Journal of Machine Learning Research 15 (2014) 3065-3105 Submitted 11/13; Revised 5/14; Published 10/14

High-Dimensional Learning of Linear Causal Networks via Inverse Covariance Estimation

Po-Ling Loh LOH@WHARTON.UPENN.EDU
Department of Statistics
The Wharton School
466 Jon M. Huntsman Hall
3730 Walnut Street
Philadelphia, PA 19104, USA

Peter Bühlmann BUHLMANN@STAT.MATH.ETHZ.CH
Seminar für Statistik
ETH Zürich
Rämistrasse 101, HG G17
8092 Zürich, Switzerland

Editor: Hui Zou

Abstract

We establish a new framework for statistical estimation of directed acyclic graphs (DAGs) when data are generated from a linear, possibly non-Gaussian structural equation model. Our framework consists of two parts: (1) inferring the moralized graph from the support of the inverse covariance matrix; and (2) selecting the best-scoring graph amongst DAGs that are consistent with the moralized graph. We show that when the error variances are known or estimated to close enough precision, the true DAG is the unique minimizer of the score computed using the reweighted squared ℓ_2-loss. Our population-level results have implications for the identifiability of linear SEMs when the error covariances are specified up to a constant multiple. On the statistical side, we establish rigorous conditions for high-dimensional consistency of our two-part algorithm, defined in terms of a "gap" between the true DAG and the next best candidate. Finally, we demonstrate that dynamic programming may be used to select the optimal DAG in linear time when the treewidth of the moralized graph is bounded.

Keywords: causal inference, dynamic programming, identifiability, inverse covariance matrix estimation, linear structural equation models

1. Introduction

Causal networks arise naturally in a wide variety of application domains, including genetics, epidemiology, and time series analysis (Hughes et al., 2000; Stekhoven et al., 2012; Aalen et al., 2012). However, inferring the graph structure of a causal network from joint observations is a rather challenging problem. Whereas undirected graphical structures may be estimated via pairwise conditional independence testing, with worst-case time scaling as the square of the number of nodes, estimation methods for directed acyclic graphs (DAGs) first require learning an appropriate permutation order of the vertices, leading to computational complexity that scales exponentially in the graph size. Greedy algorithms present an

attractive computationally efficient alternative, but such methods are not generally guaranteed to produce the correct graph (Chickering, 2002). In contrast, exact methods for causal inference that search exhaustively over the entire DAG space may only be tractable for relatively small graphs (Silander and Myllymaki, 2006).

1.1 Restricted Search Space

In practice, knowing prior information about the structure of the underlying DAG may lead to vast computational savings. For example, if a natural ordering of the nodes is known, inference may be performed by regressing each node upon its predecessors and selecting the best functional fit for each node. This yields an algorithm with runtime linear in the number of nodes and overall quadratic complexity. In the linear high-dimensional Gaussian setting, one could apply a version of the graphical Lasso, where the feasible set is restricted to matrices that are upper-triangular with respect to the known ordering (Shojaie and Michailidis, 2010). However, knowing the node order is unrealistic for many applications. If instead a conditional independence graph or superset of the skeleton is specified a priori, the number of required conditional independence tests may also be reduced dramatically. This appears to be a more reasonable assumption, and various authors have devised algorithms to compute the optimal DAG efficiently in settings where the input graph has bounded degree and/or bounded treewidth (Perrier et al., 2008; Ordyniak and Szeider, 2012; Korhonen and Parviainen, 2013).

Unfortunately, appropriate tools for inferring such superstructures are rather limited, and the usual method of using the graphical Lasso to estimate a conditional independence graph is rigorously justified only in the linear Gaussian setting (Yuan and Lin, 2007). Recent results have established that a version of the graphical Lasso may also be used to learn a conditional independence graph for variables taking values in a discrete alphabet when the graph has bounded treewidth (Loh and Wainwright, 2013), but results for more general distributions are absent from the literature. Bühlmann et al. (2014) isolate sufficient conditions under which Lasso-based linear regression could be used to recover a conditional independence graph for general distributions, and use the Lasso as a prescreening step for nonparametric causal inference in additive noise models; however, it is unclear which non-Gaussian distributions satisfy the prescribed conditions.

1.2 Our Contributions

We propose a new algorithmic strategy for inferring the DAG structure of a linear, potentially non-Gaussian structural equation model (SEM). Deviating slightly from the literature, we use the term *non-Gaussian* to refer to the fact that the variables are not jointly Gaussian; however, we do *not* require non-Gaussianity of all exogenous noise variables, as assumed by Shimizu et al. (2011). We proceed in two steps, where each step is of independent interest: First, we infer the moralized graph by estimating the inverse covariance matrix of the joint distribution. The novelty is that we justify this approach for non-Gaussian linear SEMs. Second, we find the optimal causal network structure by searching over the space of DAGs that are consistent with the moralized graph and selecting the DAG that minimizes an appropriate score function. When the score function is decomposable and the moralized graph has bounded treewidth, the second step may be performed via dynamic program-

ming in time linear in the number of nodes (Ordyniak and Szeider, 2012). Our algorithm is also applicable in a high-dimensional setting when the moralized graph is sparse, where we estimate the support of the inverse covariance matrix using a method such as the graphical Lasso (Ravikumar et al., 2011). Our algorithmic framework is summarized in Algorithm 1:

Algorithm 1 Framework for DAG estimation

1: **Input:** Data samples $\{x_i\}_{i=1}^n$ from a linear SEM
2: Obtain estimate $\widehat{\Theta}$ of inverse covariance matrix (e.g., using graphical Lasso)
3: Construct moralized graph $\widehat{\mathcal{M}}$ with edge set defined by $\mathrm{supp}(\widehat{\Theta})$
4: Compute scores for DAGs that are consistent with $\widehat{\mathcal{M}}$ (e.g., using squared ℓ_2-error)
5: Find minimal-scoring $\widehat{G}$ (using dynamic programming when score is decomposable and $\widehat{\mathcal{M}}$ has bounded treewidth)
6: **Output:** Estimated DAG $\widehat{G}$

We prove the correctness of our graph estimation algorithm by deriving new results about the theory of linear SEMs. We present a novel result showing that for almost every choice of linear coefficients, the support of the inverse covariance matrix of the joint distribution is identical to the edge structure of the moralized graph. Although a similar relationship between the support of the inverse covariance matrix and the edge structure of an undirected conditional independence graph has long been established for multivariate Gaussian models (Lauritzen, 1996), our core result in Theorem 2 does not exploit Gaussianity, and the proof technique is entirely new.

Since we do not impose constraints on the error distribution of our SEM, standard parametric maximum likelihood methods are *not* applicable to score and compare candidate DAGs. Consequently, we use the squared ℓ_2-error to score DAGs, and prove that in the case of homoscedastic errors, the true DAG uniquely minimizes this score function. As a side corollary, we establish that the DAG structure of a linear SEM is identifiable whenever the additive errors are homoscedastic, which generalizes a recent result derived only for Gaussian variables (Peters and Bühlmann, 2013). In addition, our result covers cases with Gaussian and non-Gaussian errors, whereas Shimizu et al. (2011) require all errors to be non-Gaussian (see Section 4.2). A similar result is implicitly contained under some assumptions in van de Geer and Bühlmann (2013), but we provide a more general statement and additionally quantify a regime where the errors may exhibit a certain degree of heteroscedasticity. Thus, when errors are not too heteroscedastic, the much more complicated ICA algorithm (Shimizu et al., 2006, 2011) may be replaced by a simple scoring method using squared ℓ_2-loss.

On the statistical side, we show that our method produces consistent estimates of the true DAG by invoking results from high-dimensional statistics. We note that our theoretical results only require a condition on the gap between squared ℓ_2-scores for various DAGs in the restricted search space and eigenvalue conditions on the true covariance matrix, which is a much weaker assumption than the restrictive beta-min condition from previous work (van de Geer and Bühlmann, 2013). Furthermore, the size of the gap is *not* required to scale linearly with the number of nodes in the graph, unlike similar conditions in van de Geer and Bühlmann (2013) and Peters and Bühlmann (2013), leading to genuinely high-dimensional results. Although the precise size of the gap relies heavily on the structure

of the true DAG, we include several examples providing intuition for when our condition could be expected to hold (see Sections 4.4 and 5.2 below). Finally, since inverse covariance matrix estimation and computing scores based on linear regression are both easily modified to deal with systematically corrupted data (Loh and Wainwright, 2012), we show that our methods are also applicable for learning the DAG structure of a linear SEM when data are observed subject to corruptions such as missing data and additive noise.

The remainder of the paper is organized as follows: In Section 2, we review the general theory of probabilistic graphical models and linear SEMs. Section 3 describes our results on the relationship between the inverse covariance matrix and conditional independence graph of a linear SEM. In Section 4, we discuss the use of the squared ℓ_2-loss for scoring candidate DAGs. Section 5 establishes results for the statistical consistency of our proposed inference algorithms and explores the gap condition for various graphs. Finally, Section 6 describes how dynamic programming may be used to identify the optimal DAG in linear time, when the moralized graph has bounded treewidth. Proofs of supporting results are contained in the Appendix.

2. Background

We begin by reviewing some basic background material and introducing notation for the graph estimation problems studied in this paper.

2.1 Graphical Models

In this section, we briefly review the theory of directed and undirected graphical models, also known as conditional independence graphs (CIGs). For a more in-depth exposition, see Lauritzen (1996) or Koller and Friedman (2009) and the references cited therein.

2.1.1 Undirected Graphs

Consider a probability distribution $q(x_1, \ldots, x_p)$ and an undirected graph $G = (V, E)$, where $V = \{1, \ldots, p\}$ and $E \subseteq V \times V$. We say that G is a *conditional independence graph* (CIG) for q if the following *Markov condition* holds: For all disjoint triples $(A, B, S) \subseteq V$ such that S separates A from B in G, we have $X_A \perp\!\!\!\perp X_B \mid X_S$. Here, $X_C := \{X_j : j \in C\}$ for any subset $C \subseteq V$. We also say that G *represents* the distribution q.

By the well-known Hammersley-Clifford theorem, if q is a strictly positive distribution (i.e., $q(x_1, \ldots, x_p) > 0$ for all $(x_1, \ldots, x_p)$), then G is a CIG for q if and only if we may write

$$q(x_1, \ldots, x_p) = \prod_{C \in \mathcal{C}} \psi_C(x_C), \tag{1}$$

for some potential functions $\{\psi_C : C \in \mathcal{C}\}$ defined over the set of cliques $\mathcal{C}$ of G. In particular, note that the complete graph on p nodes is always a CIG for q, but CIGs with fewer edges may exist.

2.1.2 Directed Acyclic Graphs (DAGs)

Changing notation slightly, consider a *directed* graph $G = (V, E)$, where we now distinguish between edges (j, k) and (k, j). We say that G is a *directed acyclic graph* (DAG) if there

are no directed paths starting and ending at the same node. For each node $j \in V$, let $\mathrm{Pa}(j) := \{k \in V : (k,j) \in E\}$ denote the *parent set* of j, where we sometimes write $\mathrm{Pa}_G(j)$ to emphasize the dependence on G. A DAG G *represents* a distribution $q(x_1, \ldots, x_p)$ if q factorizes as

$$q(x_1, \ldots, x_p) \propto \prod_{j=1}^{p} q(x_j \mid x_{\mathrm{Pa}(j)}). \tag{2}$$

Finally, a permutation π of the vertex set $V = \{1, \ldots, p\}$ is a *topological order* for G if $\pi(j) < \pi(k)$ whenever $(j,k) \in E$. Such a topological order exists for any DAG, but it may not be unique. The factorization (2) implies that $X_j \perp\!\!\!\perp X_{\nu(j)} \mid X_{\mathrm{Pa}(j)}$ for all j, where $\nu(j)$ is the set of all nondescendants of j (nodes that cannot be reached via a directed path from j) excluding $\mathrm{Pa}(j)$.

Given a DAG G, we may form the *moralized graph* $\mathcal{M}(G)$ by fully connecting all nodes within each parent set $\mathrm{Pa}(j)$ and dropping the orientations of directed edges. Note that moralization is a purely graph-theoretic operation that transforms a directed graph into an undirected graph. However, if the DAG G represents a distribution q, then $\mathcal{M}(G)$ is also a CIG for q. This is because each set $\{j\} \cup \mathrm{Pa}(j)$ forms a clique C_j in $\mathcal{M}(G)$, and we may define the potential functions $\psi_{C_j}(x_{C_j}) := q(x_j \mid x_{\mathrm{Pa}(j)})$ to obtain the factorization (1) from the factorization (2).

Finally, we define the *skeleton* of a DAG G to be the undirected graph formed by dropping orientations of edges in G. Note that the edge set of the skeleton is a subset of the edge set of the moralized graph, but the latter set is generally much larger. The skeleton is not in general a CIG.

2.2 Linear Structural Equation Models

We now specialize to the framework of linear structural equation models.

We say that a random vector $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ follows a *linear structural equation model* (SEM) if

$$X = B^T X + \epsilon, \tag{3}$$

where B is a strictly upper triangular matrix known as the *autoregression matrix*. We assume $\mathrm{E}[X] = \mathrm{E}[\epsilon] = 0$ and $\epsilon_j \perp\!\!\!\perp (X_1, \ldots, X_{j-1})$ for all j.

In particular, observe that the DAG G with vertex set $V = \{1, \ldots, p\}$ and edge set $E = \{(j,k) : B_{jk} \neq 0\}$ represents the joint distribution q on X. Indeed, Equation (3) implies that

$$q(X_j \mid X_1, \ldots, X_{j-1}) = q(X_j \mid X_{\mathrm{Pa}_G(j)}),$$

so we may factorize

$$q(X_1, \ldots, X_p) = \prod_{j=1}^{p} q(X_j \mid X_1, \ldots, X_{j-1}) = \prod_{j=1}^{p} q(X_j \mid X_{\mathrm{Pa}_G(j)}).$$

Given samples $\{X^i\}_{i=1}^n$, our goal is to infer the unknown matrix B, from which we may recover G (or vice versa).

3. Moralized Graphs and Inverse Covariance Matrices

In this section, we describe our main result concerning inverse covariance matrices of linear SEMs. It generalizes a result for multivariate Gaussians, and states that the inverse covariance matrix of the joint distribution of a linear SEM reflects the structure of a conditional independence graph.

We begin by noting that

$$\mathrm{E}[X_j \mid X_1, \ldots, X_{j-1}] = b_j^T X,$$

where b_j is the j^{th} column of B, and

$$b_j = \left(\Sigma_{j,1:(j-1)} \left(\Sigma_{1:(j-1),1:(j-1)}\right)^{-1}, 0, \ldots, 0\right)^T.$$

Here, $\Sigma := \mathrm{cov}[X]$. We call $b_j^T X$ the *best linear predictor* for X_j amongst linear combinations of $\{X_1, \ldots, X_{j-1}\}$. Defining $\Omega := \mathrm{cov}[\epsilon]$ and $\Theta := \Sigma^{-1}$, we see from Equation (3) that

$$\Sigma = (I - B)^{-T} \Omega (I - B)^{-1}. \tag{4}$$

(Note that $(I - B)$ is always invertible because B is strictly upper triangular.) Furthermore, we have the following lemma, proved in Appendix C.1:

Lemma 1 *The matrix of error covariances is diagonal:* $\Omega = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ *for some* $\sigma_i > 0$. *The entries of* Θ *are given by*

$$\Theta_{jk} = -\sigma_k^{-2} B_{jk} + \sum_{\ell > k} \sigma_\ell^{-2} B_{j\ell} B_{k\ell}, \qquad \forall j < k, \tag{5}$$

$$\Theta_{jj} = \sigma_j^{-2} + \sum_{\ell > j} \sigma_\ell^{-2} B_{j\ell}^2, \qquad \forall j. \tag{6}$$

In particular, Equation (5) has an important implication for causal inference, which we state in the following theorem. Recalling the notation of Section 2.1.2, the graph $\mathcal{M}(G)$ denotes the moralized DAG.

