvert = 7$	0.88 (0.03)	**0.89** (0.02)	2.20 (0.87)	**2.05** (0.80)	**14 seconds**	1.5 hours
$\lvert\mathbf{X}_{\mathcal{G}}\rvert = 8$	0.86 (0.02)	-	3.56 (0.78)	-	**17 seconds**	>100 hours
$\lvert\mathbf{X}_{\mathcal{G}}\rvert = 9$	0.85 (0.03)	-	4.11(1.52)	-	**19 seconds**	>1000 hours

Section 3.1 is highly related to Ng et al. (2024), and yet our contribution in Section 3.1 is unique, with reasons as follows.

First, to establish structure identifiability based on likelihood score and dimension, we have to assume the generalized faithfulness (the spirit of which is similar to the classical CI faithfulness). To assume the generalized faithfulness, one has to formally define several necessary notions including equality constraints, $H(\mathcal{G})$ (the set of equality constraints for a graph $\mathcal{G}$), $B(\mathcal{G})$ (the set of canonical equality constraints with respect to reduced Grobner basis), and $\mathbb{H}^n$ (the union of $B(\mathcal{G})$ over all possible graphs with n observed variables). This paper provides formal definitions of these crucial notions, which serves as the basis for a rigorous version of the generalized faithfulness Assumption 1; as a contrast, these notions are not rigorously defined in Ng et al. (2024), which gives rise to some counter examples that shows the violation set of the generalized faithfulness might not be of measure zero.

Second, the identifiability theory proposed in Ng et al. (2024) is based on $\text{score}_{\text{dim}}$, where the likelihood and dimension are entangled. This actually requires a score based method (either exact search or greedy search) to be able to explicitly characterize the dimension of any candidate graph during the search procedure. However, characterizing an arbitrary latent causal structure still remains an open challenge in the field, which limits the usage of the theory. In contrast, this work explicitly disentangled the maximum likelihood score and dimension when stating the identifiability theory in Theorem 1. This gives rise to our greedy search method that does not need to explicitly characterize the dimension of each candidate graph during the search.

This work is also closely related to the RLCD algorithm (Dong et al., 2024a). Specifically, both RLCD and LGES aims to identify the underlying causal structure involving latent variables given observational distribution. The RLCD algorithm, which is a constraint based method, makes use of rank constraints. Note that the set of all rank constraints is a super-set of the vanishing partial correlation constraints and a sub-set of all the equality constraints, and RLCD examines the rank constraints by using statistical rank tests. In contrast, LGES is a score based method, which does not rely on statistical tests. In essence, by comparing the likelihood scores and dimensions, LGES implicitly makes use of all the equality constraints, which is a super-set of the rank constraints.

The research on nested Markov model Shpitser et al. (2012; 2014); Richardson et al. (2023) is related to causal discovery in the presence of latent variables. This line of work is elegant in that it accomodates Verma-type constraints and contains marginal distributions given by a DAG model with latent variables. However, nested Markov models follow the acyclic directed mixed graph (ADMG) framework, where the effect of latent variables are simplified into bidirected edges between observed variables. That is to say, within ADMG, only structure among observed variables is concerned. On the contrary, this paper aims to identify the whole underlying causal structure among both observed and latent variables (e.g., an edge from a latent variable to another latent variable). The research on learning phylogenetic tree Felsenstein (2004); Huelsenbeck et al. (2001) is also related as it aims to infer the evolutionary relationships among a group of organisms using observed data—typically and outputs a tree-structured graph. Yet, the graphical assumption in this line of work is much stronger then the GNFM considered in this paper.

B Proofs

B.1 Proof of Theorem 1

Theorem 1 (Algebraic Equivalence by Score and Dimension)**.** *Suppose a model follows Definition 1 with $\mathcal{G}^*$ and distribution $\Sigma^*_{\mathbf{X}}$ satisfies the generalized faithfulness assumption. Given observation $\hat{\Sigma}_{\mathbf{X}}$*

Table 5: Ablation study on the sensitivity of hyper-parameter δ.

value of δ	F1 score for skeleton ↑	SHD for MEC ↓
5×10^{-4}	0.79(0.02)	10.85(0.79)
1×10^{-3}	0.81(0.02)	9.52(0.84)
2×10^{-3}	**0.82**(0.02)	**8.80**(0.70)
4×10^{-3}	0.80(0.02)	9.85(0.59)
1×10^{-2}	0.78(0.02)	11.13(0.70)

and let $\mathbb{G}^* = \arg\max_{\mathcal{G} \in \mathbb{G}^n} score_{ML}(\mathcal{G}, \hat{\Sigma}_{\mathbf{X}})$. *If* $\hat{\mathcal{G}} \in \mathbb{G}^*$ *and* $\hat{\mathcal{G}} \in \arg\min_{\mathcal{G} \in \mathbb{G}^*} \dim(\mathcal{G})$, *then* $\hat{\mathcal{G}}$ *and* $\mathcal{G}^*$ *are algebraic equivalent, i.e.,* $H(\hat{\mathcal{G}}) = H(\mathcal{G}^*)$, *in the large sample limit.*

The overall proof strategy below is inspired by Ghassami et al. (2020); Ng et al. (2024).

