PROVABLE GUARANTEES FOR LEARNING PURE LATENT HIGH-DIMENSIONAL STRUCTURAL EQUATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning Directed Acyclic Graphs (DAGs) from observational data is a fundamental challenge in causal inference, exacerbated when the true causal mechanisms occur among unobserved latent variables. In this paper, we study the identifiability and learning of pure latent high-dimensional Structural Equation Models (SEMs). While traditional causal discovery algorithms only recovering Markov equivalence classes (CPDAGs) or succumbing to latent confounding. We propose a generic solver framework to recover the exact, fully-directed latent DAG in pure latent high-dimensional SEMs. Our two-phase approach first reconstructs the latent covariance from observational data, and then iteratively identifying and marginalizing terminal vertices to recover the latent DAG structure. Under this framework, we relaxes the faithfulness assumption in standard causal discovery algorithm, and provides theoretical guarantees for topological ordering, bounded parameter estimation error, and exact support recovery via thresholding. We establishing that exact latent DAG recovery is achievable with logarithmic sample complexity of $\mathcal{O}(\log p)$ where p is the number of latent variable. Our experiments on synthetic data and a real-world fMRI brain connectivity dataset demonstrates that our method successfully recovers exact directed causal structures where standard observed or equivalence-class baselines fail.

1 INTRODUCTION

In modern statistical learning and causal inference, the identification of causal relationships from purely observational data is a foundational challenge. However, in many domains the key causal features of a system are rarely observed directly. Researchers resort to working with high-dimensional surrogate measurements generated by lower-dimensional latent variables. For example, in neuroscience and social science, cognitive processes such as inhibitory control, cue reactivity, and self-awareness are highly complex and impossible to measure directly, instead, researchers carefully design tasks to collect some surrogate data, such as recording brain activity through functional magnetic resonance imaging. Traditional causal discovery algorithms often operate under the assumption of causal sufficiency, assuming that no unobserved confounders exist in the system. When this assumption is relaxed, the structural learning problem becomes highly indeterminate, as latent variables can induce spurious correlations among observed variables that cannot be attributed to true causation.

In this paper, we study DAG learning in structural equation models, specifically focusing on the problem of latent DAG recovery. We define latent DAG recovery as the task of identifying the true DAG structure and causal ordering strictly among the unobserved latent variables, leveraging only the measurements available in the observed data space.

Existing methodologies for latent DAG recovery face severe computational bottlenecks and often rely on restrictive identifiability conditions that provide only asymptotic guarantees. Score-based global optimization techniques, such as in [Ng et al. (2024)], require exact searches over exponential discrete DAG spaces; while they offer certain finite-sample performance bounds, they identify structures only up to an algebraic equivalence class under the strict assumption that the latent noise covariance is the identity matrix. Constraint-based methods [Silva et al. (2006); Kummerfeld and Ramsey

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(2016)[, derive their identifiability strictly from the "pure measurement model" [Silva et al. (2006); Kummerfeld and Ramsey (2016); Huang et al. (2022)] and rank faithfulness [Kummerfeld and Ramsey (2016); Huang et al. (2022)] assumptions. Consequently, they offer only asymptotic consistency in the large-sample limit without finite-sample guarantees. Recent advancements using generalized rank deficiency constraints, successfully extend discovery to complex hierarchical latent structures [Huang et al. (2022)], however, they still rely on rank faithfulness and recover only the Markov Equivalence class asymptotically. Higher-Order Cumulant algorithms [Xie et al. (2022; 2024b); Cai et al. (2025); Xie et al. (2024a)] which require at least third-order statistical information can successfully identify the exact DAG; however, algorithms relying on higher-order cumulants suffer from extreme sample complexity. Finally, copula-based rank approaches designed for mixed data establish identifiability via copula normality and provide asymptotic consistency [Cui et al. (2018)], but ultimately rely on sequential conditional independence tests that compound multiple-testing errors in high dimensions.

To bypass these limitations, this paper proposes a novel, polynomial-time framework that combines scalable latent precision matrix estimation with deterministic topological sorting. We start the study by considering the pure model [Silva et al. (2006); Kummerfeld and Ramsey (2016); Huang et al. (2022)], where each observed indicator is caused by exactly one latent variable, and there are absolutely no direct causal edges or correlated errors between the observed indicators themselves. On top of this, we consider a linear structural equation model where observed variables act as proxies for latent variables, mapped via an orthogonal latent-to-observed embedding matrix. By exploiting the block structure of the full precision matrix, we demonstrate that the marginal precision matrix of the latent variables uniformly shifts the precision matrix of the underlying DAG. Our proposed pipeline leverages the recent spiked covariance estimation algorithm by [Long et al. (2026)] to robustly isolate and estimate the latent precision matrix. Subsequently, we apply the terminal vertex sorting mechanism developed by [Ghoshal and Honorio (2018)] to recover the structure of the internal latent SEM.

Our method does not rely on the restrictive faithfulness assumption to achieve exact DAG identifiability. In standard causal discovery (like the PC algorithm), the "faithfulness assumption" dictates that if two variables appear statistically uncorrelated in the data, it must mean there is no causal link between them in the true graph. However, this assumption often fails in reality. By operating strictly on second-order statistics under the orthogonal measurement model. our method circumvents the NP-hard search spaces of score-based models, and achieves rigorous finite-sample statistical efficiency. This paper establishes our two-step pipeline as a mathematically rigorous and scalable solution, with sample complexity of $\mathcal{O}(\log p)$ under sub-Gaussian noise, where p is the number of latent variables, and exact DAG identifiability against the contemporary literature.

2 NOTATION

In this section, we introduce the notations that will be used throughout the paper. We use lowercase bold letters for vectors (e.g. $\mathbf{a}$) and uppercase bold letters for matrices (e.g., $\mathbf{A}$), and subscripts of non-bold letters to denote respective entries of a vector or matrix (e.g., a_i and A_{ij}). For any matrix $\mathbf{A}$, we denote its support set of the matrix by $\mathcal{S}(\mathbf{A}) := \{(i,j)|A_{i,j} \neq 0\}$. For a vector $\mathbf{a}$, $\|\mathbf{a}\|_1$, $\|\mathbf{a}\|_\infty$ and $\|\mathbf{a}\|_2$ denote the ℓ_1 norm, ℓ_∞ norm and Euclidean norm respectively. For a matrix $\mathbf{A}$, $\|\mathbf{A}\|_1$, and $\|\mathbf{A}\|_\infty$denote the entrywise ℓ_1 norm, and entrywise ℓ_∞ norm respectively. The identity matrix with k rows and k columns is denoted as $\mathbf{I}_k$. For a matrix $\mathbf{A}$, $\mathrm{diag}(\mathbf{A})$ is the vector containing the diagonal entries of $\mathbf{A}$. For a vector $\mathbf{a}$, $\mathrm{Diag}(\mathbf{a})$ is the matrix containing $\mathbf{a}$ in its diagonal, and zero everywhere else.

Let $G = (V, E)$ be a directed acyclic graph where $V = \{1, \ldots, p\}$ are the set of vertices/nodes, E is the edge set where edge $(i, j) \in E$ implies node $i \leftarrow j$. The set of parent of node i is denoted by $\pi(i) = \{j \in V \mid (i, j) \in E\}$. The set of children of i is denoted by $\phi(i) = \{j \in V \mid (j, i) \in E\}$. A node i is called a *terminal vertex* if it has no children, i.e., $\phi(i) = \emptyset$. A *topological ordering set* $\mathcal{T}_G$ of G is a set of linear ordering τs of its vertices such that for every directed edge $(i, j) \in E$, node j appears before node i in the ordering, i.e. $\mathcal{T}_G = \{\tau \in S_V | \tau(j) < \tau(i), \forall (i, j) \in E\}$, where S_V is the set of all possible premutation of V.

