Learning linear structural equation models in polynomial time and sample complexity

Asish Ghoshal and Jean Honorio
Department of Computer Science
Purdue University
West Lafayette, IN - 47907
{aghoshal, jhonorio}@purdue.edu

The problem of learning structural equation models (SEMs) from data is a fundamental problem in causal inference. We develop a new algorithm — which is computationally and statistically efficient and works in the high-dimensional regime — for learning linear SEMs from purely observational data with arbitrary noise distribution. We consider three aspects of the problem: identifiability, computational efficiency, and statistical efficiency. We show that when data is generated from a linear SEM over p nodes and maximum degree d, our algorithm recovers the directed acyclic graph (DAG) structure of the SEM under an *identifiability condition* that is more general than those considered in the literature, and without *faithfulness* assumptions. In the population setting, our algorithm recovers the DAG structure in $\mathcal{O}(p(d^2 + \log p))$ operations. In the finite sample setting, if the estimated precision matrix is sparse, our algorithm has a smoothed complexity of $\widetilde{\mathcal{O}}\left(p^3 + pd^7\right)$, while if the estimated precision matrix is dense, our algorithm has a smoothed complexity of $\widetilde{\mathcal{O}}\left(p^5\right)$. For sub-Gaussian noise, we show that our algorithm has a sample complexity of $\mathcal{O}(\frac{d^8}{\varepsilon^2}\log(\frac{p}{\sqrt{\delta}}))$ to achieve ε element-wise additive error with respect to the true *autoregression matrix* with probability at most $1-\delta$, while for noise with bounded $(4m)$-th moment, with m being a positive integer, our algorithm has a sample complexity of $\mathcal{O}(\frac{d^8}{\varepsilon^2}(\frac{p^2}{\delta})^{1/m})$.

1. Introduction

Motivation. Elucidating causal relationship between different entities or variables is a fundamental task in various scientific disciplines such as finance, genetics, medicine, neuroscience, artificial intelligence, among others. Learning cause-effect relationships from purely observational data is often the only recourse available in situations where performing randomized experiments or interventions can be expensive, impractical, unethical, or downright impossible. For continuous-valued variables, structural equation models (SEMs) is a commonly employed formalism for performing causal inference. Conditions under which SEMs can be uniquely identified from observational data have been recently characterized. Unfortunately, for linear SEMs, identifiability conditions have been rather limited, and existing structure learning algorithms are inefficient. In this paper, we consider the problem of learning linear SEMs over p variables and bounded-degree d, from purely observational data, with arbitrary noise distributions having bounded second moment — including but not limited to the Gaussian distribution.

We generalize existing identifiability conditions for learning linear SEMs, and present computationally and statistically efficient algorithms for learning the structure of linear SEMs when identifiable. The paper makes the following contributions.

Our Contribution. We present a new identifiability condition for learning linear SEMs from observational data that generalizes the homoscedastic Gaussian noise (equal noise variance) case considered by [16]. Our algorithm also works for the case when the noise variances are known up to a constant factor — a sufficient condition under which linear SEMs are identifiable as shown by [14]. This disproves an earlier conjecture by [14] that "variance scaling or non-Gaussianity is necessary in order to guarantee identifiability" of linear SEMs. Moreover, we show that our identifiability condition is in general necessary for ensuring identifiability of linear SEMs, in the sense that there exist an exponential number of DAGs, which under uncountably many autogression matrices and noise variances, induce the same covariance and precision matrix, and specify distributions that have the same conditional independence structures.

Our method is fully non-parametric, works for both Gaussian and non-Gaussian noise, and, to the best of our knowledge, the most efficient algorithm available for learning linear SEMs with provable guarantees. Given the inverse covariance (or precision) matrix, our method, which resembles a Cholesky factorization, can recover the structure and parameters of the SEM exactly in $\mathcal{O}(p(d^2 + \log p))$ floating-point operations. In the finite sample setting, our method involves estimating the precision matrix, which can be done by solving p linear programs (LPs) and then performing p iterations to learn the structure and parameters of the SEM by identifying and removing terminal (sink) vertices. If the estimated precision matrix is sparse, then each iteration involves solving at most d linear programs in at most d^2 dimensions, leading to an overall smoothed complexity of $\widetilde{\mathcal{O}}\left(p^3 + pd^7\right)$. When the estimated precision matrix is dense our method has a smoothed complexity of $\widetilde{\mathcal{O}}\left(p^5\right)$. This is significantly better than [16]'s algorithm for learning linear Gaussian SEMs as well as [14]'s algorithm for learning SEMs with known noise variance. While the former is is exponential in p, the latter is exponential in d and the tree-width of the SEM when the estimated precision matrix is sparse and exponential in p for the dense case.

Our algorithm also works in the high-dimensional regime when $d = o(p)$, and has a sample complexity of $\mathcal{O}(\frac{d^8}{\varepsilon^2}\log(\frac{p}{\sqrt{\delta}}))$ and $\mathcal{O}(\frac{d^8}{\varepsilon^2}(\frac{p^2}{\delta})^{1/m})$ for sub-Gaussian noise and noise with bounded $4m$-th moment respectively, for recovering the autoregression matrix of the SEM up to ϵ additive error with probability at least $1 - \delta$. The sample complexity of our algorithm for sub-Gaussian noise is better than [14]'s algorithm, which has a sample complexity of $\mathcal{O}\left(p^2 \log p\right)$, and is therefore unsuitable for the high-dimensional regime. Moreover, unlike [14]'s algorithm, and other methods that use conditional independence tests, for instance, the PC algorithm for learning Gaussian SEMs [13], our algorithm does not require any faithfulness conditions, and only requires a weaker *causal minimality* condition. The faithfulness assumption requires that the distribution $\mathcal{P}(X)$ contain only those conditional independence assertions that are implied by the *d-separation* criteria of the DAG [24]. However, faithfulness cannot be tested from data in full generality [29] and algorithms that infer the DAG structure from a finite number of samples must require *strong faithfulness* [28], which is a restrictive assumption. Our results has the following significant yet hitherto known implication for learning Gaussian Bayesian networks. Given data generated from a Gaussian Bayesian network that is causal minimal to the true DAG structure, one can recover the DAG structure in polynomial time and sample complexity from a finite number of samples, under more general identifiability conditions than homoscedastic noise.

Lastly, we obtain several useful results about the theory of linear SEMs en route to developing our main algorithm for learning linear SEMs.

Our Techniques. Our algorithm for learning linear SEMs differs conceptually from previous test-based, score-based or inverse-covariance-estimation-based methods. We are therefore able to get rid of many of the shortcomings of existing methods like requirement of strict non-Gaussianity of noise [22], homoscedasticity [16], and faithfulness [14, 13]. We do so my obtaining and exploiting various properties of terminal vertices in linear SEMs. We obtain our sample complexity results by using various properties of sub-Gaussian and bounded-moment variables and using concentration results for the empirical covariance matrix under the aforementioned noise conditions. Lastly, we improve the computational complexity of our algorithm by exploiting the sparsity structure of the precision matrix to obtain solutions of "larger" LPs (size $\mathcal{O}(p)$) by solving much "smaller" LPs (size $\mathcal{O}\left(d^2\right)$).

2. Related Work

We start our discussion of existing literature by first presenting known identifiability conditions for learning SEMs and Bayesian networks. [17] proved identifiability of distributions drawn from a restricted SEM with additive noise, where in the restricted SEM the functions are assumed to be non-linear and thrice continuously differentiable. Linear SEMs are identifiable if (a) the noise variables are non-Gaussian [21], (b) the noise variances are known up to a constant factor [14], and (c) noise variables are Gaussian and have the same variance [16] (homoscedastic noise). [15] introduced Quadratic Variance Function (QVF) DAG models — a class of Bayesian networks in which the conditional variance of a variable is a quadratic function of its conditional mean — and proved identifiability of the models from observational data. However, QVF DAG models cannot be expressed as SEMs, and the quadratic variance property holds for a handful of conditional distributions which includes Binomial, Poisson, Exponential, Gamma, and a few others.

The computational and statistical complexity landscape of learning linear SEMs is peppered by inefficient algorithms. This is in part justified by various hardness results known in the literature for learning DAGs from observational data [3, 5]. Algorithms for learning DAGs can be divided into two categories: independence test based methods and score based methods. Score based methods use a score function, typically penalized log-likelihood, to find the best scoring DAG among the space of all DAGs. Since the number of DAGs and degree-bounded DAGs is exponential in p [19, 8], score-based methods are exponential time. A popular score function for learning Gaussian SEMs is the ℓ_0-penalized Gaussian log-likelihood score proposed by [26]. [16] proposed using ℓ_0-penalized Gaussian log-likelihood score for learning homoscedastic noise linear Gaussian SEMs along with a heuristic greedy search algorithm which is not guaranteed to find the correct (highest-scoring) solution. [14] showed that under a faithfulness assumption, the sparsity pattern of the precision matrix corresponds to the edge structure of the *moral graph* of the underlying DAG. They exploit this property to devise an algorithm that searches for the highest-scoring DAG, using dynamic programming, that has the same moral graph as that given by the sparsity pattern of the precision matrix. Independence test based methods on the other hand require restrictive faithfulness conditions to guarantee structure recovery. [13] proposed using the PC algorithm to learn Gaussian SEMs, which has a computational complexity of $\mathcal{O}\left(p^d\right)$ and is only efficient for learning very sparse Gaussian SEMs. Among computationally efficient algorithms, the *Direct-LiNGAM* algorithm [22], which strictly requires non-Gaussianity of the noise variables, needs an infinite number of samples to guarantee structure recovery. This is because of the use of independence testing between a variable and its residuals to detect exogenous variables (variables with no parents). For the same reason, the correctness of *RESIT* [17], which is a computationally efficient algorithm for learning *non-linear SEMs*, is only guaranteed in the population setting.

Other authors have proposed various approximation algorithms and heuristic methods for learning Bayesian networks, which can be used to learn Gaussian SEMs by using appropriate score functions.

Popular heuristic methods are max-min hill climbing (MMHC) algorithm by [25], and the Greedy Equivalence Search (GES) algorithm proposed by [4]. [11] proposed an LP-relaxation based method for learning Bayesian networks which is an approximation algorithm.

3. Preliminaries

We begin this section by introducing our notations and definitions before formalizing the problem of learning linear SEMs from observational data. We will let $[p] \stackrel{\text{def}}{=} \{1, \ldots, p\}$. Vectors and matrices are denoted by lowercase and uppercase bold faced letters respectively. Random variables (including random vectors) are denoted by uppercase letters. For any two non-empty index sets $s_r, s_c \subseteq [p]$, the matrix $\mathbf{A}_{s_r,s_c} \in \mathbb{R}^{|s_r|\times|s_c|}$ denotes the submatrix of $\mathbf{A} \in \mathbb{R}^{p\times p}$ obtained by selecting the s_r rows and s_c columns of $\mathbf{A}$. With a slight abuse of notation, we will allow the index sets s_r and s_c to be a single index, e.g., i, and we will denote the index set of all rows (or columns) by $*$. For any matrix $\mathbf{A}$ (equivalently for vectors), we will denote its support set by: $\mathcal{S}(\mathbf{A}) = \{(i,j) \in [p] \times [p] \,|\, A_{i,j} \neq 0\}$. Vector ℓ_p norms are denoted by $\|\cdot\|_p$. For matrices, $\|\cdot\|_p$ denotes the induced (or operator) ℓ_p-norm and $|\cdot|_p$ denotes the elementwise ℓ_p norm, i.e., $|\mathbf{A}|_p \stackrel{\text{def}}{=} (\sum_{i,j} |A_{i,j}|^p)^{1/p}$. For two matrices $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \circ \mathbf{B}$ denotes the Hadamard product of $\mathbf{A}$ and $\mathbf{B}$, while $\mathbf{diag}(\mathbf{A})$ denotes the vector formed by taking the diagonal of $\mathbf{A}$. For a vector $\mathbf{v}$, $\mathbf{Diag}(\mathbf{v})$ denotes the diagonal matrix with $\mathbf{v}$ in the diagonal. Finally, we define the set $-\mathsf{i} \stackrel{\text{def}}{=} [p] \setminus \{i\}$.