Proof. In the large sample limie, we have $\Sigma^*_{\mathbf{X}} = \hat{\Sigma}_{\mathbf{X}}$. By $\hat{\mathcal{G}} \in \mathbb{G}^*$, we have that $\hat{\mathcal{G}}$ can generate $\Sigma^*_{\mathbf{X}}$ and thus $\Sigma^*_{\mathbf{X}}$ contains all the equality and inequality constraints of $\hat{\mathcal{G}}$. Under the generalized faithfulness assumption, we have

$$H(\hat{\mathcal{G}}) \subseteq H(\mathcal{G}^*). \tag{4}$$

Further, we have $\hat{\mathcal{G}} \in \arg\min_{\mathcal{G} \in \mathbb{G}^*} \dim(\mathcal{G})$. As $\mathcal{G}^* \in \mathbb{G}^*$, we have $\dim(\hat{\mathcal{G}}) \leq \dim(\mathcal{G}^*)$. Suppose by contradiction that $H(\hat{\mathcal{G}}) \subsetneq H(\mathcal{G}^*)$. This implies $\dim(\hat{\mathcal{G}}) > \dim(\mathcal{G}^*)$, which contradicts with $\dim(\hat{\mathcal{G}}) \leq \dim(\mathcal{G}^*)$. Therefore, we have

$$H(\hat{\mathcal{G}}) \not\subseteq H(\mathcal{G}^*). \tag{5}$$

Taking Equations (4) and (5) together, we have $H(\hat{\mathcal{G}}) = H(\mathcal{G}^*)$. □

B.2 Proof of Theorem 2

Theorem 2 (Identifiability of Generalized N Factor Models by Equality Constraint up to MEC)**.** *For* $\mathcal{G}_1, \mathcal{G}_2 \in \mathbb{G}_{GNFM}$, *if they are algebraically equivalent, i.e.,* $H(\mathcal{G}_1) = H(\mathcal{G}_2)$, *then* $\mathcal{G}_1$ *and* $\mathcal{G}_2$ *belong to the same MEC (same skeleton and v-structures over all variables).*

Proof. We prove by showing that, by using equality constraints we can identify a graph $\mathcal{G} \in \mathbb{G}_{\text{GNFM}}$ up to MEC.

Suppose $\mathcal{G}$ satisfies Definition 2 and thus there exists a partition of all latent variables in $\mathcal{G}$ that satisfies the requirement in Definition 2, as $\{\mathbf{L}_i\}_1^P$. For each $\mathbf{L}_i$, there exist at least $|\mathbf{L}_i| * 2$ observed variables $\mathbf{X}_i$ such that for all $\mathsf{X} \in \mathbf{X}_i, \text{Pa}_{\mathcal{G}}(\mathsf{X}) = \mathbf{L}_i$. We first prove by induction that the structure from latent variables to observed variables can be identified up to MEC by equality constraints. Let $k = 1$. For those $|\mathbf{L}_i| = k$, all the pure children of $\mathbf{L}_i$ can be identified by rank constraints: We can simply check size $k+1$ combination of observed variables $\hat{\mathbf{X}}$ and we have the variables in $\hat{\mathbf{X}}$ are pure children of the same size 1 latent group, iff $\text{rank}_{\Sigma_{\hat{\mathbf{X}}, \mathbf{X}_{\mathcal{G}} \setminus \hat{\mathbf{X}}}} = k$, given the relation between rank and t-separation in Sullivant et al. (2010). Next, suppose when the pure children of all the latent groups with $|\mathbf{L}_i| \leq k$ have been found, we show the pure children of latent groups with $|\mathbf{L}_i| = k+1$ can also be found by rank constraint. In this case, check all size $k+2$ combination of observed variables $\hat{\mathbf{X}}$ such that $\text{rank}_{\Sigma_{\hat{\mathbf{X}}, \mathbf{X}_{\mathcal{G}} \setminus \hat{\mathbf{X}}}} = k+1$. Consider all such $\{\hat{\mathbf{X}}\}_i^c$ and maintain a disjoint set for them such that $\hat{\mathbf{X}}_i$ and $\hat{\mathbf{X}}_j$ belong to the same group if they have at least one common element. Take the union of such a group, say $\tilde{\mathbf{X}}$. If $\tilde{\mathbf{X}}$ share no common element with the pure children of any $\mathbf{L}_i$ that has size$\leq k$, then elements in $\tilde{\mathbf{X}}$ are pure children of the same latent group $\mathbf{L}_j$ with $|\mathbf{L}_j| = k+1$. By induction, all the pure children of each latent group can be identified.

Further, for an observed variable, say, X, that is a common child of multiple latent groups, suppose its common parents are $\mathcal{L}$, which is a set of some groups of latent variables. Let $\mathbf{A} = \mathsf{X} \cup \bigcup_{\mathbf{L}_j \in \mathcal{L}} \{\text{PCh}^1_{\mathcal{G}}(\mathbf{L}_j)\}$, where $\text{PCh}^1_{\mathcal{G}}(\mathbf{L}_j)$ is a set of $|\mathbf{L}_j|$ pure children of $\mathbf{L}_j$, we have $\text{rank}_{\Sigma_{\mathbf{A}, \mathbf{X}_{\mathcal{G}} \setminus \mathbf{A}}} = \sum_{\mathbf{L}_j \in \mathcal{L}} |\mathbf{L}_j|$, and that when any edge related to X is changed, the related observed rank constraint

will change, and thus the set of equality constraints will also change. Therefore, we have that if $H(\mathcal{G}_1) = H(\mathcal{G}_2)$, then the structure from latent to observed variables in $\mathcal{G}_1$ and that in $\mathcal{G}_2$ are the same.

