Learning Large Causal Structures from Inverse Covariance Matrix via Matrix Decomposition

Shuyu Dong, Kento Uemura, Akito Fujii, Shuang Chang, Yusuke Koyanagi, Koji Maruhashi, Michèle Sebag

Learning Large Causal Structures from Inverse Covariance Matrix via Matrix Decomposition

Shuyu Dong*, Kento Uemura†, Akito Fujii†, Shuang Chang†, Yusuke Koyanagi†, Koji Maruhashi†, and Michèle Sebag*

*LISN, INRIA, Université Paris-Saclay
first.last@inria.fr
†Fujitsu Laboratories Ltd.
first.last@fujitsu.com

Abstract

Learning causal structures from observational data is a fundamental yet highly complex problem when the number of variables is large. In this paper, we start from linear structural equation models (SEMs) and investigate ways of learning causal structures from the inverse covariance matrix. The proposed method, called $\mathcal{O}$-ICID (for *Independence-preserving* Decomposition from Oracle Inverse Covariance matrix), is based on continuous optimization of a type of matrix decomposition that preserves the nonzero patterns of the inverse covariance matrix. We show that $\mathcal{O}$-ICID provides an efficient way for identifying the true directed acyclic graph (DAG) under the knowledge of noise variances. With weaker prior information, the proposed method gives directed graph solutions that are useful for making more refined causal discovery. The proposed method enjoys a low complexity when the true DAG has bounded node degrees, as reflected by its time efficiency in experiments in comparison with state-of-the-art algorithms.

1 Introduction

Discovering causal relations from observational data emerges as an important problem for artificial intelligence [Pea00, PJS17] with fundamental and practical motivations. One notable reason is that causal models support modes of reasoning, e.g., counterfactual reasoning and algorithmic recourse [TKL+21], that are otherwise out of reach by correlation-based machine learning, as recent developments [PBM16, ABGLP19, SG21] show how causal structure learning can contribute to machine learning.

The learning of causal structures from data, also referred to as causal discovery, faces challenges in both statistical and algebraic aspects, since one needs to uncover not only correlations from data but also the underlying causal directions. In addition to difficulties related to a restricted number n of observational samples hindering the estimation process, learning a directed acyclic graph (DAG) is NP-hard [Chi96] even in the large sample limit, as the search space of DAGs increases super-exponentially with respect to the number d of variables.

To overcome the difficulties especially the fast growing space of DAGs, one common strategy in different causal discovery approaches such as [SGSH00, Mee95, Chi02, LB14, SWU21] is to conduct a search of DAGs in a restrictive way. For learning the DAG of a linear structure equation model (SEM), [LB14] show that the moral graph of the DAG coincides with the support graph of the inverse covariance matrix under a mild faithfulness assumption. This motivates a search of DAGs through dynamic programming within the support graph of the found inverse covariance matrix; the selection criterion during the search is a score function based on the log-likelihood of the causal structure. When paired with the true noise variances (up to a multiplicative scalar), the score function admits the true DAG as the unique minimizer, as demonstrated by [LB14]. If the true noise variances are assumed known or can be estimated, the restricted search of DAGs always need to be thorough since the unique minimal score value is unknown.

Another strategy for causal discovery is to formulate a continuous optimization problem based on a differentiable or subdifferentiable (e.g., ℓ_1-penalty) score function [ZARX18, AAZ19, NGZ20, NZZZ21]. In this line of work, the optimization of the causal model, defined on the set of real square matrices, is subject to a differentiable DAG constraint [ZARX18], or alternatively, a sparsity-promoting constraint on the sought matrix [NGZ20]. The computational cost of these optimization approaches, different from combinatorial methods, depends mainly on the gradient computations which scale well enough but the overall complexity may still be high due to the non-convex optimization landscape (e.g.,[ZARX18, AAZ19]) of the underlying problem.