Let $\mathsf{G} = ([p], \mathsf{E})$ be a directed acyclic graph (DAG) where $[p]$ is the vertex set and $\mathsf{E} \subset [p] \times [p]$ is the set of directed edges. An edge $(i,j) \in \mathsf{E}$ implies the edge $i \leftarrow j$. We denote by $\pi_{\mathsf{G}}(i)$ and $\phi_{\mathsf{G}}(i)$ the parent set and the set of children of the i-th node respectively, in the graph G; and drop the subscript G when the clear from context. The set of neighbors of the i-th node is denoted by $\mathsf{N}_{\mathsf{G}}(i) = \pi_{\mathsf{G}}(i) \cup \phi_{\mathsf{G}}(i)$. A node j is a *descendant* of i in G if there exists a (directed) path from i to j in G. We will denote the set of descendants of i by $\mathsf{D}_{\mathsf{G}}(i)$. Similarly, we will denote the set of ancestors of i — nodes j such that there is a path from j to i in G — by the set $\mathsf{A}_{\mathsf{G}}(i)$. A vertex $i \in [p]$ is a *terminal vertex* in G if $\phi_{\mathsf{G}}(i) = \varnothing$. For each $i \in [p]$ we have a random variable $X_i \in \mathbb{R}$, $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ is the p-dimensional vector of random variables, and $\mathbf{x} = (x_1, \ldots, x_p)$ is a joint assignment to X. Every DAG $\mathsf{G} = ([p], \mathsf{E})$ defines a set of topological orderings $\mathcal{T}_{\mathsf{G}}$ over $[p]$ that are compatible with the DAG G, i.e., $\mathcal{T}_{\mathsf{G}} = \{\tau \in \mathrm{S}_p \mid \tau(j) < \tau(i) \text{ if } (i,j) \in \mathsf{E}\}$, where S_p is the set of all possible permutations of $[p]$.

The random vector X follows a linear structural equation model (SEM), if each variable can be written as a linear combination of the variables in its parent set as follows:

$$X_i = \sum_{j \in \pi_{\mathsf{G}}(i)} B_{i,j} X_j + N_i \qquad (\forall i \in [p]), \tag{1}$$

where $\mathsf{G} = ([p], \mathsf{E})$ is a DAG, $N = (N_1, \ldots, N_p)$ are the noise variables, and $N_i \perp\!\!\!\perp X_1, \ldots, X_{i-1}$. Without loss of generality, we assume that $\mathbb{E}[X_i] = \mathbb{E}[N_i] = 0$, $\forall i \in [p]$. As is typically the case in the literature of SEMs, we further assume that the noise variables N_i have bounded second moments and are independent. Thus $\mathrm{Cov}[N] = \mathbb{E}\left[NN^T\right] = \mathbf{Diag}(\sigma_1^2, \ldots, \sigma_p^2)$. We can then write (1) in vector form as follows:

$$X = \mathbf{B}X + N, \tag{2}$$

where $\mathbf{B} = (B_{i,j})$ is referred to as the *autoregression matrix* and $\mathcal{S}(\mathbf{B}) = \mathsf{E}$. Therefore, we will denote an SEM by the triple $(\mathsf{G}, \mathbf{B}, \{\sigma_i^2\}_{i\in[p]})$, or more compactly by $(\mathsf{G}, \mathbf{B}, \{\sigma_i^2\})$.

Given an SEM $(\mathsf{G}, \mathbf{B}, \{\sigma_i^2\})$, the joint distribution $\mathcal{P}(X)$ is completely determined and factorizes according to the DAG structure G:

$$\mathcal{P}(X; \mathsf{G}) = \prod_{i=1}^{p} \mathcal{P}_i(X_i | X_{\pi_{\mathsf{G}}(i)}; \mathsf{G}), \tag{3}$$

where $\mathcal{P}_i$ is the conditional distribution of the X_i. We then say that the distribution $\mathcal{P}$ is *Markov with respect to the DAG* G, i.e., X_i satisfies the Markov condition: $X_i \perp\!\!\!\perp X_j \mid X_{\pi(i)}, \forall i \in [p], \forall j \in [p] \setminus (\mathsf{D}(i) \cup \pi(i) \cup \{i\})$. Thus an SEM is equivalent to a *Bayesian network*. Specifically, if the noise variables are Gaussian, then $\mathcal{P}$ is a *Gaussian Bayesian network* (GBN), where the joint distribution $\mathcal{P}$ and the conditional distributions $\mathcal{P}_i$ are Gaussian. We obtain our theoretical results for the class of degree-bounded DAGs $\mathcal{G}_{p,d} \stackrel{\text{def}}{=} \{\mathsf{G} \mid \mathsf{G} = ([p], \mathsf{E}) \text{ is a DAG and } |\mathsf{N}_{\mathsf{G}}(i)| \leq d, \forall i \in [p]\}$.

Next, we define the notion of *causal minimality*, introduced by [29], which is important for ensuring identifiability of linear SEMs considered in this paper.

Definition 1 (Causal Minimality)**.** *Given a DAG* G*, a distribution* $\mathcal{P}(X)$*, that is Markov with respect to* G*, is causal minimal if* $\mathcal{P}$ *is not Markov with respect to a proper subgraph of* G*.*

Our assumption of $\mathcal{S}(\mathbf{B}) = \mathsf{E}$, ensures that Lemma 4 of [17] holds for all SEMs $(\mathsf{G}, \mathbf{B}, \{\sigma_i^2\})$. This in turn implies that the joint distribution $\mathcal{P}(X)$ determined by the SEM $(\mathsf{G}, \mathbf{B}, \{\sigma_i^2\})$ is causal minimal with respect to G (see Proposition 2 in [17]). Therefore, the SEMs considered in the paper are causal minimal.

The problem of learning the structure of an SEM is as follows. Given an $n \times p$ data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, with $\mathbf{x}_i \in \mathbb{R}^n$, drawn from an SEM $(\mathsf{G}^*, \mathbf{B}^*, \{\sigma_i^2\})$ with $\mathsf{G}^* \in \mathcal{G}_{p,d}$, we want to learn an SEM $(\widehat{\mathsf{G}}, \widehat{\mathbf{B}}, \{\widehat{\sigma}_i^2\})$ from $\mathbf{X}$ such that $\mathsf{G}^* = \widehat{\mathsf{G}}$.

4. Learning SEMs with unknown error variances

We start with presenting our main results for learning SEMs when the error variances are unknown. Our algorithm for learning SEMs works by constructing the SEM in a bottom-up fashion. The algorithm has p iterations. In each iteration it identifies and removes a terminal vertex, learning its parent set and edge weights along the way. We show that, under a certain identifiability condition which generalizes other identifiability conditions known in the literature, e.g., homoscedastic errors, and without assuming faithfulness of the distribution to the DAG, each of these steps can be performed efficiently using only the precision matrix or an estimator of it.

4.1. Identifiability

The following assumption gives a sufficient condition under which the structure and parameters of an SEM can be uniquely recovered from observational data using Algorithm 1. The assumption is defined in terms of subgraphs of G obtained by removing terminal vertices sequentially. For any $\tau \in \mathcal{T}_{\mathsf{G}}$, we will consider sequence of graphs $\mathsf{G}[m, \tau] = (\mathsf{V}[m, \tau], \mathsf{E}[m, \tau])$, indexed by (m, τ), where $\mathsf{G}[m, \tau]$ is the induced subgraph of G over the first m vertices in the topological ordering τ, i.e., $\mathsf{V}[m, \tau] \stackrel{\text{def}}{=} \{i \in [p] \mid \tau(i) \leq m\}$ and $\mathsf{E}[m, \tau] \stackrel{\text{def}}{=} \{(i, j) \in \mathsf{E} \mid i \in \mathsf{V}[m, \tau] \wedge j \in \mathsf{V}[m, \tau]\}$.

Assumption 1 (Identifiability condition)**.** *Given an SEM* $(\mathsf{G}, \mathbf{B}, \{\sigma_i^2\})$ *with* $\mathsf{G} \in \mathcal{G}_{p,d}$*, then* $\forall (i, j) \in \mathsf{V}[m, \tau] \times \mathsf{V}[m, \tau], m \in [p]$*, and* $\tau \in \mathcal{T}_{\mathsf{G}}$*, such that* $\phi_{\mathsf{G}[m,\tau]}(i) = \varnothing \wedge \phi_{\mathsf{G}[m,\tau]}(j) \neq \varnothing$*:*

$$\frac{1}{\sigma_i^2} < \frac{1}{\sigma_j^2} + \sum_{l \in \phi_{\mathsf{G}[m,\tau]}(j)} \frac{B_{l,j}^2}{\sigma_l^2}, \tag{4}$$

As we will show later, Assumption 1 essentially lays down a condition under which terminal vertices, and subsequently the causal order, can be identified from the precision matrix. From Assumption 1, we immediately get the following special cases for identifiability of linear SEMs, where the first one is the homoscedastic case known in the literature, while the second case is new.

Proposition 1 (Sufficient conditions for identifiability)**.** *Let* $(\mathsf{G}, \mathbf{B}, \{\sigma_i^2\})$ *be an SEM satisfying Assumption 1, with precision matrix* $\mathbf{\Omega}$*. Then, either of the following two conditions are sufficient for uniquely identifying the autoregression matrix* $\mathbf{B}$ *and the DAG* G *from* $\mathbf{\Omega}$*:*

(i) $\forall i \in [p]$, $\sigma_i = \sigma$*, for some* $\sigma > 0$*,*

(ii) $\forall i \in [p]$, $1 < \sigma_i \leq B_{\min}$*, where* $B_{\min} \stackrel{\text{def}}{=} \min\{|B_{i,j}| \mid (i,j) \in \mathsf{E}\}$.

Next we show that the identifiability condition 1 is in general necessary, i.e, if Assumption 1 is violated, then there exists an exponential number of DAG structures that, coupled with an uncountable number of autoregression matrices and noise variances, induce the same covariance and precision matrix, and determine joint distributions $\mathcal{P}(X)$ that are causal minimal and Markov to the DAG structures. *Therefore, no algorithm based on independence testing, or that uses solely the covariance or precision matrix, can recover the true DAG structure in polynomial time*. In the following lemma we will equivalently denote an SEM by $(\mathsf{G}, \mathbf{B}, \mathbf{D})$ where $\mathbf{D}$ is a diagonal matrix with $D_{i,i} = \sigma_i^2$.

Lemma 1. *There exists* $\widetilde{\mathcal{G}}_{p,d} \subset \mathcal{G}_{p,d}$ *with* $|\widetilde{\mathcal{G}}_{p,d}| = 2^{\Theta(p)}$*, autoregression matrices* $\mathbf{B}(\beta)$ *parameterized by* β*, and diagonal matrices* $\mathbf{D}(v_1, v_2)$ *parameterized by* v_1, v_2 *such that any SEM* $(\mathsf{G}, \mathbf{B}(\beta), \mathbf{D}(v_1, v_2))$ *with* $\mathsf{G} \in \widetilde{\mathcal{G}}_{p,d}$ *does not satisfy Assumption 1, induces the same covariance and precision matrix, and distribution* $\mathcal{P}(X)$ *that has the same conditional independence structure,* $\forall \beta \in (-\infty, \infty)$*,* $v_1 \in (0, \infty)$ *and* $v_2 > v_1$.