Next, we show that the structure among latent groups can be identified up to MEC by equality constraints. By making use of Corollary 1.3 in Di (2009), we can translate d-separation between latent groups into t-separation among the pure children of these latent groups. Specifically, for $\mathbf{L}_i, \mathbf{L}_j, i \neq j$, we have $\mathbf{L}_i, \mathbf{L}_j$, are d-separated by $\mathcal{L} \subseteq \{\mathbf{L}_l\}_1^P \backslash \{\mathbf{L}_i, \mathbf{L}_j\}$, iff $\mathbf{A} = \text{PCh}^1_{\mathcal{G}}(\mathbf{L}_i) \cup \bigcup_{\mathbf{L}_l \in \mathcal{L}} \{\text{PCh}^1_{\mathcal{G}}(\mathbf{L}_l)\}$ and $\mathbf{B} = \text{PCh}^1_{\mathcal{G}}(\mathbf{L}_j) \cup \bigcup_{\mathbf{L}_l \in \mathcal{L}} \{\text{PCh}^2_{\mathcal{G}}(\mathbf{L}_l)\}$ are t-separated by $\{\mathcal{L}, \emptyset\}$ or $\{\emptyset, \mathcal{L}\}$, where $\text{PCh}^1_{\mathcal{G}}(\mathbf{L}_l)$ and $\text{PCh}^2_{\mathcal{G}}(\mathbf{L}_l)$ refer to two disjoint groups of $|\mathbf{L}_l|$ pure children of $\mathbf{L}_l$ (by definition we know such two groups must exist). Given Assumption 1 and the relation between rank and t-separation in Sullivant et al. (2010), we have that $\mathbf{L}_i$ and $\mathbf{L}_j$ are d-separated by $\mathcal{L}$, iff $\text{rank}_{\Sigma_{\mathbf{A},\mathbf{B}}} = ||\mathcal{L}||$,where $||\mathcal{L}|| = |\bigcup_{\mathbf{L}_l \in \mathcal{L}} \mathbf{L}_l|$. This means that the d-separations among latent groups can be inferred from rank constraints on observed variables, and d-separations can be used to identified the structure among latent variables up to MEC. As rank constraints are part of the equality constraints, we have that the structure among latent variables can be identified up to MEC by equality constraints. Taking that the structure from latent to observed variables can also be identified up to MEC by equality constraints (proved before) and in Generalized N factor models there is no direct edge between observed variables, we have that if $H(\mathcal{G}_1) = H(\mathcal{G}_2)$, then $\mathcal{G}_1$ and $\mathcal{G}_2$ belong to the same MEC. □

B.3 Proof of Corollary 1

Corollary 1 (Global Consistency by Score for Generalized N Factor Models)**.** *Suppose a model follows Definition 1 with $\mathcal{G}^* \in \mathbb{G}_{GNFM}$ and distribution $\Sigma^*_{\mathbf{X}}$ satisfies Assumption 1. Given observation $\hat{\Sigma}_{\mathbf{X}}$ and let $\mathbb{G}^* = \arg\max_{\mathcal{G} \in \mathbb{G}_{GNFM}} score_{ML}(\mathcal{G}, \hat{\Sigma}_{\mathbf{X}})$. If $\hat{\mathcal{G}} \in \mathbb{G}^*$ and $\hat{\mathcal{G}} \in \arg\min_{\mathcal{G} \in \mathbb{G}^*} \dim(\mathcal{G})$, then $\hat{\mathcal{G}}$ and $\mathcal{G}^*$ are Markov equivalent in the large sample limit.*

Proof. Similar to the proof of Theorem 1, we can show that $H(\hat{\mathcal{G}}) = H(\mathcal{G}^*)$. By Theorem 2, we have $\hat{\mathcal{G}}$ and $\mathcal{G}^*$ belong to the same MEC. □

B.4 Proof of Lemma 1

Lemma 1 (Properties of Initial State)**.** *Suppose a model follows Definition 1 with $\mathcal{G}^* \in \mathbb{G}_{GNFM}$ and we are given observation $\hat{\Sigma}_{\mathbf{X}}$. Then $\mathcal{S}_{init}$ is a supergraph of $\mathcal{S}(\mathcal{G}^*)$ and $\mathcal{S}_{init}$ can generate the observed distribution, i.e., $score_{ML}(\mathcal{S}_{init}, \hat{\Sigma}_{\mathbf{X}}) = \max_{\mathcal{G} \in \mathbb{G}_{GNFM}} score_{ML}(\mathcal{G}, \hat{\Sigma}_{\mathbf{X}})$ in the large sample limit.*

Proof. By the graphical assumption in Definition 2, suppose $\mathbf{L}$ is the set of all latent variables in $\mathcal{G}^*$, then there must exist at least $|\mathbf{L}| \times 2$ observed variables in $\mathcal{G}^*$. As such, the number of latent variables $|\mathbf{L}|$ should be no more than $\lfloor \frac{|\mathbf{X}|}{2} \rfloor$. Thus, by Definition 3, $\mathcal{S}_{\text{init}}$ must be a supergraph of the CPDAG of the ground truth, $\mathcal{S}(\mathcal{G}^*)$, up to permutation of latent variables, and thus $\mathcal{S}_{\text{init}}$ must be able to generate the observation equally well as $\mathcal{S}(\mathcal{G}^*)$. □