3 MAIN RESULTS

The primary motivation of this paper is to discover the causal relationships between hidden, unobserved variables form observable surrogate. Because we cannot measure these hidden variables directly, we must rely on high-dimensional proxy variables that we can observe. In this section, we address the challenge of learning pure latent high-dimensional SEMs from observed data. We first describe our model assumptions, follow by the identifiability condition, then we demonstrate our algorithms for recovering the structure of the internal latent DAG and show that it works with any precision matrix estimator that satisfies the Generic Sparse Estimator Guarantees. We finish the section by providing theoretical guarantees, where all proofs are postponed to Appendix A

3.1 MODEL DEFINITION

The data generating process is defined by the following linear structural equations:

$$\begin{bmatrix}\mathbf{x}\\ \mathbf{z}\end{bmatrix} = \begin{bmatrix}\mathbf{B} & \mathbf{C}\\ 0 & \mathbf{D}\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{z}\end{bmatrix} + \begin{bmatrix}\boldsymbol{\varepsilon}_x\\ \boldsymbol{\varepsilon}_z\end{bmatrix} \tag{1}$$

Formally,

$$\mathbf{x} = \begin{bmatrix}\mathbf{x}_i\\ \vdots\\ \mathbf{x}_p\end{bmatrix}, \mathbf{z} = \begin{bmatrix}z_i\\ \vdots\\ z_p\end{bmatrix}, \mathbf{C} = \begin{bmatrix}\mathbf{c}_1 & \mathbf{0} & \dots & \mathbf{0}\\ \mathbf{0} & \mathbf{c}_2 & \dots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{c}_p\end{bmatrix} \tag{2}$$

where p is the number of latent. The vector $\mathbf{z} \in \mathbb{R}^p$ is the hidden variables we care about (e.g., true underlying psychological states), vector $\mathbf{x} \in \mathbb{R}^Q$ with $Q := \sum_{i=1}^{p} q_i$ is the high-dimensional proxies we can actually measure (e.g., survey responses or fMRI pixels). Matrix $\mathbf{B}$ dictates how the observed variables x interact with each other. In this paper we set $\mathbf{B} = 0$ that means there is no causal link between observed variables. [Ng et al. (2024); Huang et al. (2022); Xie et al. (2022)] also assume $B = 0$ but they have impure latent where an observed variable is influenced by 2 or more latent variable, and tried to recover all cross-layer and observed edges. [Cai et al. (2025); Xie et al. (2024b)] assume fully impure Models where $B \neq 0$ and aims to recover entire integrated causal structure. Matrix $\mathbf{C} \in \mathbb{R}^{Q\times p}$ defines how the hidden variables $\mathbf{z}$ generate the observed variables $\mathbf{x}$. Matrix $\mathbf{D} \in \mathbb{R}^{p\times p}$ is the directed edges in the internal latent DAG representing the causal network of $\mathbf{z}$ (such that $\mathcal{S}(\mathbf{D}) = E$), and it is what we aim to recover from $\mathbf{x}$.

For $i = 1, \dots, p$, the observable random vector $\mathbf{x}_i \in \mathbb{R}^{q_i}$ is associated with a latent random variable $z_i \in \mathbb{R}$, where the dimension $q_i > 1$. Each observed variable $\mathbf{x}_i$ follow a linear model in which

$$\mathbf{x}_i = \mathbf{c}_i z_i + \boldsymbol{\varepsilon}_{x_i}$$

for some sparse vector constant $\mathbf{c}_i$, with $\|\mathbf{c}_i\|_2 = 1$, and random noise $\boldsymbol{\varepsilon}_{x_i} \in \mathbb{R}^{q_i}$ with $\mathbb{E}[\boldsymbol{\varepsilon}_{x_i}] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{\varepsilon}_{x_i}\boldsymbol{\varepsilon}_{x_i}^T] = \sigma_i^2\mathbf{I}_{qi_i}$. The latent scaler z_i's are zero mean and the random noise vectors $\boldsymbol{\varepsilon}_{x_1}, \dots, \boldsymbol{\varepsilon}_{x_p}$ are independent.

The noise terms are distributed as $\boldsymbol{\varepsilon}_x \sim \mathcal{N}(0, \mathbf{N}_\mathbf{x})$, where $\mathbf{N}_\mathbf{x} = \mathrm{Diag}(\sigma^2_{x_1}, \dots, \sigma^2_{x_Q})$, and $\boldsymbol{\varepsilon}_z \sim \mathcal{N}(0, \mathbf{N}_\mathbf{z})$, where $\mathbf{N}_\mathbf{z} = \mathrm{Diag}(\sigma^2_{z_1}, \dots, \sigma^2_{z_p})$.

Our full precision matrix $\boldsymbol{\Omega}$ of the joint distribution $(\mathbf{x}, \mathbf{z})$ is structured as:

$$\boldsymbol{\Omega}_{full} = \begin{bmatrix}\mathbf{N}_\mathbf{x}^{-1} & -\mathbf{N}_\mathbf{x}^{-1}\mathbf{C}\\ -\mathbf{C}^T\mathbf{N}_\mathbf{x}^{-1} & \mathbf{C}^T\mathbf{N}_\mathbf{x}^{-1}\mathbf{C} + (\mathbf{I}-\mathbf{D})^T\mathbf{N}_\mathbf{z}^{-1}(\mathbf{I}-\mathbf{D})\end{bmatrix} \tag{3}$$

Let $\boldsymbol{\Omega} := (\mathbf{I}-\mathbf{D})^T\mathbf{N}_\mathbf{z}^{-1}(\mathbf{I}-\mathbf{D})$ denote the latent precision matrix of the latent DAG structure. The bottom-right block corresponding to $\mathbf{z}$ simplifies to $\boldsymbol{\Omega}_{\mathbf{zz}} := \mathbf{C}^T\mathbf{N}_\mathbf{x}^{-1}\mathbf{C} + \boldsymbol{\Omega}$.

3.2 ISOLATION OF THE INTERNAL LATENT DAG

Before recovering the topological ordering of the internal structural equation model over $\mathbf{z}$, we must first prove that the algorithm successfully identifies and marginalizes out the observed terminal variables $\mathbf{x}$, thereby isolating the internal precision matrix without structural distortion.

Lemma 1 (Isolation of Internal SEM)**.** *Under the model definition of the full precision matrix 3 defined above, the following property hold:*

Let $\mathbf{\Omega}_{(-X)}$ denote the Schur complement of $\mathbf{\Omega}_{full}$ with respect to the entire X block. This marginalization perfectly recovers the true precision matrix of the internal DAG, denoted $\mathbf{\Omega}$. Formally:

$$\mathbf{\Omega}_{(-\mathbf{x})} = \mathbf{\Omega} = (\mathbf{I} - \mathbf{D})^T \mathbf{N}_{\mathbf{z}}^{-1} (\mathbf{I} - \mathbf{D}) \tag{4}$$

3.3 Identifiability Condition

The following assumption gives a sufficient condition under which the structure and parameters of an SEM can be uniquely recovered from observational data using our algorithm. The assumption is defined in terms of subgraphs of G obtained by removing terminal vertices sequentially. For any $\tau \in \mathcal{T}_G$, we will consider sequence of graphs $G[m, \tau] = (V[m, \tau), (E[m, \tau])]$, indexed by $[m, \tau]$, where $G[m, \tau]$ is the induced subgraph of G over the first m vertices in the topological ordering τ.

Assumption 1 (Identifiability Condition)**.** *Given the block model defined above, the latent SEM $(G, \mathbf{D}, \{\sigma_{z_i}^2\})$ over $\mathbf{z}$ must satisfy the condition that for any induced subgraph $G[m, \tau]$, the terminal vertices are strictly smaller diagonal entries in the latent precision matrix $\mathbf{\Omega}$ than any non-terminal vertices. More formally, $\forall (i, j) \in V[m, \tau] \times V[m, \tau]$, $m \in [p]$, and $\forall \tau \in \mathcal{T}_G$, such that $\phi_{G[m,\tau]}(i) = \emptyset \wedge \phi_{G[m,\tau]}(j) \neq \emptyset$:*

$$(\sigma_{z_i}^2)^{-1} < (\sigma_{z_j}^2)^{-1} + \sum_{l \in \phi_{G[m,\tau]}(j)} (\sigma_{z_l}^2)^{-1} D_{l,j}^2 \tag{5}$$

Next we show that the above condition is necessary.

Lemma 2. *There exists a family of subgraphs $\tilde{G} \subset G$ with $|G| = 2^{\Theta(p)}$, corresponding autoregression matrices $\mathbf{D}$, and diagonal noise matrices $\mathbf{N}$, such that the models do not satisfy Assumption 1, yet induce the exact same global precision matrix $\mathbf{\Omega}$ and a distribution $\mathcal{P}(\mathbf{x}, \mathbf{z})$ that shares the same conditional independence structure.*

3.4 Structure Learning Algorithm

Our approach to learning the pure latent high-dimensional SEM is to first estimate the latent precision matrix of the internal latent variables by leveraging sparse PCA and a generic sparse precision solver. After that, we sequentially peel the topological layers of the latent DAG using the precision matrix entries.

3.4.1 Latent Covariance Matrix Estimation

Because the internal variables $\mathbf{z}$ are strictly unobserved, we cannot directly compute their empirical precision matrix. Instead, we must reconstruct it from the observed variables $\mathbf{x}$. This reconstruction relies on the theoretical guarantees of pure measurement models, as established by [Long et al. (2026)].

