High-dimensional learning of linear causal networks via inverse covariance estimation

Po-Ling Loh PLOH@BERKELEY.EDU
Department of Statistics, University of California, Berkeley, CA 94720, USA

Peter Bühlmann BUHLMANN@STAT.MATH.ETHZ.CH
Seminar für Statistik, ETH Zürich, Switzerland

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 precision, the true DAG is the unique minimizer of 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. We demonstrate that dynamic programming may be used to select the optimal DAG in linear time when the moralized graph has bounded treewidth.

1. Introduction

Causal networks arise in a wide variety of applications, including genetics, epidemiology, and time series analysis (Hughes et al., 2000; Stekhoven et al., 2012; Aalen et al., 2012). The task of inferring the graph structure of a causal network from joint observations is a relevant but 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 are only tractable for small graphs (Silander & 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, one may regress each node upon its predecessors and select 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 & Michailidis, 2010). However, knowing the node order is unrealistic in many situations. If instead a conditional independence graph or superset of the skeleton is specified a priori, the number of required conditional independence tests may be reduced dramatically; various authors have devised algorithms to compute the optimal DAG efficiently in when the input graph has bounded degree and/or bounded treewidth (Perrier et al., 2008; Ordyniak & Szeider, 2012; Korhonen & Parviainen, 2013).

Unfortunately, tools for inferring such superstructures are 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 & Lin, 2007). Recent results have established that a version of the graphical Lasso may also be used to learn a conditional independence graph for discrete-valued variables (Loh &

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

Wainwright, 2013), but results for more general distributions are absent from the literature. Bühlmann et al. (2013) isolate sufficient conditions under which Lasso-based linear regression could be used to recover a conditional independence graph for general distributions, and use it as a prescreening step for nonparametric causal inference in additive noise models; however, it is unclear which non-Gaussian distributions satisfy such 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, *non-Gaussian* refers 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 programming in time linear in the number of nodes (Ordyniak & 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
3: Construct moralized graph $\widehat{\mathcal{M}}$ with edge set defined by
   $\operatorname{supp}(\widehat{\Theta})$
4: Compute scores for DAGs that are consistent with $\widehat{\mathcal{M}}$
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 the conditional independence graph is known for multivariate Gaussians (Lauritzen, 1996), our result does not exploit Gaussianity.

Since we do not impose constraints on the error distribution of our SEM, standard parametric maximum likelihood methods are *not* applicable to compare candidate DAGs. Consequently, we use the squared ℓ_2-error to score DAGs, and prove that for homoscedastic errors, the true DAG uniquely minimizes this score function. As a corollary, the DAG structure of a linear SEM is identifiable whenever the additive errors are homoscedastic, generalizing a recent result derived only for Gaussian variables (Peters & Bühlmann, 2013). Our result covers cases with Gaussian and non-Gaussian errors, whereas Shimizu et al. (2011) require all errors to be non-Gaussian. 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. Our theoretical results only require a condition on the gap between squared ℓ_2-scores for DAGs in the restricted search space and eigenvalue conditions on the true covariance matrix, which is weaker than the beta-min condition appearing in previous work (van de Geer & Bühlmann, 2013). Furthermore, the size of the gap is *not* required to scale linearly with the number of nodes, unlike similar conditions in van de Geer & Bühlmann (2013) and Peters & Bühlmann (2013). Since inverse covariance matrix estimation and computing scores based on linear regression are both easily modified to deal with systematically corrupted data (Loh & Wainwright, 2012), our methods are 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: Section 2 reviews background of graphical models and linear SEMs. Section 3 describes the relationship between the inverse covariance matrix and conditional independence graph. Section 4 discusses the use of the squared ℓ_2-loss for scoring DAGs. Section 5 presents results for statistical consistency and describes the gap condition. Finally, Section 6 explains how dynamic programming may be used to identify the optimal DAG in linear time, when the moralized graph has bounded treewidth. Proofs may be found in the arXiv version of the paper (Loh & Bühlmann, 2013).

2. Background

We begin with brief background on graphical models. For a more in-depth exposition, see Koller & Friedman (2009).