B.5 Proof of Lemma 2

Lemma 2 (Correctness of Phase 1 of LGES)**.** *Suppose a model follows Definition 1 with $\mathcal{G}^* \in \mathbb{G}_{GNFM}$ and we are given observation $\hat{\Sigma}_{\mathbf{X}}$. In the large sample limit the output $\mathcal{S}_{phase1}$ of Algorithm 1 is a CPDAG such that the number of latent variables in $\mathcal{S}_{phase1}$ is the same as that of $\mathcal{G}^*$ and the edges from $\mathbf{X}$ to $\mathbf{L}$ in $\mathcal{S}_{phase1}$ is the same as that of $\mathcal{G}^*$, up to permutation of latent variables.*

We provide a sketch of proof as follows.

Proof. We prove by induction. First consider $k = 1$. We show that all the observed variables that belong to a size 1 latent group can be identified. Specifically, $\mathcal{O}_{\mathbf{LX}}(\mathcal{S}, \mathbf{L} \setminus \mathbf{L}', \mathbf{X}_j)$ in line 17 introduces a rank constraint $\text{rank}(\Sigma_{\mathbf{X}_j, \mathbf{X} \setminus \mathbf{X}_j}) = 1$ to $\mathcal{S}_j$, if $\mathcal{S}_j$ in line 17 of Algorithm 1 can still generate the observation, then under the generalized faithfulness in Assumption 1 this rank constraint must also belong to $H(\mathcal{G}^*)$, which means that variables in $\mathbf{X}_j$ must belong to the same size 1 latent group in $\mathcal{G}^*$. As such, all observed variables that belong to the same size 1 latent group can be

identified. Now, suppose we have already identified those observed variables whose parent set has cardinality t and let the set of these observed variables as $\mathbf{X}^t_{\text{done}}$, for all $t \leq T$. We show that for $k = T+1$, all the observed variables that belong to a size k latent group can be identified. Specifically, $\mathcal{S}_j$ introduces a rank constraint $\text{rank}(\Sigma_{\mathbf{X}_j, \mathbf{X}\setminus\mathbf{X}_j}) = k$ to $\mathcal{S}_j$, if $\mathcal{S}_j$ in line 17 of Algorithm 1 can still generate the observation, then under the generalized faithfulness in Assumption 1 this rank constraint must also belong to $H(\mathcal{G}^*)$. Given that $\mathbf{X}_j$ has no common variable with any $\mathbf{X}^t_{\text{done}}$ for all $t \leq T$, we have that $\mathbf{X}_j$ must belong to the same size k latent group in $\mathcal{G}^*$. Therefore, by the end of the while in line 27 in Algorithm 1, all the parents of each observed variable can be identified. Till now, in the state considered in the algorithm, there might still exist some latent variables having no observed variable as children. These latent variables are removed from the current state in line 28 in Algorithm 1, and thus the number of latent variables in $\mathcal{S}_{\text{phase1}}$ is the same as that of the ground truth and the structure between latent variables and observed variables in $\mathcal{S}_{\text{phase1}}$ is also the same as that of the ground truth up to permutation of latent variables. □

B.6 Proof of Theorem 3

Theorem 3 (Correctness of LGES)**.** *Suppose a model follows Definition 1 with $\mathcal{G}^* \in \mathbb{G}_{GNFM}$ and we are given observation $\hat{\Sigma}_{\mathbf{X}}$. In the large sample limit the output $\mathcal{S}_{final}$ of Algorithms 1 and 2 is a CPDAG that represent the MEC of $\mathcal{G}^*$, up to permutation of latent variables.*

The proof is partially inspired by the proof of Lemma 10 in Chickering (2002).

Proof. First, we know that all the states must be able to generate the observation. Assume that Phase 2 terminates with a sub-optimal state $\mathcal{S}'$ and let $\mathcal{G}'$ be a DAG that belongs to $\mathcal{S}'$. By Theorem 4 in Chickering (2002) we know that there must exist a sequence of covered edge reversals and edge additioins that transforms $\mathcal{G}^*$ to $\mathcal{G}'$. Suppose $\mathcal{G}''$ precedes the last edge addition in the sequence. We have that $\mathcal{G}''$ must also be able to generate the observation and $\mathcal{S}(\mathcal{G}'')$ is a neighboring state of $\mathcal{S}'$, which means Phase 2 should not terminate at $\mathcal{S}'$, yielding a contradiction. Thus, by the end of Phase 2, the structure among latent variables in $\mathcal{S}_{\text{final}}$ must be the same as that of the ground truth up to permutation of latent variables. Taking Lemma 2 also into consideration, we have that in the large sample limit the output $\mathcal{S}_{\text{final}}$ of Algorithms 1 and 2 is a CPDAG that represent the MEC of $\mathcal{G}^*$, up to permutation of latent variables. □

C Additional Definitions, Implementation Details, Runtime Analysis, and Examples

C.1 Detailed Definition of Equality Constraints, Inequality Constraints, $H(\mathcal{G})$, and $\mathbb{H}^n$.