Proposition 1 (Latent Covariance Reconstruction, Theorem 2 in [Long et al. (2026))**.** *(Latent Covariance Reconstruction) Given the orthogonal mapping matrix $\mathbf{C} \in \mathbb{R}^{Q \times p}$ such that $\mathbf{C}^T\mathbf{C} = \mathbf{I}_p$, Under model 1 the covariance matrix of the internal latent variables $\mathbf{\Sigma}_{\mathbf{zz}}$ can be perfectly recovered from the observed covariance matrix via the projection:*

$$\mathbf{\Sigma}_{\mathbf{zz}} = \mathbf{C}^T \mathbf{\Sigma}_{\mathbf{xx}} \mathbf{C} - \mathbf{C}^T \mathbf{N}_{\mathbf{x}} \mathbf{C} \tag{6}$$

By applying this proposition to our finite-sample data, we apply a sparse PCA program [Lei and Vu (2015)] to the empirical sample covariance matrix of the observed variables, $\widehat{\mathbf{\Sigma}}_{\mathbf{xx}}$. By extracting the maximum eigenvectors, we identify the disjoint clusters of observed variables and estimate the vectors $\widehat{\mathbf{c}}_i$, forming the matrix $\widehat{\mathbf{C}}$. Then using the estimated mapping matrix $\widehat{\mathbf{C}}$, we project the observed covariance into the latent space to reconstruct the empirical covariance matrix of the latent variables:

$$\widehat{\mathbf{\Sigma}}_{\mathbf{zz}} = \widehat{\mathbf{C}}^T \widehat{\mathbf{\Sigma}}_{\mathbf{xx}} \widehat{\mathbf{C}} - \widehat{\mathbf{N}}_{\mathbf{x}}(i, i) 1[i = j] \tag{7}$$

where $\widehat{\mathbf{N}}_{\mathbf{x}}$ is estimated from the residual variances, and $\widehat{\mathbf{N}}_{\mathbf{x}}(i,i)1[i=j]$ is the diagonal entries of $\widehat{\mathbf{N}}_{\mathbf{x}}$. With $\widehat{\mathbf{\Sigma}}_{\mathbf{zz}}$ recovered, we apply the Graphical Lasso estimator [Ravikumar et al. (2011)] to obtain the latent precision matrix. We solve the l_1-penalized log-determinant divergence program to yield $\widehat{\mathbf{\Omega}} = \text{solve}(\widehat{\mathbf{\Sigma}}_{\mathbf{zz}}, \lambda_n)$, where λ_n is the regularization parameter.

However, since we are dealing with latent variables, there is some expected uncertainties in the sign of the non-diagonal entries of the covariance matrices. [Long et al. (2026)] proposed a sign-flip equivalence that will still produced a reliable estimator:

Definition 1. *A* sign-flip *matrix is a diagonal matrix $\mathbf{F} \in \mathbb{R}^{p\times p}$ where each entry in its diagonal is either 1 or -1. That is, $F_{ii} \in \{-1,+1\}$ for all $i = 1,\dots,p$. A* sign-flip transform *of matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$ is the product $\mathbf{FAF}$ where $\mathbf{F}$ is a sign-flip matrix. We say that two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p\times p}$ are* sign-flip equivalent*, if $\mathbf{B}$ is a sign-flip transform of $\mathbf{A}$, or equivalently if $\mathbf{A}$ is a sign-flip transform of $\mathbf{B}$.*

We define a *sign-flip* matrix $\widehat{\mathbf{F}}$ as in Definition 1 as follows. For each $i = 1,\dots,p$ we set $\widehat{F}_{ii} = 1$ if $\widehat{\mathbf{c}}_i$ and $\mathbf{c}_i$ point in the same direction, and $\widehat{F}_{ii} = -1$ if $\widehat{\mathbf{c}}_i$ and $\mathbf{c}_i$ point in opposite directions.

In Lemma 5 of [Long et al. (2026)], they provided a bound for $\|\widehat{\mathbf{F}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{F}} - \mathbf{\Sigma}\|_\infty$. Hence we will also provided bound in $\|\widehat{\mathbf{F}}\widehat{\mathbf{D}}\widehat{\mathbf{F}} - \mathbf{D}\|_\infty$ later in our theorem.

3.4.2 Sequential DAG Recovery

With the empirical latent precision matrix $\widehat{\mathbf{\Omega}}$ isolated, we adapt the sequential peeling algorithm [Ghoshal and Honorio (2018)] to recover the internal DAG structure, denoted by the weighted adjacency matrix D. The algorithm relies on two fundamental structural properties of SEM precision matrices.

Proposition 2 (Topological and Edge Weight Recovery, Proposition 2 and 4 in [Ghoshal and Honorio (2018)). *] Let Ω be the true precision matrix of the latent DAG at any valid step of the elimination sequence. For any terminal vertex i and any non-terminal vertex j, the diagonal entries satisfy:*

$$\Omega_{i,i} = \frac{1}{\sigma^2_{z_i}} \quad \textit{and} \quad \Omega_{j,j} = \frac{1}{\sigma^2_{z_j}} + \sum_{l\in\phi(j)} \frac{D^2_{l,j}}{\sigma^2_{zl}} \tag{8}$$

where $\phi(j)$ is the childern of node j. We will later show that under the Finite-Sample Identifiability Condition 2, $\Omega_{i,i} < \Omega_{j,j}$ strictly holds.

The structural edge weight $D_{i,j}$ from any parent $j \in \pi(i)$ to i is uniquely encoded in the i-th row of the precision matrix as:

$$D_{i,j} = -\frac{\Omega_{i,j}}{\Omega_{i,i}} \tag{9}$$

For any node $k \notin \pi(i)$, $\Omega_{i,k} = 0$

This proposition guarantees that we can deterministically identify a terminal vertex node by searching for the minimum diagonal entry in the precision matrix, and extract its edge weights. This is the key result which helps us avoid the faithfulness condition.

3.5 Finite-Sample Identifiability Condition

Next we state our finite sample identifiability condition. This differs from the population version 1 in that we require a gap between the diagonal entries of the precision matrix for terminal and non-terminal vertices. Condition (ii) of the below assumption also restricts how fast the minimum nondiagonal entry of the precision matrix must decay.

Assumption 2. *Let the true internal SEM over $\mathbf{z}$ be $(G, \mathbf{D}, \{\sigma^2_{z_i}\})$. Let $\mathbf{\Omega}_{(m,\tau)}$ denote the true precision matrix over the internal variables $\mathbf{z}_{V[m,\tau]}$ at any valid step of the elimination sequence. Let ϵ be the bound on the empirical covariance error such that the generic solver guarantees an L_∞ estimation error of $A\epsilon$. We assume the following conditions hold:*

(i) ***Finite-Sample Identifiability Condition:*** *To ensure the correct identification of terminal vertices from the estimated matrix, the true diagonal entry of any terminal vertex must be*

strictly smaller than that of any non-terminal vertex by a margin exceeding the worst-case estimation fluctuation.

Formally, $\forall(i,j) \in V[m,\tau] \times V[m,\tau]$, $m \in [p]$, *and* $\forall \tau \in \mathcal{T}_G$, *such that* $\phi_{G[m,\tau]}(i) = \emptyset$ *and* $\phi_{G[m,\tau]}(j) \neq \emptyset$*:*

$$\frac{1}{\sigma_i^2} < \frac{1}{\sigma_j^2} + \sum_{l \in \phi_{G[m,\tau]}(j)} \frac{D_{l,j}^2}{\sigma_l^2} - 2A\epsilon$$

(ii) ***Minimum Signal Strength:*** *To guarantee exact support recovery during the Schur complement updates, the minimum non-zero entry in any valid marginalized true precision matrix must strictly exceed the solver's error tolerance.*

Formally:

$$\min\big\{|(\mathbf{\Omega}_{(m,\tau)})_{j,k}| : (\mathbf{\Omega}_{(m,\tau)})_{j,k} \neq 0,\ (j,k) \in V[m,\tau] \times V[m,\tau], m \in [p], \forall \tau \in \mathcal{T}_G\big\} > A\epsilon$$

To decouple our structure learning algorithm from any specific precision matrix estimator (such as CLIME or Graphical Lasso), we define a broad class of admissible solvers. We assume that the chosen solver, denoted as $\mathrm{solve}(\cdot)$, satisfies standard statistical guarantees for high-dimensional sparse precision matrix estimation.