Next, we present a series of results building towards our main result for learning SEMs from precision matrix. In the following proposition we characterize the precision matrix of linear SEMs.

Proposition 2. *Let* $(\mathsf{G}, \mathbf{B}, \{\sigma_i^2\})$ *be an SEM over* X*, then the precision matrix is given as:* $\mathbf{\Omega} = (\mathbf{I} - \mathbf{B})^T \mathbf{D}^{-1} (\mathbf{I} - \mathbf{B})$*, where* $\mathbf{D} = \mathbf{Diag}(\sigma_1^2, \ldots, \sigma_p^2)$*. The entries of the precision matrix is given as:*

$$\Omega_{i,i} = \frac{1}{\sigma_i^2} + \sum_{l \in \phi(i)} \frac{B_{l,i}^2}{\sigma_l^2}, \qquad \Omega_{i,j} = -\frac{B_{i,j}}{\sigma_i^2} - \frac{B_{j,i}}{\sigma_j^2} + \sum_{l \in \phi(i) \cap \phi(j)} \frac{B_{l,i} B_{l,j}}{\sigma_l^2}. \tag{5}$$

The above characterization of the precision matrix motivates our indentifiability condition given by Assumption 1, and also provides a recipe for identifying terminal vertices from the precision matrix as is formalized by the following proposition.

Proposition 3. *Let* $(\mathsf{G}, \mathbf{B}, \{\sigma_i^2\})$ *be a SEM over* X *with precision matrix* $\mathbf{\Omega}$*, that satisfies the identifiability condition given by Assumption 1. Then,* i *is a terminal vertex in* G *if and only if* $i \in \operatorname{argmin}(\mathbf{diag}(\mathbf{\Omega}))$*. Further, if* i *is a terminal vertex then* $\sigma_i^2 = 1/\Omega_{i,i}$.

The next proposition, which follows directly from Proposition 3 and (5), states that for a terminal vertex the parent set and edge weights can be conveniently "read off" from the precision matrix. This is the key result which helps us avoid the faithfulness condition.

Proposition 4. *Let* $(\mathsf{G}, \mathbf{B}, \{\sigma_i^2\})$ *be an SEM over* X *with precision matrix* $\mathbf{\Omega}$*. If* i *is a terminal vertex in* G*, then* $\mathbf{B}_{i,*} = -\mathbf{\Omega}_{i,*}/\Omega_{i,i}$ *and* $\pi_{\mathsf{G}}(i) = \mathcal{S}(\mathbf{\Omega}_{i,*}) \setminus \{i\}$.

The following lemma is a useful result about linear SEMs with arbitrary noise distribution, that generalizes a result so far known only for the Gaussian distribution — for a terminal vertex i, the precision matrix over $X_{-\mathsf{i}}$ can be obtain by performing a Schur complement update of the precision matrix over X. While, the result for the Gaussian distribution holds for all variables, the analogous result for general SEMs holds only for terminal vertices.

Lemma 2. *Let $(\mathsf{G}, \mathbf{B}, \{\sigma_i^2\})$ be an SEM over X with precision matrix $\mathbf{\Omega}$. Let i be a terminal vertex in the G, then the precision matrix over $X_{-\mathsf{i}}$, $\mathbf{\Omega}_{(-\mathsf{i})}$, is given as:*

$$\mathbf{\Omega}_{(-\mathsf{i})} = \mathbf{\Omega}_{-\mathsf{i},-\mathsf{i}} - \frac{1}{\Omega_{i,i}} \mathbf{\Omega}_{-\mathsf{i},i} \mathbf{\Omega}_{i,-\mathsf{i}}.$$

Finally, the following lemma characterizes the entries of the precision matrix over $X_{-\mathsf{i}}$ and will be very useful in developing our finite-sample algorithm for learning SEMs.

Lemma 3. *Let $(\mathsf{G}, \mathbf{B}, \{\sigma_i^2\})$ be a SEM over X with precision matrix $\mathbf{\Omega}$. Let i be a terminal vertex in the G and let $\mathbf{\Omega}_{(-\mathsf{i})}$ denote the precision matrix over $X_{-\mathsf{i}}$. Then,*

$$\begin{aligned} (\mathbf{\Omega}_{(-\mathsf{i})})_{j,k} &= \mathbf{\Omega}_{j,k} && (\forall (j,k) \in -\mathsf{i} \times -\mathsf{i} \mid \{j,k\} \nsubseteq \pi_{\mathsf{G}}(i)), \\ \mathcal{S}((\mathbf{\Omega}_{(-\mathsf{i})})_{j,*}) &\subseteq (\mathcal{S}(\mathbf{\Omega}_{j,*}) \setminus \{i\}) \cup \pi_{\mathsf{G}}(i) && (\forall j \in \pi_{\mathsf{G}}(i)). \end{aligned}$$

With the required results in place, we are now ready to present our main algorithm, detailed in Algorithm 1, for learning SEMs from the precision matrix. The role of the diagonal matrix $\mathbf{D}$ will become clear in the next section where we focus on the problem of learning SEMs with known error variances. For now we simply set $\mathbf{D}$ to the identity matrix $\mathbf{I}$. The following theorem proves the correctness of our algorithm in the population setting.

Algorithm 1 SEM structure learning algorithm.

Input: Precision matrix $\mathbf{\Omega}$, diagonal matrix $\mathbf{D}$.
Output: $\widehat{\mathsf{G}}, \widehat{\mathbf{B}}$.

$\widehat{\mathbf{B}} \leftarrow \mathbf{0}$.
for $t \in [p]$ **do**
  $i \leftarrow \operatorname{argmin}(\mathbf{diag}(\mathbf{\Omega} \circ \mathbf{D}))$.
  $\mathbf{B}_{i,*} \leftarrow -\mathbf{\Omega}_{i,*}/\Omega_{i,i}$, $B_{i,i} \leftarrow 0$.
  $\mathbf{\Omega} \leftarrow \mathbf{\Omega} - \frac{1}{\Omega_{i,i}} \mathbf{\Omega}_{*,i} \mathbf{\Omega}_{i,*}$.
  $\Omega_{i,i} \leftarrow \infty$.
end for
$\widehat{\mathsf{G}} \leftarrow ([p], \mathcal{S}(\widehat{\mathbf{B}}))$.

Algorithm 2 Updating a precision matrix, after removing a terminal vertex, using CLIME.

function UPDATE($\widehat{\mathbf{\Omega}}, i, \lambda_n$)
  $\widehat{\pi}(i) \leftarrow \mathcal{S}(\widehat{\mathbf{\Omega}}_{i,*}) \setminus \{i\}$.
  for $j \in \widehat{\pi}(i)$ **do**
    $\widehat{\mathsf{S}}_j \leftarrow \Big(\mathcal{S}(\widehat{\mathbf{\Omega}}_{j,*}) \setminus \{i\}\Big) \cup \widehat{\pi}(i)$.
    Compute $\bar{\boldsymbol{\omega}}_j$ by solving (7) for $\mathbf{\Sigma}^n_{\widehat{\mathsf{S}}_j, \widehat{\mathsf{S}}_j}$.
    $\widehat{\mathbf{\Omega}}_{j,\widehat{\mathsf{S}}_j} = \widehat{\mathbf{\Omega}}_{\widehat{\mathsf{S}}_j,j} \leftarrow \bar{\boldsymbol{\omega}}_j$
  end for
  $\widehat{\mathbf{\Omega}}_{i,*} \leftarrow \mathbf{0}$ and $\widehat{\mathbf{\Omega}}_{*,i} \leftarrow \mathbf{0}$.
  return $\widehat{\mathbf{\Omega}}$.
end function

Theorem 1. *Let $(\mathsf{G}, \mathbf{B}, \{\sigma_i^2\})$ be an SEM over X, with precision matrix $\mathbf{\Omega}$, satisfying Assumption 1. Then, given $(\mathbf{\Omega}, \mathbf{I})$ as input, Algorithm 1 returns a unique $(\widehat{\mathsf{G}}, \widehat{\mathbf{B}})$ such that $\widehat{\mathsf{G}} = \mathsf{G}$ and $\widehat{\mathbf{B}} = \mathbf{B}$.*

As a consequence of the above theorem we have the following corollary about identifiability of linear SEMs.

Corollary 1. *An SEM $(\mathsf{G}, \mathbf{B}, \{\sigma_i^2\})$ satisfying Assumption 1 is identifiable, and can be uniquely identified from the precision matrix $\mathbf{\Omega}$.*

4.2. Statistical guarantees for estimation

Algorithm 1 can be used to learn a SEM given an estimate of the precision matrix, computed from a finite number of samples, with a slight modification. In line 5 instead of using the Schur complement update, we use Algorithm 2 to update the precision matrix after a terminal vertex has been identified (and removed). The rationale behind this is that even if the estimated precision matrix is close to the true precision matrix, the Schur updates could still result in errors accumulating in the precision matrix. In order to ensure that our algorithm is statistically efficient, we need more control over those errors, which in turns calls for some sort of penalization for estimating from a finite number of samples.

Inverse covariance matrix estimation. A key step of our algorithm is estimating the inverse covariance matrix over X or a subset of X. Due in part to its role in undirected graphical model selection, the problem of inverse covariance matrix estimation has received significant attention over the years. A popular approach for inverse covariance estimation, under high-dimensional settings, is the ℓ_1-penalized Gaussian maximum likelihood estimate (MLE) studied by [27], [1], and [7], among others. The ℓ_1-penalized Gaussian MLE estimate of the inverse covariance matrix has attractive theoretical guarantees as shown by [18]. However, the elementwise ℓ_∞ guarantees for the inverse covariance estimate obtained by [18] require an edge-based mutual incoherence condition that is quite restrictive. Many algorithms have been developed in the recent past for solving the ℓ_1-penalized Gaussian MLE problem [10, 9, 20, 12]. While, technically, these algorithms can be used in conjunction with our algorithm for learning SEMs, in this paper we use the method called CLIME, developed by [2]. The primary motivation behind using CLIME is that the theoretical guarantees obtained by [2] does not require the edge-based mutual incoherence condition. Further, CLIME is computationally attractive because it computes $\widehat{\mathbf{\Omega}}$ columnwise by solving p independent linear programs. Even though the CLIME estimator $\widehat{\mathbf{\Omega}}$ is not guaranteed to be positive-definite (it is positive-definite with high probability) it is suitable for our purpose. Next, we briefly describe the CLIME method for inverse covariance estimation and instantiate the theoretical results of [2] for our purpose.