In this paper, we consider a matrix decomposition approach taking into account the strengths of the two different strategies above. It is known that, in the learning of linear SEMs, the inverse covariance matrix Θ is related to the causal structure B (the weighted adjacency matrix of a DAG) via a matrix equation of the form

$$\Theta = (I - B)\Omega^{-1}(I - B)^{\mathrm{T}},$$

where Ω is the diagonal matrix of noise variances, and that the unmixing from Θ to a causal structure B via this matrix equation is not unique. We show that such an unmixing is unique under the knowledge of the true noise variances. This property is similar to the aforementioned uniqueness result by [LB14]. A main difference with this previous work, however, is that we not only use the support graph of Θ but also the exact matrix equation for computing eligible matrix decompositions. This results in a specific type of matrix decomposition that we call $\mathcal{O}$-ICID, where the nonzero pattern of B is restricted within the support graph of Θ.

Solutions of $\mathcal{O}$-ICID, in absence of a DAG constraint, generally contain cycles. We consider an optimization of $\mathcal{O}$-ICID that selects matrix solutions with as few nonzeros as possible, noticing that sparsity is shown to be an effective constraint for learning DAGs under mild, sparsity-related assumptions on the true causal structure [RU18, AAZ19, NGZ20]. For this purpose, the ℓ_1-penalty of B is used as a continuous relaxation of the number of nonzeros for the objective function. The underlying problem is an equality-constrained ℓ_1-minimization, analogous to the Dantzig selector [CT07]. An algorithm based on the augmented Lagrangian method (ALM) is proposed for solving this problem. When Θ is sparse, the gradient computation in the ALM algorithm of $\mathcal{O}$-ICID scales as $O(d^2)$, which is an improvement over NOTEARS [ZARX18] and GOLEM [NGZ20]. We combine $\mathcal{O}$-ICID with an empirical inverse covariance estimator and compare with these algorithms on causal discovery tasks; significant time efficiency gains by the proposed method are observed.

In addition to continuous optimization techniques, $\mathcal{O}$-ICID is related to Cholesky decomposition and causal order-based methods [SIS$^+$11, GH18, SAU20, RSW21]. Under mild assumptions on the noise variance distribution, there are simple rules for estimating a node ordering that is consistent with the true causal structure [GH18, RSW21]. Subsequently, given a consistent node ordering, $\mathcal{O}$-ICID can be reduced to Cholesky decomposition and [GH18, Alg. 1], both of which have even lower complexity. With weaker assumptions on the noise variances, however, there is no longer reason for Cholesky decomposition or [GH18, Alg. 1] to learn the causal structure accurately using the same ordering estimator (e.g., [GH18, Lemma 1]); in such cases, we observe empirically that the solutions of $\mathcal{O}$-ICID are much more relevant than the two former algorithms.

2 Background

2.1 Definitions and Notation

A graph $G := (V, E)$ consists of a set of nodes V and a set of edges $E \subset V \times V$. Unless specified otherwise, all graphs (respectively, edges) are directed. The binary adjacency matrix of a graph G is such that its (i, j)-th entry is 1 if and only if $(i, j) \in E$. Conversely, any matrix $B \in \mathbb{R}^{d\times d}$ determines an unique edge set through the nonzero entries, $\mathrm{E}(B) := \{(i, j) : B_{ij} \neq 0\}$, which we also refer to as the support of B. From the definition of edge set $\mathrm{E}(B)$, we assume by convention that all graphs are ordered (according to the indexing of matrix B). The number of nonzeros of matrix B is denoted as $\|B\|_0$ or $\mathrm{nnz}(B)$ indifferently.

Given a directed acyclic graph (DAG) G, the moralization of G is defined as the undirected graph $\mathcal{M}(G)$, with all directed nodes in G made undirected, and new undirected edges (i, j) created for all pairs (i, j) parents of a same node k.

An undirected graph is a *chordal graph* if every cycle of length greater than three has a chord, i.e., a shortcut that triangulates the cycle. For simplicity, a matrix A is

said to be a DAG (respectively, chordal) if the graph by $\mathrm{E}(A)$ is a DAG (resp. chordal).

The set of $d \times d$ symmetric positive definite matrices is denoted by $\mathcal{S}^d_{++}$. The positive definiteness of a symmetric matrix Θ is denoted as $\Theta \succ 0$.