2.1. Graphical models

Consider a probability distribution $q(x_1, \dots, x_p)$ and an undirected graph $G = (V, E)$, where $V = \{1, \dots, p\}$ 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 Hammersley-Clifford theorem, if q is a strictly positive distribution, then G is a CIG for q if and only if

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

for potential functions $\{\psi_C : C \in \mathcal{C}\}$ defined over the set of cliques $\mathcal{C}$ of G.

Now consider a *directed* graph $G = (V, E)$, where we 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 $q(x_1, \dots, x_p)$ if q factorizes as

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

A permutation π of the vertex set V is a *topological order* for G if $\pi(j) < \pi(k)$ whenever $(j, k) \in E$. The factorization (2) implies $X_j \perp\!\!\!\perp X_{\nu(j)} \mid X_{\mathrm{Pa}(j)}$ for all j, where $\nu(j)$ is the set of 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 if the DAG G represents a distribution q, then $\mathcal{M}(G)$ is also a CIG for q. Finally, we define the *skeleton* of a DAG G to be the undirected graph formed by dropping orientations of edges in G. 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 say that a random vector $X = (X_1, \dots, X_p) \in \mathbb{R}^p$ follows a *linear structural equation model* (SEM) if

$$X = B^T X + \epsilon, \qquad (3)$$

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

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

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

so we may factorize

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

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

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

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

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

Here, $\Sigma := \mathrm{cov}[X]$. We call $b_j^T X$ the *best linear predictor* for X_j among linear combinations of $\{X_1, \dots, X_{j-1}\}$. Let $\Omega := \mathrm{cov}[\epsilon]$ and $\Theta := \Sigma^{-1}$. We have the following lemma:

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

$$\Theta_{jk} = -\sigma_k^{-2} B_{jk} + \sum_{\ell > k} \sigma_\ell^{-2} B_{j\ell} B_{k\ell}, \qquad \forall j < k, \qquad (4)$$

$$\Theta_{jj} = \sigma_j^{-2} + \sum_{\ell > j} \sigma_\ell^{-2} B_{j\ell}^2, \qquad \forall j. \qquad (5)$$

Equation (4) has an important implication for causal inference, which we state in the following theorem:

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* $\mathcal{M}(G)$.

We will assume that the converse of Theorem 2 also holds:

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 $\mathcal{M}(G)$. 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 *Theorem 2 may be viewed as an extension of the canonical result for Gaussian graphical models. Indeed, 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. Theorem 2 states that* $\text{supp}(\Theta)$ *is always a subset of the edge set of $\mathcal{M}(G^\pi)$. Assumption 1 is a type of* faithfulness *assumption (Koller & Friedman, 2009; Spirtes et al., 2000).*

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 it enjoys the nice property that the population-level version is maximized only under a correct parametrization of the model class. However, such maximum likelihood methods presuppose a fixed parametrization. In the case of linear SEMs, this translates into an appropriate parametrization 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 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 when 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 also show that the true linear SEM is identifiable.

4.1. 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 $\Omega = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$ and a candidate matrix B with columns $\{b_j\}_{j=1}^p$, define

$$\text{score}_\Omega(B) := \mathbb{E}_X \left[\|\Omega^{-1/2} (I - B)^T X\|_2^2 \right]. \qquad (6)$$

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, and expectations are taken with respect to X.

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

$$\text{score}_\Omega(G) := \min_{B \in \mathcal{U}_G} \{\text{score}_\Omega(B)\}, \qquad (7)$$

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 consistent with the structure of G.

Remark 4 *Examining the form of the score function* (6)*, we see that if $\{\text{Pa}_G(j)\}_{j=1}^p$ denotes the parent sets of nodes in G, then the matrix*

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

is unique, and the columns of B_G are equal to the coefficients of the best linear predictor of X_j regressed upon $X_{\text{Pa}_G(j)}$. The value of B_G does not depend on Ω.

The following lemma relates the score of the underlying DAG G_0 to the score of the true autoregression matrix B_0:

Lemma 5 *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*

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

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

We now turn to the main theorem of this section, in which we consider the problem of minimizing $\text{score}_\Omega(B)$ 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 $\mathcal{U} := \bigcup_{G \in \mathcal{D}} \mathcal{U}_G$.