Definition 6 (Definition of Equality Constraints, Inequality Constraints, and $H(\mathcal{G})$.)**.** *Let $\mathcal{G}$ be the DAG structure of a Partially Observed Linear Causal Model with m latent variables and n observed variables (as in Definition 1). The entries of the observed covariance matrix $\Sigma_{\mathbf{X}}$ are polynomial functions of model parameters $\theta = (F^T, \Omega_\epsilon)$, where F^T is the edge coefficient matrix and Ω_ϵ is the diagonal noise variance matrix. This induces a parametric map under $\mathcal{G}$:*

$$\phi_{\mathcal{G}} : \mathbb{R}^{|\theta|} \to \mathbb{R}^{n(n+1)/2},$$

from parameters to observed covariance matrix, defining the system of equations:

$$\{\Sigma_{\mathbf{X}ij} - \phi_{\mathcal{G},ij}(\theta) = 0 | 1 \leq i \leq j \leq n\}.$$

Each equality constraint is just a polynomial equality equation consists of entries of $\Sigma_{\mathbf{X}}$ and θ, and each inequality constraint is just a polynomial inequality equation consists of entries of $\Sigma_{\mathbf{X}}$ and θ. As our objective is to find information from $\Sigma_{\mathbf{X}}$ to infer the causal structure, we should focus on those equality equations that consists of only entries of $\Sigma_{\mathbf{X}}$, which can be achieved as follows.

Let $\mathbb{R}[\theta, \Sigma_{\mathbf{X}}]$ be the polynomial ring which contains all variables for all model parameters θ and all distinct covariance entries $\Sigma_{\mathbf{X}ij}$, while $\mathbb{R}[\Sigma_{\mathbf{X}}]$ be the polynomial ring on $\Sigma_{\mathbf{X}}$ only. Define the ideal $I_{\mathcal{G}} \subseteq \mathbb{R}[\theta, \Sigma_{\mathbf{X}}]$ generated by the above equations as:

$$I_{\mathcal{G}} = \langle \{\Sigma_{\mathbf{X}ij} - \phi_{\mathcal{G},ij}(\theta) = 0 | 1 \leq i \leq j \leq n\} \rangle.$$

Then, $H(\mathcal{G})$ is the elimination ideal obtained by intersecting $I_{\mathcal{G}}$ with $\mathbb{R}[\Sigma_{\mathbf{X}}]$, i.e.,

$$H(\mathcal{G}) := I_{\mathcal{G}} \cap \mathbb{R}[\Sigma_{\mathbf{X}}].$$

In other words, $H(\mathcal{G})$ contains all equality constraints implied by $\mathcal{G}$ on the observed covariance matrix.

Definition 7 (Definition of $B(\mathcal{G})$ and $\mathbb{H}^n$.)**.** *Let $>$ be a fixed lexicographic monomial ordering on $\mathbb{R}[\theta, \Sigma_{\mathbf{X}}]$ such that all parameter variables in θ are greater than all covariance variables in $\Sigma_{\mathbf{X}}$. Define $B(\mathcal{G})$ as follows. (i) Compute the reduced Gröbner basis of $I_{\mathcal{G}}$ following ordering $>$, i.e., $G_B(I_{\mathcal{G}}, >)$. (ii) Retain only those polynomials that involve only variables in $\Sigma_{\mathbf{X}}$. Formally,*

$$B(\mathcal{G}) := G_B(I_{\mathcal{G}}, >) \cap \mathbb{R}[\Sigma_{\mathbf{X}}].$$

Then $\mathbb{H}^n$ is defined as

$$\mathbb{H}^n := \bigcup_{\mathcal{G} \in \mathbb{G}^n} B(\mathcal{G}).$$

In essence, $H(\mathcal{G})$ contains all equality constraints implied by $\mathcal{G}$ on the observed covariance matrix, while $B(\mathcal{G})$ consists of a canonical and minimal (owing to reduced Gröbner basis) set of polynomial constraints among the observed covariances that must vanish for any distribution consistent with structure $\mathcal{G}$. Since the reduced Gröbner basis is unique (given a fixed monomial order), $B(\mathcal{G})$ serves as a standard representative of these constraints. The vanishing set of $B(\mathcal{G})$ defines the smallest algebraic variety that contains all observed covariance matrices generated by the model.

C.2 Definition of One Factor Model by Silva et al. (2003) and Comparison

Definition 8 (One Factor Model (Silva et al., 2003))**.** *DAG $\mathcal{G}$ satisfies the definition of One Factor Model if each measured variable has a single latent parent, and each latent variable has at least three measured variables as children.*

First, the generalized N factor model takes one factor model as a special case. An example of one factor model can be found in Figure 7 (b). As a comparison, an example of generalized N factor model can be found in Figure 7 (a). Specifically, (a) differs from (b) in that, (i) (a) allows latent variables to form a group and share observed variables as children, e.g., $\{\mathsf{L}_4, \mathsf{L}_5\}$ in (a) compared to L_4 in (b), (ii) (a) allows some observed variables to be common children of muiltiple groups of latent variables, e.g., X_{14} in (a), while (b) does not.