2.2 Structural Equation Models

Structural equation models (SEMs) are defined on a set of random variables $\mathbf{X} = (X_1, \ldots, X_d)$. A linear SEM (B, Ω) expresses the causal relations among the variables as:

$$\mathbf{X} = B^{\mathrm{T}}\mathbf{X} + \mathbf{E} \tag{1}$$

where matrix $B \in \mathbb{R}^{d\times d}$ is supported on a DAG and $\mathbf{E} = (\epsilon_1, \ldots, \epsilon_d)$ is a vector of d noise variables. Here, the variables $\mathbf{X}$ and $\mathbf{E}$ are assumed to be centered and that $E_j \perp\!\!\!\perp X_i$ for all $i \in \mathrm{PA}^G_j$.

Let $G := (\mathbf{X}, E)$ be the support graph of B, i.e., $B_{ij} \neq 0$ if and only if $(i, j) \in E$. Given the linear SEM (1) and the underlying DAG G, the joint distribution $P_{\mathbf{X}}$ satisfies the Markov property with respect to G, and its density function can be factorized as $p(\mathbf{x}) = \Pi_{i=1}^d p(x_i|x_{\mathrm{PA}^G_i})$, where PA^G_i is the set of parent nodes of X_i in G.

Since X and $\mathbf{E}$ are centered and $E_j \perp\!\!\!\perp X_i$ for all $i \in \mathrm{PA}^G_j$, the covariance of $\mathbf{E}$ is a diagonal matrix $\Omega := \mathrm{diag}(\omega_1^2, \ldots, \omega_d^2)$, where ω_i^2 denotes the variance of the i-th noise variable. The linear SEM (1) relates the variables $\mathbf{X}$ to $\mathbf{E}$ via a linear transformation $\mathbf{X} = (I - B)^{-1}\mathbf{E}$ since $(I - B)$ is invertible given that B is supported on a DAG. Consequently, the covariance matrix of $\mathbf{X}$ is $\mathrm{cov}(\mathbf{X}) = (I - B)^{-\mathrm{T}}\Omega(I - B)^{-1}$, and the inverse covariance matrix (or the precision matrix) of $\mathbf{X}$ reads:

$$\Theta := \varphi(B, \Omega^{-1}) := (I - B)\Omega^{-1}(I - B)^{\mathrm{T}}. \tag{2}$$

The property above can be summarized as the following lemma.

Lemma 1 ([LB14])**.** *Let $\mathbf{X}$ be a random variable following the SEM (1) of (B, Ω). Then the coefficients of the inverse covariance matrix Θ of $\mathbf{X}$ are as follows, for all i and $j \neq i$ in $[\![d]\!]$:*

$$\Theta_{ij} = -\frac{B_{ij}}{\omega_j^2} - \frac{B_{ji}}{\omega_i^2} + \sum_{\ell=1}^{d} \frac{B_{i\ell}B_{j\ell}}{\omega_\ell^2}, \tag{3}$$

$$\Theta_{ii} = \frac{1}{\omega_i^2} + \sum_{\ell=1}^{d} \frac{B_{i\ell}^2}{\omega_\ell^2}. \tag{4}$$

From Lemma 1, the nonzero patterns (or edge set $\mathrm{E}(\Theta)$) of the inverse covariance matrix is in fact a subset of the moralized graph of B, as the next theorem shows.

Theorem 2 ([LB14])**.** *The inverse covariance matrix Θ (2) reflects the graph structure of the moralization $\mathcal{M}(B)$ through inclusion: for $i \neq j$, (i, j) is an edge in $\mathcal{M}(B)$ if $\Theta_{ij} \neq 0$.*

The converse of the theorem above is not always true but is considered as a mild assumption, stating that any edge (i, j) in Θ yields either a directed edge between i and j in B, or the existence of a common child between i and j in B.

Assumption 3 ([LB14])**.** *Let* $\mathbf{X}$ *be a random variable following the SEM of* (B, Ω)*. The inverse covariance matrix* Θ (2) *of* $\mathbf{X}$ *satisfies, for all* $i \neq j$*,* $\Theta_{ij} = 0$ *only if* $B_{ji} = B_{ij} = 0$ *and* $B_{i\ell}B_{j\ell} = 0$ *for all* ℓ*.*

The condition above is a type of faithfulness assumption [KF09, SGSH00].

3 Matrix decomposition for learning linear SEMs

Taking inspiration from [LB14, GH18], we consider the causal discovery problem with the linear SEM in a two-step framework: i) estimating the inverse covariance matrix Θ (2) of $\mathbf{X}$ from observational data (statistical part); ii) recovering a causal structure matrix B from Θ (structural learning part). We study the second, structural learning part assuming Θ is given, which in the case of linear SEMs consists of discovering causal relations from Θ through the matrix equation (2).