The CLIME estimator $\widehat{\mathbf{\Omega}}$ is obtained as follows. First, we compute a potentially non-symmetric estimate $\bar{\mathbf{\Omega}} = (\bar{\omega}_{i,j})$ by solving the following:

$$\bar{\mathbf{\Omega}} = \underset{\mathbf{\Omega} \in \mathbb{R}^{p \times p}}{\operatorname{argmin}} |\mathbf{\Omega}|_1 \text{ s.t. } |\mathbf{\Sigma}^n \mathbf{\Omega} - \mathbf{I}|_\infty \leq \lambda_n, \tag{6}$$

where $\lambda_n > 0$ is the regularization parameter, $\mathbf{\Sigma}^n \overset{\text{def}}{=} (1/n)\mathbf{X}^T\mathbf{X}$ is the empirical covariance matrix, and $|\cdot|_1$ (respectively $|\cdot|_\infty$) denotes elementwise ℓ_1 (respectively ℓ_∞) norm. Finally, the symmetric estimator is obtained by selecting the smaller entry among $\bar{\omega}_{i,j}$ and $\bar{\omega}_{j,i}$, i.e., $\widehat{\mathbf{\Omega}} = (\widehat{\omega}_{i,j})$, where $\widehat{\omega}_{i,j} = \bar{\omega}_{i,j}\mathbf{1}\left[|\bar{\omega}_{i,j}| < |\bar{\omega}_{j,i}|\right] + \bar{\omega}_{j,i}\mathbf{1}\left[|\bar{\omega}_{j,i}| \leq |\bar{\omega}_{i,j}|\right]$. It is easy to see that (6) can be decomposed into p linear programs as follows. Let $\bar{\mathbf{\Omega}} = (\bar{\boldsymbol{\omega}}_1, \ldots, \bar{\boldsymbol{\omega}}_p)$, then

$$\bar{\boldsymbol{\omega}}_i = \underset{\boldsymbol{\omega} \in \mathbb{R}^p}{\operatorname{argmin}} \|\boldsymbol{\omega}\|_1 \text{ s.t. } |\mathbf{\Sigma}^n \boldsymbol{\omega} - \mathbf{e}_i|_\infty \leq \lambda_n, \tag{7}$$

where $\mathbf{e}_i = (e_{i,j})$ such that $e_{i,j} = 1$ for $j = i$ and $e_{i,j} = 0$ otherwise. The main result about the CLIME estimator that we use from [2] is given by the following lemma, which is a minor reformulation of Theorem 6 in [2]:

Lemma 4 ([2])**.** *Let $(\mathsf{G}, \mathbf{B}, \{\sigma_i^2\})$ be an SEM over X, with covariance and precision matrix $\mathbf{\Sigma}$ and $\mathbf{\Omega}$ respectively. Let $\widehat{\mathbf{\Omega}}$ be the estimator of $\mathbf{\Omega}$ obtained by solving the optimization problem given by 7. Then if $\lambda_n \geq \|\mathbf{\Omega}\|_1 |\mathbf{\Sigma} - \mathbf{\Sigma}^n|_\infty$, then $|\mathbf{\Omega} - \widehat{\mathbf{\Omega}}|_\infty \leq 4\|\mathbf{\Omega}\|_1 \lambda_n$. Further, if*

$$\min\{|\Omega_{i,j}| \mid (i,j) \in [p] \times [p] \wedge |\Omega_{i,j}| \neq 0\} > 4\|\mathbf{\Omega}\|_1 \lambda_n,$$

then $\mathcal{S}(\mathbf{\Omega}) \subseteq \mathcal{S}(\widehat{\mathbf{\Omega}})$.

Next we state out finite sample identifiability condition. This differs from the population version in that we require a "gap" between the diagonal entries of the precision matrix for terminal and non-terminal vertices. This gap, as we show later, must scale as $\Omega\left(d^2\sqrt{\frac{\log p}{n}}\right)$ and $\Omega\left(\frac{d^2(p)^{1/m}}{\sqrt{n}}\right)$ for sub-Gaussian noise and bounded moment noise respectively. Condition (ii) of the below assumption also restricts how fast the "minimum" non-diagonal entry of the precision matrix must decay. Note that our conditions are weaker than those of [14] due to which we are able to achieve better sample complexity than their algorithm.

Assumption 2 (Finite Sample Identifiability Condition)**.** *Let* $(\mathsf{G}, \mathbf{B}, \{\sigma_i^2\})$ *be an SEM with inverse covariance matrix* $\mathbf{\Omega}$*. Let* $\mathbf{\Omega}_{(m,\tau)}$ *denote the inverse covariance matrix over* $X_{\mathsf{V}[m,\tau]}$*, and*

$$M \stackrel{\text{def}}{=} \max\{\|\mathbf{\Omega}_{(m,\tau)}\|_1 \mid m \in [p], \tau \in \mathcal{T}_{\mathsf{G}}\}. \tag{8}$$

Then, we have that

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

$$\frac{1}{\sigma_i^2} < \frac{1}{\sigma_j^2} + \sum_{l \in \phi_{\mathsf{G}[m,\tau]}(j)} \frac{B_{l,j}^2}{\sigma_l^2} - 8M\lambda_n,$$

(ii) $\min\{|(\Omega_{(m,\tau)})_{i,j}| \mid (\Omega_{(m,\tau)})_{i,j} \neq 0, (i,j) \in \mathsf{V}[m,\tau] \times \mathsf{V}[m,\tau], m \in [p], \tau \in \mathcal{T}_{\mathsf{G}}\} > 4M\lambda_n$*,*

(iii) for all $i \in [p]$*,* $\sigma_i^2 \in o(1/4M\lambda_n)$*.*

The following lemma proves the correctness of Algorithm 2 which updates the precision matrix, after removing a terminal vertex.

Lemma 5. *Let* $(\mathsf{G}, \mathbf{B}, \{\sigma_i^2\})$ *be an SEM over* X *with precision matrix* $\mathbf{\Omega}$*. Let* $\widehat{\mathbf{\Omega}}$ *be an estimator of* $\mathbf{\Omega}$ *such that* $|\mathbf{\Omega} - \widehat{\mathbf{\Omega}}|_\infty \leq 4M\lambda_n$*, and* $\mathcal{S}(\mathbf{\Omega}) \subseteq \mathcal{S}(\widehat{\mathbf{\Omega}})$*, where* M *is defined in* (8)*. Let* i *be a terminal vertex in the* G*,* $\mathbf{\Omega}_{(-\mathsf{i})}$ *be the true precision matrix over* $X_{-\mathsf{i}}$*, and let* $\widehat{\mathbf{\Omega}}'$ *be the matrix returned by the function* UPDATE*. Then,* $|\mathbf{\Omega}_{(-\mathsf{i})} - \widehat{\mathbf{\Omega}}'_{-\mathsf{i},-\mathsf{i}}|_\infty \leq 4M\lambda_n$ *and* $\mathcal{S}(\mathbf{\Omega}_{(-\mathsf{i})}) \subseteq \mathcal{S}(\widehat{\mathbf{\Omega}}')$*.*

Theorem 2. *Let* $(\mathsf{G}^*, \mathbf{B}^*, \{\sigma_i^2\})$ *be the true SEM, with covariance and precision matrix* $\mathbf{\Sigma}^*$ *and* $\mathbf{\Omega}^*$*, respectively, from which a data set* $\mathbf{X}$ *of* n *samples is drawn. If the regularization parameter satisfies* $\lambda_n \geq M|\mathbf{\Sigma}^n - \mathbf{\Sigma}^*|$*, then under Assumption 2, the Algorithm 1, with* $\mathbf{D}$ *set to* $\mathbf{I}$*, returns an estimator* $\widehat{\mathbf{B}}$ *such that* $|\mathbf{B}^* - \widehat{\mathbf{B}}| \leq c4M(1 + B_{\max})\sigma_{\max}^2\lambda_n$*,* $\mathcal{S}(\mathbf{B}^*) \subseteq \mathcal{S}(\widehat{\mathbf{B}})$*, and* $\mathcal{T}_{\widehat{\mathsf{G}}} \subseteq \mathcal{T}_{\mathsf{G}^*}$*, where* $c \leq \sigma_{\min}^2/(1 - 4M\lambda_n\sigma_{\min}^2)$ *is a constant.*

Theorem 3 (Sub-Gaussian noise)**.** *Given an SEM* $(\mathsf{G}^*, \mathbf{B}^*, \{\sigma_i^2\})$ *with* $\mathsf{G}^* \in \mathcal{G}_{p,d}$ *satisfying Assumptions 2 such that* N_i/σ_i *is sub-Gaussian with parameter* ν*; if the regularization parameter and number of samples satisfy the following conditions:*

$$\lambda_n \geq MC_1\sqrt{\frac{2}{n}\log\left(\frac{2p}{\sqrt{\delta}}\right)}, \qquad n \geq \frac{2(cC_1 4M^2(1 + B_{\max})\sigma_{\max}^2)^2}{\varepsilon^2}\log\left(\frac{2p}{\sqrt{\delta}}\right),$$

then $|\mathbf{B}^* - \widehat{\mathbf{B}}|_\infty \leq \varepsilon$ *with probability at least* $1 - \delta$*, where* $C_1 = \sqrt{128}(1 + 4\nu^2)(\max_i \mathbf{\Sigma}_{i,i}^*)$*,* c *is defined in Theorem 2, and* M *is given by* (8)*. Further, thresholding* $\widehat{\mathbf{B}}$ *at the level* ε *we get that* $\mathcal{S}(\widehat{\mathbf{B}}) = \mathcal{S}(\mathbf{B}^*)$ *and* $\widehat{\mathsf{G}} = \mathsf{G}^*$*.*

Theorem 4 (Bounded moment noise)**.** *Given an SEM* $(\mathsf{G}^*, \mathbf{B}^*, \{\sigma_i^2\})$ *with* $\mathsf{G}^* \in \mathcal{G}_{p,d}$ *satisfying Assumption 2 such that* $(\mathbb{E}[N_i]/\sigma_i)^{4m} \leq K_m$, $\forall i \in [p]$, *where* m *is a positive integer and* $K_m \in \mathbb{R}^+$ *is a constant. If the regularization parameter and number of samples satisfy the following conditions:*

$$\lambda_n \geq MC_2 \left(\frac{p^2}{n^m \delta}\right)^{1/2m}, \qquad n \geq \frac{(cC_2 4M^2(1+B_{\max})\sigma_{\max}^2)^2}{\varepsilon^2}\left(\frac{p^2}{\delta}\right)^{1/m},$$

then $|\mathbf{B}^* - \widehat{\mathbf{B}}|_\infty \leq \varepsilon$ *with probability at least* $1-\delta$, *where* $C_2 = 2(\max_i \mathbf{\Sigma}^*_{i,i})(C_m(C_m(K_m+1)+1))^{1/2m}$, C_m *is a constant that depends only on* m, c *is defined in Theorem 2, and* M *is given by* (8)*. Further, thresholding* $\widehat{\mathbf{B}}$ *at the level* ε *we get that* $\mathcal{S}(\widehat{\mathbf{B}}) = \mathcal{S}(\mathbf{B}^*)$ *and* $\widehat{\mathsf{G}} = \mathsf{G}^*$.

5. Learning SEMs with known error variances

In this section we focus our attention on the problem of learning SEMs when the error variances are known upto a constant factor. We will consider SEMs $(\mathsf{G}, \mathbf{B}, \{\alpha\sigma_i^2\})$ where $\{\sigma_i^2\}_{i=1}^p$ are known (to the learner) and $\alpha > 0$ is some unknown constant. Identifiability of this class of SEMs was proved by [14] under a *faithfulness* assumption. However, we will merely assume that $(\mathsf{G}, \mathbf{B}, \{\alpha\sigma_i^2\})$ is causal minimal, i.e., $\mathcal{S}(\mathbf{B}) = \mathsf{E}$ — this ensures that the distribution $\mathcal{P}(X)$ defined by the SEM is causal minimal to the DAG $\mathsf{G} = ([p], \mathsf{E})$. An immediate consequence of Proposition 2 is the following observation about terminal vertices:

Proposition 5. *Let* $(\mathsf{G}, \mathbf{B}, \{\alpha\sigma_i^2\})$ *be an SEM over* X *with precision matrix* $\mathbf{\Omega}$, $\{\sigma_i^2\}_{i=1}^p$ *known and* $\alpha > 0$ *is some unknown constant. Then,* i *is a terminal vertex in* G *if and only if* $i \in \operatorname{argmin} \mathbf{diag}(\mathbf{\Omega} \circ \mathbf{D})$, *where* $\mathbf{D} = \mathbf{Diag}(\sigma_1^2, \ldots, \sigma_p^2)$.