Theorem 2 *Suppose* X *is generated from the linear structural equation model* (3). *Then* Θ *reflects the graph structure of the moralized DAG; i.e., for* $j \neq k$, *we have* $\Theta_{jk} = 0$ *if* (j, k) *is not an edge in* $\mathcal{M}(G)$.

Proof Suppose $j \neq k$ and (j, k) is not an edge in $\mathcal{M}(G)$, and assume without loss of generality that $j < k$. Certainly, $(j, k) \notin E$, implying that $B_{jk} = 0$. Furthermore, j and k cannot share a common child, or else (j, k) would be an edge in $\mathcal{M}(G)$. This implies that either $B_{j\ell} = 0$ or $B_{k\ell} = 0$ for each $\ell > k$. The desired result then follows from Equation (5). ■

Note that Theorem 2 may be viewed as an extension of the canonical result for Gaussian graphical models, although we do *not* require ϵ to follow a Gaussian distribution, so the class of linear SEMs covered by Theorem 2 is much broader. A multivariate Gaussian distribution may be written as a linear SEM with respect to any permutation order π of the

variables, giving rise to a DAG G^π. In that case, Theorem 2 states that $\mathrm{supp}(\Theta)$ is always a subset of the edge set of $\mathcal{M}(G^\pi)$.

In the results to follow, we will assume that the converse of Theorem 2 holds, as well. This is stated in the following Assumption:

Assumption 1 *Let (B, Ω) be the matrices of the underlying linear SEM. For every $j < k$, we have*

$$-\sigma_k^{-2} B_{jk} + \sum_{\ell > k} \sigma_\ell^{-2} B_{j\ell} B_{k\ell} = 0$$

only if $B_{jk} = 0$ and $B_{j\ell} B_{k\ell} = 0$ for all $\ell > k$.

Combined with Theorem 2, Assumption 1 implies that $\Theta_{jk} = 0$ if and only if (j, k) is not an edge in $\mathcal{M}(G)$. (Since $\Theta \succ 0$, the diagonal entries of Θ are always strictly positive.) Note that when the nonzero entries of B are independently sampled continuous random variables, Assumption 1 holds for all choices of B except on a set of Lebesgue measure zero.

Remark 3 *Assumption 1 is a type of* faithfulness *assumption (Koller and Friedman, 2009; Spirtes et al., 2000). By selecting different topological orders π, one may then derive the familiar result that $X_j \perp\!\!\!\perp X_k \mid X_{\setminus\{j,k\}}$ if and only if $\Theta_{jk} = 0$, in the Gaussian setting. Note that this conditional independence assertion may not always hold for linear SEMs, however, since non-Gaussian distributions are not necessarily expressible as a linear SEM with respect to an arbitrary permutation order. Indeed, we only require Assumption 1 to hold with respect to a single (fixed) order.*

4. Score Functions for DAGs

Having established a method for reducing the search space of DAGs based on estimating the moralized graph, we now move to the more general problem of scoring candidate DAGs. As before, we assume the setting of a linear SEM.

Parametric maximum likelihood is often used as a score function for statistical estimation of DAG structure, since the likelihood enjoys the nice property that the population-level version is maximized only under a correct parameterization of the model class. This follows from the relationship between maximum likelihood and KL divergence:

$$\arg\max_\theta \mathrm{E}_{\theta_0}[\log p_\theta(X)] = \arg\min_\theta \mathrm{E}_{\theta_0}\left[\log\left(\frac{p_{\theta_0}(X)}{p_\theta(X)}\right)\right] = \arg\min_\theta D_{KL}(p_{\theta_0} \| p_\theta),$$

and the latter quantity is minimized exactly when $p_{\theta_0} \equiv p_\theta$, almost everywhere. If the model is identifiable, this happens if and only if $\theta = \theta_0$.

However, such maximum likelihood methods presuppose a fixed parameterization for the model class. In the case of linear SEMs, this translates into an appropriate parameterization of the error vector ϵ. For comparison, note that minimizing the squared ℓ_2-error for ordinary linear regression may be viewed as a maximum likelihood approach when errors are Gaussian, but the ℓ_2-minimizer is still statistically consistent for estimation of the regression vector when errors are *not* Gaussian. When our goal is recovery of the autoregression matrix B of the DAG, it is therefore natural to ask whether squared ℓ_2-error could be used in place of maximum likelihood as an appropriate metric for evaluating DAGs.

We will show that in settings where the noise variances $\{\sigma_j\}_{j=1}^p$ are specified up to a constant (e.g., homoscedastic error), the answer is affirmative. In such cases, the true DAG uniquely minimizes the ℓ_2-loss. As a side result, we will also show that the true linear SEM is identifiable.

Remark 4 *Nowzohour and Bühlmann (2014) study the use of nonparametric maximum likelihood methods for scoring candidate DAGs. We remark that such methods could also be combined with the framework of Sections 3 and 6 to select the optimal DAG for linear SEMs with nonparametric error distributions: First, estimate the moralized graph via the inverse covariance matrix, and then find the DAG with minimal score using a method such as dynamic programming. Similar statistical guarantees would hold in that case, with parametric rates replaced by nonparametric rates. However, our results in this section imply that in settings where the error variances are known or may be estimated accurately, the much simpler method of squared ℓ_2-loss may be used in place of a more complicated nonparametric approach.*

4.1 Weighted Squared ℓ_2-Loss

Suppose X is drawn from a linear SEM (3), where we now use B_0 to denote the true autoregression matrix and Ω_0 to denote the true error covariance matrix. For a fixed diagonal matrix $\Omega = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ and a candidate matrix B with columns $\{b_j\}_{j=1}^p$, define the *score of B with respect to Ω* according to

$$\mathrm{score}_\Omega(B) = \mathrm{E}\left[\|\Omega^{-1/2}(I-B)^T X\|_2^2\right] = \sum_{j=1}^p \frac{1}{\sigma_j^2} \cdot \mathrm{E}[(X_j - b_j^T X)^2]. \tag{7}$$

This is a weighted squared ℓ_2-loss, where the prediction error for the j^{th} coordinate is weighted by the diagonal entry σ_j^2 coming from Ω, and expectations are taken with respect to the true distribution on X.

It is instructive to compare the score function (7) to the usual parametric maximum likelihood when $X \sim N(0, \Sigma)$. For a distribution parameterized by the pair (B, Ω), the inverse covariance matrix is $\Theta = (I-B)\Omega^{-1}(I-B)^T$, using Equation (4), so the expected log likelihood is

$$\begin{aligned}
\mathrm{E}_{X\sim N(0,\Sigma)}[\log p_{B,\Omega}(X)] &= -\operatorname{tr}[(I-B)\Omega^{-1}(I-B)^T\Sigma] + \log\det[(I-B)\Omega^{-1}(I-B)^T] \\
&= -\operatorname{tr}[(I-B)\Omega^{-1}(I-B)^T\Sigma] + \log\det(\Omega^{-1}) \\
&= -\,\mathrm{score}_\Omega(B) + \log\det(\Omega^{-1}).
\end{aligned}$$

Hence, minimizing the score over B for a fixed Ω is identical to maximizing the likelihood. For non-Gaussians, however, the convenient relationship between minimum score and maximum likelihood no longer holds.

Now let $\mathcal{D}$ denote the class of DAGs. For $G \in \mathcal{D}$, define the score of G to be

$$\mathrm{score}_\Omega(G) := \min_{B \in \mathcal{U}_G} \{\mathrm{score}_\Omega(B)\}, \tag{8}$$

where

$$\mathcal{U}_G := \{B \in \mathbb{R}^{p\times p} : B_{jk} = 0 \text{ when } (j,k) \notin E(G)\}$$

is the set of matrices that are consistent with the structure of G.

Remark 5 *Examining the form of the score function* (7)*, we see that if* $\{\mathrm{Pa}_G(j)\}_{j=1}^p$ *denotes the parent sets of nodes in* G*, then the matrix*

$$B_G := \arg\min_{B \in \mathcal{U}_G} \{\mathrm{score}_\Omega(B)\}$$

is unique, and the columns of B_G *are equal to the coefficients of the best linear predictor of* X_j *regressed upon* $X_{\mathrm{Pa}_G(j)}$*. Furthermore, the value of* B_G *does not depend on* Ω*, since the minimizing value of* b_j *in the argument of Equation* (7) *is unaffected by the weighting factor* σ_j^2*.*

The following lemma relates the score of the underlying DAG G_0 to the score of the true autoregression matrix B_0. In fact, the score of any DAG containing G_0 has the same score. The proof is contained in Appendix C.2.

Lemma 6 *Suppose* X *follows a linear SEM with autoregression matrix* B_0*, and let* G_0 *denote the underlying DAG. Consider any* $G \in \mathcal{D}$ *such that* $G_0 \subseteq G$*. Then for any diagonal weight matrix* Ω*, we have*

$$\mathrm{score}_\Omega(G) = \mathrm{score}_\Omega(B_0),$$

and B_0 *is the unique minimizer of* $\mathrm{score}_\Omega(B)$ *over* $\mathcal{U}_G$*.*

We now turn to the main theorem of this section, in which we consider the problem of minimizing $\mathrm{score}_\Omega(B)$ with respect to all matrices B that are permutation similar to upper triangular matrices. Such a result is needed to validate our choice of score function, since when the DAG structure is not known a priori, the space of possible autoregression matrices must include all $\mathcal{U} := \bigcup_{G \in \mathcal{D}} \mathcal{U}_G$. Note that $\mathcal{U}$ may be equivalently defined as the set of all matrices that are permutation similar to upper triangular matrices. We have the following vital result:

Theorem 7 *Given a linear SEM* (3) *with error covariance matrix* $\alpha\Omega_0$ *and autoregression matrix* B_0*, where* $\alpha > 0$*, we have*

$$\mathrm{score}_{\Omega_0}(B) \geq \mathrm{score}_{\Omega_0}(B_0) = \alpha p, \qquad \forall B \in \mathcal{U}, \tag{9}$$

with equality if and only if $B = B_0$*.*

The proof of Theorem 7, which is based on matrix algebra, is contained in Section 4.5. In particular, Theorem 7 implies that the squared ℓ_2-loss function (7) is indeed an appropriate measure of model fit when the components are correctly weighted by the diagonal entries of Ω_0.

Note, however, that Theorem 7 requires the score to be taken with respect to (a multiple of) the true error covariance matrix Ω_0. The following example gives a cautionary message that if the weights Ω are chosen incorrectly, minimizing $\mathrm{score}_\Omega(B)$ may produce a structure that is *inconsistent* with the true model:

Example 1 *Suppose* (X_1, X_2) *is distributed according to the following linear SEM:*

$$X_1 = \epsilon_1, \quad \textit{and} \quad X_2 = -\frac{X_1}{2} + \epsilon_2,$$

so the autoregression matrix is given by $B_0 = \begin{pmatrix} 0 & -\frac{1}{2} \\ 0 & 0 \end{pmatrix}$. *Let* $\Omega_0 = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{4} \end{pmatrix}$. *Consider* $B_1 = \begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix}$. *Then*

$$\text{score}_I(B_1) < \text{score}_I(B_0), \tag{10}$$

so using squared ℓ_2*-loss weighted by the identity will select an inappropriate model.*

Proof To verify Equation (10), we first compute

$$\Sigma = (I - B_0)^{-T}\Omega_0(I - B_0) = \begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Then

$$\mathrm{E}[\|X - B_1^T X\|_2^2] = \mathrm{tr}\left[(I - B_1)^T\Sigma(I - B_1)\right] = \mathrm{tr}\left[\begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}\right] = 1,$$
$$\mathrm{E}[\|X - B_0^T X\|_2^2] = \mathrm{tr}\left[(I - B_0)^T\Sigma(I - B_0)\right] = \mathrm{tr}\left[\begin{pmatrix} 1 & 0 \\ 0 & 1/4 \end{pmatrix}\right] = \frac{5}{4},$$

implying Inequality (10). ■

4.2 Identifiability of Linear SEMs

Theorem 7 also has a useful consequence in terms of identifiability of a linear SEM, which we state in the following corollary:

Corollary 8 *Consider a fixed diagonal covariance matrix* Ω_0*, and consider the class of linear SEMs parameterized by the pair* $(B, \alpha\Omega_0)$*, where* $B \in \mathcal{U}$ *and* $\alpha > 0$ *is a scale factor. Then the true model* $(B_0, \alpha_0\Omega_0)$ *is identifiable. In particular, the class of homoscedastic linear SEMs is identifiable.*

Proof By Theorem 7, the matrix B_0 is the unique minimizer of $\text{score}_{\Omega_0}(B)$. Since $\alpha_0 \cdot (\Omega_0)_{11} = \mathrm{var}[X_1]$, the scale factor α_0 is also uniquely identifiable. The statement about homoscedasticity follows by taking $\Omega_0 = I$. ■

Corollary 8 should be viewed in comparison to previous results in the literature regarding identifiability of linear SEMs. Theorem 1 of Peters and Bühlmann (2013) states that when X is Gaussian and ϵ is an i.i.d. Gaussian vector with $\mathrm{cov}[\epsilon] = \alpha\Omega_0$, the model is identifiable. Indeed, our Corollary 8 implies that result as a special case, but it does not impose any additional conditions concerning Gaussianity. Shimizu et al. (2006) establish identifiability

of a linear SEM when ϵ is a vector of independent, non-Gaussian errors, by reducing to ICA, but our result does not require errors to be non-Gaussian.

The significance of Corollary 8 is that it supplies an elegant proof showing that the model is still identifiable even in the presence of both Gaussian and non-Gaussian components, provided the error variances are specified up to a scalar multiple. Since any multivariate Gaussian distribution may be written as a linear SEM with respect to an arbitrary ordering, some constraint such as variance scaling or non-Gaussianity is necessary in order to guarantee identifiability.

4.3 Misspecification of Variances

Theorem 7 implies that when the diagonal variances of Ω_0 are known up to a scalar factor, the weighted ℓ_2-loss (7) may be used as a score function for linear SEMs. Example 1 shows that when Ω is misspecified, we may have $B_0 \notin \arg\min_{B \in \mathcal{U}} \{\text{score}_\Omega(B)\}$. In this section, we further study the effect when Ω is misspecified. Intuitively, provided Ω is close enough to Ω_0 (or a multiple thereof), minimizing $\text{score}_\Omega(B)$ with respect to B should still yield the correct B_0.

Consider an arbitrary diagonal weight matrix Ω_1. We first provide bounds on the ratio between entries of Ω_0 and Ω_1 which ensure that $B_0 = \arg\min_{B \in \mathcal{U}} \{\text{score}_{\Omega_1}(B)\}$, even though the model is misspecified. Let

$$a_{\max} := \lambda_{\max}(\Omega_0 \Omega_1^{-1}) \quad \text{and} \quad a_{\min} := \lambda_{\min}(\Omega_0 \Omega_1^{-1})$$

denote the maximum and minimum ratios between corresponding diagonal entries of Ω_1 and Ω_0. Now define the additive gap between the score of G_0 and the next best DAG, given by

$$\xi := \min_{G \in \mathcal{D},\, G \not\supseteq G_0} \{\text{score}_{\Omega_0}(G) - \text{score}_{\Omega_0}(G_0)\} = \min_{G \in \mathcal{D}, G \not\supseteq G_0} \{\text{score}_{\Omega_0}(G)\} - p. \tag{11}$$

By Theorem 7, we know that $\xi > 0$. The following theorem provides a sufficient condition for correct model selection in terms of the gap ξ and the ratio $\frac{a_{\max}}{a_{\min}}$, which are both invariant to the scale factor α. It is a measure of robustness for how roughly the entries of Ω_0 may be approximated and still produce B_0 as the unique minimizer. The proof of the theorem is contained in Appendix C.3.