3.1 Support-constrained decomposition

Given an inverse covariance matrix Θ, the matrix equation (2) generally admits multiple DAG solutions, and the set of solutions becomes even larger without the DAG constraint on B. From the preliminary results in Section 2, however, it is possible to reduce the vast solution set, in the absence of the DAG constraint, by considering a restriction on the candidate support graph. Assumption 3 ensures that the non-zero pattern (or edge set) of Θ coincides with the moralization of B. This allows us to define a matrix decomposition model more specific than equation (2), which we call a *support-constrained* decomposition.

Definition 4 (Support-constrained decomposition)**.** Let $\Theta \in \mathcal{S}^d_{++}$ be a positive definite matrix. Consider the following set:

$$\mathcal{S}(\Theta) := \big\{(B, D) : \Theta = (I - B)D(I - B)^{\mathrm{T}}, \mathrm{diag}(B) = 0,\ D \succ 0 \text{ is diagonal}, \\ \mathrm{E}(B) \subset \mathrm{E}(\Theta)\big\}. \tag{5}$$

We call a pair of $d \times d$ matrices (B, D) a (solution to the) *support-constrained* decomposition of Θ if it belongs to $\mathcal{S}(\Theta)$.

The support-constrained decomposition is always well-defined in the case with chordal matrices. In fact, when a positive definite matrix Θ is supported on a chordal graph, the set $\mathcal{S}(\Theta)$ is nonempty. More precisely, a support-constrained decomposition (B, D) of Θ is permutation similar to the Cholesky decomposition of Θ under a

perfect elimination order of the chordal graph $\mathrm{E}(\Theta)$ [Ros70, VA15]; see Proposition 8 (Appendix A.1).

In more general cases where Θ is not necessarily a chordal matrix, we show that Assumption 3 is sufficient for the solution set $\mathcal{S}(\Theta)$ to be relevant to causal discovery. Proofs of the following results are given in Appendix A.

Proposition 5. *Let Θ be the inverse covariance matrix of $\mathbf{X}$ obeying the linear SEM with (B, Ω), i.e., $\Theta = \varphi(B, \Omega)$ (2). Suppose that this SEM satisfies Assumption 3, then the set $\mathcal{S}(\Theta)$ contains at least (B, Ω^{-1}).*

Other than the DAG G of B, for any DAG G' Markov equivalent to G, there exists B' supported on G' and a diagonal matrix $\Omega' \succ 0$ such that $(B', \Omega'^{-1}) \in \mathcal{S}(\Theta)$. The reason is that there exists a rotation matrix Q that maps $(I - B)$ to $(I - B')$; the existence and characterization of such rotations are studied in [GYKZ20]. Moreover, since B' is Markov equivalent to B, $\mathcal{M}(B) = \mathcal{M}(B')$ which coincide with $\mathrm{E}(\Theta)$. Hence B' also satisfies the support constraint in (5). The existence of $\Omega' \succ 0$ follows from Ω and the rotation Q.

In particular, when paired with the true noise variances Ω, a matrix B such that (B, Ω) belongs to the set (5) recovers the causal structure of the true SEM, as we show in the next theorem.

Theorem 6. *Let $(B^\star, \Omega^\star)$ be a linear SEM satisfying Assumption 3 and let $\Theta^\star = \varphi(B^\star, \Omega^\star)$ be the inverse covariance of $\mathbf{X}$ following this SEM. Suppose that $\Omega^\star$ is known and that $B \in \mathbb{R}^{d \times d}$ is supported on a DAG such that $(B, \Omega^{\star -1}) \in \mathcal{S}(\Theta^\star)$, then $B = B^\star$.*

The result of Theorem 6, similar to [LB14, Theorem 7], requires the knowledge of the noise variances. The difference with [LB14, Theorem 7] lies in its implication for the proposed causal discovery method that we will introduce next. More precisely, [LB14, Theorem 7] shows that the sought solution $B^\star$ is the unique minimizer of a least squares score function when paired with the true noise variances $\Omega^\star$. When a score-based method proceeds by enumerating candidate DAGs and selecting the one with lowest score, the challenge is in the enumeration of DAGs, since the search space of eligible DAGs can be vast (with large number d of variables) and *the* unique minimum of the score function is not known a priori. In contrast to score-based methods, we consider seeking solutions to the decomposition of the inverse covariance matrix with respect to the feasible conditions specified in the solution set (5).