Thus, when the error variances are known upto a constant factor, Algorithm 1 can be used to learn SEMs, under the assumption of causal minimality, by setting $\mathbf{D} = \mathbf{Diag}(\sigma_1^2, \ldots, \sigma_p^2)$. Consequently, we have the following result about learning SEMs with known error variances:

Theorem 5. *Let* $(\mathsf{G}, \mathbf{B}, \{\alpha\sigma_i^2\})$ *be an SEM over* X, *with precision matrix* $\mathbf{\Omega}$ *and* $\{\sigma_i^2\}_{i=1}^p$ *known. Then, if* $(\mathsf{G}, \mathbf{B}, \{\alpha\sigma_i^2\})$ *is causal minimal and given* $\mathbf{\Omega}$, $\mathbf{D} = \mathbf{Diag}(\sigma_1^2, \ldots, \sigma_p^2)$ *as input, Algorithm 1 returns a unique* $(\widehat{\mathsf{G}}, \widehat{\mathbf{B}})$ *such that* $\widehat{\mathsf{G}} = \mathsf{G}$ *and* $\widehat{\mathbf{B}} = \mathbf{B}$.

Misspecified error variances. Our algorithm can also be used to learn SEMs with misspecified error variances as considered by [14]. For instance, if the true SEM is $(\mathsf{G}, \mathbf{B}, \{\sigma_i^2\})$ while the diagonal matrix passed to Algorithm 1 is $\mathbf{D} = \mathbf{Diag}((\sigma_1')^2, \ldots, (\sigma_p')^2)$, then it is straightforward to verify that the following condition is sufficient to ensure that Algorithm 1 still recovers the structure and parameters of the SEM correctly:

$$\sum_{l \in \phi_{\mathsf{G}[m,\tau]}(j)} B_{l,j}^2 > \frac{\alpha_{\max}}{\alpha_{\min}} - 1, \qquad (\forall j \in \mathsf{V}[m,\tau] \wedge \phi_{\mathsf{G}[m,\tau]}(j) \neq \varnothing, m \in [p], \tau \in \mathcal{T}),$$

where $\alpha_{\max} \stackrel{\text{def}}{=} \max\{(\sigma_i')^2/\sigma_i^2 \mid i \in [p]\}$ (similarly $\alpha_{\min}$). Next, we obtain statistical guarantees for our algorithm for learning SEMs with known error variances.

5.1. Statistical guarantees for estimation

In order to learn SEMs with known error variances from a finite number of samples, we make the following assumptions:

Assumption 3. *Given an SEM* $(\mathsf{G}, \mathbf{B}, \{\alpha\sigma_i^2\})$ *with precision matrix* $\mathbf{\Omega}$ *and* $\{\sigma_i^2\}_{i=1}^p$ *known, let* $\mathbf{\Omega}_{(m,\tau)}$ *denote the inverse covariance matrix over* $X_{\mathsf{V}[m,\tau]}$*. Then,*

(i) $\forall i \in \mathsf{V}[m,\tau], m \in [p]$*, and* $\tau \in \mathcal{T}_{\mathsf{G}}$*, such that* $\phi_{\mathsf{G}[m,\tau]}(i) \neq \varnothing$*:*

$$\sum_{l \in \phi_{\mathsf{G}[m,\tau]}(i)} \left(\frac{\sigma_i^2}{\sigma_l^2}\right) B_{l,i}^2 > 8\alpha M \lambda_n,$$

(ii) $\min\{|(\Omega_{(m,\tau)})_{i,j}| \mid (\Omega_{(m,\tau)})_{i,j} \neq 0, (i,j) \in \mathsf{V}[m,\tau] \times \mathsf{V}[m,\tau], m \in [p], \tau \in \mathcal{T}_{\mathsf{G}}\} > 4M\lambda_n$*,*

(iii) for all $i \in [p]$*,* $\sigma_i^2 \in o(1/4\alpha M \lambda_n)$*.*

Using CLIME to estimate and update the precision matrix, it is easy to verify that Theorems 3 and 4 hold for SEMs with known error variances satisfying Assumption 3, with $\sigma_{\max}^2$ and $\sigma_{\min}^2$ replaced by $\alpha\sigma_{\max}^2$ and $\alpha\sigma_{\min}^2$, respectively. Thus, given a data set of n samples drawn from an SEM satisfying Assumption 3, with autoregression matrix $\mathbf{B}^*$ and DAG structure $\mathsf{G}^* = ([p], \mathsf{E}^*)$, we have the following results about sub-Gaussian and bounded-moment noise:

Remark 1. *For sub-Gaussian noise, if* $\lambda_n = \Omega\left(\frac{d^2}{\sqrt{n}}\sqrt{\log(\frac{p}{\sqrt{\delta}})}\right)$*, and the number of samples* $n = \Omega\left(\frac{d^8}{\varepsilon^2}\log(\frac{p}{\sqrt{\delta}})\right)$*, then Algorithm 1 with* $\mathbf{D} = \mathbf{Diag}(\sigma_1^2, \ldots, \sigma_p^2)$ *returns an estimator* $\widehat{\mathbf{B}}$ *such that* $|\widehat{\mathbf{B}} - \mathbf{B}^*|_\infty \leq \varepsilon$*, with probability at least* $1 - \delta$*. Thresholding* $\widehat{\mathbf{B}}$ *at the level* ε*, we have* $\mathcal{S}(\widehat{\mathbf{B}}) = \mathsf{E}^*$*.*

Remark 2. *For noise with bounded* $(4m)$*-th moment, with* m *being a positive integer, if the regularization parameter* $\lambda_n = \Omega\left(\frac{d^2}{\sqrt{n}}(\frac{p}{\sqrt{\delta}})^{1/m}\right)$*, and the number of samples* $n = \Omega\left(\frac{d^8}{\varepsilon^2}(\frac{p^2}{\delta})^{1/m}\right)$*, then Algorithm 1 with* $\mathbf{D} = \mathbf{Diag}(\sigma_1^2, \ldots, \sigma_p^2)$ *returns an estimator* $\widehat{\mathbf{B}}$ *such that* $|\widehat{\mathbf{B}} - \mathbf{B}^*|_\infty \leq \varepsilon$*, with probability at least* $1 - \delta$*. Thresholding* $\widehat{\mathbf{B}}$ *at the level* ε*, we have* $\mathcal{S}(\widehat{\mathbf{B}}) = \mathsf{E}^*$*.*

The above remarks use the fact that $M = \mathcal{O}\left(d^2\right)$, which follows from Proposition 6 given in Appendix.

6. Computational complexity

In the population setting, i.e., given the true precision matrix, our algorithm can be implemented by storing the diagonal of the precision matrix separately and sorting it once which takes $\mathcal{O}\left(p \log p\right)$ time. In each iteration, updating the precision matrix in line 5 takes $\mathcal{O}\left(d^2\right)$ time since $\mathbf{\Omega}_{*,i}$ and $\mathbf{\Omega}_{i,*}$ are d-sparse. Updating the diagonal takes $\mathcal{O}\left(d \log p\right)$ time, while searching for the minimum diagonal element takes $\mathcal{O}\left(\log p\right)$ time. Therefore, Algorithm 1 computes the $\widehat{\mathbf{B}}$ matrix in $\mathcal{O}\left(p(d^2 + d\log p)\right)$ time. In the population setting, the computational complexity of [14]'s algorithm is $\mathcal{O}\left(p2^{2(w+1)(w+d)}\right)$, where w is the tree-width of the DAG structure of the true SEM. Note that the population version of our algorithm can still be used in the finite sample setting if the precision matrix is estimated accurately enough.

In the finite sample setting, the computational complexity of our algorithm is dominated by the steps for estimating and updating the precision matrix — the latter depends on how well the sparsity pattern of the precision matrix is estimated. First, we analyze the computational complexity of our algorithm assuming exact support recovery, then we analyze the worst-case performance of our algorithm without assuming sparsity of the estimated precision matrix. Estimating the precision matrix can be done by solving p linear programs in $2p$-dimension and with $4p$ constraints. The smoothed complexity of this step is $\mathcal{O}\left(p^3 \log(p/\sigma)\right)$ when using interior point LP solvers [6], where σ^2 is variance of the Gaussian

perturbations [1]. Next observe that $|\mathbf{\Omega}^* - \widehat{\mathbf{\Omega}}|_\infty \leq |\mathbf{B}^* - \widehat{\mathbf{B}}|_\infty \leq \varepsilon$. By thresholding $\widehat{\mathbf{\Omega}}$ at the level ε, each time the precision matrix is updated, we can ensure exact support recovery in each iteration. Thus, in the UPDATE function $\widehat{\pi}(i) = \pi_{\mathsf{G}^*}(i)$ and $|\widehat{\mathsf{S}}_j| \leq d^2 \leq p$. Therefore, the UPDATE function takes $\mathcal{O}\left(d^7 \log(d/\sigma)\right)$ operations, leading to an overall complexity of $\widetilde{\mathcal{O}}\left(p^3 + pd^7\right)$. In the worst case, i.e., without any thresholding, $\widehat{\mathbf{\Omega}}$ can be dense. Therefore, the UPDATE function might re-estimate the full precision matrix over $p - t$ variables in iteration t, which takes $\mathcal{O}\left((p-t)^4 \log((p-t)/\sigma)\right)$ operations, leading to an overall complexity of $\widetilde{\mathcal{O}}\left(p^5\right)$. Thus, in the finite sample setting the complexity of our algorithm is between $\widetilde{\mathcal{O}}\left(p^3 + pd^7\right)$ and $\widetilde{\mathcal{O}}\left(p^5\right)$. Note that [14]'s analysis of the computational complexity of their algorithm assumes perfect support recovery of the precision matrix. In this regime, the computational complexity of their method is $\mathcal{O}\left(p2^{2(w+1)(w+d)} + p^3\right)$, including the step to estimate the precision matrix using graphical Lasso [7], where w is the tree-width of the true DAG. However, without thresholding the output of graphical Lasso can be dense leading to a worst-case computational complexity that is exponential in p.

7. Appendix

A. Detailed Proofs

Proof of Proposition 1. When $\sigma_i^2 = \sigma^2$ for all $i \in [p]$, then (4) reduces to:

$$\sum_{l \in \phi_{\mathsf{G}[m,\tau]}(j)} \frac{B_{l,j}^2}{\sigma_l^2} > 0,$$

which holds trivially by causal minimality since $B_{l,j}^2 > 0$ for $(l,j) \in \mathsf{E}$. This proves part (i).

Now under (ii), $1/\sigma_i^2 - 1/\sigma_j^2 < 1, \forall i,j \in [p]$. Also, $B_{l,j}^2/\sigma_l^2 \geq 1$ for all $(l,j) \in \mathsf{E}$. Thus (4) is satisfied. □

Proof of Lemma 1. Consider the following two SEMs over three nodes, where the noise variances are shown within braces below each node, and the edge weights are shown on the edges.