Theorem 9 *Suppose*

$$\frac{a_{\max}}{a_{\min}} \leq 1 + \frac{\xi}{p}. \tag{12}$$

Then $B_0 \in \arg\min_{B \in \mathcal{U}} \{\text{score}_{\Omega_1}(B)\}$. *If Inequality* (12) *is strict, then* B_0 *is the unique minimizer of* $\text{score}_{\Omega_1}(B)$.

Remark 10 *Theorem 9 provides an error allowance concerning the accuracy to which we may specify the error covariances and still recover the correct autoregression matrix* B_0 *from an improperly weighted score function. In the case* $\Omega_1 = \alpha\Omega_0$, *we have* $a_{\max} = a_{\min} = 1$, *so the condition* (12) *is always strictly satisfied, which is consistent with our earlier result in Theorem 7 that* $B_0 = \arg\min_{B \in \mathcal{U}} \{\text{score}_{\alpha\Omega_0}(B)\}$.

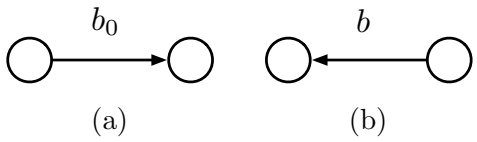

Figure 1: Two-variable DAG. (a) The forward model. (b) The backward model.

Naturally, the error tolerance specified by Theorem 9 is a function of the gap ξ between the true DAG G_0 and the next best candidate: If ξ is larger, the score is more robust to misspecification of the weights Ω. Note that if we restrict our search space from the full set of DAGs $\mathcal{D}$ to some smaller space $\mathcal{D}'$, so $B \in \bigcup_{G \in \mathcal{D}'} \mathcal{U}_G$, we may restate the condition in Theorem 9 in terms of the gap

$$\xi(\mathcal{D}') := \min_{G \in \mathcal{D}',\, G \not\supseteq G_0} \{\text{score}_{\Omega_0}(G) - \text{score}_{\Omega_0}(G_0)\}, \tag{13}$$

which may be considerably larger than ξ when $\mathcal{D}'$ is much smaller than $\mathcal{D}$. See Equation (22) and Section 5.2 on weakening the gap condition below.

Specializing to the case when $\Omega_1 = I$, we may interpret Theorem 9 as providing a window of variances around which we may treat a heteroscedastic model as homoscedastic, and use the simple (unweighted) squared ℓ_2-score to recover the correct model. See Lemma 11 in the next section for a concrete example.

4.4 Example: 2- and 3-Variable Models

In this section, we develop examples illustrating the gap ξ introduced in Section 4.3. We study two cases, involving two and three variables, respectively.

4.4.1 Two Variables

We first consider the simplest example with a two-variable directed graph. Suppose the forward model is defined by

$$B_0 = \begin{pmatrix} 0 & b_0 \\ 0 & 0 \end{pmatrix}, \qquad \Omega_0 = \begin{pmatrix} d_1^2 & 0 \\ 0 & d_2^2 \end{pmatrix},$$

and consider the backward matrix defined by the autoregression matrix

$$B = \begin{pmatrix} 0 & 0 \\ b & 0 \end{pmatrix}. \tag{14}$$

The forward and backward models are illustrated in Figure 1.

A straightforward calculation shows that

$$\text{score}_{\Omega_0}(B) = 2 + b^2 b_0^2 + \left(b \frac{d_2}{d_1} - b_0 \frac{d_1}{d_2} \right)^2,$$

which is minimized for $b = \frac{b_0}{b_0^2 + \frac{d_2^2}{d_1^2}}$, implying that

$$\xi = \min_b \{\text{score}_{\Omega_0}(B) - \text{score}_{\Omega_0}(B_0)\} = \frac{b_0^4}{\frac{d_2^4}{d_1^4} + b_0^2 \frac{d_2^2}{d_1^2}}. \tag{15}$$

We see that the gap ξ grows with the strength of the true edge b_0, when $|b_0| > 1$, and is symmetric with respect to the sign of b_0. The gap also grows with the magnitude of the ratio $\frac{d_1}{d_2}$.

To gain intuition for the interplay between b_0 and $\frac{d_1}{d_2}$, we derive the following lemma, a corollary of Theorem 9 specialized to the case of the two-variable DAG:

Lemma 11 *Consider the two-variable DAG defined by Equation* (14)*. Let* $\Omega_1 = I_2$ *and define* $r := \frac{d_2}{d_1}$*. Suppose the following conditions hold:*

$$b_0^2 \geq \begin{cases} r^2 \left((r^2 - 1) + \sqrt{r^4 - 1} \right), & \text{if } r \geq 1, \\ (1 - r^2) + \sqrt{1 - r^4}, & \text{if } r \leq 1. \end{cases} \tag{16}$$

Then $B_0 = \arg\min_{B \in \mathcal{U}} \{\text{score}_{\Omega_1}(B)\}$*; i.e.,* B_0 *is the unique minimizer of the score function under the unweighted squared-*ℓ_2 *loss.*

Lemma 11 is proved in Appendix C.4.

Remark 12 *Note that the two right-hand expressions in Inequality* (16) *are similar, although the expression in the case* $r^2 \geq 1$ *contains an extra factor of* r^2*, so the sufficient condition is stronger. Both lower bounds in Equation* (16) *increase with* $|r-1|$*, which agrees with intuition: If the true model is more non-homoscedastic, the strength of the true edge must be stronger in order for the unweighted squared-*ℓ_2 *score to correctly identify the model. When* $r = 1$*, we have the vacuous condition* $b_0^2 \geq 0$*, since* $\Omega_1 = \alpha\Omega_0$ *and the variances are correctly specified, so Theorem 7 implies* $B_0 = \arg\min_{B \in \mathcal{U}} \{\text{score}_{\Omega_1}(B)\}$ *for any choice of* b_0*.*

4.4.2 v-Structure

We now examine a three-variable graph. Suppose the actual graph involves a v-structure, as depicted in Figure 2, and is parameterized by the matrices

$$B_0 = \begin{pmatrix} 0 & 0 & b_{13} \\ 0 & 0 & b_{23} \\ 0 & 0 & 0 \end{pmatrix}, \qquad \Omega = \begin{pmatrix} d_1^2 & 0 & 0 \\ 0 & d_2^2 & 0 \\ 0 & 0 & d_3^2 \end{pmatrix}. \tag{17}$$

We have the following lemma, proved in Appendix C.5:

Lemma 13 *Consider the three-variable DAG characterized by Figure 2 and Equations* (17)*. The gap* ξ *defined by Equation* (11) *is given by*

$$\xi = \min \left\{ \frac{b_{23}^4}{\frac{d_3^4}{d_2^4} + b_{23}^2 \frac{d_3^2}{d_2^2}}, \ \frac{b_{13}^4}{\frac{d_3^4}{d_1^4} + b_{13}^2 \frac{d_3^2}{d_1^2}} \right\}.$$

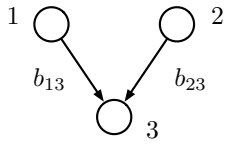

Figure 2: Three-variable DAG with v-structure.

A key reduction in the proof of Lemma 13 is to note that we only need to consider a relatively small number of DAGs, given by Figure 3 in Appendix C.5. Indeed, for $G_1 \subseteq G_2$, we have $\text{score}_{\Omega_0}(G_2) \leq \text{score}_{\Omega_0}(G_1)$, so it suffices to consider maximal elements in the poset of DAGs not containing the true DAG G_0.

Remark 14 *Note that the form of the gap in Lemma 13 is very similar to the form for the two-variable model, and the individual ratios scale with the strength of the edge and the ratio of the corresponding error variances. Indeed, we could derive a version of Lemma 11 for the three-variable model, giving lower bounds on the edge strengths b_{23}^2 and b_{13}^2 that guarantee the accuracy of the unweighted squared ℓ_2-loss; however, the conditions would be more complicated. It is interesting to note from our calculations in Appendix C.5 that the gap between models accumulates according to the number of edge reversals from the misspecified model: Reversing the directions of edges $(2,3)$ and $(1,3)$ in succession leads to an additional term in the expressions for ξ_1 and ξ_2 in Equations (43) and (44) below. We will revisit these observations in Section 5.2, where we define a version of the gap function rescaled by the number of nodes that differ in their parents sets.*

4.5 Proof of Theorem 7

First note that for a constant $\alpha > 0$, we have

$$\text{score}_{\alpha\Omega}(B) = \frac{1}{\alpha} \cdot \text{score}_{\Omega}(B),$$

so minimizing $\text{score}_{\alpha\Omega}(B)$ is equivalent to minimizing $\text{score}_{\Omega}(B)$. Furthermore, it suffices to prove the statement for $\alpha = 1$; the statement for general $\alpha > 0$ follows by a simple rescaling.

Recalling Equation (4), we may write

$$\begin{aligned}
\text{score}_{\Omega_0}(B) &= \mathrm{E}_{B_0}[\|\Omega_0^{-1/2}(I-B)^T X\|_2^2] \\
&= \text{tr}\left[\Omega_0^{-1/2}(I-B)^T \cdot \text{cov}_{B_0}[X] \cdot (I-B)\Omega_0^{-1/2}\right] \\
&= \text{tr}\left[\Omega_0^{-1/2}(I-B)^T \cdot (I-B_0)^{-T}\Omega_0(I-B_0)^{-1} \cdot (I-B)\Omega_0^{-1/2}\right].
\end{aligned}$$

Now note that

$$\begin{aligned}
(I-B)\Omega_0^{-1/2} &= \Omega_0^{-1/2}(I - \Omega_0^{1/2}B\Omega_0^{-1/2}) := \Omega_0^{-1/2}(I-\widetilde{B}), \\
\Omega_0^{1/2}(I-B_0)^{-1} &= (I - \Omega_0^{1/2}B_0\Omega_0^{-1/2})^{-1}\Omega_0^{1/2} := (I-\widetilde{B}_0)^{-1}\Omega_0^{1/2},
\end{aligned}$$

where $\widetilde{B}, \widetilde{B}_0 \in \mathcal{U}$. Hence, we may rewrite

$$\begin{aligned}\text{score}_{\Omega_0}(B) &= \text{tr}\left[(I-\widetilde{B})^T\Omega_0^{-1/2}\cdot\Omega_0^{1/2}(I-\widetilde{B}_0)^{-T}\cdot(I-\widetilde{B}_0)^{-1}\Omega_0^{1/2}\cdot\Omega_0^{-1/2}(I-\widetilde{B})\right]\\ &= \text{tr}\left[(I-\widetilde{B})^T(I-\widetilde{B}_0)^{-T}(I-\widetilde{B}_0)^{-1}(I-\widetilde{B})\right]\\ &= \text{tr}\left[(I-\widetilde{B})(I-\widetilde{B})^T(I-\widetilde{B}_0)^{-T}(I-\widetilde{B}_0)^{-1}\right].\end{aligned}$$

Since $\widetilde{B}, \widetilde{B}_0 \in \mathcal{U}$, the matrices $I-\widetilde{B}$ and $I-\widetilde{B}_0$ are both permutation similar to lower triangular matrices with 1's on the diagonal. Hence, Lemma 27 in Appendix B implies

$$\text{score}_{\Omega_0}(B) \geq p,$$

with equality if and only if

$$I-\widetilde{B} = I-\widetilde{B}_0,$$

or equivalently, $B = B_0$, as claimed.

5. Consequences for Statistical Estimation

The population-level results in Theorems 2 and 7 provide a natural avenue for estimating the DAG of a linear SEM from data. In this section, we outline how the true DAG may be estimated in the presence of fully-observed or systematically corrupted data. Our method is applicable also in the high-dimensional setting, assuming the moralized DAG is sufficiently sparse.

Our inference algorithm consists of two main components:

1. Estimate the moralized DAG $\mathcal{M}(G_0)$ using the inverse covariance matrix of X.
2. Search through the space of DAGs consistent with $\mathcal{M}(G_0)$, and find the DAG that minimizes $\text{score}_\Omega(B)$.

Theorem 2 and Assumption 1 ensure that for almost every choice of autoregression matrix B_0, the support of the true inverse covariance matrix Θ_0 exactly corresponds to the edge set of the moralized graph. Theorem 7 ensures that when the weight matrix Ω is chosen appropriately, B_0 will be the unique minimizer of $\text{score}_\Omega(B)$.

5.1 Fully-Observed Data

We now present concrete statistical guarantees for the correctness of our algorithm in the usual setting when $\{x_i\}_{i=1}^n$ are fully-observed and i.i.d. Recall that a random variable X is *sub-Gaussian* with parameter σ^2 if

$$\text{E}[\exp(\lambda(X-\text{E}[X]))] \leq \exp\left(\frac{\sigma^2\lambda^2}{2}\right), \qquad \forall\lambda\in\mathbb{R}.$$

If $X \in \mathbb{R}^p$ is a random vector, it is sub-Gaussian with parameter σ^2 if v^TX is a sub-Gaussian random variable with parameter σ^2 for all unit vectors $v \in \mathbb{R}^p$.

5.1.1 Estimating the Inverse Covariance Matrix

We first consider the problem of inferring Θ_0. Let

$$\Theta_0^{\min} := \min_{j,k} \left\{ |(\Theta_0)_{jk}| : (\Theta_0)_{jk} \neq 0 \right\}$$

denote the magnitude of the minimum nonzero element of Θ_0. We consider the following two scenarios:

Low-dimensional setting. If $n \geq p$, the sample covariance matrix $\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n x_i x_i^T$ is invertible with high probability, and we use the estimator

$$\widehat{\Theta} = (\widehat{\Sigma})^{-1}.$$

We have the following lemma, which follows from standard bounds in random matrix theory:

Lemma 15 *Suppose the x_i's are i.i.d. sub-Gaussian vectors with parameter σ^2. With probability at least $1 - c_1 \exp(-c_2 p)$, we have*

$$\|\widehat{\Theta} - \Theta_0\|_{\max} \leq c_0 \sigma^2 \sqrt{\frac{p}{n}},$$

and thresholding $\widehat{\Theta}$ at level $\tau = c_0 \sigma^2 \sqrt{\frac{p}{n}}$ succeeds in recovering $\operatorname{supp}(\Theta_0)$*, if $\Theta_0^{\min} > 2\tau$.*

For the proof, see Appendix D.1. Here, we use to $\|\cdot\|_{\max}$ denote the elementwise ℓ_∞-norm of a matrix.

High-dimensional setting. If $p > n$, we assume each row of the true inverse covariance matrix Θ_0 is d-sparse. Then we use the graphical Lasso:

$$\widehat{\Theta} \in \arg\min_{\Theta \succeq 0} \left\{ \operatorname{tr}(\Theta \widehat{\Sigma}) - \log\det(\Theta) + \lambda \sum_{j \neq k} |\Theta_{jk}| \right\}. \tag{18}$$

Standard results (Ravikumar et al., 2011) establish the statistical consistency of the graphical Lasso (18) as an estimator for the inverse covariance matrix in the setting of sub-Gaussian observations; consequently, we omit the proof of the following lemma.

Lemma 16 *Suppose the x_i's are i.i.d. sub-Gaussian vectors with parameter σ^2. Suppose the sample size satisfies $n \geq Cd \log p$. With probability at least $1 - c_1 \exp(-c_2 \log p)$, we have*

$$\|\widehat{\Theta} - \Theta_0\|_{\max} \leq c_0 \sigma^2 \sqrt{\frac{\log p}{n}},$$

and thresholding $\widehat{\Theta}$ at level $\tau = c_0 \sigma^2 \sqrt{\frac{\log p}{n}}$ succeeds in recovering $\operatorname{supp}(\Theta_0)$*, if $\Theta_0^{\min} > 2\tau$.*

Alternatively, we may perform nodewise regression with the ordinary Lasso (Meinshausen and Bühlmann, 2006) to recover the support of Θ_0, with similar rates for statistical consistency.