3.2 Algorithms

The results in Section 3.1 suggest the learning of the causal DAG of a linear SEM or its Markov equivalence class can be realized by finding a DAG B such that (B, D) (for some $D \succ 0$) constitute a support-constraint decomposition of the inverse covariance matrix Θ. To simplify the search space of the underlying matrix decomposition problem, we relax the acyclicity of the candidate matrix B. Such a simplification is at the cost

of enlarging the set of eligible solutions to the entire set $\mathcal{S}(\Theta)$ (5). As a remedy to the absence of the DAG constraint, we set the objective of the underlying matrix decomposition problem as a minimization of the number of nonzeros in the matrix B. This leads to the following matrix decomposition program:

$$\underset{B \in E_\Theta,\ D \succ 0}{\text{minimize}}\ \ell_1(B) \quad \text{subject to} \ \ \Theta - \varphi(B, D) = \mathbf{0} \tag{6}$$

where the function $\varphi(B, D) = (I - B)D(I - B)^{\mathrm{T}}$ is defined from (2), the search space of B, defined according to the support constraint of (5), is

$$E_\Theta := \{B \in \mathbb{R}^{d \times d} : \operatorname{diag}(B) = \mathbf{0},\ \ \mathrm{E}(B) \subset \mathrm{E}(\Theta)\}, \tag{7}$$

and the term $\ell_1(B) := \sum_{i,j} |B_{ij}|$ is a continuous relaxation of $\ell_0(B)$ (number of nonzeros of B) for limiting the number of nonzeros in the solution.

The optimization problem (6) is analogous to the Dantzig selector [CT07] for high-dimensional statistical problems. For causal discovery, problem (6) is the second subproblem of the aforementioned two-step framework. We call this framework ICID, for Inverse Covariance estimation and Independence-preserving Decomposition, and refer to problem (6) as $\mathcal{O}$-ICID, for Oracle-ICID, since the input matrix Θ of (6) is considered as an oracle inverse covariance matrix. We use the term "independence-preserving decomposition" instead of support-constrained decomposition given the context of causal discovery, noting that the nonzero/zero pattern $\mathrm{E}(\Theta)$ in (6) come from the independence relations between variables of $\mathbf{X}$ underlying this causal model.

An augmented Lagrangian method. $\mathcal{O}$-ICID (6) is a nonconvex, nonsmooth problem since the feasible set of the matrix equation $\varphi(B, D) = \Theta$ is nonconvex and the ℓ_1-term in the objective is nonsmooth. We consider the augmented Lagrangian method (ALM) [Ber99] to solve (6).

In Algorithm 1, we detail the procedure for optimizing the augmented Lagrangian for a fixed diagonal matrix $D = I$. This algorithm addresses $\mathcal{O}$-ICID (6) partially by assuming that the noise variances of the true SEM are equal or vary mildly around a constant.

The matrix equation in (6) consists of $\frac{d(d+1)}{2}$ equalities, and results in a sum of inner products with Lagrange multipliers as follows: $\langle \Delta, \Theta - \varphi(B, D)\rangle$ for $\Delta \in \mathbb{R}^{d \times d}$. Hence, the augmented Lagrangian of (6) is

$$L^\rho(B; \Delta) = \ell_1(B) + \langle \Delta, \Theta - \varphi(B, D)\rangle + \frac{\rho}{2}\|\Theta - \varphi(B, D)\|_{\mathrm{F}}^2. \tag{8}$$

In line 3, we use the FISTA algorithm [BT09] for solving the primal descent step (9). Details are given in Appendix B.