Assumption 3. *(Sign-Flip Equivalent Generic Estimator Guarantees) Let* $\mathbf{\Sigma}$ *and* $\mathbf{\Omega}$ *be the true covariance and precision matrices of the latent variables. Let* $\widehat{\mathbf{\Sigma}}$ *be an empirical covariance matrix estimated from data. We assume there exists a diagonal sign-flip matrix* $\widehat{\mathbf{F}}$ *(where* $\widehat{F}_{i,i} \in \{-1, +1\}$*) such that the empirical covariance satisfies the concentration bound:*

$$\|\widehat{\mathbf{F}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{F}} - \mathbf{\Sigma}\|_\infty \leq \epsilon$$

Under this condition, we assume a generic estimator $\widehat{\mathbf{\Omega}} = \mathrm{solve}(\widehat{\mathbf{\Sigma}})$ *that requires a tuning parameter (or uniform error bound)* $\epsilon > 0$ *which provides an estimation error bounded up to the same sign-flip equivalence:*

$$\|\widehat{\mathbf{F}}\widehat{\mathbf{\Omega}}\widehat{\mathbf{F}} - \mathbf{\Omega}\|_\infty \leq A\epsilon$$

where $A > 0$ *is a solver-specific constant, and* $\|\cdot\|_\infty$ *denotes the element-wise maximum absolute value norm.*

3.6 Generalized Finite-Sample Recovery Guarantees

In this section we provide the theorotical grarentee of our algorithm for the correct support recovery of $\mathbf{D}$, and the topological ordering for the inernal latent DAG. The following lemma proves the correctness of our algorithm which updates the precision matrix, after removing a terminal vertex.

Lemma 3. *Let* $(G, \mathbf{D}, \{\sigma_{z_i}^2\})$ *be an SEM over* $\mathbf{x}$ *with true latent precision matrix* $\mathbf{\Omega}$ *and true latent covariance matrix* $\mathbf{\Sigma}$*. Given a good estimator of* $\mathbf{\Sigma}$ *that follows Assumption 3, applying the estimator to the empirical covariance matrix yields the estimated precision matrix* $\widehat{\mathbf{\Omega}} = \mathrm{solve}(\widehat{\mathbf{\Sigma}})$ *which satisfies:*

$$\|\widehat{\mathbf{F}}\widehat{\mathbf{\Omega}}\widehat{\mathbf{F}} - \mathbf{\Omega}\|_\infty \leq A\epsilon \text{ (for some constant} A > 0)$$

Let i *be a terminal vertex in* G*. Let* $\mathbf{\Omega}_{(-i)}$ *be the true marginal precision matrix over* X_{-i}*, and let* $\mathbf{\Sigma}_{-i,-i}$ *be the true marginal covariance matrix over* X_{-i}*.*

Under minimal signal strength condition in Assumption 2, if we apply the solver to the empirical sub-covariance matrix to obtain $\widehat{\mathbf{\Omega}}_{(-i)} = \mathrm{solve}(\widehat{\mathbf{\Sigma}}_{-i,-i})$*, then:*

$$\|\widehat{\mathbf{F}}_{(-i)}\widehat{\mathbf{\Omega}}_{(-i)}\widehat{\mathbf{F}}_{(-i)} - \mathbf{\Omega}_{(-i)}\|_\infty \leq A\epsilon \quad \textit{and} \quad \mathcal{S}(\mathbf{\Omega}_{(-i)}) \subseteq \mathcal{S}(\widehat{\mathbf{\Omega}}_{(-i)})$$

Building upon the step-wise guarantees of Lemma 3, we now establish the guarantees for the structural equation model using a generic sparse precision estimator.

Theorem 1. *Let the true data-generating process be defined by the block model over observed variables* $\mathbf{x}$ *and internal DAG variables* $\mathbf{z}$*, where the internal SEM is* $(G, \mathbf{D}, \{\sigma_i^2\})$*. Let* $D_{\max} \triangleq \max\{|D_{i,j}| : (i,j) \in E\}$ *be the maximum true edge weight in* $\mathbf{D}$*, and let* $\sigma^2_{\max}$ *and* $\sigma^2_{\min}$ *be the maximum and minimum true noise variances of the internal DAG variables* $\mathbf{z}$*. Assume we have a estimator of* Σ_{latent} *that follows Assumption 3,*

Under Assumption 2 and lemma 3, the sequential structure learning algorithm deterministically removes the X *variables and subsequently returns an estimator* $\widehat{\mathbf{D}}$ *for the internal DAG such that: 1.* $\|\widehat{\mathbf{F}}\widehat{\mathbf{D}}\widehat{\mathbf{F}} - \mathbf{D}\|_\infty \leq c \cdot A\epsilon(1 + D_{\max})\sigma^2_{\max}$*, 2.* $\mathcal{S}(\mathbf{D}) \subseteq \mathcal{S}(\widehat{\mathbf{D}})$*, 3.* $\mathcal{T}_{\widehat{G}} \subseteq \mathcal{T}_G$*, where* $c \leq \sigma^2_{\min}/(1 - A\epsilon\sigma^2_{\min})$ *is a strictly positive constant. Furthermore, through a thresholding step that filters* $\widehat{\mathbf{D}}$ *with error* ϵ*, we have* $\mathcal{S}(\widehat{\mathbf{D}}) = E$

3.7 Sample Complexity for CLIME

To instantiate our generalized theorem for the CLIME estimator [Tony Cai and Luo (2011)], we must bound the column-wise density of the precision matrices.

The CLIME estimator aims to optimize the following program:

$$\min \|\mathbf{\Omega}\|_1 \textbf{ subject to: } |\mathbf{\Sigma\Omega} - \mathbf{I}|_\infty \leq \lambda_n, \mathbf{\Omega} \in \mathbb{R}^{p \times p} \tag{10}$$

According to the properties of sign flip equivalence in [Long et al. (2026)] Lemma 1, it is showed that for sign-flip equivalent matrices the magnitudes of entries are the same and therefore their l_1 norm $\|\mathbf{F\Omega F}\|_1 = \|\mathbf{\Omega}\|_1$, and for the constraint because the absolute value of every single entry is preserved, the maximum absolute value is also preserved. Therefore $|\mathbf{F\Sigma F F\Omega F} - \mathbf{I}|_\infty = |\mathbf{F\Sigma\Omega F} - \mathbf{I}|_\infty = |\mathbf{F}(\mathbf{\Sigma\Omega} - \mathbf{I})\mathbf{F}|_\infty = |\mathbf{\Sigma\Omega} - \mathbf{I}|_\infty$.

Assumption 4. *(CLIME Regularity) We assume the data follows a sub-Gaussian distribution. Furthermore, for any valid elimination sequence of terminal vertices, the true marginalized precision matrix* $\mathbf{\Omega}_{(-i)}$ *satisfies* $\|\mathbf{\Omega}_{(-i)}\|_{L_1} \leq M$*, where* $\|\cdot\|_{L_1}$ *denotes the maximum absolute column sum.*

Corollary 1. *(Sample Complexity with CLIME) Suppose Assumption 3 holds, and we solve the sub-problems using the CLIME estimator [Tony Cai and Luo (2011)]. Based on the theoretical guarantees of CLIME, the generic solver constants are instantiated as* $\epsilon = \mathcal{O}\left(\sqrt{\frac{\log p}{n}}\right)$ *and* $A = 4M$*.*

Let $\mathcal{K}_{SEM} = \frac{2c(1+D_{\max})\sigma^2_{\max}}{|D_{\min}|}$*, where* $D_{\min} = \min i, j|D_{i,j}|$*, be a structural constant of the underlying SEM. If the sample size satisfies:*

$$n = \mathcal{O}\left(M^2 \cdot \mathcal{K}^2_{SEM} \log p\right) \tag{11}$$

then with high probability, thresholding the estimated DAG parameter matrix $\widehat{\mathbf{D}}$ *at the level* $|D_{\min}|/2$ *yields exact structural recovery:* $\mathcal{S}(\widehat{\mathbf{D}}) = \mathcal{S}(\mathbf{D})$*.*

3.8 Sample Complexity for Graphical Lasso

To instantiate our generalized theorem for the l_1-penalized log-determinant divergence (Graphical Lasso), we require structural assumptions on the Hessian of the objective function, $\mathbf{\Gamma}^* = \mathbf{\Omega}^{-1} \otimes \mathbf{\Omega}^{-1}$[ ]Ravikumar et al. (2011)]. Let S index the non-zero off-diagonal entries of the true precision matrix, and S^c index the zero entries [Ravikumar et al. (2011)].

Assumption 5. *(Graphical Lasso Irrepresentability) We assume the data follows a sub-Gaussian distribution. Let* $K_{\mathbf{\Gamma}^*} = \|(\mathbf{\Gamma}^*_{SS})^{-1}\|_\infty$*. Furthermore, we assume the mutual incoherence condition holds for the Hessian of any marginalized true precision matrix: there exists an* $\alpha \in (0, 1]$ *such that* $\|\mathbf{\Gamma}^*_{S^cS}(\mathbf{\Gamma}^*_{SS})^{-1}\|_\infty \leq (1 - \alpha)$ *[Ravikumar et al. (2011)].*

It is shown in [Long et al. (2026)] that the optimization in [Ravikumar et al. (2011)] is preserved under sign-flip equivalence.

