Learning linear structural equation models in polynomial time and sample complexity

Asish Ghoshal
aghoshal@purdue.edu

Jean Honorio
jhonorio@purdue.edu

Department of Computer Science
Purdue University, West Lafayette, IN, U.S.A. - 47907.

Abstract

The problem of learning structural equation models (SEMs) from observational 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 Markov blanket size 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 $\mathcal{O}(p(d + \log p))$ operations. In the finite sample setting, if the estimated precision matrix is sparse, our algorithm has a smoothed complexity of $\widetilde{\mathcal{O}}\left(p^3 + pd^4\right)$, 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 and bounded ($4m$-th, m being a positive integer) moment noise, our algorithm has a sample complexity of $\mathcal{O}(\frac{d^4}{\varepsilon^2}\log(\frac{p}{\sqrt{\delta}}))$ and $\mathcal{O}(\frac{d^4}{\varepsilon^2}(\frac{p^2}{\delta})^{1/m})$ resp., to achieve ε element-wise additive error with respect to the true *autoregression matrix* with probability at least $1 - \delta$.

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

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. Structural equation models (SEMs) is a commonly employed mathematical machinery 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 contributions. 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 [PB14]. 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 [LB13]. This disproves an earlier conjecture by [LB13] that "variance scaling or non-Gaussianity is necessary in order to guarantee identifiability" of linear SEMs. Moreover, we show that our identifiability condition is necessary for ensuring identifiability of linear SEMs, in the sense that, if the identifiability condition is violated then there exist an exponential number of DAGs which induce the same covariance and precision matrix, and specify distributions that have the same conditional independence structures.

To the best of our knowledge, ours is the first method for

learning SEMs with element-wise ℓ_∞ guarantees for recovering the autoregression matrix — the matrix of (directed) edge weights of the SEM. In contrast, score based approaches [VDGB13, LB13] have guarantees on the score of the learned DAG structure. An unfortunate consequence of this is that, in order for these methods to recover the true DAG structure by finding the highest scoring DAG structure on the sample data set, the "score gap" between the true structure and the next best structure must scale as $\Omega(p)$ (see Equation 27 in [LB13]), which is unreasonable since the best DAG structure and the next best DAG structure might only differ on a constant number of edges, in which case the scores might differ by $o(p)$.

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 $\mathcal{O}(p(d+\log p))$ floating-point operations. In contrast [LB13]'s algorithm takes $\mathcal{O}\left(p2^{2(w+1)(w+d)}\right)$ time in the population setting, where w is the tree-width and d is the maximum degree of the graph. 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 dimensions, leading to an overall smoothed complexity of $\widetilde{\mathcal{O}}\left(p^3+pd^4\right)$. 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 [PB14]'s algorithm for learning linear Gaussian SEMs as well as [LB13]'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 $n \ll p$ and $d = o(p)$, and has a sample complexity of $\mathcal{O}(\frac{d^4}{\varepsilon^2}\log(\frac{p}{\sqrt{\delta}}))$ and $\mathcal{O}(\frac{d^4}{\varepsilon^2}(\frac{p^2}{\delta})^{1/m})$ 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 [LB13]'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 [LB13]'s algorithm, and other methods that use conditional independence tests, for instance, the PC algorithm for learning Gaussian SEMs [KP07], our algorithm does not require any faithfulness conditions, and only requires a weaker *causal minimality* condition. The PC algorithm and [LB13]'s algorithm can fail to recover the correct DAG for distributions that are not faithful to the DAG structure. Our results have the following significant yet hitherto unknown 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.