Both the SEMs make the following conditional independence assertion: $X_1 \perp\!\!\!\perp X_3 \mid X_2$, and are therefore Markov and causal minimal to $\mathcal{P}(X)$. Set $b_2 = \sqrt{1 - \frac{v_1}{v_2}}$. Then using the formulas derived in Proposition 2 it can be verified that the precision matrix for both the SEMs is:

$$\mathbf{\Omega} = \frac{1}{v_1} \times \begin{bmatrix} 1 & -\beta & 0 \\ -\beta & 1+\beta^2 & -b_2 \\ 0 & -b_2 & 1 \end{bmatrix}. \tag{9}$$

The SEM on the left does not satisfy Assumption 1 because vertex 3 is a non-terminal vertex but $3 \in \operatorname{argmin}(\mathbf{\Omega})$. The SEM on the right does not satisfy Assumption 1 because after the vertex 1 is removed

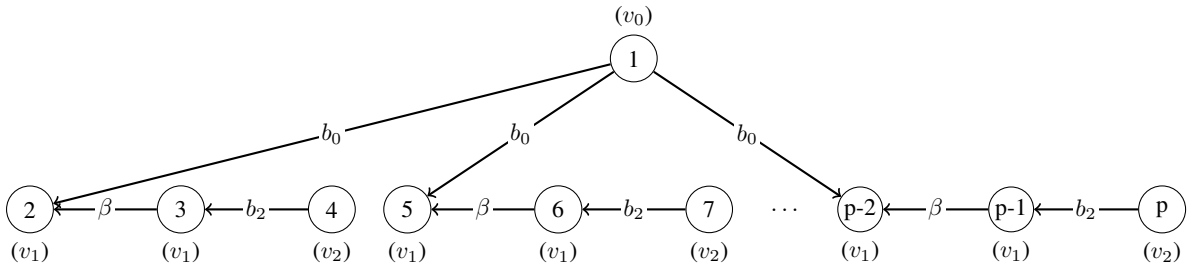
[1] The worst-case complexity of interior point methods for solving LPs is $\mathcal{O}\left(p^3 L\right)$ where L " is a parameter measuring the precision needed to perform the arithmetic operations exactly" and grows as $\Omega(p)$ [23]. However, interior-point methods work much more efficiently in practice and have an average complexity of $\mathcal{O}\left(p^3 \log p\right)$ (see [23] and the references therein).

we have that vertex 3 is a non-terminal vertex but satisfies $3 \in \operatorname{argmin}(\mathbf{\Omega}_{(-1)})$, where $\mathbf{\Omega}_{(-1)}$ is the precision matrix over vertices $\{2, 3\}$.

Now we construct the subset $\widetilde{\mathcal{G}}_{p,d}$ with $p = 3k$ for $k = 1, 2, \ldots$, as follows. We randomly set the DAG structure over nodes $(3i-1), (3i)$ and $(3i+1)$ to one of the two configurations shown in the above figure. Therefore we have, $|\widetilde{\mathcal{G}}_{p,d}| = 2^{(p-1)/3}$. We generate matrices $\mathbf{B}(\beta)$ and $\mathbf{D}(v_1, v_2)$ as prescribed. The precision matrix block over the nodes $(3i-1), (3i)$, and $(3i+1)$, for $i \in [(p-1)/3]$, is given by (9), and all the other entries of the precision matrix are zeros. This proves our claim.

While the above constructions constructs a family of disconnected DAGs, with $d = 1$, it is easy to come up with subsets of DAGs that are connected and still satisfy the statement of the lemma. One such construction is shown below where $d = (p-1)/3$. The entries of the first row (and also the first column) of the precision matrix, for $i \in [(p-1)/3]$, are as follows:

$$\Omega_{1,1} = \frac{1}{v_0} + \frac{(p-1)b_0^2}{3v_1},\ \Omega_{1,3i-1} = -\frac{b_0}{v_1},\ \Omega_{1,3i} = \frac{b_0\beta}{v_1}.$$



As shown before, each triplet of nodes $(3i-1) \leftarrow (3i) \leftarrow (3i+1)$, for $i \in [(p-1)/3]$, can be oriented as $(3i-1) \leftarrow (3i) \rightarrow (3i+1)$ without changing the block of the precision matrix over the nodes $(3i-1), (3i)$ and $(3i+1)$, and the entries $\mathbf{\Omega}_{1,*}$ or $\mathbf{\Omega}_{*,1}$. □

Proof of Proposition 2. From (2) we have that $(\mathbf{I} - \mathbf{B})X = N$, and since $(\mathbf{I} - \mathbf{B})$ is invertible, $X = (\mathbf{I} - \mathbf{B})^{-1}N$. Therefore:

$$\mathbf{\Sigma} = \mathbb{E}\left[XX^T\right] = \mathbb{E}\left[(\mathbf{I} - \mathbf{B})^{-1}NN^T(\mathbf{I} - \mathbf{B})^{-T}\right] = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{D}(\mathbf{I} - \mathbf{B})^{-T}.$$

From which it follows that $\mathbf{\Omega} = (\mathbf{I} - \mathbf{B})^T\mathbf{D}^{-1}(\mathbf{I} - \mathbf{B})$, where $\mathbf{D}^{-1} = \mathbf{Diag}(1/\sigma_1^2, \ldots, 1/\sigma_p^2)$. From this the result for the entries of the precision matrix follows by sparsity pattern of $\mathbf{B}$. □

Proof of Proposition 3. From (5) we have that for a terminal vertex i, $\Omega_{i,i} = 1/\sigma_i^2$, while for a non-terminal vertex j, $\Omega_{j,j} = 1/\sigma_j^2 + \sum_{l \in \phi(j)} B_{l,j}^2/\sigma_l^2$. Therefore, by Assumption 1 we have that for all non-terminal vertices j and terminal vertices i, $\Omega_{j,j} > \Omega_{i,i}$.

Now since every DAG has at least one terminal vertex, if $i \in \operatorname{argmin}(\mathbf{diag}(\mathbf{\Omega}))$, then once again by Assumption 1, we have that i must be a terminal vertex. □

Proof of Lemma 2. First note that since i is a terminal vertex, the autoregression matrix over $X_{-\mathsf{i}}$ is simply $\mathbf{B}_{-\mathsf{i},-\mathsf{i}}$. Therefore, denoting $\mathbf{D}' \stackrel{\text{def}}{=} \mathbf{Diag}(\sigma_1^2, \ldots, \sigma_{i-1}^2, \sigma_{i+1}^2, \sigma_p^2)$ and by Proposition 2 we have:

$$\begin{aligned}
\mathbf{\Omega}_{(-\mathsf{i})} &= (\mathbf{I} - \mathbf{B}_{-\mathsf{i},-\mathsf{i}})^T(\mathbf{D}')^{-1}(\mathbf{I} - \mathbf{B}_{-\mathsf{i},-\mathsf{i}}) = \sum_{j \in -\mathsf{i}} \frac{1}{\sigma_j^2}((\mathbf{e}_j)_{-\mathsf{i}} - \mathbf{B}_{j,-\mathsf{i}}^T)((\mathbf{e}_j)_{-\mathsf{i}}^T - \mathbf{B}_{-\mathsf{i},j}) \\
&= \sum_{j \in [p]} \frac{1}{\sigma_j^2}\left((\mathbf{e}_j - \mathbf{B}_{j,*}^T)(\mathbf{e}_j^T, -\mathbf{B}_{j,*})\right)_{-\mathsf{i},-\mathsf{i}} - \frac{1}{\sigma_i^2}\left((\mathbf{e}_i - \mathbf{B}_{i,*}^T)(\mathbf{e}_i^T - \mathbf{B}_{i,*})\right)_{-\mathsf{i},-\mathsf{i}} \\
&= \mathbf{\Omega}_{-\mathsf{i},-\mathsf{i}} - \frac{1}{\sigma_i^2}\left(\mathbf{B}_{i,-\mathsf{i}}^T\mathbf{B}_{i,-\mathsf{i}}\right) = \mathbf{\Omega}_{-\mathsf{i},-\mathsf{i}} - \Omega_{i,i}\frac{\mathbf{\Omega}_{i,-\mathsf{i}}^T}{\Omega_{i,i}}\frac{\mathbf{\Omega}_{i,-\mathsf{i}}}{\Omega_{i,i}} = \mathbf{\Omega}_{-\mathsf{i},-\mathsf{i}} - \frac{1}{\Omega_{i,i}}\mathbf{\Omega}_{-\mathsf{i},i}\mathbf{\Omega}_{i,-\mathsf{i}},
\end{aligned}$$

where in the last line we used the fact that for a terminal vertex $\Omega_{i,i} = 1/\sigma_i^2$ (Proposition 3), and $\mathbf{B}_{i,-\mathsf{i}} = -\mathbf{\Omega}_{i,-\mathsf{i}}/\Omega_{i,i}$ (Proposition 4). □

Proof of Lemma 3. First consider the case when $j \notin \pi_{\mathsf{G}}(i)$. Then, for any $k \in [p] \setminus \{i, j\}$, $i \notin (\phi_{\mathsf{G}}(j) \cap \phi_{\mathsf{G}}(k))$. Therefore, by Proposition 2, $(\Omega_{(-\mathsf{i})})_{j,k} = \Omega_{j,k}$, and by symmetry of the precision matrix $(\Omega_{(-\mathsf{i})})_{k,j} = \Omega_{k,j}$. Thus, we have that for any (j, k) if at least one of $\{j, k\}$ is not in $\pi_{\mathsf{G}}(i)$, then $(\Omega_{(-\mathsf{i})})_{j,k} = \Omega_{j,k}$, which proves our first claim. Thus, the only remaining case to consider is when both $j, k \in \pi_{\mathsf{G}}(i)$. The are two ways is which the set $\mathcal{S}((\mathbf{\Omega}_{(-\mathsf{i})})_{j,*})$ can be larger than the set $\mathcal{S}(\mathbf{\Omega}_{j,*})$, i.e., the support set of the j-th node can increase after deleting the terminal node i. The first being when $j, k \in \pi_{\mathsf{G}}(i)$ and either $(j, k) \in \mathsf{E}$ or $(k, j) \in \mathsf{E}$ but $\Omega_{j,k} = 0$, in which case we have:

$$\sum_{l \in \phi(j) \cap \phi(k)} (B_{l,j} B_{l,k})/\sigma_l^2 = B_{j,k}/\sigma_j^2 + B_{k,j}/\sigma_k^2.$$

Then, after removing the terminal node i, we have

$$(\Omega_{(-\mathsf{i})})_{j,k} = -B_{j,k}/\sigma_j^2 - B_{k,j}/\sigma_k^2 + \sum_{l \in (\phi(j) \cap \phi(k) \setminus \{i\})} (B_{l,j} B_{l,k})/\sigma_l^2 \neq 0.$$

The other case is when $j, k \in \pi_{\mathsf{G}}(i)$, $(j, k) \notin \mathsf{E}$, $(k, j) \notin \mathsf{E}$ but $\Omega_{j,k} = 0$, in which case we have:

$$\sum_{l \in \phi(j) \cap \phi(k)} (B_{l,j} B_{l,k})/\sigma_l^2 = 0.$$

Therefore, after removing the terminal node we have:

$$(\Omega_{(-\mathsf{i})})_{j,k} = \sum_{l \in (\phi(j) \cap \phi(k) \setminus \{i\})} (B_{l,j} B_{l,k})/\sigma_l^2 \neq 0.$$

Thus, $\mathcal{S}((\mathbf{\Omega}_{(-\mathsf{i})})_{j,*}) \subseteq (\mathcal{S}(\mathbf{\Omega}_{j,*}) \setminus \{i\}) \cup \pi_{\mathsf{G}}(i)$. □

Proof of Theorem 1. Let i_t be the terminal vertex identified in iteration t, $\mathcal{I}_t \stackrel{\text{def}}{=} \{i_1, \ldots, i_t\}$ and $\mathcal{R}_t \stackrel{\text{def}}{=} [p] \setminus \mathcal{I}_t$. Let $\mathbf{\Omega}_{(i)}$ be the precision matrix after iteration i. The correctness of the algorithm follows from the following loop invariants:

(i) By Lemma 2 we have that, $(\mathbf{\Omega}_{(t)})_{\mathcal{R}_t, \mathcal{R}_t}$ is the correct precision matrix over $X_{\mathcal{R}_t}$.

(ii) The algorithm identifies a correct terminal vertex in iteration t, since $(\mathbf{\Omega}_{(t-1)})_{\mathcal{R}_{t-1}, \mathcal{R}_{t-1}}$ is the correct precision matrix over $X_{\mathcal{R}_{t-1}}$, the SEM over $X_{\mathcal{R}_{t-1}}$ satisfies Assumption 1 by definition, and $\forall i \in \mathcal{I}_{t-1}$, $\Omega_{i,i} = \infty$.