Corollary 2. *(Sample Complexity with Graphical Lasso) Suppose* **??** *holds under Assumption 3, and we solve the sub-problems using the* l_1*-penalized log-determinant estimator. Based on the theoretical*

guarantees of this estimator, the generic solver constants are instantiated as $\epsilon = \mathcal{O}\left(\sqrt{\frac{\log p}{n}}\right)$ *and* $A = 2(1 + 8\alpha^{-1})K_{\Gamma^*}$ *[Ravikumar et al. (2011)].*

Let d *be the maximum node degree, and let* $\mathcal{K}_{SEM} = \frac{2c(1+D_{\max})\sigma^2_{\max}}{|D_{\min}|}$*. If the sample size satisfies:*

$$n = \mathcal{O}\left(d^2(1+\alpha^{-1})^2 K^2_{\Gamma^*} \cdot \mathcal{K}^2_{SEM} \log p\right) \tag{12}$$

then with high probability, thresholding the estimated DAG parameter matrix $\widehat{\mathbf{D}}$ *at the level* $|D_{\min}|/2$ *yields exact structural recovery:* $\mathcal{S}(\widehat{\mathbf{D}}) = \mathcal{S}(\mathbf{D})$.

4 Experimental Results

In this section, we conduct experiments to validate our theoretical sample complexity bounds and demonstrate the scalability of the proposed algorithm in high-dimensional settings. We also show the unique advantages of our method in recovering directed causal edges compared to other baselines. We also ran experiment on real world cocaine addition study which is show in the appendix.

Sample Complexity and Scalability. To rigorously test our framework, we generate synthetic observational data drawn from pure latent high-dimensional SEMs. We parameterize the environment by the number of latent variables p, the dimensionality of observed variables per latent block q_i, and the sample size n. The underlying latent structure is generated as a DAG with a maximum degree of $k = 3$. To ensure data generation aligns with our theoretical assumptions regarding sparsity, the measurement model is constructed such that each latent variable maps to exactly $s = 3$ active indicators, while the remaining $q_i - s$ variables are pure sub-Gaussian noise ($\sigma^2 = 0.1$). We control the sample size using a multiplier c_{sample}, defining the sample size $n = \lceil c_{sample}(\log p + \log q_i)\rceil$.

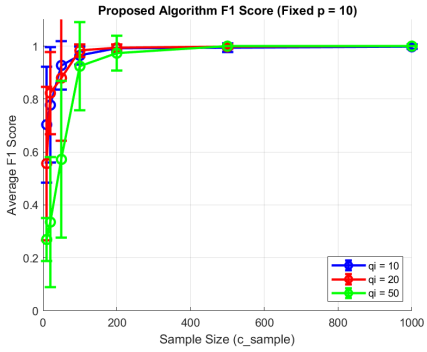
For the proposed algorithm, the regularization hyperparameters must satisfy the established finite-sample concentration bounds. Consequently, we set the precision matrix penalty parameter $\lambda \propto \sqrt{(\log p)/n}$ for the chosen solver [Ravikumar et al. (2011)] and the covariance thresholding parameter $\rho \propto \sqrt{(\log q_i)/n}$ for the sparse PCA program [Lei and Vu (2015)] in estimating $\mathbf{C}$. Edges recovered with absolute weights below a threshold of 0.2 are pruned to suppress spurious numerical noise.

To evaluate the scalability and convergence of our proposed algorithm, we compute the Average F1 Score of the recovered structural DAG over 10 independent repetitions across varying dimensions and sample sizes.

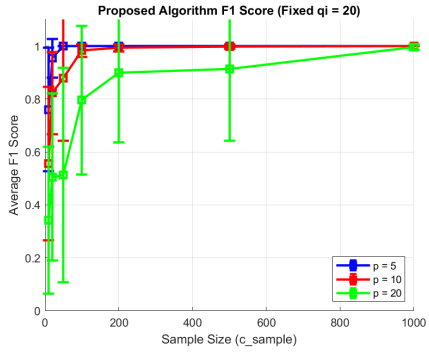
Figure 1a illustrates the performance for a fixed latent dimension ($p = 10$) while varying the observed dimensionality ($q_i \in \{10, 20, 50\}$). Because the number of active indicators is fixed ($s = 3$), increasing q_i significantly increases the sparsity of the $\mathbf{C}$ matrix, flooding the environment with pure noise variables. Despite this challenging signal-to-noise ratio, the algorithm rapidly converges to perfect structural recovery (F1 = 1.0) across all configurations. Notably, the algorithm requires only a marginal increase in c_{sample} to perfectly filter out the massive influx of noise variables when $q_i = 50$, demonstrating exceptional robustness to high-dimensional sparsity.

Figure 1b shows the algorithm's scalability for a fixed dimension of observed variable ($q_i = 20$) while varying the size of the latent DAG ($p \in \{5, 10, 20\}$). As the complexity of the internal causal structure grows, the required sample size gracefully scales to maintain high accuracy. The algorithm consistently achieves an F1 score of 1.0 for all tested dimensions given sufficient samples. These results empirically validate our theoretical guarantee that structural recovery is strictly bounded by a sample complexity of $\mathcal{O}(\sqrt{(\log p)/n})$.

Baseline Comparison and Structural Recovery. This section evaluates the structural recovery performance of the proposed algorithm against baselines algorithms Mimbuild [Silva et al. (2006)], the Copula Factor Model [Cai et al. (2025)], and the Generalized Independent Noise (GIN) [Xie et al. (2024a)] condition—across varying sample sizes. To isolate the structural learning capabilities, all algorithms were provided with the true measurement clusters. The evaluation is split into three metrics: directed edge recovery, causal ordering accuracy, and undirected skeleton recovery in the appendix. Error bars represent the 2 standard divination over 10 independent repetitions.

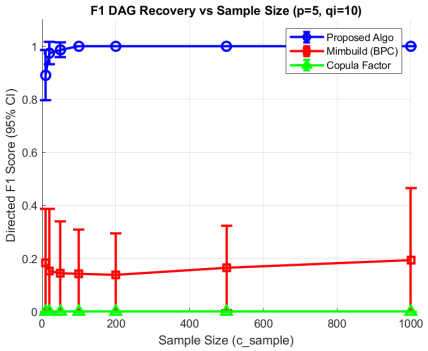


(a) The average F1 score for the support recovery of $\mathbf{D}$ with varying number of dimension q_i in observed variable $\mathbf{x}$

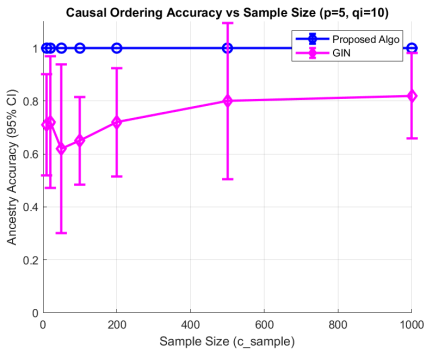


(b) The average F1 score for the support recovery of $\mathbf{D}$ with varying number of latent variable p within the internal latent DAG

Figure 1



(a) The average F1 score for the support recovery of $\mathbf{D}$ compared to baselines Mimbuild and Copula Factor Model



(b) The average ancestry accuracy for the internal latent DAG $\mathbf{D}$ compared to GIN

Figure 2

Directed Edge Recovery (F1 DAG). The F1 DAG graph 2a exposes the critical vulnerability of traditional causal discovery algorithms: the inability to break Markov equivalence. Our method maintain a flawless Directed F1 score of 1.0. Mimbuild largely fails to orient the internal edges. Its F1 score remains at approximately 0.15 to 0.20. Crucially, it proves that simply providing more data does not overcome of Markov equivalence. Copula Factor Model flatlines at 0. Because this Bayesian approach strictly estimates an undirected CPDAG, it is fundamentally incapable of assigning causal arrows in this setting, resulting in a total failure on the directed metric.

Causal Ordering (Ancestry Accuracy). The Ordering graph 2b compares the proposed method against GIN in recovering the true topological sequence of the latent variables. Our method captures the perfect topological sequence immediately, achieving a 1.0 ancestry accuracy with zero variance. While GIN: exhibits severe instability. While its average accuracy trends upward with larger sample sizes. At smaller sample sizes, GIN's performance is highly erratic, occasionally dropping near random chance, making it an unreliable structural solver for these dimensionalities.