Theorem 6 *Given a linear SEM* (3) *with error covariance matrix $\alpha\Omega_0$ and autoregression matrix B_0, where $\alpha > 0$, we have*

$$\text{score}_{\alpha\Omega_0}(B) \geq \text{score}_{\alpha\Omega_0}(B_0) = p, \qquad \forall B \in \mathcal{U}, \quad (8)$$

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

Theorem 6 implies that the squared ℓ_2-loss function (6) is an appropriate measure of model fit when the components are correctly weighted by the diagonal entries of Ω_0.

4.2. Identifiability of linear SEMs

Theorem 6 also has a useful consequence in terms of identifiability of a linear SEM:

Corollary 7 *Consider a fixed diagonal covariance Ω_0, and consider the class of linear SEMs parametrized 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.*

Corollary 7 may be compared to previous results in the literature regarding identifiability of linear SEMs. Theorem 1 of Peters & Bühlmann (2013) states that when X is Gaussian and ϵ is an i.i.d. Gaussian vector with $\text{cov}[\epsilon] = \alpha\Omega_0$, the model is identifiable. Our Corollary 7 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 Gaussian errors.

4.3. Misspecification of variances

Theorem 6 implies that when the diagonal variances of Ω_0 are known up to a scalar factor, the weighted ℓ_2-loss (6) may be used as a score function for linear SEMs. In this section, we study the effect when Ω is misspecified.

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. Define the additive gap between the score of G_0 and the next best DAG:

$$\xi :== \min_{G \in \mathcal{D}, G \not\supseteq G_0} \{\text{score}_{\Omega_0}(G)\} - p. \tag{9}$$

By Theorem 6, 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.

Theorem 8 *Suppose*

$$\frac{a_{\max}}{a_{\min}} \leq 1 + \frac{\xi}{p}. \tag{10}$$

Then $B_0 \in \arg\min_{B \in \mathcal{U}} \{\text{score}_{\Omega_1}(B)\}$. If inequality (10) is strict, then B_0 is the unique minimizer of $\text{score}_{\Omega_1}(B)$.

Specializing to the case when $\Omega_1 = I$, we may interpret Theorem 8 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.

5. Consequences for statistical estimation

The population-level results in Theorems 2 and 6 provide a natural avenue for estimating the DAG of a linear SEM from data. In this section, we provide statistical guarantees for the success of our inference algorithm.

Our 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 6 ensures that when the weight matrix Ω is chosen appropriately, B_0 will be the unique minimizer of $\text{score}_\Omega(B)$.

5.1. 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, and we use the estimator $\widehat{\Theta} = (\widehat{\Sigma})^{-1}$. We have the following lemma, which follows from standard random matrix theory:

Lemma 9 *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 $\text{supp}(\Theta_0)$, if $\Theta_0^{\min} > 2\tau$.

Here, we use $\|\cdot\|_{\max}$ to denote the elementwise ℓ_∞-norm.

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:

$$\widehat{\Theta} \in \arg\min_{\Theta \succeq 0} \left\{ \operatorname{tr}(\Theta\widehat{\Sigma}) - \log\det(\Theta) + \lambda \sum_{j \neq k} |\Theta_{jk}| \right\}. \tag{11}$$

Standard results (Ravikumar et al., 2011) establish the statistical consistency of the graphical Lasso (11) as an estimator for the inverse covariance matrix for sub-Gaussian observations, resulting in the following lemma:

Lemma 10 *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 & Bühlmann, 2006) to recover $\operatorname{supp}(\Theta_0)$, with similar rates for consistency.

5.2. Scoring candidate DAGs

Moving on to the second step of the algorithm, we need to estimate the score functions $\operatorname{score}_\Omega(B)$ of candidate DAGs and choose the minimally scoring candidate. 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 dynamic programming procedures 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. 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

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

denote the set of DAGs with skeleton contained in the graph defined by $\operatorname{supp}(\Theta)$. By Theorem 2 and Assumption 1, we have $G_0 \in \mathcal{D}_{\Theta_0}$, so if $\operatorname{supp}(\widehat{\Theta}) \supseteq \operatorname{supp}(\Theta_0)$, which occurs with high probability under the conditions of Section 5.1, it suffices to search over the reduced DAG space $\mathcal{D}_{\widehat{\Theta}}$.