2 Related Work

We start our discussion of existing literature by first presenting known identifiability conditions for learning SEMs and Bayesian networks. [PMJS14] 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 [SHHK06], (b) the noise variances are known up to a constant factor [LB13], and (c) noise variables are Gaussian and have the same variance [PB14] (homoscedastic noise). [PR17] 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 in general, 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 [Chi96, Das99]. 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 [Rob77, GH17a] brute force methods, and existing 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 [VDGB13]. [PB14] 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. [LB13] showed that under a faithfulness assumption, the sparsity pattern of the preci-

sion 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. [KP07] proposed using the PC algorithm, which was originally proposed by [SGS00] and has a computational complexity of $\mathcal{O}\left(p^d\right)$, for learning Gaussian SEMs and proved asymptotic uniform consistency of the algorithm for recovering the Markov equivalence class, i.e., a CPDAG. However the PC algorithm is only efficient for learning very sparse Gaussian SEMs. Among computationally efficient algorithms, the *Direct-LiNGAM* algorithm [SIS$^+$11], 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* [PMJS14], which is a computationally efficient algorithm for learning *non-linear SEMs*, is only guaranteed in the population setting. [GH17b] proposed a polynomial time algorithm, similar to the one proposed in this paper, for learning Gaussian SEMs (or Gaussian Bayesian networks) with a sample complexity of $\mathcal{O}\left(d^4 \log p\right)$. However, their method, theoretical guarantees and proofs crucially rely on the Gaussianity of the data distribution.

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 [TBA06], and the Greedy Equivalence Search (GES) algorithm proposed by [Chi03]. [JSG$^+$10] 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. 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 $\mathsf{A}_{\mathsf{G}}(i)$. The Markov blanket of a node is defined as: $\mathsf{MB}_{\mathsf{G}}(i) = \mathsf{N}_{\mathsf{G}}(i) \cup \{k \in \pi_{\mathsf{G}}(j) \mid j \in \phi_{\mathsf{G}}(i)\}$.

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 $(\mathsf{G}, \mathbf{B}, \{\sigma_i^2\})$ [1].

[1] An SEM is fully characterized by G, $\mathbf{B}$ and the distribution of the exogenous variables. However, since we are concerned with learning SEMs using second moments only, our notation captures all the required information.

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 DAGs with Markov blanket at most d: $\mathcal{G}_{p,d} \stackrel{\text{def}}{=} \{\mathsf{G} \mid \mathsf{G} = ([p], \mathsf{E})$ is a DAG and $|\mathsf{MB}_{\mathsf{G}}(i)| \leq d, \forall i \in [p]\}$.

Next, we define the notion of *causal minimality*, introduced by [ZS08], 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 [PMJS14] 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 [PMJS14]). Therefore, the SEMs considered in the paper are causal minimal. Causal minimality is much weaker than faithfulness which requires that the distribution $\mathcal{P}(X)$ contain only those conditional independence assertions that are implied by the *d-separation* criteria of the DAG [SGS00]. However, faithfulness cannot be tested from data in full generality [ZS08] and algorithms that infer the DAG structure from a finite number of samples must require *strong faithfulness* [ZS02], which is a restrictive assumption.

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

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

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

For detailed proofs, see Appendix A. At this point one might ask if the above assumption is *necessary* for identifiability of linear SEMs. We answer this question in affirmative in the following lemma, which states that if Assumption 1 is violated, then there exists an exponential number of DAGs structures that induce the same covariance and precision matrix, and determine joint distributions $\mathcal{P}(X)$

that are causal minimal and Markov to the DAG structures. 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 for each $\beta \in (-\infty, \infty)$ and $v_1 \in (0, \infty)$ and $v_2 > v_1$, the SEMs $\{(\mathsf{G}, \mathbf{B}(\beta), \mathbf{D}(v_1, v_2)) \mid G \in \widetilde{\mathcal{G}}_{p,d}\}$, do not satisfy Assumption 1, induce the same covariance and precision matrix, and distribution $\mathcal{P}(X)$ that has the same conditional independence structure.*

Given that the true SEM can come from the aforementioned family, no algorithm, that uses only conditional independence tests and second moments, can consistently recover the true DAG structure if Assumption 1 is not satisfied. 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:*

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

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

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

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,*

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

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**
    $\widehat{\mathsf{S}}_j \leftarrow \left(\mathcal{S}(\widehat{\mathbf{\Omega}}_{j,*}) \setminus \{i\}\right) \cup \widehat{\pi}(i)$.
    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. In the finite sample setting, our algorithm involves estimating the inverse covariance matrix, and subsequently updating the inverse covariance matrix after removing a terminal vertex. Due in part to its role in undirected graphical model selection, the problem of inverse covariance matrix estimation has received significant attention and many algorithms have been developed for the problem. In this paper we use the CLIME algorithm, proposed by [CLL11], to estimate the inverse covariance matrix and propose a modification of the CLIME algorithm for efficiently computing the inverse covariance matrix over the variables remaining after eliminating a terminal vertex in Algorithm 2. For a discussion on why CLIME was preferred over other methods, see Appendix D.