Concluding Remarks. In this paper, we introduced a solver-agnostic theoretical framework for learning pure latent high-dimensional Structural Equation Models, proving that exact DAG recovery is achievable with logarithmic sample complexity. Future research will focus on establishing theoretical minimax lower bounds to confirm algorithmic optimality, alongside relaxing the pure measurement assumption to accommodate overlapping latent children. Furthermore, extending this static framework to handle time-varying latent causal structures for future applicability.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for [assist in the writing of proofs,]. We have not used generative AI tools for [Generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods], and [assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results] are not applicable to this work. Additionally, we used generative AI tools for [tasks with recommended disclosure]. We have reviewed all AI-assisted work. [Elaborate. For example, "we checked LLM-generated research ideas for potential plagiarism through a manual literature survey", "LLM-generated code was verified and tested for correctness by 2 authors", etc.]. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Ming Cai, Penggang Gao, and Hisayuki Hara. Causal discovery for linear DAGs with dependent latent variables via higher-order cumulants. *CoRR*, abs/2510.14780, 2025. doi: 10.48550/ARXIV.2510.14780. URL https://doi.org/10.48550/arXiv.2510.14780.

Ruifei Cui, Perry Groot, Moritz Schauer, and Tom Heskes. Learning the causal structure of copula models with latent variables. In Amir Globerson and Ricardo Silva, editors, *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 188–197. AUAI Press, 2018. URL http://auai.org/uai2018/proceedings/papers/65.pdf.

Asish Ghoshal and Jean Honorio. Learning linear structural equation models in polynomial time and sample complexity. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1466–1475. PMLR, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/ghoshal18a.html.

Biwei Huang, Charles Low, Feng Xie, Clark Glymour, and Kun Zhang. Latent hierarchical causal structure discovery with rank constraints. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=lIeuKiTZsLY.

Erich Kummerfeld and Joseph Ramsey. Causal clustering for 1-factor measurement models. volume 2016, pages 1655–1664, 08 2016. doi: 10.1145/2939672.2939838.

Jing Lei and Vincent Q. Vu. Sparsistency and agnostic inference in sparse PCA. *The Annals of Statistics*, 43(1):299 – 322, 2015.

Haichi Long, Qifan Song, and Jean Honorio. Provable guarantees for estimating covariances between latent variables with application to precision matrix estimation. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026. URL https://openreview.net/forum?id=LfVHb2WWIB.

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. URL https://openreview.net/forum?id=ZdSe1qnuia.

P. Ravikumar, M. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Ricardo Silva, Richard Scheine, 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.

Weidong Liu Tony Cai and Xi Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

Feng Xie, Biwei Huang, Zhengming Chen, Yangbo He, Zhi Geng, and Kun Zhang. Identification of linear non-Gaussian latent hierarchical structure. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 24370–24387. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/xie22a.html.

Feng Xie, Biwei Huang, Zhengming Chen, Ruichu Cai, Clark Glymour, Zhi Geng, and Kun Zhang. Generalized independent noise condition for estimating causal structure with latent variables. *Journal of Machine Learning Research*, 25(191):1–61, 2024a. URL http://jmlr.org/papers/v25/23-1052.html.

Feng Xie, Biwei Huang, Zhengming Chen, Ruichu Cai, Clark Glymour, Zhi Geng, and Kun Zhang. Generalized independent noise condition for estimating causal structure with latent variables. *Journal of Machine Learning Research*, 25(191):1–61, 2024b. URL http://jmlr.org/papers/v25/23-1052.html.

A Detailed Proofs

A.1 Proof of Lemma 1

Proof. Let $\mathbf{\Omega} = (\mathbf{I}-\mathbf{D})^T\mathbf{N}_\mathbf{z}^{-1}(\mathbf{I}-\mathbf{D})$ be the true internal precision matrix SEM. The joint precision matrix is thus partitioned as:

$$\mathbf{\Omega}_{full} = \begin{bmatrix} \mathbf{\Omega_{xx}} & \mathbf{\Omega_{xz}} \\ \mathbf{\Omega_{zx}} & \mathbf{\Omega_{zz}} \end{bmatrix} = \begin{bmatrix} \mathbf{N}_X^{-1} & -\mathbf{N}_X^{-1}\mathbf{C} \\ -\mathbf{C}^T\mathbf{N}_X^{-1} & \mathbf{C}^T\mathbf{N}_X^{-1}\mathbf{C}+\mathbf{\Omega} \end{bmatrix}$$

When the algorithm completely marginalizes out the $\mathbf{x}$ block, the updated precision matrix for the remaining variables is the Schur complement $\mathbf{\Omega}_{(-\mathbf{x})} = \mathbf{\Omega_{zz}} - \mathbf{\Omega_{zx}}\mathbf{\Omega_{xx}^{-1}}\mathbf{\Omega_{xz}}$. Substituting the blocks derived above:

$$\begin{aligned} \mathbf{\Omega}_{(-\mathbf{x})} &= \left[\mathbf{C}^T\mathbf{N}_X^{-1}\mathbf{C}+\mathbf{\Omega}\right] - \left(-\mathbf{C}^T\mathbf{N}_X^{-1}\right)\left(\mathbf{N}_X^{-1}\right)^{-1}\left(-\mathbf{N}_X^{-1}\mathbf{C}\right) \\ &= \left[\mathbf{C}^T\mathbf{N}_X^{-1}\mathbf{C}+\mathbf{\Omega}\right] - \left(\mathbf{C}^T\mathbf{N}_X^{-1}\mathbf{N}_X\mathbf{N}_X^{-1}\mathbf{C}\right) \\ &= \mathbf{C}^T\mathbf{N}_X^{-1}\mathbf{C}+\mathbf{\Omega}-\mathbf{C}^T\mathbf{N}_X^{-1}\mathbf{C} \\ &= \mathbf{\Omega} \end{aligned}$$

The structural mapping matrix from latent to observed $\mathbf{C}$ perfectly cancels out, leaving the true precision matrix of the internal DAG. This concludes the proof. □

A.2 Proof of Lemma 2

Proof. By isolating out the observed $\mathbf{x}$ in lemma 1, the global identifiability relies entirely on the identifiability of $\mathbf{\Omega} = (\mathbf{I}-\mathbf{D})^T\mathbf{N}_\mathbf{z}^{-1}(\mathbf{I}-\mathbf{D})$.

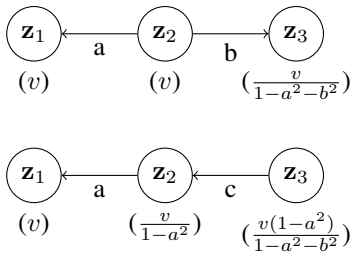


Figure 3

Consider the following two SEMs as shown in Figure 3 over three nodes, where the noise variances are shown within braces below each node, and the edge weights are shown on the edges. Let $\delta = 1/\sigma_x^2$, $a^2+b^2<1$, $c=\frac{b}{1-a^2}$. we have Both the SEMs make the following conditional independence assertion: $\mathbf{z}_1 \perp\!\!\!\perp \mathbf{z}_3|\mathbf{z}_2$, and the exact same $\mathbf{z}$-block precision matrix, denoted $\mathbf{\Omega_{zz}}$:

$$\mathbf{\Omega_{zz}} = \begin{bmatrix} \delta+\frac{1}{v} & -\frac{a}{v} & 0 \\ -\frac{a}{v} & \delta+\frac{1}{v} & -\frac{b}{v} \\ 0 & -\frac{b}{v} & \delta+\frac{1}{v} \end{bmatrix}$$

Let $\Delta = \delta+\frac{1}{v}$, Assumption 1 requires that the true terminal vertex has a strictly smaller diagonal entry than any non-terminal vertex, In the first graph, $\mathbf{z}_2$ is a parent, but its diagonal is Δ, perfectly tied with terminal vertex $\mathbf{z}_1$ and $\mathbf{z}_3$. In the second graph, $\mathbf{z}_3$ is the root, but its diagonal is Δ, perfectly tied with the terminal vertex $\mathbf{z}_1$. Because of the tie, the algorithm cannot distinguish the terminal vertex from the parents.

If we remove the terminal vertex $\mathbf{z}_1$, We calculate the Schur complement of $\mathbf{\Omega_{zz}}$ with respect to node $\mathbf{z}_1$. Lets call the reduced matrix $\mathbf{\Omega}_{-i}$, $\mathbf{\Omega}_{-i}$ has diagonal:

$$\begin{aligned} (\mathbf{\Omega}_{-i})_{2,2} &= \Delta - \frac{a^2}{v^2\Delta} \\ (\mathbf{\Omega}_{-i})_{3,3} &= \Delta \end{aligned}$$

Because $\frac{a^2}{v^2\Delta}$ is a positive number, thus $(\mathbf{\Omega}_{-i})_{2,2} < (\mathbf{\Omega}_{-i})_{3,3}$.

Since the true DAG is constructed using $p/3$ independent, disjoint triplets of this form, and each triplet possesses 2 indistinguishable configurations that preserve the global $\mathbf{\Omega}$ matrix, there are $2^{p/3}$ total valid graph structures. Since $2^{p/3} = 2^{\Theta(p)}$, an exponential number of equivalent DAGs exist when Assumption 1 is violated, concluding the proof. □

A.3 Proof of Lemma 3

Proof. By definition, the empirical covariance matrix for the remaining variables, $\widehat{\mathbf{\Sigma}}_{-i,-i}$, is a principal submatrix of the full empirical covariance matrix $\widehat{\mathbf{\Sigma}}$. Similarly, $\mathbf{\Sigma}_{-i,-i}$ is a principal submatrix of the full true covariance matrix $\mathbf{\Sigma}$.