(iii) By proposition 3 we have that at the end of round t, the sub-matrix $\mathbf{B}_{\mathcal{I}_t, *}$ has been correctly set and that $\forall i \in \mathcal{I}_t$, $\pi_{\mathsf{G}}(i) = \mathcal{S}(\mathbf{B}_{i,*})$.

To see that the algorithm returns a unique autoregression matrix $\widehat{\mathbf{B}}$, consider the following. If at iteration t there is a unique minimizer of $\mathbf{diag}(\mathbf{\Omega}_{(t-1)})$, which implies a single terminal vertex, then the algorithm selects it and the incoming edge weights of the node is uniquely determined. While, in iteration t if there are multiple terminal vertices, leading to multiple minimizers of $\mathbf{diag}(\mathbf{\Omega}_{(t-1)})$, then the order in which they are eliminated does not matter. Or in other words, once a vertex becomes a terminal vertex, for instance after deletion of its children, its edge weights do not change. To see this, assume that there are two terminal vertices, i and j after iteration $t-1$. Then i and j are not in each other's parent sets. Therefore, if node i is eliminated in iteration t, then by Lemma 3 we have that $(\Omega_{(t)})_{j,k} = (\Omega_{(t-1)})_{j,k}$, $\forall k \in \pi_{\mathsf{G}}(j)$. Hence, we have that $\mathbf{B}$ is the unique autoregression matrix returned by the algorithm. □

Proof of Lemma 5. Let $\mathbf{\Omega}_{(-\mathsf{i})} = (\boldsymbol{\omega}_j)_{j\in-\mathsf{i}}$ be the true precision matrix over $X_{-\mathsf{i}}$ and let $\widehat{\mathbf{\Omega}}' = (\boldsymbol{\omega}'_j)_{j\in[p]}$ be the matrix returned by the function UPDATE. The estimator $\widehat{\mathbf{\Omega}}_{(-\mathsf{i})} = (\widehat{\boldsymbol{\omega}}_j)_{j\in-\mathsf{i}}$ of $\mathbf{\Omega}_{(-\mathsf{i})}$ can be obtained by solving (7) using $\mathbf{\Sigma}^n_{-\mathsf{i},-\mathsf{i}}$. By Lemma 4, and the facts that $|\mathbf{\Sigma}^n_{-\mathsf{i},-\mathsf{i}} - \mathbf{\Sigma}_{-\mathsf{i},-\mathsf{i}}|_\infty \leq |\mathbf{\Sigma}^n - \mathbf{\Sigma}|_\infty$ and $\|\mathbf{\Omega}_{(\mathsf{i})}\|_1 \leq M$, we have that $|\mathbf{\Omega}_{(-\mathsf{i})} - \widehat{\mathbf{\Omega}}_{(-\mathsf{i})}| \leq 4M\lambda_n$. Since i is a terminal vertex, by Proposition 4 we have $\pi_{\mathsf{G}}(i) = \mathcal{S}(\mathbf{\Omega}_{i,*}) \setminus \{i\}$. Further, since $\mathcal{S}(\mathbf{\Omega}_{j,*}) \subseteq \mathcal{S}(\widehat{\mathbf{\Omega}}_{j,*})$, $\forall j \in [p]$, we have by Assumption 2 (ii) that, $\pi_{\mathsf{G}}(i) \subseteq \widehat{\pi}(i) = \mathcal{S}(\widehat{\mathbf{\Omega}}_{i,*}) \setminus \{i\} \subseteq \widehat{\mathsf{S}}$. By Lemma 3 and Assumption 2 (ii) we have that $\forall j \in \widehat{\mathsf{S}}_j$, $\mathcal{S}(\boldsymbol{\omega}_j) \subseteq \mathcal{S}\left(\mathbf{\Omega}_{j,*} \setminus \{i\}\right) \cup \pi_{\mathsf{G}}(i) \subseteq \mathcal{S}\left(\widehat{\mathbf{\Omega}}_{j,*} \setminus \{i\}\right) \cup \widehat{\pi}(i) \stackrel{\text{def}}{=} \widehat{\mathsf{S}}_j$. Or in other words we have $\left(\mathbf{\Omega}_{(i)}\right)_{j,\widehat{\mathsf{S}}_j^c} = \left(\mathbf{\Omega}_{(i)}\right)_{\widehat{\mathsf{S}}_j^c,j} = \mathbf{0}$. Now for $j \in -\mathsf{i}$ we set $(\boldsymbol{\omega}'_j)_{\widehat{\mathsf{S}}_j} = \bar{\boldsymbol{\omega}}_j$ and $(\boldsymbol{\omega}'_j)_{\widehat{\mathsf{S}}_j^c} = \mathbf{0}$, where $\bar{\boldsymbol{\omega}}_j$ is obtained by solving:

$$\begin{array}{ll} \underset{\boldsymbol{\omega}\in\mathbb{R}^{|\widehat{\mathsf{S}}_j|}}{\text{argmin}} & \|\boldsymbol{\omega}\|_1, \\ \text{sub. to} & \left|\mathbf{\Sigma}^n_{k,\widehat{\mathsf{S}}_j}\boldsymbol{\omega}\right| \leq \lambda_n,\ \forall k \notin \{i,j\}, \\ & \left|\mathbf{\Sigma}^n_{j,\widehat{\mathsf{S}}_j}\boldsymbol{\omega} - 1\right| \leq \lambda_n. \end{array}$$

Since $\bar{\boldsymbol{\omega}}_j$ is a solution to the above linear program, we have that $|\mathbf{\Sigma}^n_{-\mathsf{i},-\mathsf{i}}\boldsymbol{\omega}'_j - \mathbf{e}_j| \leq \lambda_n$ and $\|\boldsymbol{\omega}'_j\|_1 \leq \|\widehat{\boldsymbol{\omega}}_j\|_1$. Therefore, $|\mathbf{\Omega}_{(-\mathsf{i})} - \widehat{\mathbf{\Omega}}'_{-\mathsf{i},-\mathsf{i}}| \leq 4M\lambda_n$. Moreover, by Assumption 2 (ii), and the fact that $\widehat{\mathbf{\Omega}}'_{i,*} = \widehat{\mathbf{\Omega}}'_{*,i} = \mathbf{0}$, we get: $\mathcal{S}(\mathbf{\Omega}_{(-\mathsf{i})}) \subseteq \mathcal{S}(\widehat{\mathbf{\Omega}}')$. □

Proof of Theorem 2. Let i_t denote the terminal vertex identified in iteration t and let $\mathcal{I}_t \stackrel{\text{def}}{=} \{i_1, \ldots, i_t\}$. Let $\mathcal{R}_t \stackrel{\text{def}}{=} [p] \setminus \mathcal{I}_t$ denote the vertices remaining after iteration t. Let $\widehat{\mathbf{\Omega}}_{(t)}$ denote the precision matrix at the end of iteration t, $\widehat{\mathbf{\Omega}}_{(\mathcal{R}_t)} \stackrel{\text{def}}{=} (\widehat{\mathbf{\Omega}}_{(t)})_{\mathcal{R}_t,\mathcal{R}_t}$, and $\mathbf{\Omega}^*_{(\mathcal{R}_t)}$ be the true precision matrix over $X_{\mathcal{R}_t}$. Since $\|\mathbf{\Omega}^*\|_1 \leq M$, where M is defined in (8), we have that $\lambda_n \geq M|\mathbf{\Sigma}^n - \mathbf{\Sigma}^*|_\infty \geq \|\mathbf{\Omega}^*\|_1|\mathbf{\Sigma}^n - \mathbf{\Sigma}^*|_\infty$. Therefore, by Lemma 4 and Assumption 2 (ii), we have that $\left|\widehat{\mathbf{\Omega}}_{(\mathcal{R}_0)} - \mathbf{\Omega}^*_{(\mathcal{R}_0)}\right|_\infty = |\widehat{\mathbf{\Omega}} - \mathbf{\Omega}^*|_\infty \leq 4M\lambda_n$, and $\mathcal{S}(\mathbf{\Omega}^*_{(\mathcal{R}_0)}) \subseteq \mathcal{S}(\widehat{\mathbf{\Omega}}_0)$. Therefore, by Assumption 2 we have that the Algorithm 1 identifies the correct terminal vertex in iteration 1. Therefore, by Lemma 5 we have that $\left|\mathbf{\Omega}^*_{(\mathcal{R}_{t_1})} - \widehat{\mathbf{\Omega}}_{(\mathcal{R}_{t_1})}\right| \leq 4M\lambda_n$ and $\mathcal{S}(\mathbf{\Omega}^*_{(\mathcal{R}_{t_1})}) \subseteq \widehat{\mathbf{\Omega}}_{(t_1)}$.

Let $\mathbf{E} = (\varepsilon_{i,j})$, where $\varepsilon_{i,j} = \Omega^*_{i,j} - \widehat{\Omega}_{i,j}$. To simplify notation in this paragraph, we will denote the i_1 vertex by simply i. Then, for any $j \neq i$, we have that

$$\begin{aligned} |\widehat{B}_{i,j} - B^*_{i,j}| &= \left|\frac{\widehat{\Omega}_{i,j}}{\widehat{\Omega}_{i,i}} - \frac{\Omega^*_{i,j}}{\Omega^*_{i,i}}\right| = \left|\frac{\Omega^*_{ii}(\Omega^*_{i,j} - \varepsilon_{i,j}) - (\Omega^*_{i,i} - \varepsilon_{i,i})\Omega^*_{i,j}}{(\Omega^*_{i,i} - \varepsilon_{i,i})\Omega^*_{i,i}}\right| \\ &= \left|\frac{\Omega^*_{i,i}\varepsilon_{i,j} - \Omega^*_{i,j}\varepsilon_{i,i}}{(\Omega^*_{i,i} - \varepsilon_{i,i})\Omega^*_{i,i}}\right| = \left|\frac{\varepsilon_{i,i} - \sigma_i^2\Omega^*_{i,j}\varepsilon_{i,i}}{1/\sigma_i^2 - \varepsilon_{i,i}}\right| \\ &= \left|\frac{\varepsilon_{i,i} - B^*_{i,j}\varepsilon_{i,i}}{1/\sigma_i^2 - \varepsilon_{i,i}}\right| \\ &\leq \frac{4M\lambda_n(1 + |B^*_{i,j}|)}{|1/\sigma_i^2 - \varepsilon_{i,i}|} \leq 4cM(1 + |B^*_{i,j}|)\sigma_i^2\lambda_n, \end{aligned}$$

where the second and third lines follow from the fact that i is a terminal vertex and therefore, $\Omega^*_{i,i} = 1/\sigma_i^2$ and $\Omega_{i,j} = -B_{i,j}/\sigma_i^2$. Therefore, we have that $|\mathbf{B}^*_{i_1,*} - \widehat{\mathbf{B}}_{i_1,*}|_\infty = 4cM(1 + B_{\max})\sigma^2_{\max}\lambda_n$.