C.3 Definition of V-Structure and CPDAG

Definition 9 (V-Structure / Collider)**.** *Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a directed graph, where $\mathbf{V}$ is a set of vertices (nodes) and $\mathbf{E}$ is a set of directed edges. A v-structure is an ordered triplet of distinct nodes $(\mathsf{X}, \mathsf{Y}, \mathsf{Z})$ where $\mathsf{X}, \mathsf{Y}, \mathsf{Z} \in \mathbf{V}$, that satisfies the following two conditions:*

- *The graph $\mathcal{G}$ contains the directed edges $\mathsf{X} \to \mathsf{Y}$ and $\mathsf{Z} \to \mathsf{Y}$.*
- *X and Z are not adjacent in $\mathcal{G}$.*

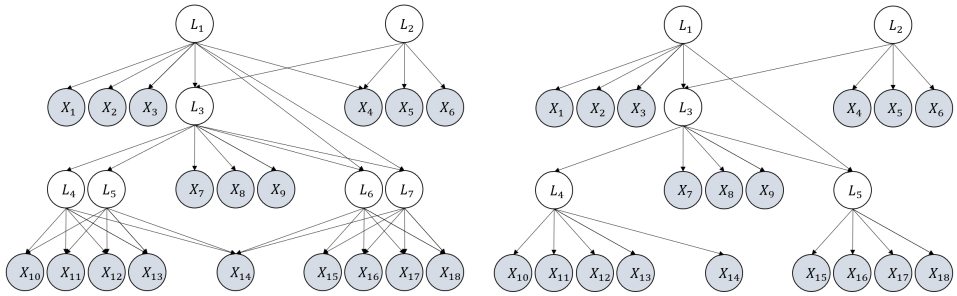
The node Y is referred to as the collider node of the v-structure.

Definition 10 (CPDAG (also called essential graph or maximally oriented graphs) (Spirtes et al., 2001; Chickering, 2002))**.** *Let $\mathcal{E}$ be the Markov equivalence class of a DAG $\mathcal{G}$. The CPDAG that represents $\mathcal{E}$ is the unique graph $\mathcal{G}^*$ (consisting of both directed and undirected edges) such that:*

- ***Directed Edge:** An edge $\mathsf{X} \to \mathsf{Y}$ is in $\mathcal{G}^*$ if and only if the edge $\mathsf{X} \to \mathsf{Y}$ exists in every DAG $\mathcal{G}_1 \in \mathcal{E}$.*
- ***Undirected Edge:** An edge $\mathsf{X} - \mathsf{Y}$ is in $\mathcal{G}^*$ if and only if there exists at least one DAG $\mathcal{G}_1 \in \mathcal{E}$ containing the edge $\mathsf{X} \to \mathsf{Y}$ and at least one other DAG $\mathcal{G}_2 \in \mathcal{E}$ containing the edge $\mathsf{X} \leftarrow \mathsf{Y}$.*

C.4 Goodness-of-model-fit Measures

RMSEA (Steiger, 1980; 1990) measures the discrepancy due to the approximation per degree of freedom. It is actually a badness-of-fit measure and thus the lower value the better fit of the model.



(a) An illustrative graph that satisfies generalized N factor model.

(b) An illustrative graph that satisfies one factor model.

Figure 7: Illustrative examples to compare two graphical assumptions, generalized N factor model v.s. one factor model.

The sample RMSEA is estimated as follows.

$$\hat{\text{RMSEA}} = \sqrt{\frac{\max(\chi^2 - df, 0)}{df(N-1)}}, \tag{6}$$

where χ^2 is the chi-square statistic of the concerned model and df is the degree of freedom of the chi-square statistic.

The CFI (Bentler, 1990) measures the relative improvement in terms of fit from the baseline model to the proposed model. The sample CFI is estimated as follows:

$$\hat{\text{CFI}} = 1 - \frac{\max(\chi_k^2 - df_k, 0)}{\max(\chi_0^2 - df_0, 0)}, \tag{7}$$

where χ_k^2 and df_k corresponds to the concerned model while χ_0^2 and df_0 corresponds to the baseline independent model that can only parameterize a diagonal covariance matrix.

The TLI (Tucker & Lewis, 1973; Bentler & Bonett, 1980) measures a relative reduction in misfit per degree of freedom. The sample estimator of TLI can be given as follows:

$$\hat{\text{TLI}} = \frac{\chi_0^2/df_0 - \chi_k^2/df_k}{\chi_0^2/df_0 - 1}. \tag{8}$$

C.5 Implementation Details and Discussion on the design of δ

Our code is based on Python3.7 and PyTorch (Paszke et al., 2017) and the optimization problem in Equation (2) is solved by Adam (Kingma & Ba, 2014) and LBFGS. Data is standardized to have zero mean and unit variance. The hyper parameter δ in Algorithms 1 and 2 is set as $\delta = 0.25 \times \frac{\log(N)}{N}$, where N is the sample size. This design follows the spirit of BIC score such that $\delta \to 0$ when $N \to \infty$. In practice, we found that the result is only influenced marginally by a small change of δ.

Bellow we provide a further discussion on δ and the criterion of keeping edge removal in LGES.

Specifically, in LGES an edge removal is kept only when $|\text{LogL\_curr} - \text{LogL\_prev}| \leq k\frac{\log N}{N}$, where $\delta \coloneqq k\frac{\log N}{N}$ is the tolerance level, N is the sample size, and k is a hyper-parameter that controls sparsity. This design follows the spirit of BIC score in GES ($\text{score\_BIC} = \text{LogL} - 0.5\frac{\log N}{N}\text{dim}$). Thus, the behavior of LGES is quite similar to GES - both of them accommodates asymptotic consistency and finite sample performance at the same time.