Remark 7. Algorithm 1 subsumes the process of solving the following ℓ_1-regularized problem

$$\underset{B \in E_\Theta}{\text{minimize}}\ \frac{1}{2}\|\Theta - \varphi(B, D)\|_{\mathrm{F}}^2 + \lambda \ell_1(B) \tag{10}$$

Algorithm 1 ($\mathcal{O}$-ICID) $\mathcal{O}$-ICID using an ALM

Input: Inverse covariance matrix Θ, parameter $\beta \in (0,1)$, tolerance ϵ
Output: $B_t \in E_\Theta$

1: Initialize: $B_0 = \mathbf{0}$, $\Delta_0 = \mathbf{0}$, $\rho = \rho_0 = 1$, $D = I$.
2: **for** $t = 1, \ldots,$ **do**
3: Primal descent: for $L^\rho(B; \Delta)$ defined in (8), compute

$$B_t = \arg\min_{B \in E_\Theta} L^\rho(B; \Delta), \tag{9}$$

with $\rho \geq \rho_0$ such that $\mathcal{I}_t := \|\Theta - \varphi(B, D)\|_{\mathrm{F}} \leq \beta \mathcal{I}_{t-1}$
4: Dual ascent: $\Delta_t = \Delta_{t-1} + \rho(\Theta - \varphi(B_t, D))$.
5: If $\|\Theta - \varphi(B_t, D)\|_{\mathrm{F}} \leq \epsilon$, return B_t as solution
6: **end for**

for a certain $\lambda > 0$. In fact, through the initial states $\Delta = \mathbf{0}$ and $\rho = \rho_0 > 0$ of the ALM procedure (Algorithm 1, line 1), the primal descent (9) at the first iteration corresponds to solving (10). Moreover, the subsequent ALM iterations of Algorithm 1 enforces the feasibility (regarding (5)) of (B, D) in a more systematical way than the optimization and parameter selection of (10).

Special case of $\mathcal{O}$-ICID using an esitmated node ordering. The support-constrained decomposition of an inverse covariance matrix $\Theta^\star$ does not require the knowledge of a causal order underlying the true DAG $B^\star$. However, when such an ordering (denoted $\hat{\sigma}$) is known or can be estimated accurately enough , $\mathcal{O}$-ICID (6) becomes much simplified since the Cholesky decomposition of $\Theta^\star$ under the node ordering $\hat{\sigma}$ recovers the unit lower-triangular matrix that is permutation similar to $(I - B^\star)$. In this special case, the complexity of $\mathcal{O}$-ICID (6) reduces to that of Cholesky decomposition.

We implement this special case algorithm using the CVXOPT toolbox along with Chompack [VA15] for Cholesky decomposition of sparse symmetric positive definite matrices. In Section 4, this special case algorithm is represented as Chol($\hat{\sigma}$) where $\hat{\sigma}$ is an estimated node ordering given by the ordering of the variances of X_i's.

4 Experiments

We test our algorithms for $\mathcal{O}$-ICID and ICID in causal discovery tasks. The experiments are conducted on synthetic data and on a real-world dataset [SPP$^+$05] (Section 4.4).

The synthetic data are generated from linear SEMs on two types of random graphs: (i) Erdős–Rényi (ER) graphs and (ii) Scale-free (SF) [BA99] graphs, following the settings of [ZARX18, NGZ20] for the support DAG of B, the type of noise distributions (Gaussian or exponential), and the distribution of the noise variances Ω (equal and non-

equal noise variances (EV and NV)). The edge weights encoded in B are i.i.d. samples from the uniform distribution $\mathrm{Unif}([-2,-0.5]\cup[0.5,2])$; see details in Appendix C.1.

Two studies are conducted for evaluating $\mathcal{O}$-ICID and ICID respectively. The first study (Section 4.1) is to assess the performance of $\mathcal{O}$-ICID given the inverse covariance matrix Θ of the true linear SEM. The second study (Sections 4.2–4.3) is to assess ICID in terms of its robustness against errors in the statistical estimation of the inverse covariance matrix, and its scalability with respect to the number of variables. The implementation of ICID consists of Algorithm 4 for inverse covariance estimation followed by Algorithm 1 for $\mathcal{O}$-ICID.

In the experiments, $\mathcal{O}$-ICID is tested in comparison with [GH18, Alg. 1] denoted as $\mathcal{O}$-Ghoshal and the Cholesky decomposition-based algorithm $\mathrm{Chol}(\hat{\sigma})$. On the other hand, ICID is tested in comparison with NOTEARS [ZARX18] and GOLEM [NGZ20]. Results of all methods but GOLEM are obtained on one single CPU of Intel(R) Xeon(R) Gold 5120 14 cores @ 2.2GHz. The results of GOLEM are obtained on a GPU of Tesla V100-PCIE-32GB.