Because the element-wise infinity norm $\|\cdot\|_\infty$ measures the maximum absolute value among all entries in a matrix, the maximum error in any submatrix cannot exceed the maximum error of the full matrix from which it is derived. the sign-flip equivalence bound strictly holds for the sub-matrix:

$$\|\widehat{\mathbf{F}}_{-i}\widehat{\mathbf{\Sigma}}_{-i,-i}\widehat{\mathbf{F}}_{-i} - \mathbf{\Sigma}_{-i,-i}\|_\infty = \max_{j\neq i, k\neq i} |\widehat{\mathbf{F}}_{j,k}\widehat{\Sigma}_{j,k}\widehat{\mathbf{F}}_{j,k} - \Sigma_{j,k}| \leq \max_{j,k} |\widehat{\mathbf{F}}_{j,k}\widehat{\Sigma}_{j,k}\widehat{\mathbf{F}}_{j,k} - \Sigma_{j,k}| = \|\widehat{\mathbf{F}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{F}} - \mathbf{\Sigma}\|_\infty$$

Since $\|\widehat{\mathbf{F}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{F}} - \mathbf{\Sigma}\|_\infty \leq \epsilon$, it directly follows that:

$$\|\widehat{\mathbf{F}}_{-i}\widehat{\mathbf{\Sigma}}_{-i,-i}\widehat{\mathbf{F}}_{-i} - \mathbf{\Sigma}_{-i,-i}\|_\infty \leq \epsilon$$

The marginal distribution over the subset X_{-i} possesses the true covariance matrix $\mathbf{\Sigma}_{-i,-i}$ and the true precision matrix $\mathbf{\Omega}_{(-i)}$.

We have established that the input to the solver, $\widehat{\mathbf{\Sigma}}_{-i,-i}$, satisfies the required error bound condition ($\leq \epsilon$) with respect to its true target $\mathbf{\Sigma}_{-i,-i}$. Because this sub-problem meets the strict condition of the generic solver's "contract," the solver's theoretical guarantees automatically activate for the output $\widehat{\mathbf{\Omega}}_{(-i)} = \text{solve}(\widehat{\mathbf{\Sigma}}_{-i,-i})$. Therefore,

$$\|\widehat{\mathbf{F}}_{-i}\widehat{\mathbf{\Omega}}_{(-i)}\widehat{\mathbf{F}}_{-i} - \mathbf{\Omega}_{(-i)}\|_\infty \leq A\epsilon$$

To prove support containment ($\mathcal{S}(\mathbf{\Omega}_{(-i)}) \subseteq \mathcal{S}(\widehat{\mathbf{\Omega}}_{(-i)})$), let (j,k) be any edge in the true support of the marginal precision matrix, meaning $(j,k) \in \mathcal{S}(\Omega_{(-i)})$. Under the Assumption 2 the magnitude of every non-zero entry in the true precision matrix is strictly bounded away from zero by a factor greater than the estimation error:

$$|\Omega_{(-i),j,k}| > A\epsilon$$

We define the estimated entry error as $\delta_{j,k}$. We know $|\delta_{j,k}| \leq A\epsilon$ from above. The sign-flipped estimated entry is related to the true entry by:

$$\widehat{F}_{(-i),j,j}\widehat{\Omega}_{(-i),j,k}\widehat{F}_{(-i),k,k} = \Omega_{(-i),j,k} + \delta_{j,k}$$

Because the true signal $|\Omega_{(-i),j,k}|$ is strictly greater than the maximum possible error $|\delta_{j,k}|$, the right-hand side of the equation cannot be zero. Consequently, the left-hand side cannot be zero. Because the diagonal entries of $\widehat{F}_{(-i)}$ are strictly ± 1, they cannot force the product to zero. Thus, it must be that:

$$\widehat{\Omega}_{(-i),j,k} \neq 0$$

Since every non-zero entry in the true precision matrix results in a strictly non-zero entry in the estimated precision matrix, Therefore:

$$\mathcal{S}(\mathbf{\Omega}_{(-i)}) \subseteq \mathcal{S}(\widehat{\mathbf{\Omega}}_{-i})$$

□

A.4 PROOF OF THEOREM 1

Proof. The proof proceeds by induction over the elimination steps of the algorithm.

As demonstrated in Lemma 1, the uncertainty lies within the bottom-right block of the full precision matrix $\mathbf{\Omega}_{full}$. The Schur complement update isolates the empirical estimate of the internal precision matrix $\widehat{\mathbf{\Omega}}$.

We assume that at step t, the working empirical covariance matrix satisfies the L_∞ error bound ϵ. By Lemma 3, the generic solver outputs an estimated precision matrix for the remaining variables satisfying $\|\widehat{\mathbf{F}}_{(-i)}\widehat{\mathbf{\Omega}}_{(-i)}\widehat{\mathbf{F}}_{(-i)} - \mathbf{\Omega}_{(-i)}\|_\infty \leq A\epsilon$.

Because the Assumption 2guarantees a gap greater than $2A\epsilon$ between true terminal and non-terminal diagonal entries, the element-wise estimation error ($\pm A\epsilon$) cannot alter the relative ordering of the minimum diagonal entry. Thus, the algorithm correctly identifies a true terminal vertex at every step, preserving the topological ordering ($\mathcal{T}_{\widehat{G}} \subseteq \mathcal{T}_G$).

Upon identifying a terminal vertex i, the parents are identified from the non-zero entries of the i-th row. By Lemma 3 since $\mathcal{S}(\mathbf{\Omega}_{(-i)}) \subseteq \mathcal{S}(\widehat{\mathbf{\Omega}}_{(-i)})$, no true parents are missed, ensuring $\mathcal{S}(\mathbf{D}) \subseteq \mathcal{S}(\widehat{\mathbf{D}})$.

Finally, the edge weights are estimated via $\widehat{D}_{i,j} = -\frac{\widehat{\Omega}_{i,j}}{\widehat{\Omega}_{i,i}}$, and the terminal vertex $\Omega_{i,i} = 1/\sigma_{z_i}^2$ as shown in Proposition 2. The absolute error of this ratio is bounded by applying the L_∞ error bound $A\epsilon$ to both the numerator and the denominator. Let $\widetilde{\mathbf{\Omega}} = \widehat{\mathbf{F}}\widehat{\mathbf{\Omega}}\widehat{\mathbf{F}}$ be the sign-flipped empirical precision matrix. By Assumption 3, we know that the element-wise error is bounded by $A\varepsilon$. Therefore, we can define the entry-wise error $\delta_{i,j}$ such that:

$$\widetilde{\Omega}_{i,j} = \Omega_{i,j} + \delta_{i,j} \quad \text{where} \quad |\delta_{i,j}| \leq A\varepsilon$$

Crucially, because $\widehat{F}_{i,i} \in \{-1, 1\}$, its square is 1. This preserves the diagonal entries exactly:

$$\widetilde{\Omega}_{i,i} = \widehat{F}_{i,i}\widehat{\Omega}_{i,i}\widehat{F}_{i,i} = \widehat{F}_{i,i}^2\widehat{\Omega}_{i,i} = \widehat{\Omega}_{i,i}$$

We define the sign-flipped estimated edge weight as $\widetilde{D}_{i,j} = \widehat{F}_{i,i}\widehat{D}_{i,j}\widehat{F}_{j,j}$. Substituting the empirical calculation $\widehat{D}_{i,j} = -\frac{\widehat{\Omega}_{i,j}}{\widehat{\Omega}_{i,i}}$, we obtain:

$$\begin{aligned}
\widetilde{D}_{i,j} &= \widehat{F}_{i,i}\left(-\frac{\widehat{\Omega}_{i,j}}{\widehat{\Omega}_{i,i}}\right)\widehat{F}_{j,j} \\
&= -\frac{\widehat{F}_{i,i}\widehat{\Omega}_{i,j}\widehat{F}_{j,j}}{\widehat{\Omega}_{i,i}} \\
&= -\frac{\widetilde{\Omega}_{i,j}}{\widetilde{\Omega}_{i,i}}
\end{aligned}$$