Next, assume that the algorithm correctly identifies terminal vertices upto round t. Then $|\widehat{\mathbf{\Omega}}_{(\mathcal{R}_t)} - \mathbf{\Omega}^*_{(\mathcal{R}_t)}|_\infty \leq 4M\lambda_n$, $\mathcal{S}(\mathbf{\Omega}^*_{(\mathcal{R}_t)}) \subseteq \mathcal{S}(\widehat{\mathbf{\Omega}}_{(t)})$, and $|\mathbf{B}^*_{\mathcal{I}_t,\mathcal{I}_t} - \widehat{\mathbf{B}}_{\mathcal{I}_t,\mathcal{I}_t}| \leq 4cM(1+B_{\max})\sigma^2_{\max}\lambda_n$. Therefore, once again by Assumption 2, it follows that the algorithm identifies the correct terminal vertex in round $t+1$, $|\widehat{\mathbf{\Omega}}_{(\mathcal{R}_{t+1})} - \mathbf{\Omega}^*_{(\mathcal{R}_{t+1})}|_\infty \leq 4M\lambda_n$, $\mathcal{S}(\mathbf{\Omega}^*_{(\mathcal{R}_{t+1})}) \subseteq \mathcal{S}(\widehat{\mathbf{\Omega}}_{(t+1)})$, and $|\mathbf{B}^*_{\mathcal{I}_{t+1},\mathcal{I}_{t+1}} - \widehat{\mathbf{B}}_{\mathcal{I}_{t+1},\mathcal{I}_{t+1}}| \leq 4cM(1+B_{\max})\sigma^2_{\max}\lambda_n$. Hence, the final claim follows by induction. The claim that $\mathcal{S}(\mathbf{B}^*) \subseteq \mathcal{S}(\widehat{\mathbf{B}})$ follows from the fact that $\mathcal{S}(\mathbf{\Omega}^*) \subseteq \mathcal{S}(\widehat{\mathbf{\Omega}})$. Finally, since $\mathcal{S}(\mathbf{B}^*) \subseteq \mathcal{S}(\widehat{\mathbf{B}})$ implies that $\mathcal{T}_{\widehat{\mathsf{G}}} \subseteq \mathcal{T}_{\mathsf{G}^*}$. □

Proof of Theorem 3. Given that the data was generated by the SEM $(\mathsf{G}^*, \mathbf{B}^*, \{\sigma_i^2\})$, each X_i can be written as follows:

$$X_i = \sum_{j \in \mathsf{A}_{\mathsf{G}^*}(i)} w_{i,j} N_j,$$

for some $w_{i,j} \geq 0$. N_i is sub-Gaussian with parameter $\sigma_i \nu$, X_i is sub-Gaussian with parameter $\nu\sqrt{\sum_{j \in \mathsf{A}_{\mathsf{G}^*}(i)} w_{i,j}^2 \sigma_i^2}$ and $\Sigma^*_{i,i} = \sum_{j \in \mathsf{A}_{\mathsf{G}^*}(i)} w_{i,j}^2 \sigma_i^2$. Therefore, it follows that ${}^{X_i}/_{\sqrt{\Sigma^*_{i,i}}}$ is sub-Gaussian with parameter ν. From Lemma 1 of [18] and Theorem 2 we have that the regularization parameter λ_n need to satisfy the following bound in order to guarantee that $|\widehat{\mathbf{B}} - \mathbf{B}^*|_\infty \leq \varepsilon$:

$$MC_1\sqrt{\frac{2}{n}\log\left(\frac{2p}{\sqrt{\delta}}\right)} \leq \lambda_n \leq \frac{\varepsilon}{c4M(1+B_{\max})\sigma^2_{\max}}. \tag{10}$$

The above holds in the regime where the number of samples scales as given in the statement of the Theorem. □

Proof of Theorem 4. Given that the data was generated by the SEM $(\mathsf{G}^*, \mathbf{B}^*, \{\sigma_i^2\})$, each X_i can be written as follows:

$$X_i = \sum_{j \in \mathsf{A}_{\mathsf{G}^*}(i)} w_{i,j} N_j,$$

for some $w_{i,j} \geq 0$. Now,

$$\left(\sqrt{\mathbf{\Sigma}^*_{i,i}}\right)^{4m} = \left(\sum_{j \in \mathsf{A}_{\mathsf{G}^*}(i)} w_{i,j}^2 \sigma_i^2\right)^{2m} \geq \sum_{j \in \mathsf{A}_{\mathsf{G}^*}(i)} (w_{i,j}\sigma_i)^{4m} \tag{11}$$

Now, by Rosenthal's inequality we have:

$$\begin{aligned}
\mathbb{E}\left[(X_i)^{4m}\right] &\leq C_m \left\{ \sum_{j \in \mathsf{A}_{\mathsf{G}^*}(i)} w_{i,j}^{4m} \mathbb{E}\left[N_j^{4m}\right] + \sum_{j \in \mathsf{A}_{\mathsf{G}^*}(i)} w_{i,j}^{4m} \mathrm{Var}\left[N_i\right]^{2m} \right\} \\
&\leq C_m \left\{ \sum_{j \in \mathsf{A}_{\mathsf{G}^*}(i)} w_{i,j}^{4m} \sigma_i^{4m} K_m + \sum_{j \in \mathsf{A}_{\mathsf{G}^*}(i)} w_{i,j}^{4m} \sigma_i^{4m} \right\} \\
&= C_m(K_m+1) \sum_{j \in \mathsf{A}_{\mathsf{G}^*}(i)} (w_{i,j}\sigma_i)^{4m}
\end{aligned} \tag{12}$$

Combining (11) and (12) we have

$$\mathbb{E}\left[\left(\frac{X_i}{\sqrt{\Sigma^*_{i,i}}}\right)^{4m}\right] \leq C_m(K_m+1). \tag{13}$$

From the above and invoking Lemma 2 of [18] we get:

$$|\mathbf{\Sigma}^n - \mathbf{\Sigma}^*|_\infty < C_2 \left(\frac{p^2}{n^m \delta}\right)^{1/2m}, \tag{14}$$

with probability at least $1-\delta$. From Theorem 2 and (14) we have that the regularization parameter λ should satisfy the following for $|\widehat{\mathbf{B}} - \mathbf{B}^*|_\infty \leq \varepsilon$ to hold:

$$MC_2 \left(\frac{p^2}{n^m \delta}\right)^{1/2m} \leq \lambda_n \leq \frac{\varepsilon}{c4M(1+B_{\max})\sigma^2_{\max}}. \tag{15}$$

The above holds in the regime where the number of samples scales as given in the statement of the Theorem. □

Proposition 6. *Let* $(\mathsf{G}, \mathbf{B}, \{\sigma_i^2\})$ *be an SEM over* X *with* $\mathsf{G} \in \mathcal{G}_{p,d}$ *and precision matrix* $\mathbf{\Omega}$*. Then,* $|\mathcal{S}(\mathbf{\Omega}_{i,*}) \setminus \{i\}| \leq d^2$, $\forall i \in [p]$.

Proof of Proposition 6. For any node i, we will define the following set: $\mathsf{S}_\mathsf{G}(i) = \{j \in -\mathsf{i} \mid (i,j) \notin \mathsf{E} \wedge (j,i) \notin \mathsf{E} \wedge |\Omega_{i,j}| \neq 0\}$. Then, from Proposition 2, we have: if $j \in \mathsf{S}_\mathsf{G}(i)$ then $\Omega_{i,j} = \sum_{l \in \phi(i) \cap \phi(j)} {}^{(B_{l,i}B_{l,j})}/_{\sigma_l^2} \neq 0$. In other words, if $j \in \mathsf{S}_\mathsf{G}(i)$ then i and j have at least one common child, i.e., $\phi_\mathsf{G}(i) \cap \phi_\mathsf{G}(j) \neq \varnothing$. Node i can have at most d children, and each child $k \in \phi_\mathsf{G}(i)$ can have at most $d-1$ parents other than i making them all members of $\mathsf{S}(i)$. Thus, $\mathsf{S}(i) \leq d(d-1)$. Therefore, we have that $\mathcal{S}(\mathbf{\Omega}_{i,*}) \setminus \{i\} \subseteq \mathsf{N}_\mathsf{G}(i) \cup \mathsf{S}_\mathsf{G}(i)$. Then, using the inclusion-exclusion principle we have that:

$$|\mathcal{S}(\mathbf{\Omega}_{i,*}) \setminus \{i\}| \leq |\mathsf{N}_\mathsf{G}(i)| + |\mathsf{S}_\mathsf{G}(i)| - |\mathsf{N}_\mathsf{G}(i) \cap \mathsf{S}_\mathsf{G}(i)| = |\mathsf{N}_\mathsf{G}(i)| + |\mathsf{S}_\mathsf{G}(i)| \leq d + d(d-1) = d^2.$$

The SEM which achieves the above upper bound is precisely the one constructed in the proof, i.e., there exists a node i with exactly d children, each child in turn has $d-1$ "other parents" which are all members of $\mathsf{S}_\mathsf{G}(i)$. □

References

[1] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9(Mar):485–516, 2008.

[2] Tony Cai, Weidong Liu, and Xi Luo. A Constrained L1 Minimization Approach to Sparse Precision Matrix Estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

[3] David Maxwell Chickering. Learning bayesian networks is np-complete. In *Learning from data*, pages 121–130. Springer, 1996.

[4] David Maxwell Chickering. Optimal Structure Identification with Greedy Search. *J. Mach. Learn. Res.*, 3:507–554, March 2003.

[5] Sanjoy Dasgupta. Learning polytrees. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 134–141. Morgan Kaufmann Publishers Inc., 1999.

[6] John Dunagan, Daniel A. Spielman, and Shang-Hua Teng. Smoothed analysis of condition numbers and complexity implications for linear programming. *Mathematical Programming*, 126(2):315–350, February 2011.

[7] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[8] Asish Ghoshal and Jean Honorio. Information-theoretic limits of Bayesian network structure learning. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 767–775, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR.

[9] Cho-Jui Hsieh, Arindam Banerjee, Inderjit S Dhillon, and Pradeep K Ravikumar. A divide-and-conquer method for sparse inverse covariance estimation. In *Advances in Neural Information Processing Systems*, pages 2330–2338, 2012.

[10] Cho-Jui Hsieh, Màtyàs A Sustik, Inderjit S Dhillon, Pradeep Ravikumar, and Russell Poldrack. BIG & QUIC : Sparse Inverse Covariance Estimation for a Million Variables. In *Advances in Neural Information Processing Systems*, volume 26, pages 3165–3173, 2013.

[11] Tommi S. Jaakkola, David Sontag, Amir Globerson, Marina Meila, and others. Learning Bayesian Network Structure using LP Relaxations. In *AISTATS*, pages 358–365, 2010.

[12] Christopher C Johnson, Ali Jalali, and Pradeep Ravikumar. High-dimensional sparse inverse covariance estimation using greedy methods. In *AISTATS*, volume 22, pages 574–582, 2012.

[13] Markus Kalisch and Bühlmann Peter. Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. *Journal of Machine Learning Research*, 8:613–636, 2007.

[14] Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *arXiv:1311.3492 [math, stat]*, November 2013. arXiv: 1311.3492.

[15] Gunwoong Park and Garvesh Raskutti. Learning Quadratic Variance Function (QVF) DAG models via OverDispersion Scoring (ODS). *arXiv:1704.08783 [cs, stat]*, April 2017. arXiv: 1704.08783.

[16] J. Peters and P. Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

[17] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal Discovery with Continuous Additive Noise Models. *Journal of Machine Learning Research*, 15(June):2009–2053, 2014.

[18] Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5(0):935–980, 2011.

[19] R W Robinson. Counting unlabeled acyclic digraphs. *Combinatorial Mathematics V*, 622:28–43, 1977.

[20] Benjamin Rolfs, Bala Rajaratnam, Dominique Guillot, Ian Wong, and Arian Maleki. Iterative thresholding algorithm for sparse inverse covariance estimation. In *Advances in Neural Information Processing Systems*, pages 1574–1582, 2012.

[21] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

[22] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth Bollen. DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model. *Journal of Machine Learning Research*, 12(Apr):1225–1248, 2011.

[23] Daniel A. Spielman and Shang-Hua Teng. Smoothed analysis of termination of linear programming algorithms. *Mathematical Programming*, 97(1):375–404, 2003.

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

[25] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

[26] Sara Van De Geer and Peter Bühlmann. L0-Penalized maximum likelihood for sparse directed acyclic graphs. *Annals of Statistics*, 41(2):536–567, 2013.

[27] Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[28] Jiji Zhang and Peter Spirtes. Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 632–639. Morgan Kaufmann Publishers Inc., 2002.

[29] Jiji Zhang and Peter Spirtes. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18(2):239–271, 2008.