In the asymptotic case, $0.5\frac{\log N}{N}\text{dim} \to 0$, and thus the likelihood term dominates the BIC score. Therefore, what GES favors is exactly the graph that (i) has the best likelihood, (ii) at the premise of (i) the dimension should be as small as possible. Similarly, in the asymptotic case $\delta \to 0$, and thus in LGES what is performed is precisely "Only when the likelihood after the deletion is still the best do we keep the deletion" in the asymptotic case. Combined with other designs as discussed in our last response, LGES also achieves the goal of finding the graph that (i) has the best likelihood, (ii) at

the premise of (i) the dimension should be as small as possible, and thus guarantees the asymptotic consistency.

In the finite sample case, there exists a problem that a supermodel always has a better likelihood. To address this problem, in GES the BIC score sensibly encourages an edge removal by the term $0.5\frac{\log N}{N}$dim. As in GES each edge removal results in exactly 1 dimension decrease, the criterion in GES is equivalent to keeping an edge removal as long as LogL_curr $\geq$ LogL_prev $- 0.5\frac{\log N}{N}$. In LGES, an edge removal is kept when $|\text{LogL\_curr} - \text{LogL\_prev}| \leq k\frac{\log N}{N}$, and thus it is also encouraged sensibly in LGES with finite samples.

C.6 Runtime Analysis

Next we discuss the time complexity of LGES. Similar to the classical PC and GES, our method has a worst-case complexity exponential in the number of observed variables. However, if the underlying graph is sparse, which is a common and reasonable assumption Kalisch & Bühlman (2007), the complexity becomes polynomial. The intuition is as follows. Similar to PC and GES, during the process, LGES enumerates different combinations of variables and check the score to decide whether to delete some edges. Although the number of all combinations is exponential (also in GES and PC), if the underlying graph is sparse, e.g., maximum degree of a node is P, the algorithm will successfully find the correct combination to delete the edge before enumerating all the combinations. Thus the number of combinations that are actually enumerated only depends on the constant P instead of number of variables. Therefore, the time complexity will become a term polynomial in N, times a term polynomial in constant P, and thus polynomial in N.

In our implementation, the computational cost is almost irrelevant to sample size, as we only need to calculate the sample covariance once and cache it. Further, lines 5,6,13 in Alg 1 and lines 5,6,9 in Alg 2 can be executed in parallel. Owe to these designs, in practice LGES is fairly efficient: on average it takes only one minute to handle a graph with 20 variables.

We conduct all the experiments with single Intel(R) Xeon(R) CPU E5-2470. Thanks to the fact that lines 5,6,13 in Algorithm 1 and lines 5,6,9 in Algorithm 2 can be executed in parallel, we employ the package joblib in python to conduct them by multi-processing. As such, on average it takes only one minute to handle a graph with 20 variables. We note that the computational cost is almost irrelevant to sample size, as we only need to calculate the sample covariance matrix once and cache it for further use. Compared to RLCD and GIN, LGES is faster than RLCD but slower than GIN. Specifically, it takes RLCD around 2 minutes and GIN around 20 seconds to handle a graph with 20 variables, while around one minute for LGES. The reason why GIN is faster than LGES, is that GIN only focuses on the structure between latent variables and observed variables and does not identify structure among latent variables, while RLCD and LGES identify the whole underlying structure involving both observed and latent variables.

C.7 Whether the Identifiability Theory Can be Extended to Non-Gaussian or Nonlinear Models?

The identifiability result can be extended to both linear non-Gaussian and certain kinds of nonlinear models. (i) The key role of the score in our identifiability theory is to check whether the constraints on the observed covariance matrix is a subset of the constraints entailed by a candidate structure. For any structure, the constraints entailed by in the non-Gaussian case is exactly the same as that of the Gaussian case. Therefore, the proposed identifiability theory still holds in the non-Gaussian scenario. (ii) For certain kinds of nonlinearity, the proposed identifiability theory still works. For example, in Nonparanormal models where there exist smooth, monotonic transformations for each variable to transform variables to be jointly Gaussian, as the monotonic transformations can be identified up to trivial indeterminacy, we can still use the proposed score-based theory for structure identification.

C.8 Further Discussion About Error Propagation

Our score-based greedy search method is similar to a constraint-based method in the sense that both starts from a complete graph and deletes edges according to some criterion. In this sense, why score-based methods are expected to suffer less from error propagation? The main reason lies in

the difference of the used criteria. Score-based methods rely on MLE score and dimension, while constraint-based methods rely on statistical tests. We provide our further analysis as follows.

Using Score as the Criterion Can Have Fewer Algorithm Steps. The core theoretical basis of constraint-based methods and score-based methods are the same - checking whether the equality constraints on the observational distribution are aligned with the constraints entailed by a candidate graph. Thus, roughly speaking, given a dataset the total number of constraints that need to be checked is the same for both score-based and constraint-based methods (if we want to guarantee asymptotic correctness). However, constraint-based methods only examine one constraint in each step by a single test, while score-based methods can in essence examine multiple constraints together at the same time. More specifically, in score-based methods, we can introduce multiple constraints in a step and examine them together at the same time, by checking whether the likelihood is still maximal. As a consequence, score-based methods require fewer algorithm steps, which mitigates the problem of error propagation. This is also aligned with our empirical observation that, although the score-based GES and the constraint-based PC require basically the same assumptions, in practice GES often has fewer steps to finish, and can often handle 10 times more variables than PC.