5.1.2 Scoring Candidate DAGs

Moving on to the second step of the algorithm, we need to estimate the score functions $\text{score}_\Omega(B)$ of candidate DAGs and choose the minimally scoring candidate. In this section, we focus on methods for estimating an empirical version of the score function and derive rates for statistical estimation under certain models. If the space of candidate DAGs is sufficiently small, we may evaluate the empirical score function for every candidate DAG and select the optimum. In Section 6, we describe computationally efficient procedures based on dynamic programming to choose the optimal DAG when the candidate space is too large for naive search.

The input of our algorithm is the sparsely estimated inverse covariance matrix $\widehat{\Theta}$ from Section 5.1.1. For a matrix Θ, define the candidate neighborhood sets

$$N_\Theta(j) := \{k : k \neq j \text{ and } \Theta_{jk} \neq 0\}, \qquad \forall j,$$

and let

$$\mathcal{D}_\Theta := \{G \in \mathcal{D} : \text{Pa}_G(j) \subseteq N_\Theta(j), \quad \forall j\}$$

denote the set of DAGs with skeleton contained in the graph defined by $\text{supp}(\Theta)$. By Theorem 2 and Assumption 1, we have $G_0 \in \mathcal{D}_{\Theta_0}$, so if $\text{supp}(\widehat{\Theta}) \supseteq \text{supp}(\Theta_0)$, which occurs with high probability under the conditions of Section 5.1.1, it suffices to search over the reduced DAG space $\mathcal{D}_{\widehat{\Theta}}$.

Remark 17 *In fact, we could reduce the search space even further to only include DAGs with moralized graph equal to the undirected graph defined by* $\text{supp}(\Theta)$*. The dynamic programming algorithm to be described in Section 6 only requires as input a superset of the skeleton; for alternative versions of the dynamic programming algorithm taking as input a superset of the moralized graph, we would indeed restrict* $\mathcal{D}_\Theta$ *to DAGs with the correct moral structure.*

We now consider an arbitrary d-sparse matrix Θ, with $d \leq n$, and take $G \in \mathcal{D}_\Theta$. By Remark 5, we have

$$\text{score}_\Omega(G) = \sum_{j=1}^{p} f_{\sigma_j}(\text{Pa}_G(j)), \tag{19}$$

where

$$f_{\sigma_j}(S) := \frac{1}{\sigma_j^2} \cdot \text{E}[(x_j - b_j^T x_S)^2],$$

and $b_j^T x_S$ is the best linear predictor for x_j regressed upon x_S. In order to estimate $\text{score}_\Omega(G)$, we use the empirical functions

$$\widehat{f}_{\sigma_j}(S) := \frac{1}{\sigma_j^2} \cdot \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - x_{i,S}^T \widehat{b}_j)^2 = \frac{1}{\sigma_j^2} \cdot \frac{1}{n} \|X_j - X_S \widehat{b}_j\|_2^2, \tag{20}$$

where

$$\widehat{b}_j := (X_S^T X_S)^{-1} X_S^T X_j$$

is the usual ordinary least squares solution for linear regression of X_j upon X_S. We will take $S \subseteq N_\Theta(j)$, so since $|N_\Theta(j)| \leq d \leq n$, the matrix $X_S^T X_S$ is invertible with high probability. The following lemma, proved in Appendix D.2, provides rates of convergence for the empirical score function:

Lemma 18 *Suppose the x_i's are i.i.d. sub-Gaussian vectors with parameter σ^2. Suppose $d \leq n$ is a parameter such that $|N_\Theta(j)| \leq d$ for all j. Then*

$$|\widehat{f}_{\sigma_j}(S) - f_{\sigma_j}(S)| \leq \frac{c_0 \sigma^2}{\sigma_j^2} \sqrt{\frac{\log p}{n}}, \qquad \forall j \text{ and } S \subseteq N_\Theta(j), \tag{21}$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$.

In particular, we have the following result, proved in Appendix D.3, which provides a sufficient condition for the empirical score functions to succeed in selecting the true DAG. Here,

$$\xi_\Omega(\mathcal{D}_\Theta) := \min_{G \in \mathcal{D}_\Theta, G \not\supseteq G_0} \{\text{score}_\Omega(G) - \text{score}_\Omega(G_0)\} \tag{22}$$

is the gap between G_0 and the next best DAG in $\mathcal{D}_\Theta$. Note that the expression in Equation (22) is reminiscent of Equation (13) in Section 4.3, but we now allow Ω to be arbitrary.

Lemma 19 *Suppose Inequality* (21) *holds, and suppose*

$$c_0 \sigma^2 \sqrt{\frac{\log p}{n}} \cdot \sum_{j=1}^{p} \frac{1}{\sigma_j^2} < \frac{\xi_\Omega(\mathcal{D}_\Theta)}{2}. \tag{23}$$

Then

$$\widehat{\text{score}}_\Omega(G_0) < \widehat{\text{score}}_\Omega(G), \qquad \forall G \in \mathcal{D}_\Theta : G \not\supseteq G_0. \tag{24}$$

Remark 20 *Lemma 19 does not explicitly assume that Ω is equal to Ω_0, the true matrix of error variances. However, Inequality* (23) *can only be satisfied when $\xi_\Omega(\mathcal{D}_G) > 0$; hence, Ω should be chosen such that $G_0 = \arg\min_{G \in \mathcal{D}_\Theta, G \not\supseteq G_0} \{\text{score}_\Omega(G)\}$. As discussed in Section 4.3, this condition holds for a wider range of choices for Ω.*

Note that the conclusion (24) in Lemma 19 is not quite the same as the condition

$$G_0 = \arg \min_{G \in \mathcal{D}_\Theta, G \not\supseteq G_0} \left\{\widehat{\text{score}}_\Omega(G)\right\}, \tag{25}$$

which is what we would need for exact recovery of our score-minimizing algorithm. The issue is that $\text{score}_\Omega(G)$ is equal for all $G \supseteq G_0$; however the empirical scores $\widehat{\text{score}}_\Omega(G)$ may differ among this class, so Equation (25) may not be satisfied. However, it is easily seen from the proof of Lemma 19 that in fact,

$$\arg \min_{G \in \mathcal{D}_\Theta} \left\{\widehat{\text{score}}_\Omega(G)\right\} \subseteq \{G \in \mathcal{D}_\Theta : G \supseteq G_0\}. \tag{26}$$

By applying a thresholding procedure to the empirical score minimizer $\widehat{G} \supseteq G_0$ selected by our algorithm, we could then recover the true G_0. In other words, since $\text{Pa}_{G_0}(j) \subseteq \text{Pa}_{\widehat{G}}(j)$

for each j, we could use standard sparse regression techniques to recover the parent set of each node in the true DAG.

To gain some intuition for the condition (23), consider the case when $\sigma_j^2 = 1$ for all j. Then the condition becomes

$$c_0\sigma^2\sqrt{\frac{\log p}{n}} < \frac{\xi(\mathcal{D}_\Theta)}{2p}. \tag{27}$$

If $\xi(\mathcal{D}_\Theta) = \Omega(1)$, which might be expected based on our calculations in Section 4.4, we require $n \geq Cp^2 \log p$ in order to guarantee statistical consistency, which is not a truly high-dimensional result. On the other hand, if $\xi(\mathcal{D}_\Theta) = \Omega(p)$, as is assumed in similar work on score-based DAG learning (van de Geer and Bühlmann, 2013; Bühlmann et al., 2014), our method is consistent provided $\frac{\log p}{n} \to 0$. In Section 5.2, we relax the condition (27) to a slightly weaker condition that is more likely to hold in settings of interest.

5.2 Weakening the Gap Condition

Motivated by our comments from the previous section, we establish a sufficient condition for statistical consistency that is slightly weaker than the condition (23), which still guarantees that Equation (26) holds.

For two DAGs $G, G' \in \mathcal{D}$, define

$$H(G, G') := \{j : \mathrm{Pa}_G(j) \neq \mathrm{Pa}_{G'}(j)\}$$

to be the set of nodes on which the parent sets differ between graphs G and G', and define the ratio

$$\gamma_\Omega(G, G') := \frac{\mathrm{score}_\Omega(G) - \mathrm{score}_\Omega(G')}{|H(G, G')|},$$

a rescaled version of the gap between the score functions. Consider the following condition:

Assumption 2 *There exists $\xi' > 0$ such that*

$$\gamma_\Omega(G_0) := \min_{G \in \mathcal{D}_\Theta, G \not\supseteq G_0} \left\{ \max_{G_1 \supseteq G_0} \{\gamma_\Omega(G, G_1)\} \right\} \geq \xi'. \tag{28}$$

Note that in addition to minimizing over DAGs in the class $\mathcal{D}_\Theta$, the expression (28) defined in Assumption 2 takes an inner maximization over DAGs containing G_0. As established in Lemma 6, we have $\mathrm{score}_\Omega(G_1) = \mathrm{score}_\Omega(G_0)$ whenever $G_0 \subseteq G_1$. However, $|H(G, G_1)|$ may be appreciably different from $|H(G, G_0)|$, and we are only interested in computing the gap ratio between a DAG $G \not\supseteq G_0$ and the closest DAG containing G_0.

We then have the following result, proved in Appendix D.4:

Lemma 21 *Under Assumption 2, suppose*

$$|\widehat{f}_{\sigma_j}(S) - f_{\sigma_j}(S)| \leq \frac{\xi'}{2}, \qquad \forall j \text{ and } S \subseteq N_\Theta(j). \tag{29}$$

Then the containment (26) *holds.*

Combining with Lemma 18, we have the following corollary:

Corollary 22 *Suppose the x_i's are i.i.d. sub-Gaussian with parameter σ^2, and $|N_\Theta(j)| \leq d$ for all j. Also suppose Assumption 2 holds. Then with probability at least $1-c_1\exp(-c_2\log p)$, the condition* (26) *is satisfied.*

We now turn to the question of what values of ξ' might be expected to give condition (28) for various DAGs. Certainly, we have

$$\gamma_\Omega(G, G') \geq \frac{\text{score}_\Omega(G) - \text{score}_\Omega(G')}{p},$$

so the condition holds when

$$p \cdot \xi' < \xi(\mathcal{D}_\Theta).$$

However, for $\xi' = \mathcal{O}(\xi(\mathcal{D}_\Theta)/p)$, Corollary 22 yields a scaling condition similar to Inequality (27), which we wish to avoid. As motivated by our computations of the score functions for small DAGs (see Remark 14 in Section 4.4), the difference $\{\text{score}_\Omega(G) - \text{score}_\Omega(G_0)\}$ seems to increase linearly with the number of edge reversals needed to transform G_0 to G. Hence, we might expect $\gamma_\Omega(G, G_0)$ to remain roughly constant, rather than decreasing linearly with p. The following lemma verifies this intuition in a special case. For a review of junction tree terminology, see Appendix A.1.

Lemma 23 *Suppose the moralized graph $\mathcal{M}(G_0)$ admits a junction tree representation with only singleton separator sets. Let $C_1, \ldots, C_k$ denote the maximal cliques in $\mathcal{M}(G_0)$, and let $\{G_0^\ell\}_{\ell=1}^k$ denote the corresponding restrictions of G_0 to the cliques. Then*

$$\gamma_\Omega(G_0) \geq \min_{1 \leq \ell \leq k} \gamma_\Omega(G_0^\ell),$$

where

$$\gamma_\Omega(G_0^\ell) := \min_{G^\ell \in \mathcal{D}_\Theta|_{C_\ell}, G^\ell \not\supseteq G_0^\ell} \left\{ \max_{G_1^\ell \supseteq G_0^\ell} \left\{ \frac{\text{score}_\Omega(G^\ell) - \text{score}_\Omega(G_1^\ell)}{|H(G^\ell, G_1^\ell)|} \right\} \right\}$$

is the gap ratio computed over DAGs restricted to clique C_ℓ that are consistent with the moralized graph.

The proof is contained in Appendix D.5.

We might expect the gap ratio $\gamma_\Omega(G_0^\ell)$ to be a function of the size of the clique. In particular, if the treewidth of $\mathcal{M}(G_0)$ is bounded by w and we have $\gamma_\Omega(G_0^\ell) \geq \xi_w$ for all ℓ, Lemma 23 implies that

$$\gamma_\Omega(G_0) \geq \xi_w,$$

and we only need the parameter ξ' appearing in Assumption 2 to be larger than ξ_w, rather than scaling as the inverse of p. We expect a version of Lemma 23 to hold for graphs with bounded treewidth even when the separator sets have larger cardinality, but a full generalization of Lemma 23 and a more accurate characterization of $\gamma_\Omega(G_0)$ for arbitrary graphs is beyond the scope of this paper.

5.3 Systematically Corrupted Data

We now describe how our algorithm for DAG structure estimation in linear SEMs may be extended easily to accommodate systematically corrupted data. This refers to the setting where we observe noisy surrogates $\{z_i\}_{i=1}^n$ in place of $\{x_i\}_{i=1}^n$. Two common examples include the following:

(a) *Additive noise.* We have $z_i = x_i + w_i$, where $w_i \perp\!\!\!\perp x_i$ is additive noise with known covariance Σ_w.

(b) *Missing data.* This is one instance of the more general setting of multiplicative noise. For each $1 \le j \le p$, and independently over coordinates, we have

$$z_{ij} = \begin{cases} x_{ij}, & \text{with probability} \quad 1-\alpha, \\ \star, & \text{with probability} \quad \alpha, \end{cases}$$

where the missing data probability α is either estimated or known.

We again divide our discussion into two parts: estimating Θ_0 and computing score functions based on corrupted data.

5.3.1 Inverse Covariance Estimation

Following the observation of Loh and Wainwright (2013), the graphical Lasso (18) may still be used to estimate the inverse covariance matrix Θ_0 in the high-dimensional setting, where we plug in a suitable estimator $\widehat{\Gamma}$ for the covariance matrix $\Sigma = \mathrm{cov}[x_i]$, based on the corrupted observations $\{z_i\}_{i=1}^n$. For instance, in the additive noise scenario, we may take

$$\widehat{\Gamma} = \frac{Z^T Z}{n} - \Sigma_w, \tag{30}$$

and in the missing data setting, we may take

$$\widehat{\Gamma} = \frac{Z^T Z}{n} \odot M, \tag{31}$$

where $\odot$ denotes the Hadamard product and M is the matrix with diagonal entries equal to $\frac{1}{1-\alpha}$ and off-diagonal entries equal to $\frac{1}{(1-\alpha)^2}$.

Assuming conditions such as sub-Gaussianity, the output $\widehat{\Theta}$ of the modified graphical Lasso (18) is statistically consistent under similar scaling as in the uncorrupted setting (Loh and Wainwright, 2013). For instance, in the additive noise setting, where the z_i's are sub-Gaussian with parameter σ_z^2, Lemma 16 holds with σ^2 replaced by σ_z^2. Analogous results hold in the low-dimensional setting, when the expressions for $\widehat{\Gamma}$ in Equations (30) and (31) are invertible with high probability, and we may simply use $\widehat{\Theta} = (\widehat{\Gamma})^{-1}$.