Consider an arbitrary d-sparse matrix Θ, with $d \leq n$, and take $G \in \mathcal{D}_\Theta$. By Remark 4, we have

$$\operatorname{score}_\Omega(G) = \sum_{j=1}^{p} f_{\sigma_j}(\operatorname{Pa}_G(j)), \tag{12}$$

where

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

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

$$\widehat{f}_{\sigma_j}(S) := \frac{1}{\sigma_j^2} \cdot \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - x_{i,S}^T \widehat{b}_j)^2 = \frac{1}{\sigma_j^2} \cdot \frac{1}{n} \|X_j - X_S \widehat{b}_j\|_2^2, \tag{13}$$

where

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

is the 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 w.h.p. The following lemma provides rates of convergence for the empirical score function:

Lemma 11 *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*

$$|\widehat{f}_{\sigma_j}(S) - f_{\sigma_j}(S)| \leq \frac{c_0\sigma^2}{\sigma_j^2} \sqrt{\frac{\log p}{n}}, \qquad \forall j \text{ and } S \subseteq N_\Theta(j), \tag{14}$$

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

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

$$\xi_\Omega(\mathcal{D}_\Theta) := \min_{G \in \mathcal{D}_\Theta, G \not\supseteq G_0} \{\operatorname{score}_\Omega(G) - \operatorname{score}_\Omega(G_0)\} \tag{15}$$

is the gap between G_0 and the next best DAG in $\mathcal{D}_\Theta$.

Lemma 12 *Suppose inequality* (14) *holds, and suppose*

$$c_0\sigma^2 \sqrt{\frac{\log p}{n}} \cdot \sum_{j=1}^{p} \frac{1}{\sigma_j^2} < \frac{\xi_\Omega(\mathcal{D}_\Theta)}{2}. \tag{16}$$

Then

$$\widehat{\operatorname{score}}_\Omega(G_0) < \widehat{\operatorname{score}}_\Omega(G), \qquad \forall G \in \mathcal{D}_\Theta : G \not\supseteq G_0. \tag{17}$$

Remark 13 *Lemma 12 does not explicitly assume $\Omega = \Omega_0$. However, inequality* (16) *is only satisfiable when $\xi_\Omega(\mathcal{D}_G) > 0$; hence, Ω should be chosen such that $G_0 = \arg\min_{G \in \mathcal{D}_\Theta, G \not\supseteq G_0} \{\operatorname{score}_\Omega(G)\}$. As discussed in Section 4.3, this condition holds for a wider range of Ω.*

Note that the conclusion (17) in Lemma 12 is not quite the same as the condition

$$G_0 = \arg \min_{G \in \mathcal{D}_\Theta, G \not\supseteq G_0} \left\{ \widehat{\text{score}}_\Omega(G) \right\}, \qquad (18)$$

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 (18) may not be satisfied. However, it is easily seen from the proof of Lemma 12 that in fact,

$$\arg \min_{G \in \mathcal{D}_\Theta} \left\{ \widehat{\text{score}}_\Omega(G) \right\} \subseteq \{G \in \mathcal{D}_\Theta : G \supseteq G_0\}. \quad (19)$$

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.

To gain some intuition for the condition (16), consider the case when $\sigma_j^2 = 1$ for all j. Then the condition becomes

$$c_0 \sigma^2 \sqrt{\frac{\log p}{n}} < \frac{\xi(\mathcal{D}_\Theta)}{2p}. \qquad (20)$$

If $\xi(\mathcal{D}_\Theta) = \Omega(1)$, 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 & Bühlmann, 2013; Bühlmann et al., 2013), our method is consistent provided $\frac{\log p}{n} \to 0$. In Section 5.3, we relax the condition (20) to a slightly weaker condition that is more likely to hold in settings of interest.