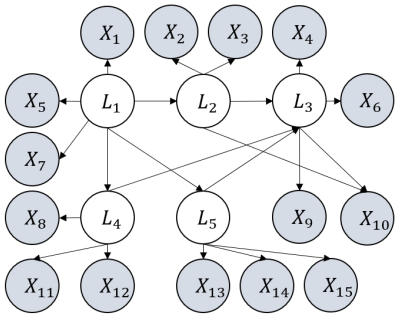
Using Test as the Criterion Suffers More From Small Sample Sizes. Statistical tests (except for some permutation-based) reply on an approximation of the asymptotic null distribution of the test statistic to calculate the p-value for controlling the type-I error. When the sample size is small, e.g., N=100, the approximation (often based on the central limit theorem) could be far away from the true null distribution. As a consequence, with small sample size the type-I error in each step of the constraint-based methods cannot be properly controlled, let along the type-II error. This is also consistent with our empirical results: in Table 1 and Table 2, LGES still performs well with a very small sample size (N=100) both with and without model mis-specification, while constraint-based methods such as RLCD and GIN does not work well in these scenarios.

D LIMITATIONS

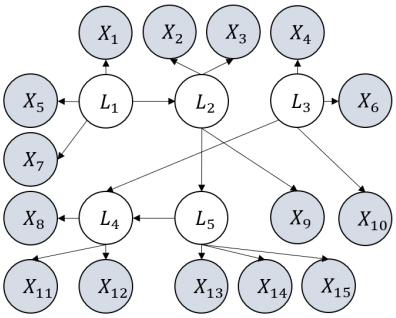
One limitation of this work is that our theoretical results are based on the assumption of linear causal models. When data is not linear, we have also conducted experiments to see the performance and it can be shown that our method still performs well. Yet, theoretical analysis and identifiability guarantee for the nonlinear case are to be developed and will be the focus of future work.

E BROADER IMPACTS

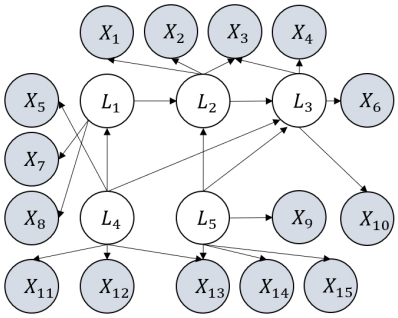
The goal of this paper is to advance the field of machine learning. We do not see any potential negative societal impacts of the work.



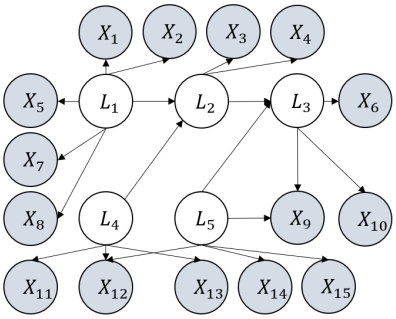
(a) $\mathcal{G}_1$.



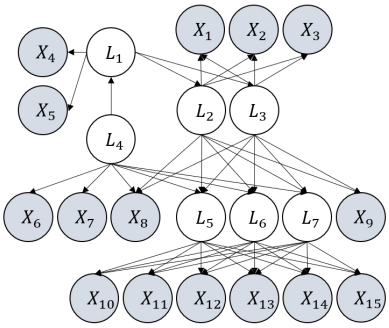
(b) $\mathcal{G}_2$.



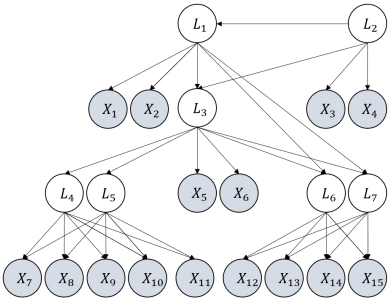
(c) $\mathcal{G}_3$.



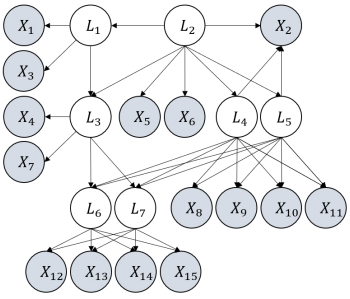
(d) $\mathcal{G}_4$.



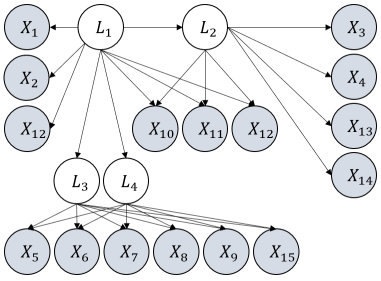
(e) $\mathcal{G}_5$.



(f) $\mathcal{G}_6$.



(g) $\mathcal{G}_7$.



(h) $\mathcal{G}_8$.

Figure 8: Examples of graphs considered in our experiments. They satisfy Definition 2.