5.3.2 Computing DAG Scores

We now describe how to estimate score functions for DAGs based on corrupted data. By Equation (19), this reduces to estimating

$$f_{\sigma_j}(S) = \frac{1}{\sigma_j^2} \cdot \mathrm{E}[(x_j - b_j^T x_S)^2],$$

for a subset $S \subseteq \{1, \ldots, p\} \backslash \{j\}$, with $|S| \leq n$. Note that

$$\sigma_j^2 \cdot f_{\sigma_j}(S) = \Sigma_{jj} - 2b_j^T \Sigma_{S,j} + b_j^T \Sigma_{SS} b_j = \Sigma_{jj} - \Sigma_{j,S} \Sigma_{SS}^{-1} \Sigma_{S,j},$$

since $b_j = \Sigma_{SS}^{-1} \Sigma_{S,j}$.

Let $\widehat{\Gamma}$ be the estimator for Σ based on corrupted data used in the graphical Lasso; e.g., Equations (30) and (31). We then use the estimator

$$\widetilde{f}_{\sigma_j}(S) = \frac{1}{\sigma_j^2} \cdot \left(\widehat{\Gamma}_{jj} - \widehat{\Gamma}_{j,S} \widehat{\Gamma}_{SS}^{-1} \widehat{\Gamma}_{S,j} \right). \tag{32}$$

Note in particular that Equation (32) reduces to the expression in Equation (20) in the fully-observed setting. We establish consistency of the estimator in Equation (32), under the following deviation condition on $\widehat{\Gamma}$:

$$\mathbb{P}\left(\left| \left| \left| \widehat{\Gamma}_{SS} - \Sigma_{SS} \right| \right| \right|_2 \geq \sigma^2 \left(\sqrt{\frac{d}{n}} + t \right) \right) \leq c_1 \exp(-c_2 n t^2), \quad \text{for any } |S| \leq d. \tag{33}$$

For instance, such a condition holds in the case of the sub-Gaussian additive noise model (cf. Lemma 29 in Appendix E), with $\widehat{\Gamma}$ given by Equation (30), where $\sigma^2 = \sigma_z^2$.

We have the following result, an extension of Lemma 18 applicable also for corrupted variables:

Lemma 24 *Suppose $\widehat{\Gamma}$ satisfies the deviation condition (33). Suppose $|N_\Theta(j)| \leq d$ for all j. Then*

$$|\widetilde{f}_{\sigma_j}(S) - f_{\sigma_j}(S)| \leq \frac{c_0 \sigma^2}{\sigma_j^2} \sqrt{\frac{\log p}{n}}, \qquad \forall j \text{ and } S \subseteq N_\Theta(j),$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$.

The proof is contained in Appendix D.6. In particular, Corollary 22, providing guarantees for statistical consistency, also holds.

6. Computational Considerations

In practice, the main computational bottleneck in inferring the DAG structure comes from having to compute score functions over a large number of DAGs. The simplest approach of searching over all possible permutation orderings of indices gives rise to $p!$ candidate DAGs, which scales exponentially with p. In this section, we describe how the result of Theorem 2 provides a general framework for achieving vast computational savings for finding the best-scoring DAG when data are generated from a linear SEM. We begin by reviewing existing methods, and describe how our results may be used in conjunction with dynamic programming to produce accurate and efficient DAG learning.

6.1 Decomposable Score Functions

Following the literature, we call a score function over DAGs *decomposable* if it may be written as a sum of score functions over individual nodes, each of which is a function of

only the node and its parents:

$$\text{score}(G) = \sum_{j=1}^{p} \text{score}_j(\text{Pa}_G(j)).$$

Note that we allow the score functions to differ across nodes. Consistent with our earlier notation, the goal is to find the DAG $G \in \mathcal{D}$ that minimizes score(G).

Some common examples of decomposable scores that are used for DAG inference include maximum likelihood, BDe, BIC, and AIC (Chickering, 1995). By Equation (19), the squared ℓ_2-score is clearly decomposable, and it gives an example where score$_j$ differs over nodes. Another interesting example is the nonparametric maximum likelihood, which extends the ordinary likelihood method for scoring DAGs (Nowzohour and Bühlmann, 2014).

Various recent results have focused on methods for optimizing a decomposable score function over the space of candidate DAGs in an efficient manner. Some methods include exhaustive search (Silander and Myllymaki, 2006), greedy methods (Chickering, 2002), and dynamic programming (Ordyniak and Szeider, 2012; Korhonen and Parviainen, 2013). We will focus here on a dynamic programming method that takes as input an undirected graph and outputs the best-scoring DAG with skeleton contained in the input graph.

6.2 Dynamic Programming

In this section, we detail a method due to Ordyniak and Szeider (2012) that will be useful for our purposes. Given an input undirected graph G_I and a decomposable score function, the dynamic programming algorithm finds a DAG with minimal score that has skeleton contained in G_I. Let $\{N_I(j)\}_{j=1}^p$ denote the neighborhood sets of G_I. The runtime of the dynamic programming algorithm is exponential in the treewidth w of G_I; hence, the algorithm is only tractable for bounded-treewidth graphs. For further characterizations of graphs with bounded treewidth, including some relevant examples, see Bodlaender (1998).

The main steps of the dynamic programming algorithm are as follows. For a review of basic terminology of graph theory, including treewidth and tree decompositions, see Appendix A; for further details and a proof of correctness, see Ordyniak and Szeider (2012).

1. Construct a *tree decomposition* of G_I with minimal treewidth.

2. Construct a *nice tree decomposition* of the graph. Let $\chi(t)$ denote the subset of $\{1, \ldots, p\}$ associated to a node t in the nice tree decomposition.

3. Starting from the leaves of the nice tree decomposition up to the root, compute the *record* for each node t. The record $\mathcal{R}(t)$ is the set of tuples (a, p, s) corresponding to minimal-scoring DAGs defined on the vertices $\chi^*(t)$ in the subtree attached to t, with skeleton contained in G_I. For each such DAG, s is the score, a lists the parent sets of vertices in $\chi(t)$, such that $a(v) \subseteq N_I(v)$ for each $v \in \chi(t)$, and $a(v)$ restricted to $\chi^*(t)$ agrees with the partial DAG; and p lists the directed paths between vertices in $\chi(t)$. The records $\mathcal{R}(t)$ may computed recursively over the nice tree decomposition as follows:

   - **Join node:** Suppose t has children t_1 and t_2. Then $\mathcal{R}(t)$ is the union of tuples (a, p, s) formed by tuples $(a_1, p_1, s_1) \in \mathcal{R}(t_1)$ and $(a_2, p_2, s_2) \in \mathcal{R}(t_2)$, where (1)

$a = a_1 = a_2$; (2) p is the transitive closure of $p_1 \cup p_2$; (3) p contains no cycles; and (4) $s = s_1 + s_2$.

- **Introduce node:** Suppose t is an introduce node with child t', such that $\chi(t) = \chi(t') \cup \{v_0\}$. Then $\mathcal{R}(t)$ is the set of tuples (a, p, s) formed by pairs $P \subseteq N_I(v_0)$ and $(a', p', s') \in \mathcal{R}(t')$, such that (1) $a(v_0) = P$; (2) for every $v \in \chi(t')$, we have $a(v) = a'(v)$; (3) p is the transitive closure of $p' \cup \{(u, v_0) : u \in P\} \cup \{(v_0, u) : v_0 \in a'(u), u \in \chi(t')\}$; (4) p contains no cycles; and (5) $s = s'$.
- **Forget node:** Suppose t is a forget node with child t', such that $\chi(t') = \chi(t) \cup \{v_0\}$. Then $\mathcal{R}(t)$ is the set of tuples (a, p, s) formed from tuples $(a', p', s') \in \mathcal{R}(t')$, such that (1) $a(u) = a'(u),\ \forall u \in \chi(t)$; (2) $p = \{(u, v) \in p' : u, v \in \chi(t)\}$; and (3) $s = s' + \text{score}_{v_0}(a'(v_0))$.

Note that Korhonen and Parviainen (2013) present a variant of this dynamic programming method, also using a nice tree decomposition, which is applicable even for graphs with unbounded degree but bounded treewidth. They assume that the starting undirected graph G_I is a superset of the moralized DAG. Their algorithm runs in time linear in p and exponential in w. Since $\text{supp}(\Theta_0)$ exactly corresponds to the edge set of $\mathcal{M}(G_0)$, the alternative method will also lead to correct recovery. In practice, the relative efficiency of the two dynamic programming algorithms will rely heavily on the structure of $\mathcal{M}(G_0)$, and it is an interesting direction of future work to investigate the behavior of the two dynamic programming algorithms for different graph structures.

6.3 Runtime

We first review the runtime of various components of the dynamic programming algorithm described in Section 6.2. This is mentioned briefly in Ordyniak and Szeider (2012), but we include details for completeness before comparing the runtime of our overall procedure with other causal inference methods. In our calculations, we assume the treewidth w of G is bounded and treat w as a constant.

The first step involves constructing a tree decomposition of minimal treewidth w, which may be done in time $\mathcal{O}(p)$. The second step involves constructing a nice tree decomposition. Given a tree decomposition of width w, a nice tree decomposition with $\mathcal{O}(p)$ nodes and treewidth w may be constructed in $\mathcal{O}(p)$ time (see Appendix A.2). Finally, the third step involves computing records for nodes in the nice tree decomposition. We consider the three different types of nodes in succession. Note that

$$|\mathcal{R}(t)| \leq 2^{(w+1)(w+d)}, \qquad \forall t, \tag{34}$$

where $d = \max_j |N_I(j)|$. This is because the number of choices of parent sets of any vertex in $\chi(t)$ is bounded by 2^d, leading to a factor of $2^{d(w+1)}$, and the number of possible pairs that are connected by a path is bounded by $2^{(w+1)w}$.

- If t is a join node with children t_1 and t_2, we may compute $\mathcal{R}(t)$ by comparing pairs of records in $\mathcal{R}(t_1)$ and $\mathcal{R}(t_2)$; by Inequality (34), this may be done in time $\mathcal{O}(2^{2(w+1)(w+d)})$.

- If t is an introduce node with child t', we may compute $\mathcal{R}(t)$ by considering records in $\mathcal{R}(t')$ and parent sets of the introduced node v_0. Since the number of choices for the latter is bounded by 2^d, we conclude that $\mathcal{R}(t)$ may be computed in time $\mathcal{O}(2^{(w+1)(w+d)+d})$.

- Clearly, if t is a forget node, then $\mathcal{R}(t)$ may be computed in time $\mathcal{O}(2^{(w+1)(w+d)})$.

Altogether, we conclude that all records of nodes in the nice tree decomposition may be computed in time $\mathcal{O}(p \cdot 2^{2(w+1)(w+d)})$. Combined with the graphical Lasso preprocessing step for estimating $\mathcal{M}(G_0)$, this leads to an overall complexity of $\mathcal{O}(p^2)$. This may be compared to the runtime of other standard methods for causal inference, including the PC algorithm (Spirtes et al., 2000), which has computational complexity $\mathcal{O}(p^w)$, and (direct) LiNGAM (Shimizu et al., 2006, 2011), which requires time $\mathcal{O}(p^4)$. It has been noted that both the PC and LiNGAM algorithms may be expedited when prior knowledge about the DAG space is available, further highlighting the power of Theorem 2 as a preprocessing step for any causal inference algorithm.

7. Discussion

We have provided a new framework for estimating the DAG corresponding to a linear SEM. We have shown that the inverse covariance matrix of linear SEMs always reflects the edge structure of the moralized graph, even in non-Gaussian settings, and the reverse statement also holds under a mild faithfulness assumption. Furthermore, we have shown that when the error variances are known up to close precision, a simple weighted squared ℓ_2-loss may be used to select the correct DAG. As a corollary, we have established identifiability for the class of linear SEMs with error variances specified up to a constant multiple. We have proved that our methods are statistically consistent, under reasonable assumptions on the gap between the score of the true DAG and the next best DAG in the model class. A characterization of this gap parameter for various graphical structures is the topic of future work.

We have also shown how dynamic programming may be used to select the best-scoring DAG in an efficient manner, assuming the treewidth of the moralized graph is small. Our results relating the inverse covariance matrix to the moralized DAG provide a powerful method for reducing the DAG search space as a preprocessing step for dynamic programming, and are the first to provide rigorous guarantees for when the graphical Lasso may be used in non-Gaussian settings. Note that the dynamic programming algorithm discussed in this paper only uses the information that the true DAG has skeleton lying in the input graph, and does *not* incorporate any information about (a) the fact that the data comes from a linear SEM; or (b) the fact that the input graph exactly equals the moralized DAG. Intuitively, both types of information should place significant constraints on the restricted DAG space, leading to further speedups in the dynamic programming algorithm. Perhaps these restrictions would make it possible to establish a version of dynamic programming for DAGs where the moralized graph has bounded degree but large treewidth.

An important open question concerns scoring candidate DAGs when the diagonal matrix Ω_0 of error variances is unknown. As we have seen, using the weighted squared ℓ_2-loss to score DAGs may produce a graph that is far from the true DAG when Ω_0 is misspecified.

Alternatively, it would be useful to have a checkable condition that would allow us to verify whether a given matrix Ω will correctly select the true DAG, or to be able to select the true Ω_0 from among a finite collection of candidate matrices.

Acknowledgments

We acknowledge all the members of the Seminar für Statistik for providing an immensely hospitable and fruitful environment when PL was visiting ETH, and the Forschungsinstitut für Mathematik at ETH Zürich for financial support. We also thank Varun Jog for helpful discussions. PL was additionally supported by a Hertz Fellowship and an NSF Graduate Research Fellowship. We thank the AE and anonymous reviewers for helpful feedback.

Appendix A. Graph-Theoretic Concepts

In this Appendix, we review some fundamental concepts in graph theory that we use in our exposition. We begin by discussing junction trees, and then move to the related notion of tree decompositions. Note that these are purely graph-theoretic operations that may be performed on an arbitrary undirected graph.

A.1 Junction Trees

We begin with the basic junction tree framework. For more details, see Lauritzen (1996) or Koller and Friedman (2009).

For an undirected graph $G = (V, E)$, a *triangulation* is an augmented graph $\widetilde{G} = (V, \widetilde{E})$ that contains no chordless cycles of length greater than three. By classical graph theory, any triangulation $\widetilde{G}$ gives rise to a *junction tree* representation of G, where nodes in the junction tree are subsets of V corresponding to maximal cliques of $\widetilde{G}$, and the intersection of any two adjacent cliques C_1 and C_2 in the junction tree is referred to as a *separator set* $S = C_1 \cap C_2$. Furthermore, any junction tree must satisfy the *running intersection property*, meaning that for any two nodes in the junction tree, say corresponding to cliques C_j and C_k, the intersection $C_j \cap C_k$ must belong to every separator set on the unique path between C_j and C_k in the junction tree. The *treewidth* of G is defined to be one less than the size of the largest clique in any triangulation $\widetilde{G}$ of G, minimized over all triangulations.

As a classic example, note that if G is a tree, then G is already triangulated, and the junction tree parallels the tree structure of G. The maximal cliques in the junction tree are equal to the edges of G and the separator sets correspond to singleton vertices. The treewidth of G is consequently equal to 1.

A.2 Tree Decompositions

We now highlight some basic concepts of tree decompositions and nice tree decompositions used in our dynamic programming framework. Our exposition follows Kloks (1994).

Let $G = (V, E)$ be an undirected graph. A *tree decomposition* of G is a tree T with node set W such that each node $t \in W$ is associated with a subset $V_t \subseteq V$, and the following properties are satisfied:

(a) $\bigcup_{t \in T} V_t = V$;

(b) for all $(u, v) \in E$, there exists a node $t \in W$ such that $u, v \in V_t$;

(c) for each $v \in V$, the set of nodes $\{t : v \in V_t\}$ forms a subtree of T.

The *width* of the tree decomposition is $\max_{t \in T} |V_t| - 1$. The *treewidth* of G is the minimal width of any tree decomposition of G; this quantity agrees with the treewidth defined in terms of junction trees in the previous section. If G has bounded treewidth, a tree decomposition with minimum width may be constructed in time $\mathcal{O}(|V|)$ (cf. Chapter 15 of Kloks 1994).