5.3. Weakening the gap condition

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

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

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

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

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

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

$$\gamma_\Omega(G_0) := \min_{G \in \mathcal{D}_\Theta, G \not\supseteq G_0} \left\{ \max_{G_1 \supseteq G_0} \{\gamma_\Omega(G, G_1)\} \right\} \geq \xi'. \qquad (21)$$

Note that in addition to minimizing over DAGs in the class $\mathcal{D}_\Theta$, the expression (21) defined in Assumption 2 takes an inner maximization over DAGs containing G_0. As established in Lemma 5, $\text{score}_\Omega(G_1) = \text{score}_\Omega(G_0)$ 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 have the following result:

Lemma 14 *Under Assumption 2, suppose*

$$|\widehat{f}_{\sigma_j}(S) - f_{\sigma_j}(S)| \leq \frac{\xi'}{2}, \qquad \forall j \text{ and } S \subseteq N_\Theta(j). \quad (22)$$

Then the containment (19) *holds.*

Combining with Lemma 11, we have the following:

Corollary 15 *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 $1 - c_1 \exp(-c_2 \log p)$, condition* (19) *is satisfied.*

We now turn to the question of what values of ξ' give condition (21) for various DAGs. As motivated by preliminary computations, 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 (cf. Section 4.4 of the technical report (Loh & Bühlmann, 2013)). 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:

Lemma 16 *Suppose the moralized graph $\mathcal{M}(G_0)$ admits a junction tree representation with only singleton separator sets. Let $C_1, \dots, C_k$ 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 $\gamma_\Omega(G_0^\ell)$ is defined as

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

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

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 16 implies $\gamma_\Omega(G_0) \geq \xi_w$, and we only need the parameter ξ' appearing in Assumption 2 to be larger than ξ_w, rather than scaling as $\frac{1}{p}$.

6. Computational considerations

In practice, the main computational bottleneck 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 Theorem 2 provides a general framework for achieving vast computational savings for finding the best-scoring DAG when data come from a linear SEM, and discuss the runtime for bounded-treewidth graphs.

6.1. Decomposable score functions

Recall that a score function over DAGs is *decomposable* if it may be written in the following manner:

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

Some common examples of decomposable scores that are used for DAG inference include maximum likelihood, BDe, BIC, and AIC (Chickering, 1995). By equation (12), the squared ℓ_2-score is clearly decomposable, and it gives an example where score_j differs over nodes. 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 & Myllymaki, 2006), greedy methods (Chickering, 2002), and dynamic programming (Ordyniak & Szeider, 2012; Korhonen & Parviainen, 2013).

6.2. Dynamic programming

We now review a method due to Ordyniak & Szeider (2012) that is 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.

The main steps of the dynamic programming algorithm are as follows. For further details and a proof of correctness, see Ordyniak & 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 $\{1, \dots, p\}$ 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 $\chi(t)$.

6.3. Runtime

The runtime of the dynamic programming algorithm is discussed in Korhonen & Parviainen (2013). Assuming the treewidth w of G is bounded, the overall runtime is $\mathcal{O}(p \cdot 2^{2(w+1)(w+d)})$. 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.

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, and are the first to provide rigorous guarantees for when the graphical Lasso may be used for preprocessing in non-Gaussian settings.

Acknowledgments

We acknowledge the members of the SfS for providing a tremendously hospitable and fruitful environment when PL was visiting ETH, and the FIM 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.

References

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

Bühlmann, P., Peters, J., and Ernest, J. CAM: Causal additive models, high-dimensional order search, and penalized regression. *arXiv e-prints*, October 2013. Available at `http://arxiv.org/abs/1310.1533`.

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

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

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

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

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

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

Loh, P. and Bühlmann, P. High-dimensional learning of linear causal networks via inverse covariance estimation. *arXiv e-prints*, November 2013. Available at `http://arxiv.org/abs/1311.3492`.

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

Loh, P. and Wainwright, M.J. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *Annals of Statistics*, 2013. To appear.

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

Nowzohour, C. and Bühlmann, P. Score-based causal learning in additive noise models. 2013. In preparation.

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

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

Peters, J. and Bühlmann, P. Identifiability of gaussian structural equation models with equal error variances. *arXiv e-prints*, August 2013. Available at `http://arxiv.org/abs/1205.2536`. To appear in *Biometrika*.

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

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

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

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

Silander, T. and Myllymaki, P. 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)*, pp. 445–452, Arlington, VA, 2006. AUAI Press.

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

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

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

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