Because the sign-flipped edge weight is exactly equal to the ratio of the sign-flipped precision matrix entries, we can directly bound the error relative to the true edge weight $D_{i,j} = -\frac{\Omega_{i,j}}{\Omega_{i,i}}$. The absolute

error is strictly bounded as follows:

$$\begin{aligned}
|\widehat{F}_{i,i}\widehat{D}_{i,j}\widehat{F}_{j,j} - D_{i,j}| &= |\widetilde{D}_{i,j} - D_{i,j}| \\
&= \left| \frac{\widetilde{\Omega}_{i,j}}{\widetilde{\Omega}_{i,i}} - \frac{\Omega_{i,j}}{\Omega_{i,i}} \right| \\
&= \left| \frac{\Omega_{i,j} + \delta_{i,j}}{\Omega_{i,i} + \delta_{i,i}} - \frac{\Omega_{i,j}}{\Omega_{i,i}} \right| \\
&= \left| \frac{(\Omega_{i,j} + \delta_{i,j})\Omega_{i,i} - (\Omega_{i,i} + \delta_{i,i})\Omega_{i,j}}{(\Omega_{i,i} + \delta_{i,i})\Omega_{i,i}} \right| \\
&= \left| \frac{\delta_{i,j}\Omega_{i,i} - \delta_{i,i}\Omega_{i,j}}{(\Omega_{i,i} + \delta_{i,i})\Omega_{i,i}} \right| \\
&\leq \frac{|\delta_{i,j}||\Omega_{i,i}| + |\delta_{i,i}||\Omega_{i,j}|}{|\Omega_{i,i} + \delta_{i,i}||\Omega_{i,i}|} \\
&= \frac{|\delta_{i,j}| + |\delta_{i,i}|\frac{|\Omega_{i,j}|}{|\Omega_{i,i}|}}{|\Omega_{i,i} + \delta_{i,i}|}
\end{aligned}$$

We substitute $|D_{i,j}| = |\Omega_{i,j}|/|\Omega_{i,i}|$. Because i is a terminal vertex, its true diagonal entry is strictly its base noise precision, $\Omega_{i,i} = 1/\sigma_{zi}^2$. Applying the reverse triangle inequality to the denominator ($|x + y| \geq |x| - |y|$) and bounding the error terms by $A\varepsilon$:

$$\begin{aligned}
&\leq \frac{A\varepsilon + A\varepsilon|D_{i,j}|}{|1/\sigma_{zi}^2| - A\varepsilon} \\
&= \frac{A\varepsilon(1 + |D_{i,j}|)}{1/\sigma_{zi}^2 - A\varepsilon} \\
&= \frac{A\varepsilon\sigma_{zi}^2(1 + |D_{i,j}|)}{1 - A\varepsilon\sigma_{zi}^2} \\
&\leq c \cdot A\varepsilon(1 + |D_{\max}|)\sigma_{\max}^2
\end{aligned}$$

where $c \leq \frac{1}{1 - A\varepsilon\sigma_{\max}^2}$, which concludes the proof. □

A.5 Proof of Corollary 1

Proof. Under sub-Gaussian noise, the empirical covariance matrix satisfies $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_\infty \leq \epsilon$ with high probability when $\epsilon = c_0\sqrt{\log p/n}$. The CLIME estimator satisfies Assumption 3 with $A = 4M$. Substituting these into the error bound of Theorem 1 yields an element-wise parameter error bounded by $c \cdot 4M\epsilon(1 + D_{\max})\sigma_{\max}^2$. To achieve exact support recovery via thresholding, this maximum estimation error must be strictly less than half of the minimum true signal ($|D_{\min}|/2$). Setting the bound strictly less than $|D_{\min}|/2$ and solving for n yields the required sample complexity $n = \mathcal{O}(M^2\mathcal{K}_{SEM}^2 \log p)$. □

A.6 Proof of Corollary 2

Proof. Under sub-Gaussian noise, the empirical covariance matrix achieves the concentration bound $\epsilon = \mathcal{O}(\sqrt{\log p/n})$ [Ravikumar et al. (2011)]. The Graphical Lasso satisfies Assumption 3 with $A = 2(1 + 8\alpha^{-1})K_{\Gamma^*}$ [Ravikumar et al. (2011)], provided the sample size is sufficiently large to maintain strict dual feasibility. This base feasibility requires $n = \mathcal{O}(d^2(1 + \alpha^{-1})^2 \log p)$ [Ravikumar et al. (2011)].

Substituting A and ϵ into the error bound of Theorem 1 yields an element-wise parameter error bounded by $c \cdot 2(1 + 8\alpha^{-1})K_{\Gamma^*}\epsilon(1 + D_{\max})\sigma_{\max}^2$. To achieve exact support recovery, this error must be bounded by $|D_{\min}|/2$. Constraining the error and solving for n yields $n = \mathcal{O}(K_{\Gamma^*}^2(1 + \alpha^{-1})^2\mathcal{K}_{SEM}^2 \log p)$. Combining this with the baseline dual-feasibility requirement yields the final sample complexity.

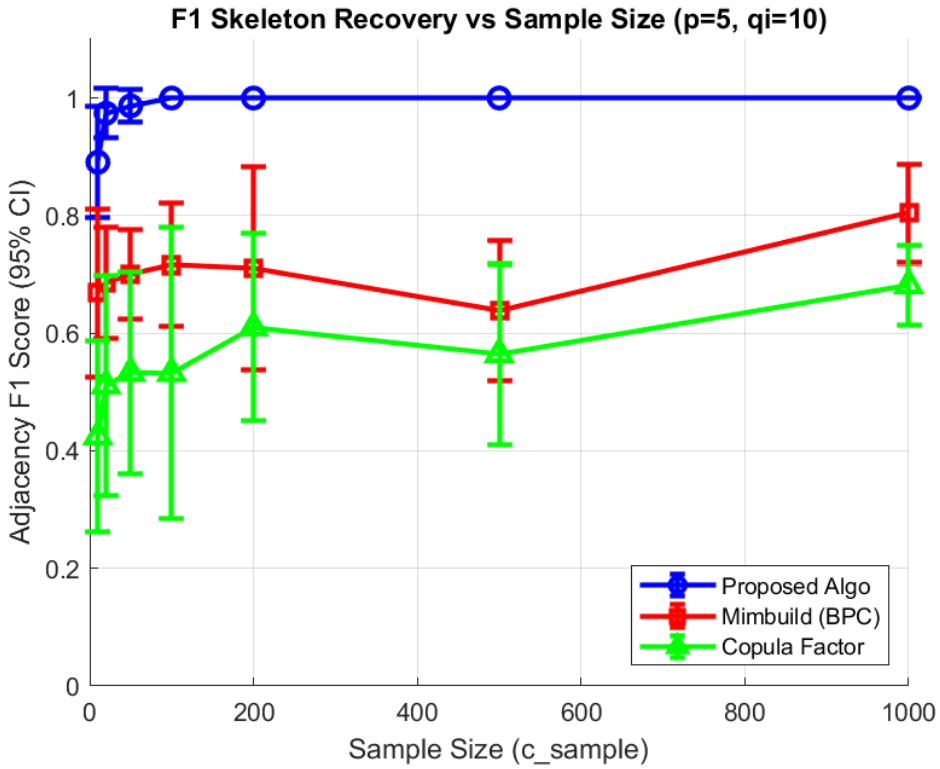


Figure 4: The average F1 score for the recovered undirected representation $\mathbf{D}$ compared to baselines Mimbuild and Copula Factor Model

□

B Extra Experiment

Undirected Skeleton Recovery (Adjacency F1). The core theoretical contribution of this paper is the exact, fully directed recovery of the latent structural DAG. Copula Factor Model and Mimbuild only recover up CPDAG equivalence classes. Measuring in CPDAG equivalence classes would be unfair to our algorithm. But we still carried out experiments measures in CPDAG equivalence classes hence measures the recovery of the undirected skeleton.

The F1 Skeleton graph 4 isolates the algorithms' ability to identify the presence of a causal connection, regardless of its direction. Our method demonstrates rapid convergence, achieving a perfect Adjacency F1 score of 1.0 and maintaining absolute stability as sample size increases. While Mimbuild: Exhibits moderate success in identifying the skeleton. However, the performance plateau across the entire sample size range indicate persistent instability. Copula Factor Model consistently underperforms Mimbuild, starting near an F1 of 0.40 and slowly climbing to approximately 0.65. The high variance further underscores its struggle to cleanly extract the skeleton in high-dimensional noise.

Real World Dataset We run experiment on a cocaine addition study, by runing our algorithm throught th fmri data and interpreted the recoverd $\mathbf{D}$, we learned that In healthy control subjects, inhibitory control affects cue reactivity. In cocaine addicted individuals, cue reactivity affects inhibitory control. The directionality of these edges are meaningful from the neuroscientific viewpoint. Healthy individuals control their behavior, and thus a higher inhibitory control produces a lower cue reactivity. Cocaine addicted subjects do not control their behavior well and react impulsively (e.g., cravings). In that case, we expect that a higher cue reactivity induces a lower inhibitory control.