A *nice tree decomposition* is rooted tree decomposition satisfying the following properties:

(a) every node has at most two children;

(b) if a node t has two children r and s, then $V_t = V_r = V_s$;

(c) if a node t has one child s, then either

(i) $|V_t| = |V_s| + 1$ and $V_s \subseteq V_t$, or

(ii) $|V_s| = |V_t| + 1$ and $V_t \subseteq V_s$.

Nodes of the form (b), (c)(i), and (c)(ii) are called *join* nodes, *introduce* nodes, and *forget* nodes, respectively. Given a tree decomposition of G with width w, a nice tree decomposition with width w and at most $4|V|$ nodes may be computed in time $\mathcal{O}(|V|)$ (cf. Lemma 13.1.3 of Kloks 1994).

Appendix B. Matrix Derivations

In this section, we present a few matrix results that are used to prove Theorem 7.

Define a *unit lower triangular (LT)* matrix to be a lower triangular matrix with 1's on the diagonal. Recall that matrices A and B are *permutation similar* if there exists a permutation matrix P such that $A = PBP^T$. Call a matrix *permutation unit LT* if it is permutation similar to a unit lower triangular matrix. We have the following lemma:

Lemma 25 *Suppose A and B are permutation unit LT matrices, and suppose $AA^T = BB^T$. Then $A = B$.*

Proof Under the appropriate relabeling, we assume without loss of generality that A is unit LT. There exists a permutation matrix P such that $C := PBP^T$ is also unit LT. We have

$$PAA^TP^T = CC^T. \tag{35}$$

Let π be the permutation on $\{1, \ldots, n\}$ such that $P_{i,\pi(i)} = 1$ for all i, and P has 0's everywhere else. Define the notation

$$\widetilde{a}_{ij} := (PA)_{ij}, \qquad \text{and} \qquad m_{ij} := (CC^T)_{ij},$$

and let $\{c_{ij}\}$ denote the entries of C. We will make use of the following equalities, which follow from Equation (35) and the fact that C is unit LT:

$$\sum_k \widetilde{a}_{ik}\widetilde{a}_{jk} \;=\; m_{ij} \;=\; \sum_{k<j} c_{ik}c_{jk} + c_{ij}, \qquad \forall i > j, \tag{36}$$

and

$$\sum_k \widetilde{a}_{ik}^2 \;=\; m_{ii} \;=\; 1 + \sum_{k<i} c_{ik}^2. \tag{37}$$

We now derive the following equality:

$$\widetilde{a}_{i,\pi(j)} = c_{ij}, \qquad \forall i, j. \tag{38}$$

Note that Equation (38) implies $(PA)P^T = C$, from which it follows that $A = B$.

If $j = i$, we have $\widetilde{a}_{i,\pi(i)} = 1$ trivially, since A has 1's on the diagonal. For the remaining cases, we induct on i. When $i = 1$, we need to show that $\widetilde{a}_{1,\pi(1)} = 1$ and all other entries in the first row are 0. By Equation (37), we have

$$\sum_k \widetilde{a}_{1k}^2 = m_{11} = 1.$$

Since $\widetilde{a}_{1,\pi(1)} = 1$, it is clear that $\widetilde{a}_{1k} = 0$ for all $k \neq \pi(1)$, establishing the base case.

For the induction step, consider $i > 1$. We first show that $\widetilde{a}_{i,\pi(j)} = c_{ij}$ for all $j < i$ by a sub-induction on j. For $j = 1$, we have by Equation (36) and the base result for $i = 1$ that

$$\widetilde{a}_{i,\pi(1)} = m_{i,1} = c_{i,1},$$

which is exactly what we want. For the sub-induction step, consider $1 < j < i$, and suppose $\widetilde{a}_{i,\pi(\ell)} = c_{i\ell}$ for all $\ell < j$. Note that $\widetilde{a}_{j,\pi(\ell)} = 0$ for all $\ell > j$ by the outer induction hypothesis. Hence, Equation (36) and the fact that $\widetilde{a}_{j,\pi(j)} = 1$ gives

$$\sum_{\ell<j} \widetilde{a}_{i,\pi(\ell)}\widetilde{a}_{j,\pi(\ell)} + \widetilde{a}_{i,\pi(j)} \;=\; m_{ij} \;=\; \sum_{k<j} c_{ik}c_{jk} + c_{ij}. \tag{39}$$

Since also $\widetilde{a}_{j,\pi(\ell)} = c_{j\ell}$ for $\ell < j$ by the outer induction hypothesis, Equation (39) condenses to

$$\sum_{\ell<j} c_{i\ell}c_{j\ell} + \widetilde{a}_{i,\pi(j)} = \sum_{k<j} c_{ik}c_{jk} + c_{ij},$$

from which it follows that $\widetilde{a}_{i,\pi(j)} = c_{ij}$, as wanted. This completes the inner induction and shows that $\widetilde{a}_{i,\pi(j)} = c_{ij}$, for all $j < i$. Finally, note that by Equation (37), we have

$$m_{ii} = \sum_k \widetilde{a}_{ik}^2 = 1 + \sum_{j\neq i} \widetilde{a}_{i,\pi(j)}^2 \geq 1 + \sum_{j<i} \widetilde{a}_{i,\pi(j)}^2 = 1 + \sum_{j<i} c_{ij}^2 = m_{ii},$$

implying that we must have $\widetilde{a}_{i,\pi(j)} = 0$, for all $j > i$. This establishes Equation (38).

■

We also need the following known result (cf. Exercise 7.8.19 in Horn and Johnson 1990). We include a proof for completeness.

Lemma 26 *Suppose $A \in \mathbb{R}^{n\times n}$ is positive definite with $\det(A) = 1$. Then*

$$\min\{\mathrm{tr}(AB) : B \succ 0 \text{ and } \det(B) = 1\} = n.$$

Proof Consider the singular value decomposition $A = U\Lambda U^*$, and note that $\mathrm{tr}(AB) = \mathrm{tr}(\Lambda(U^*BU))$. Denote $b_{ij} := (U^*BU)_{ij}$ and $\lambda_i := \Lambda_{ii}$. Then by the AM-GM inequality and Hadamard's inequality, we have

$$\frac{1}{n}\cdot \mathrm{tr}(\Lambda(U^*BU)) \geq \left(\prod_i \lambda_i b_{ii}\right)^{1/n} = \left(\det(A)\cdot \prod_i b_{ii}\right)^{1/n} \geq (\det(U^*BU))^{1/n} = 1,$$

implying the result. ■

Building upon Lemmas 25 and 26, we obtain the following result:

Lemma 27 *Suppose A and B are $n \times n$ permutation unit LT matrices. Then*

$$\min_B \mathrm{tr}(AA^TB^TB) \geq n, \tag{40}$$

with equality achieved if and only if $B = A^{-1}$.

Proof Write $A' = AA^T$ and $B' = B^TB$, and note that since $\det(A) = \det(B) = 1$, we also have $\det(A') = \det(B') = 1$. Then Inequality (40) holds by Lemma 26.

To recover the conditions for equality, note that equality holds in Hadamard's inequality if and only if some $b_{ii} = 0$ or the matrix U^*BU is diagonal. Note that the first case is not possible, since $U^*BU \succ 0$. In the second case, we see that in addition, we need $b_{ii} = \frac{1}{\lambda_i}$ for all i in order to achieve equality in the AM-GM inequality. It follows that $U^*BU = \Lambda^{-1}$, so $AA^T = B^{-1}B^{-T}$.

Since A and B are permutation unit LT, Lemma 25 implies that the last equality can only hold when $B = A^{-1}$. ■

Appendix C. Proofs for Population-Level Results

In this section, we provide proofs for the remaining results in Sections 3 and 4.

C.1 Proof of Lemma 1

We first show that Ω is a diagonal matrix. Consider $j < k$; we will show that $\epsilon_j \perp\!\!\!\perp \epsilon_k$, from which we conclude that

$$\mathrm{E}[\epsilon_j\epsilon_k] = \mathrm{E}[\epsilon_j]\cdot \mathrm{E}[\epsilon_k] = 0.$$

Indeed, we have $\epsilon_k \perp\!\!\!\perp (X_1, \ldots, X_{k-1})$ by assumption. Since $\epsilon_j = X_j - b_j^TX$ is a deterministic function of $(X_1, \ldots, X_j)$, it follows that $\epsilon_k \perp\!\!\!\perp \epsilon_j$, as claimed.

Turning to Equations (5) and (6), note that Equation (4) implies

$$\Theta = \Sigma^{-1} = (I - B)\Omega^{-1}(I - B)^T.$$

Then expanding and using the fact that B is upper triangular and Ω is diagonal, we obtain Equations (5) and (6).

C.2 Proof of Lemma 6

Since $G_0 \subseteq G$, we have $\mathrm{Pa}_{G_0}(j) \subseteq \mathrm{Pa}_G(j)$, for each j. Furthermore, no element of $\mathrm{Pa}_G(j)$ may be a descendant of j in G_0, since this would contradict the fact that G contains no cycles. By the Markov property of G_0, we therefore have

$$X_j \perp\!\!\!\perp X_{\mathrm{Pa}_G(j) \setminus \mathrm{Pa}_{G_0}(j)} \mid X_{\mathrm{Pa}_{G_0}(j)}.$$

Thus, the linear regression coefficients for X_j regressed upon $X_{\mathrm{Pa}_G(j)}$ are simply the linear regression coefficients for X_j regressed upon $X_{\mathrm{Pa}_{G_0}(j)}$ (and the remaining coefficients for $X_{\mathrm{Pa}_G(j) \setminus \mathrm{Pa}_{G_0}(j)}$ are zero). By Remark 5, we conclude that

$$B_0 = B_G = \arg\min_{B \in \mathcal{U}_G} \{\mathrm{score}_\Omega(B)\},$$

and the uniqueness of B_0 follows from the uniqueness of B_G.

C.3 Proof of Theorem 9

From the decomposition (7), it is easy to see that for any $B \in \mathcal{U}$, we have

$$a_{\min} \leq \frac{\mathrm{score}_{\Omega_1}(B)}{\mathrm{score}_{\Omega_0}(B)} \leq a_{\max}, \tag{41}$$

simply by comparing individual terms; e.g.,

$$\begin{aligned}\frac{1}{(\Omega_1)_{jj}} \cdot \mathrm{E}[(X_j - b_j^T X)^2] &\leq \max_j \left\{ \frac{1/(\Omega_1)_{jj}}{1/(\Omega_0)_{jj}} \right\} \cdot \frac{1}{(\Omega_0)_{jj}} \cdot \mathrm{E}[(X_j - b_j^T X)^2] \\ &= a_{\max} \left(\frac{1}{(\Omega_0)_{jj}} \cdot \mathrm{E}[(X_j - B_j^T X)^2] \right).\end{aligned}$$

Note that if $G \supseteq G_0$, then by Lemma 6, the matrix B_0 is the unique minimizer of $\mathrm{score}_{\Omega_1}(B)$ among the class $\mathcal{U}_G$. Now consider $G \not\supseteq G_0$ and $B \in \mathcal{U}_G$. We have

$$\left(1 + \frac{\xi}{p}\right) \cdot \mathrm{score}_{\Omega_0}(B_0) = \min_{G' \in \mathcal{D},\, G' \not\supseteq G_0} \{\mathrm{score}_{\Omega_0}(G')\} \leq \mathrm{score}_{\Omega_0}(G) \leq \mathrm{score}_{\Omega_0}(B), \tag{42}$$

where we have used the definition of the gap (11) and the fact that $\mathrm{score}_{\Omega_0}(B_0) = p$ by Theorem 7 in the first inequality. Hence,

$$\mathrm{score}_{\Omega_1}(B_0) \leq a_{\max} \cdot \mathrm{score}_{\Omega_0}(B_0) \leq \frac{a_{\max}}{1 + \xi/p} \cdot \mathrm{score}_{\Omega_0}(B) \leq \frac{a_{\max}}{a_{\min}(1 + \xi/p)} \cdot \mathrm{score}_{\Omega_1}(B),$$

where the first and third inequalities use Inequality (41), and the second inequality uses Inequality (42). By the assumption (12), it follows that

$$\mathrm{score}_{\Omega_1}(B_0) \leq \mathrm{score}_{\Omega_1}(B),$$

as wanted. The statement regarding strict inequality is clear.

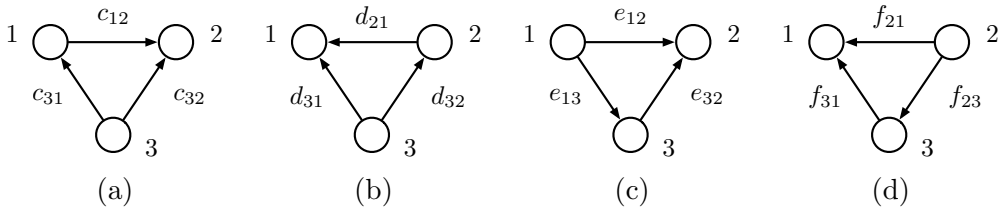

Figure 3: Alternative DAGs.

C.4 Proof of Lemma 11

We first consider the case when $r \geq 1$. Then $\frac{a_{\max}}{a_{\min}} = r^2$, so combining Theorem 9 with the expression (15), we have the sufficient condition

$$r^2 \leq 1 + \frac{b_0^4}{2(r^4 + b_0^2 r^2)}.$$

Rearranging gives

$$b_0^4 - 2r^2(r^2-1)b_0^2 - 2r^4(r^2-1) \geq 0,$$

which is equivalent to

$$b_0^2 \geq \frac{2r^2(r^2-1) + \sqrt{(4r^4(r^2-1)^2 + 8r^4(r^2-1))}}{2}.$$

Simplifying yields the desired expression.

If instead $r \leq 1$, we have $\frac{a_{\max}}{a_{\min}} = \frac{1}{r^2}$, so the sufficient condition becomes

$$\frac{1}{r^2} \leq 1 + \frac{b_0^4}{2(r^4 + b_0^2 r^2)},$$

which is equivalent to

$$b_0^4 - 2(1-r^2)b_0^2 - 2r^2(1-r^2) \geq 0,$$

or

$$b_0^2 \geq \frac{2(1-r)^2 + \sqrt{4(1-r^2)^2 + 8r^2(1-r^2)}}{2}.$$

Simplifying further yields the expression.