The CLIME estimator, $\widehat{\mathbf{\Omega}}$, of the inverse covariance matrix $\mathbf{\Omega}$ 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, $\mathbf{\Sigma}^n \stackrel{\text{def}}{=} (1/n)\mathbf{X}^T\mathbf{X}$ 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 [CLL11] is given by the following lemma, which is a minor reformulation of Theorem 6 in [CLL11]:

Lemma 4 ([CLL11])**.** *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 $\Omega\left(d\sqrt{\log p/n}\right)$ and $\Omega\left(d(p)^{1/m}/\sqrt{n}\right)$ 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 [LB13] 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 $|\mathbf{B}^* - \widehat{\mathbf{B}}| \leq c4M(1 + B_{\max})\sigma^2_{\max}\lambda_n$, $\mathcal{S}(\mathbf{B}^*) \subseteq \mathcal{S}(\widehat{\mathbf{B}})$, and $\mathcal{T}_{\widehat{\mathsf{G}}} \subseteq \mathcal{T}_{\mathsf{G}^*}$, where $c \leq {\sigma^2_{\min}}/{(1-4M\lambda_n\sigma^2_{\min})}$ is a constant.*

Next, we use known concentration results for the empirical covariance matrix to obtain finite sample results for noise distributions satisfying the following conditions.

Assumption 3 (Noise conditions)**.** *For all $i \in [p]$, we have*
(i) ***Sub-Gaussian noise:*** *${N_i}/{\sigma_i}$ is sub-Gaussian with parameter ν.*
(ii) ***Bounded-moment noise:*** *$(\mathbb{E}\left[{N_i}/{\sigma_i}\right])^{4m} \leq K_m$, for a positive integer m and positive constant K_m.*

Theorem 3 (Sample complexity)**.** *If $\lambda_n \geq \eta_1(n, p, \delta)$ and $n \geq \eta_2(p, \varepsilon, \delta)$, then $|\widehat{\mathbf{B}} - \mathbf{B}^*| \leq \varepsilon$, with probability $1 - \delta$, where*

(i) for sub-Gaussian noise (Assumption 3(i)):

$$\eta_1(n, p, \delta) = MC_1\sqrt{({2}/{n})\log\left({2p}/{\sqrt{\delta}}\right)}$$
$$\eta_2(p, \varepsilon, \delta) = 2({C_1C}/{\varepsilon})^2\log\left({2p}/{\sqrt{\delta}}\right),$$

(ii) for bounded moment noise (Assumption 3(ii)):

$$\eta_1(n, p, \delta) = MC_2\left({p^2}/{(n^m\delta)}\right)^{1/2m}$$
$$\eta_2(p, \varepsilon, \delta) = ({C_2C}/{\varepsilon})^2({p^2}/{\delta})^{1/m}$$

with $C = c4M^2(1 + B_{\max})\sigma^2_{\max}$, $C_1 = \sqrt{128}(1 + 4\nu^2)(\max_i \mathbf{\Sigma}^*_{i,i})$, $C_2 = 2(\max_i \mathbf{\Sigma}^*_{i,i})(C_m(C_m(K_m+1)+1))^{1/2m}$, and c is defined in Theorem 2. 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

Next, 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 [LB13] 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 4. *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 [LB13]. 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:

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

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 4. *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 $\phi_{\mathsf{G}[m,\tau]}(i) \neq \varnothing$:

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

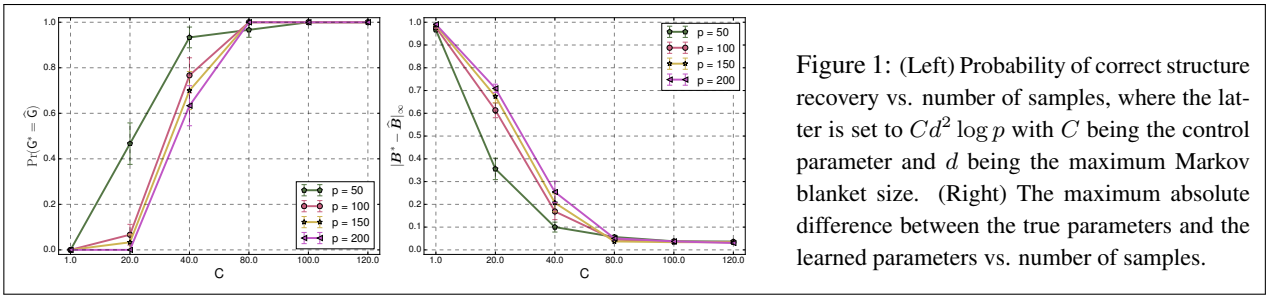

Figure 1: (Left) Probability of correct structure recovery vs. number of samples, where the latter is set to $Cd^2 \log p$ with C being the control parameter and d being the maximum Markov blanket size. (Right) The maximum absolute difference between the true parameters and the learned parameters vs. number of samples.

(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 Theorem 3 holds for SEMs with known error variances satisfying Assumption 4, with $\sigma^2_{\max}$ and $\sigma^2_{\min}$ replaced by $\alpha\sigma^2_{\max}$ and $\alpha\sigma^2_{\min}$, respectively. Thus, given a data set of n samples drawn from an SEM satisfying Assumption 4, with autoregression matrix $\mathbf{B}^*$ and DAG structure $\mathsf{G}^* = ([p], \mathsf{E}^*)$, we have the following results for sub-Gaussian and bounded-moment noise:

Remark 1. *If* $\lambda_n \geq \eta_1(n,p,\delta)$*, and* $n \geq \eta_2(p,\varepsilon,\delta)$*, then, under Assumption 4, Algorithm 1 with* $\mathbf{D} = \mathbf{Diag}(\{\sigma_i^2\})$ *returns an estimator* $\widehat{\mathbf{B}}$ *such that* $|\widehat{\mathbf{B}} - \mathbf{B}^*|_\infty \leq \varepsilon$*, with probability at least* $1-\delta$*, where for sub-Gaussian noise* $\eta_1(n,p,\delta) = \mathcal{O}\left((d/\sqrt{n})\sqrt{\log(p/\sqrt{\delta})}\right)$ *and* $\eta_2(p,\varepsilon,\delta) = \mathcal{O}\left((d^4/\varepsilon^2)\log(p/\sqrt{\delta})\right)$*, while for bounded moment noise* $\eta_1(n,p,\delta) = \mathcal{O}\left((d/\sqrt{n})(p/\sqrt{\delta})^{1/m}\right)$ *and* $\eta_2(p,\varepsilon,\delta) = \mathcal{O}\left((d^4/\varepsilon^2)(p^2/\delta)^{1/m}\right)$*. Further, thresholding* $\widehat{\mathbf{B}}$ *at the level* ε*, we have* $\mathcal{S}(\widehat{\mathbf{B}}) = \mathsf{E}^*$.

The above remark follows from the fact that $M = \mathcal{O}(d)$ which follows from Proposition 7 in Appendix A.

6 Experiments

Simulation Experiments. In this section, we validate our theoretical results through simulation experiments. We generate random SEMs by first sampling Erdős-Rényi random DAGs and then set all the noise variances to $\sigma^2 = 0.8$. Note this is a sufficient condition for ensuring identifiability (Proposition 1). We sample edge weights from the uniform distribution over $[-1,-0.5] \cup [0.5,1]$. To generate sub-Gaussian noise, we set the noise variables $N_i = \sigma_i R_i$, where R_i's are independent Rademacher random variables. We set the regularization parameter according to Theorem 3 and varied the number of samples as $Cd^2 \log p$, with C being the control parameter. Figure 1 shows the probability of correct structure recovery and the maximum absolute difference between the true edge weights and the learned edge weights, across 30 randomly sampled SEMs. Note that Theorem 3 indeed bears out in practice, and the results show a phase transition behavior for structure recovery.