C.5 Proof of Lemma 13

To compute the gap ξ, it is sufficient to consider the graphs in Figure 3. Indeed, we have $\text{score}_{\Omega_0}(G) \leq \text{score}_{\Omega_0}(G')$ whenever $G' \subseteq G$, so we only need to consider maximal elements in the poset of DAGs not containing G_0. Consider the graphs given by autoregression matrices

$$C = \begin{pmatrix} 0 & c_{12} & 0 \\ 0 & 0 & 0 \\ c_{31} & c_{32} & 0 \end{pmatrix}, \qquad E = \begin{pmatrix} 0 & e_{12} & e_{13} \\ 0 & 0 & 0 \\ 0 & e_{32} & 0 \end{pmatrix},$$

corresponding to the DAGs in panels (a) and (c) of Figure 3. A simple calculation shows that

$$\text{score}_{\Omega_0}(C) = 3 + c_{31}^2 b_{13}^2 + c_{32}^2 b_{23}^2 + c_{31}^2 b_{23}^2 \frac{d_2^2}{d_1^2} + \left(c_{12}\frac{d_1}{d_2} + c_{32} b_{13}\frac{d_1}{d_2}\right)^2 + \left(c_{31}\frac{d_3}{d_1} - b_{13}\frac{d_1}{d_3}\right)^2 + \left(c_{32}\frac{d_3}{d_2} - b_{23}\frac{d_2}{d_3}\right)^2,$$

which is minimized for

$$c_{12} = -b_{13}c_{32}, \qquad c_{31} = \frac{b_{13}}{\frac{d_3^2}{d_1^2} + b_{23}^2\frac{d_2^2}{d_1^2} + b_{13}^2}, \qquad c_{32} = \frac{b_{23}}{\frac{d_3^2}{d_2^2} + b_{23}^2},$$

leading to

$$\xi_1 = \min_{c_{12},c_{31},c_{32}} \{\text{score}_{\Omega_0}(C) - \text{score}_{\Omega_0}(B_0)\} = \frac{b_{23}^4}{\frac{d_3^4}{d_2^4} + b_{23}^2\frac{d_3^2}{d_2^2}} + \frac{b_{13}^4 + b_{13}^2 b_{23}^2\frac{d_2^2}{d_1^2}}{\frac{d_3^4}{d_1^4} + b_{13}^2\frac{d_3^2}{d_1^2} + b_{23}^2\frac{d_2^2 d_3^2}{d_1^4}}. \tag{43}$$

Similarly, we may compute

$$\text{score}_{\Omega_0}(E) = 3 + \left(e_{12}\frac{d_1}{d_2} + e_{32}b_{13}\frac{d_1}{d_2}\right)^2 + \left(e_{13}\frac{d_1}{d_3} - b_{13}\frac{d_1}{d_3}\right)^2 + \left(e_{32}\frac{d_3}{d_2} - b_{23}\frac{d_2}{d_3}\right)^2,$$

which is minimized for

$$e_{12} = -b_{13}, \qquad e_{13} = b_{13}, \qquad e_{32} = \frac{b_{23}}{\frac{d_3^2}{d_2^2} + b_{23}^2},$$

leading to

$$\xi_2 = \min_{e_{12},e_{13},e_{32}} \{\text{score}_{\Omega_0}(E) - \text{score}_{\Omega_0}(B_0)\} = \frac{b_{23}^4}{\frac{d_3^4}{d_2^4} + b_{23}^2\frac{d_3^2}{d_2^2}}. \tag{44}$$

Finally, note that the graphs in panels (b) and (d) of Figure 3 are mirror images of the graphs in panels (a) and (c), respectively. Hence, we obtain

$$\xi_3 = \min_{d_{21},d_{31},d_{32}} \{\text{score}_{\Omega_0}(D) - \text{score}_{\Omega_0}(B_0)\} = \frac{b_{13}^4}{\frac{d_3^4}{d_1^4} + b_{13}^2\frac{d_3^2}{d_1^2}} + \frac{b_{23}^4 + b_{13}^2 b_{23}^2\frac{d_1^2}{d_2^2}}{\frac{d_3^4}{d_2^4} + b_{23}^2\frac{d_3^2}{d_2^2} + b_{13}^2\frac{d_1^2 d_3^2}{d_2^4}},$$

$$\xi_4 = \min_{f_{21},f_{23},f_{31}} \{\text{score}_{\Omega_0}(F) - \text{score}_{\Omega_0}(B_0)\} = \frac{b_{13}^4}{\frac{d_3^4}{d_1^4} + b_{13}^2\frac{d_3^2}{d_1^2}},$$

simply by swapping the roles of nodes 1 and 2. Taking $\xi = \min\{\xi_1, \xi_2, \xi_3, \xi_4\}$ then yields the desired result.

Appendix D. Proofs for Statistical Consistency

In this Appendix, we provide the proofs for the lemmas on statistical consistency stated in Section 5.

D.1 Proof of Lemma 15

This result follows from the fact that

$$\|\widehat{\Theta} - \Theta_0\|_{\max} \leq \left|\!\left|\!\left| \widehat{\Theta} - \Theta_0 \right|\!\right|\!\right|_2,$$

together with results on the spectral norm of sub-Gaussian covariances and their inverses (see Lemma 29 in Appendix E).

D.2 Proof of Lemma 18

First consider a fixed pair (j, S) such that $S \subseteq N_\Theta(j)$. We may write

$$x_j = b_j^T x_S + e_j, \tag{45}$$

where e_j has zero mean and is uncorrelated with x_S (and also depends on the choice of S). In matrix notation, we have

$$\widehat{b}_j = (X_S^T X_S)^{-1}(X_S^T X_j) = (X_S^T X_S)^{-1} X_S^T (X_S b_j + E_j) = b_j + (X_S^T X_S)^{-1} X_S^T E_j,$$

where the second equality follows from Equation (45). Hence,

$$\begin{aligned}
\sigma_j^2 \cdot \widehat{f}_{\sigma_j}(S) &= \frac{1}{n}\|X_j - X_S \widehat{b}_j\|_2^2 \\
&= \frac{1}{n}\|X_S(b_j - \widehat{b}_j) + E_j\|_2^2 \\
&= \frac{1}{n}\|(I - (X_S^T X_S)^{-1} X_S^T) E_j\|_2^2. \qquad (46)
\end{aligned}$$

By the triangle inequality, we have

$$\left| \|(I - (X_S^T X_S)^{-1} X_S^T) E_j\|_2 - \|E_j\|_2 \right| \leq \|(X_S^T X_S)^{-1} X_S^T E_j\|_2 \leq \left|\!\left|\!\left| (X_S^T X_S)^{-1} X_S^T \right|\!\right|\!\right|_2 \cdot \|E_j\|_2. \tag{47}$$

Furthermore,

$$\begin{aligned}
\left|\!\left|\!\left| (X_S^T X_S)^{-1} X_S^T \right|\!\right|\!\right|_2 &= \left|\!\left|\!\left| X_S (X_S^T X_S)^{-1} \right|\!\right|\!\right|_2 \\
&= \sup_{\|v\|_2 \leq 1} \left\{ v^T (X_S^T X_S)^{-1} X_S^T X_S (X_S^T X_S)^{-1} v \right\}^{1/2} \\
&= \sup_{\|v\|_2 \leq 1} \left\{ v^T (X_S^T X_S)^{-1} v \right\}^{1/2} \\
&= \frac{1}{\sqrt{n}} \left|\!\left|\!\left| \left(\frac{X_S^T X_S}{n} \right)^{-1} \right|\!\right|\!\right|_2^{1/2} \\
&\leq \frac{C}{\sqrt{n}} \left(\left|\!\left|\!\left| \Sigma_{SS}^{-1} \right|\!\right|\!\right|_2 + 2\sigma^2 \left|\!\left|\!\left| \Sigma_{SS}^{-1} \right|\!\right|\!\right|_2^2 \cdot \max\{\delta, \delta^2\} \right)^{1/2},
\end{aligned}$$

with probability at least $1 - 2\exp(-cnt^2)$, where $\delta = c'\sqrt{\frac{|S|}{n}} + c''t$, by Lemma 29. Taking a union bound over all 2^d choices for S and p choices for j, and setting $t = c\sqrt{\frac{d + \log p}{n}}$, we have

$$\left|\!\left|\!\left| (X_S^T X_S)^{-1} X_S^T \right|\!\right|\!\right|_2 \leq \frac{C'}{\sqrt{n}}, \qquad \forall S \text{ s.t. } S \subseteq N_\Theta(j) \text{ for some } j, \tag{48}$$

with probability at least

$$1 - c_1 \exp(-c_2 n t^2) = 1 - c_1 \exp(-c_2 n t^2 + d \log 2 + \log p) \geq 1 - c_1 \exp(-c_2'(d + \log p)).$$

Combining Inequalities (47) and (48), we have the uniform bound

$$\left(1 - \frac{C'}{\sqrt{n}}\right)^2 \|E_j\|_2^2 \leq \|I - (X_S^T X_S)^{-1} X_S^T) E_j\|_2^2 \leq \left(1 + \frac{C'}{\sqrt{n}}\right)^2 \|E_j\|_2^2,$$

with high probability, which together with Equation (46) implies that

$$\left| \sigma_j^2 \cdot \widehat{f}_{\sigma_j}(S) - \frac{1}{n} \|E_j\|_2^2 \right| \leq \frac{3C'}{\sqrt{n}} \cdot \frac{1}{n} \|E_j\|_2^2, \tag{49}$$

using the fact that

$$\max\{1 - (1-a)^2, (1+a)^2 - 1\} = \max\{2a - a^2, 2a + a^2\} = 2a + a^2 \leq 3a,$$

for $a = \frac{C'}{\sqrt{n}}$ sufficiently small. Furthermore,

$$\frac{1}{n} \operatorname{E}[\|E_j\|_2^2] = \sigma_j^2 \cdot f_{\sigma_j}(S).$$

Note that the e_j's are i.i.d. sub-Gaussians with parameter at most $c\sigma^2$, since we may write $e_j = \widetilde{b}_j^T x$ for the appropriate $\widetilde{b}_j \in \mathbb{R}^p$, and $\|\widetilde{b}\|_2$ is bounded in terms of the eigenvalues of Σ. Applying the usual sub-Gaussian tail bounds, we then have

$$\mathbb{P}\left(\left| \frac{1}{n} \|E_j\|_2^2 - \frac{1}{n} \operatorname{E}[\|E_j\|_2^2] \right| \geq c\sigma^2 t \right) \leq c_1 \exp(-c_2 n t^2), \qquad \forall j,$$

and taking a union bound over j and setting $t = c' \sqrt{\frac{\log p}{n}}$ gives

$$\max_j \left| \frac{1}{n} \|E_j\|_2^2 - \frac{1}{n} \operatorname{E}[\|E_j\|_2^2] \right| \leq c_0 \sigma^2 \sqrt{\frac{\log p}{n}}, \tag{50}$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$. Combining Inequalities (49) and (50), it follows that

$$\begin{aligned}
\sigma_j^2 |\widehat{f}_{\sigma_j}(S) - f_{\sigma_j}(S)| &\leq \left| \sigma_j^2 \cdot \widehat{f}_{\sigma_j}(S) - \frac{1}{n} \|E_j\|_2^2 \right| + \left| \frac{1}{n} \|E_j\|_2^2 - \frac{1}{n} \operatorname{E}[\|E_j\|_2^2] \right| \\
&\leq \frac{3C'}{\sqrt{n}} \left(\frac{1}{n} \operatorname{E}[\|E_j\|_2^2] + c_0 \sigma^2 \sqrt{\frac{\log p}{n}} \right) + c_0 \sigma^2 \sqrt{\frac{\log p}{n}} \\
&\leq c_0' \sigma^2 \sqrt{\frac{\log p}{n}},
\end{aligned}$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$.

D.3 Proof of Lemma 19

Combining Inequalities (21) and (23) and using the triangle inequality, we have

$$\left|\widehat{\text{score}}_\Omega(G) - \text{score}_\Omega(G)\right| \leq \sum_{j=1}^{p} \left|\widehat{\text{score}}_{\sigma_j}(\text{Pa}_G(j)) - \text{score}_{\sigma_j}(\text{Pa}_G(j))\right| < \frac{\xi(\mathcal{D}_\Theta)}{2}, \tag{51}$$

for all $G \in \mathcal{D}_\Theta$. In particular, for $G_1 \in \mathcal{D}_\Theta$ such that $G_1 \not\supseteq G_0$, we have

$$\begin{aligned}\widehat{\text{score}}_\Omega(G_0) &< \text{score}_\Omega(G_0) + \frac{\xi(\mathcal{D}_\Theta)}{2} \\ &\leq (\text{score}_\Omega(G_1) - \xi(\mathcal{D}_\Theta)) + \frac{\xi(\mathcal{D}_\Theta)}{2} \\ &< \widehat{\text{score}}_\Omega(G_1),\end{aligned}$$

where the first and third inequalities use Inequality (51) and the second inequality uses the definition of the gap $\xi(\mathcal{D}_\Theta)$. This implies Inequality (24).

D.4 Proof of Lemma 21

Consider $G \in \mathcal{D}_\Theta$ with $G \not\supseteq G_0$, and consider $G_1 \supseteq G_0$ such that $\gamma_\Omega(G, G_1)$ is maximized. Note that if $\text{Pa}_G(j) = \text{Pa}_{G_1}(j)$, then certainly,

$$\widehat{f}_{\sigma_j}(\text{Pa}_G(j)) - \widehat{f}_{\sigma_j}(\text{Pa}_{G_1}(j)) = 0 = f_{\sigma_j}(\text{Pa}_G(j)) - f_{\sigma_j}(\text{Pa}_{G_1}(j)).$$

Hence,

$$\left|\left(\widehat{\text{score}}_\Omega(G) - \widehat{\text{score}}_\Omega(G_1)\right) - \left(\text{score}_\Omega(G) - \text{score}_\Omega(G_1)\right)\right| \leq |H(G, G_1)| \cdot \frac{\xi'}{2}, \tag{52}$$

using Inequality (29) and the triangle inequality. Furthermore, by Inequality (28),

$$|H(G, G_1)| \cdot \frac{\xi'}{2} \leq \frac{\text{score}_\Omega(G) - \text{score}_\Omega(G_1)}{\xi'} \cdot \frac{\xi'}{2} = \frac{\text{score}_\Omega(G) - \text{score}_\Omega(G_1)}{2}. \tag{53}$$

Combining Inequalities (52) and (53) gives

$$\widehat{\text{score}}_\Omega(G) - \widehat{\text{score}}_\Omega(G_1) \geq \frac{\text{score}_\Omega(G) - \text{score}_\Omega(G_1)}{2} = \frac{\text{score}_\Omega(G) - \text{score}_\Omega(G_0)}{2} > 0,$$

where the last inequality holds because of the assumption $\xi' > 0$. Hence,

$$G \notin \arg\min_{G \in \mathcal{D}_\Theta} \{\widehat{\text{score}}_\Omega(G)\},$$

implying the desired result.

D.5 Proof of Lemma 23

We begin with a simple lemma:

Lemma 28 *Suppose $\mathcal{M}(G)$ admits a junction tree representation with only singleton separators, and let $C_1, \ldots, C_k$ denote the maximal cliques. If X follows a linear SEM over G, then the marginal distribution of X over the nodes in any clique C_ℓ also follows a linear SEM over C_ℓ, with DAG structure specified by G_ℓ, the restriction of G to C_ℓ. In addition, the autoregression matrix for the marginal SEM is simply the autoregression matrix for the full SEM restricted to the nodes in C_ℓ.*

Proof We relabel the nodes of G so that the natural ordering on $\{1, \ldots, p\}$ is a topological order. Clearly, this induces a topological order over the nodes of G_ℓ, as well. Recall that we have Equation (3); i.e., for each j,

$$X_j = b_j^T X_{1:j-1} + \epsilon_j, \qquad \text{where } \epsilon_j \perp\!\!\!\perp (X_1, \ldots, X_{j-1}). \tag{54}$$

For each $j \in C_\ell$, we define

$$\epsilon'_j := \epsilon_j + \sum_{k<j,\, k \notin \mathrm{Pa}_{G_\ell}(j)} b_{kj} X_k,$$

and note that

$$X_j = b_j^T X_{\mathrm{Pa}_{G_\ell}(j)} + \epsilon'_j,$$

where we have abused notation slightly and used b_j to denote the same vector restricted to $\mathrm{Pa}_{G_\ell}(j)$. We claim that

$$\epsilon'_j \perp\!\!\!\perp X_{\mathrm{Pa}_{G_\ell}(j)}, \tag{55}$$

for each j, implying that the marginal distribution of X over C_ℓ follows a linear SEM with the desired properties.

First consider the case when j is not contained in a separator set of the junction tree. Then all neighbors of j must be contained in C_ℓ, implying that $\mathrm{Pa}_{G_\ell}(j) = \mathrm{Pa}_G(j)$. Since $b_{kj} \neq 0$ only when $k < j$ and $k \in \mathrm{Pa}_G(j)$, this means $\epsilon'_j = \epsilon_j$. The desired independence (55) follows from Equation (54) and the simple fact that $\mathrm{Pa}_{G_\ell}(j) \subseteq \{1, \ldots, j-1\}$. If instead j is a separator node, then either $\mathrm{Pa}_G(j) \subseteq C_\ell$ or $\mathrm{Pa}_G(j) \cap C_\ell = \emptyset$. In the first case, we again have $\mathrm{Pa}_{G_\ell}(j) = \mathrm{Pa}_G(j)$, so the argument proceeds as before. In the second case, we have $\mathrm{Pa}_{G_\ell}(j) = \emptyset$, so the independence relation (55) is vacuous; indeed, we have $\epsilon'_j = \epsilon_j + b_j^T X_{\mathrm{Pa}_G(j)} = X_j$. Hence, condition (55) holds in every case. ■

Now consider any $G \in \mathcal{D}_\Theta$ such that $G \not\supseteq G_0$. Let $\{G^\ell\}_{\ell=1}^k$ denote the restrictions of G to the cliques. By Lemma 28, X follows a linear SEM when restricted to the nodes of C_ℓ; hence, by Lemma 6 and Theorem 7, we have

$$\mathrm{score}_\Omega(G_0^\ell) \leq \mathrm{score}_\Omega(G^\ell), \tag{56}$$

with equality if and only if $G_0^\ell \subseteq G^\ell$. Consider the graph G_1 constructed such that $G_1^\ell = G^\ell$ on cliques C_ℓ such that Inequality (56) holds with equality, and $G_1^\ell = G_0^\ell$ otherwise. In particular, we have $G_0^\ell \subseteq G_1^\ell$, for each ℓ, and

$$\mathrm{score}_\Omega(G_1^\ell) = \mathrm{score}_\Omega(G_0^\ell), \qquad \forall \ell, \tag{57}$$

by construction. Note that G_1 is always a DAG, but possibly $\mathcal{M}(G_1) \neq \mathcal{M}(G_0)$. However, since $G_0 \subseteq G_1$, we have

$$\text{score}_\Omega(G_0) = \text{score}_\Omega(G_1). \tag{58}$$

We also have

$$\text{score}_\Omega(G) = \sum_{\ell=1}^{k} \text{score}_\Omega(G^\ell) - \sum_{r=1}^{k'} (m_r - 1) f_{\sigma_{s_r}}(\emptyset), \tag{59}$$

$$\text{score}_\Omega(G_0) = \sum_{\ell=1}^{k} \text{score}_\Omega(G_0^\ell) - \sum_{r=1}^{k'} (m_r - 1) f_{\sigma_{s_r}}(\emptyset), \tag{60}$$

where $\{s_r\}_{r=1}^{k'}$ denote the indices of the $k' < k$ separator nodes, and $m_r := |\{\ell : s_r \in C_\ell\}|$. This is because both G and G_0 have the property that separator nodes only have parents contained in a single clique, so we include an extra term $f_{\sigma_{s_r}}(\emptyset)$ from each adjacent clique not containing $\text{Pa}(s_r)$ in computing the sum. Combining Equation (60) with Equations (57) and (58), we must also have

$$\text{score}_\Omega(G_1) = \sum_{\ell=1}^{k} \text{score}_\Omega(G_1^\ell) - \sum_{r=1}^{k'} (m_r - 1) f_{\sigma_{s_r}}(\emptyset). \tag{61}$$

Together with Equation (59), this implies

$$\max_{G_1 \supseteq G_0} \{\gamma_\Omega(G, G_1)\} = \frac{\sum_{\ell=1}^{k} \left(\text{score}_\Omega(G^\ell) - \text{score}_\Omega(G_1^\ell)\right)}{|H(G, G_1)|}. \tag{62}$$

Also note that by Lemma 28 and Theorem 7, we have

$$\text{score}_\Omega(G_1^\ell) \leq \text{score}_\Omega(G^\ell), \qquad \forall \ell,$$

and by assumption,

$$\frac{\text{score}_\Omega(G^\ell) - \text{score}_\Omega(G_1^\ell)}{|H(G^\ell, G_1^\ell)|} \geq \gamma_\Omega(G_0^\ell), \qquad \forall \ell. \tag{63}$$

Finally, reindexing the cliques so that $\{C_1, \ldots, C_{k''}\}$ are the cliques such that $G^\ell \neq G_1^\ell$, we have

$$H(G, G_1) \subseteq \bigcup_{\ell=1}^{k''} H(G^\ell, G_1^\ell),$$

implying that

$$|H(G, G_1)| \leq \sum_{\ell=1}^{k''} |H(G^\ell, G_1^\ell)|. \tag{64}$$

Using the simple fact that $\frac{a_\ell}{b_\ell} \geq \xi$ for all ℓ, with $a_\ell, b_\ell > 0$, implies $\frac{\sum_\ell a_\ell}{\sum_\ell b_\ell} > \xi$, we conclude from Equation (62) and Inequalities (63) and (64) that

$$\max_{G_1 \supseteq G_0} \{\gamma_\Omega(G, G_1)\} \geq \frac{\sum_{\ell=1}^{k''} \left(\text{score}_\Omega(G^\ell) - \text{score}_\Omega(G_1^\ell)\right)}{\sum_{\ell=1}^{k''} |H(G^\ell, G_1^\ell)|} \geq \min_{1 \leq \ell \leq k} \gamma_\Omega(G_0^\ell).$$

Since this result holds uniformly over all G, we have $\gamma_\Omega(G_0) \geq \min_{1 \leq \ell \leq k} \gamma_\Omega(G_0^\ell)$, as well.

D.6 Proof of Lemma 24

This proof is quite similar to the proof for the fully-observed case, so we only mention the high-level details here.

We write

$$\begin{aligned}\sigma_j^2|\widetilde{f}_{\sigma_j}(S) - f_{\sigma_j}(S)| &= \left|\left(\widehat{\Gamma}_{jj} - \widehat{\Gamma}_{j,S}\widehat{\Gamma}_{SS}^{-1}\widehat{\Gamma}_{S,j}^{-1}\right) - \left(\Sigma_{jj} - \Sigma_{j,S}\Sigma_{SS}^{-1}\Sigma_{S,j}\right)\right| \\ &\leq \left|\widehat{\Gamma}_{jj} - \Sigma_{jj}\right| + \underbrace{\left|\widehat{\Gamma}_{j,S}\widehat{\Gamma}_{SS}^{-1}\widehat{\Gamma}_{S,j} - \Sigma_{j,S}\Sigma_{SS}^{-1}\Sigma_{S,j}\right|}_{A}. \end{aligned} \tag{65}$$

The first term may be bounded directly using Inequality (33) and a union bound over j:

$$\mathbb{P}\left(\max_j \left|\widehat{\Gamma}_{jj} - \Sigma_{jj}\right| \geq c\sigma^2\sqrt{\frac{\log p}{n}}\right) \leq c_1' \exp(-c_2' \log p). \tag{66}$$

To bound the second term, we use the following expansion:

$$\begin{aligned} A &\leq \left|\widehat{\Gamma}_{j,S}\left(\widehat{\Gamma}_{SS}^{-1} - \Sigma_{SS}^{-1}\right)\widehat{\Gamma}_{S,j}\right| + \left|\widehat{\Gamma}_{j,S}\Sigma_{SS}^{-1}\left(\widehat{\Gamma}_{S,j} - \Sigma_{S,j}\right)\right| + \left|\left(\widehat{\Gamma}_{j,S} - \Sigma_{j,S}\right)\Sigma_{SS}^{-1}\Sigma_{S,j}\right| \\ &\leq \left|\left|\left|\widehat{\Gamma}_{SS}^{-1} - \Sigma_{SS}^{-1}\right|\right|\right|_2 \|\widehat{\Gamma}_{S,j}\|_2^2 + \left|\left|\left|\Sigma_{SS}^{-1}\right|\right|\right|_2 \left(\|\widehat{\Gamma}_{S,j}\|_2\|\widehat{\Gamma}_{S,j} - \Sigma_{S,j}\|_2 + \|\widehat{\Gamma}_{S,j} - \Sigma_{S,j}\|_2\|\Sigma_{S,j}\|_2\right). \end{aligned}$$

As in the proof of Lemma 29 in Appendix E, we may obtain a bound of the form

$$\mathbb{P}\left(\left|\left|\left|\widehat{\Gamma}_{SS}^{-1} - \Sigma_{SS}^{-1}\right|\right|\right|_2 \leq c\sigma^2\left(\sqrt{\frac{d}{n}} + t\right)\right) \leq c_1 \exp(-c_2 n t^2),$$

by inverting the deviation condition (33). Furthermore,

$$\|\widehat{\Gamma}_{S,j} - \Sigma_{S,j}\|_2 \leq \left|\left|\left|\widehat{\Gamma}_{S'S'} - \Sigma_{S'S'}\right|\right|\right|_2,$$

where $S' := S \cup \{j\}$, which may in turn be bounded using the deviation condition (33). We also have

$$\|\widehat{\Gamma}_{S,j}\|_2 \leq \|\Sigma_{S,j}\|_2 + \left|\left|\left|\widehat{\Gamma}_{S'S'} - \Sigma_{S'S'}\right|\right|\right|_2.$$

Combining these results and taking a union bound over the 2^d choices for S and p choices for j, we arrive at a uniform bound of the form

$$\mathbb{P}\left(A \leq c'\sigma^2\sqrt{\frac{\log p}{n}}\right) \geq 1 - c_1' \exp(-c_2' \log p).$$

Together with Inequality (66) and the expansion (65), we then obtain the desired result.

Appendix E. Matrix Concentration Results

This Appendix contains matrix concentration results that are used to prove our technical lemmas. We use $|||\cdot|||_2$ to denote the spectral norm of a matrix.

Lemma 29 *Suppose $\{x_i\}_{i=1}^n \subseteq \mathbb{R}^p$ are i.i.d. sub-Gaussian vectors with parameter σ^2 and covariance Σ. Then for all $t \geq 0$, we have*

$$\mathbb{P}\left(\left|\left|\left|\frac{X^TX}{n} - \Sigma\right|\right|\right|_2 \leq \sigma^2 \cdot \max\{\delta, \delta^2\}\right) \geq 1 - 2\exp(-cnt^2), \tag{67}$$

where $\delta = c'\sqrt{\frac{p}{n}} + c''t$. Furthermore, if $\frac{X^TX}{n}$ is invertible and

$$\sigma^2 \left|\left|\left|\Sigma^{-1}\right|\right|\right|_2 \cdot \max\{\delta, \delta^2\} \leq \frac{1}{2},$$

we have

$$\mathbb{P}\left(\left|\left|\left|\left(\frac{X^TX}{n}\right)^{-1} - \Sigma^{-1}\right|\right|\right|_2 \leq 2\sigma^2 \left|\left|\left|\Sigma^{-1}\right|\right|\right|_2^2 \cdot \max\{\delta, \delta^2\}\right) \geq 1 - 2\exp(-cnt^2). \tag{68}$$

Proof For Inequality (67), see Remark 5.40 of Vershynin (2012). For Inequality (68), we use the matrix expansion

$$(A+\Delta)^{-1} = (A(I + A^{-1}\Delta))^{-1} = (I + A^{-1}\Delta)^{-1}A^{-1} = A^{-1} + \sum_{k=1}^{\infty}(-1)^k(A^{-1}\Delta)^kA^{-1},$$

valid for any matrices A and Δ such that A and $A+\Delta$ are both invertible and the series converges. By the triangle inequality and multiplicativity of the spectral norm, we then have

$$\begin{aligned}
\left|\left|\left|(A+\Delta)^{-1} - A^{-1}\right|\right|\right|_2 &\leq \sum_{k=1}^{\infty}\left|\left|\left|(A^{-1}\Delta)^kA^{-1}\right|\right|\right|_2 \\
&\leq \left|\left|\left|A^{-1}\right|\right|\right|_2 \cdot \sum_{k=1}^{\infty}\left|\left|\left|A^{-1}\Delta\right|\right|\right|_2^k \\
&= \frac{\left|\left|\left|A^{-1}\right|\right|\right|_2 \cdot \left|\left|\left|A^{-1}\Delta\right|\right|\right|_2}{1 - \left|\left|\left|A^{-1}\Delta\right|\right|\right|_2} \\
&\leq \frac{\left|\left|\left|A^{-1}\right|\right|\right|_2^2 \cdot \left|\left|\left|\Delta\right|\right|\right|_2}{1 - \left|\left|\left|A^{-1}\Delta\right|\right|\right|_2}.
\end{aligned}$$

We now take $A = \Sigma$ and $\Delta = \frac{X^TX}{n} - \Sigma$. By the assumption and Inequality (67), we have

$$\left|\left|\left|A^{-1}\Delta\right|\right|\right|_2 \leq \left|\left|\left|A^{-1}\right|\right|\right|_2 \cdot \left|\left|\left|\Delta\right|\right|\right|_2 \leq \frac{1}{2},$$

implying that

$$\left|\left|\left|(A+\Delta)^{-1} - A^{-1}\right|\right|\right|_2 \leq 2\left|\left|\left|A^{-1}\right|\right|\right|_2^2 \cdot \left|\left|\left|\Delta\right|\right|\right|_2.$$

This gives the result. ■

References

O. O. Aalen, K. Røysland, J. M. Gran, and B. Ledergerber. Causality, mediation and time: A dynamic viewpoint. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(4):831–861, 2012.

H. L. Bodlaender. A partial k-arboretum of graphs with bounded treewidth. *Theoretical Computer Science*, 209(12):1 – 45, 1998.

P. Bühlmann, J. Peters, and J. Ernest. CAM: Causal additive models, high-dimensional order search, and penalized regression. *Annals of Statistics*, 2014. To appear.

D. M. Chickering. A transformational characterization of equivalent Bayesian network structures. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 87–98, 1995.

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

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, et al. Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126, 2000.

T. Kloks. *Treewidth: Computations and Approximations*, volume 842. Springer, 1994.

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

J. H. Korhonen and P. Parviainen. Exact learning of bounded tree-width Bayesian networks. In *Artificial Intelligence and Statistics (AISTATS 2013)*, pages 370–378. JMLR, 2013.

S. L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

P. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Annals of Statistics*, 40(3):1637–1664, 2012.

P. Loh and M. J. Wainwright. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *Annals of Statistics*, 41(6):3022–3049, 12 2013.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462, 2006.

C. Nowzohour and P. Bühlmann. Score-based causal learning in additive noise models. *arXiv e-prints*, April 2014. Available at `http://arxiv.org/abs/1311.6359v2`.

S. Ordyniak and S. Szeider. Algorithms and complexity results for exact Bayesian structure learning. *CoRR*, abs/1203.3501, 2012.

E. Perrier, S. Imoto, and S. Miyano. Finding optimal Bayesian network given a super-structure. *Journal of Machine Learning Research*, 9(2):2251–2286, 2008.

J. Peters and P. Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, pages 1–10, 2013.

P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 4:935–980, 2011.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer, and K. Bollen. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12:1225–1248, 2011.

A. Shojaie and G. Michailidis. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538, 2010.

T. Silander and P. Myllymaki. A simple approach for finding the globally optimal Bayesian network structure. In *Proceedings of the Twenty-Second Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 445–452, Arlington, VA, 2006. AUAI Press.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*, volume 81. The MIT Press, 2000.

D. J. Stekhoven, I. Moraes, G. Sveinbjörnsson, L. Hennig, M. H. Maathuis, and P. Bühlmann. Causal stability ranking. *Bioinformatics*, 28(21):2819–2823, 2012.

S. van de Geer and P. Bühlmann. ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *Annals of Statistics*, 41(2):536–567, 2013.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2012.

M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.