Comparison with State-of-the-art Methods on Synthetic Data. We also compared the performance of our algorithm against three other state-of-the-art methods for learning SEMs, viz. MMHC [TBA06], GES [Chi03], and the PC algorithm [SGS00] on randomly generated SEMs. In the identifiable regime our method achieved perfect structure recovery (100% accuracy and recall). In the non-identifiable regime, where noise variances are uniformly generated between $[0.5, 1]$, our method manages to recover the structure almost perfectly achieving accuracy and recall values of around 96%. The next best method, which is the PC algorithm, only manages around 50% accuracy in both the regimes. See Appendix B.1 for more details.

Comparison with State-of-the-art Methods on Real-world Data Sets. Lastly, we also compared the performance of our method against the aforementioned methods on 7 real-world gene-expression data sets. Our method achieves the lowest average negative-log-likelihood, on the test set, across 10 bootstrap runs for all 7 data sets. Our method is also significantly faster. See Appendix B.2 for more details.

7 Discussion and Concluding Remarks

In the population setting, i.e., given the true precision matrix, our algorithm computes the $\widehat{\mathbf{B}}$ matrix in $\mathcal{O}(p(d + d\log p))$. In the finite sample setting, if the estimated precision matrix is sparse, which can be accomplished by thresholding, then our algorithm has a smoothed complexity of $\widetilde{\mathcal{O}}\left(p^3 + pd^4\right)$ using interior-point methods. In the dense case the complexity is $\widetilde{\mathcal{O}}\left(p^5\right)$. See Appendix C for more details.

One interesting possible line of future work would be to explore if some of the ideas developed herein can be extended to binary or discrete Bayesian networks. We believe our strategy of identifying terminal vertices can be incorporated into score-based methods to restrict the search space when the objective is to find the highest scoring structure on the sample data set.

References

[BGd08] 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.

[Chi96] David Maxwell Chickering. Learning bayesian networks is np-complete. In *Learning from data*, pages 121–130. Springer, 1996.

[Chi03] David Maxwell Chickering. Optimal Structure Identification with Greedy Search. *J. Mach. Learn. Res.*, 3:507–554, March 2003.

[CLL11] 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.

[Das99] Sanjoy Dasgupta. Learning polytrees. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 134–141. Morgan Kaufmann Publishers Inc., 1999.

[DST11] 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.

[FHT08] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[GH17a] 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.

[GH17b] Asish Ghoshal and Jean Honorio. Learning identifiable gaussian bayesian networks in polynomial time and sample complexity. *arXiv preprint arXiv:1703.01196*, 2017.

[HBDR12] 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.

[HSD+13] 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.

[JJR12] 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.

[JSG+10] Tommi S. Jaakkola, David Sontag, Amir Globerson, Marina Meila, and others. Learning Bayesian Network Structure using LP Relaxations. In *AISTATS*, pages 358–365, 2010.

[KP07] Markus Kalisch and Bühlmann Peter. Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. *Journal of Machine Learning Research*, 8:613–636, 2007.

[LB13] 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.

[PB14] J. Peters and P. Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

[PMJS14] 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.

[PR17] 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.

[Rob77] R W Robinson. Counting unlabeled acyclic digraphs. *Combinatorial Mathematics V*, 622:28–43, 1977.

[RRG+12] 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.

[RWRY11] 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.

[SGS00] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

[SHHK06] 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.

[SIS+11] 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.

[ST03] Daniel A. Spielman and Shang-Hua Teng. Smoothed analysis of termination of linear programming algorithms. *Mathematical Programming*, 97(1):375–404, 2003.

[TBA06] 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.

[VDGB13] Sara Van De Geer and Peter Bühlmann. L0-Penalized maximum likelihood for sparse directed acyclic graphs. *Annals of Statistics*, 41(2):536–567, 2013.

[YL07] Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[ZS02] 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.

[ZS08] Jiji Zhang and Peter Spirtes. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18(2):239–271, 2008.