All methods are evaluated by the usual metrics (SHD, TPR, FDR and FPR) for (0-1) edge prediction of directed graphs (details in Appendix C.2).

The implementation of the proposed algorithms is made available at `https://github.com/shuyu-d/icid-exp`.

4.1 Learning causal structures from the inverse covariance matrix

In this first study, we examine $\mathcal{O}$-ICID with input matrix defined as the inverse covariance matrix (2) of a linear SEM $(B^\star, \Omega^\star)$. The generation process of the linear SEMs on random DAGs is described in Appendix C.1. We distinguish two scenarios according to whether the true noise variances are all equal (EV) or not (NV). We set the NV case by transforming the data and the underlying inverse covariance matrix of the EV case through standardization [RSW21].

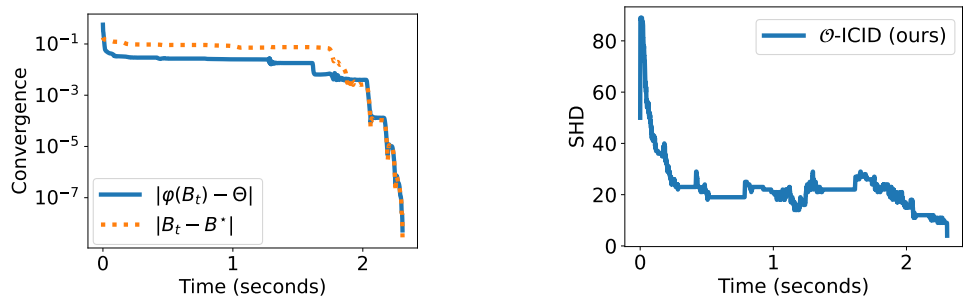
Convergence behavior. In the EV case, we examine the convergence behavior of $\mathcal{O}$-ICID depending on the average node degree, ranging from 0.5 to 2, of the true DAG. Representative learning curves of $\mathcal{O}$-ICID are plotted in Figure 1. In this figure, the following infeasibility measure

$$\mathrm{Infeasi} = \frac{1}{d}\|\varphi(B_t) - \Theta\|_{\mathrm{F}} \tag{11}$$

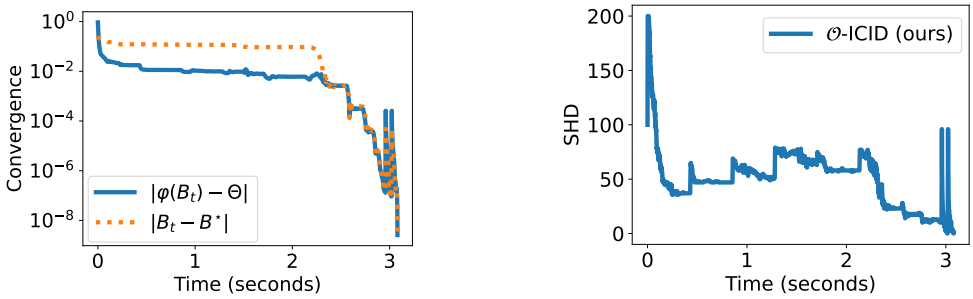
and the distance of B_t to the true causal structure $B^\star$ in SHD are presented along the learning process by Algorithm 1.

These curves illustrate that the decay to zero of the causal discovery error coincides with the decay of the infeasibility measure. This is an instance when solving $\mathcal{O}$-ICID achieves exact recovery of the true causal structure. In repeated random tests on

graphs with low average node degrees, we observe that the convergence indicators of $\mathcal{O}$-ICID, mainly Infeasi (11), are good enough indicators about the accuracy of the found model; see Table 6 (Appendix C.5). For example, on $N = 10$ repeated random tests with ER1 graphs, it is shown that (i) 7 out of 10 instances return a solution with Infeasi (11) below 10^{-5}, and at the same time, an exactly zero causal discovery error in SHD; (ii) 9 out of 10 instances return a solution with Infeasi (11) below 10^{-5} and a relative causal discovery error ($\frac{\text{SHD}(B,B^\star)}{\|B^\star\|_0}$) below 10%.



(a) ER1



(b) ER2

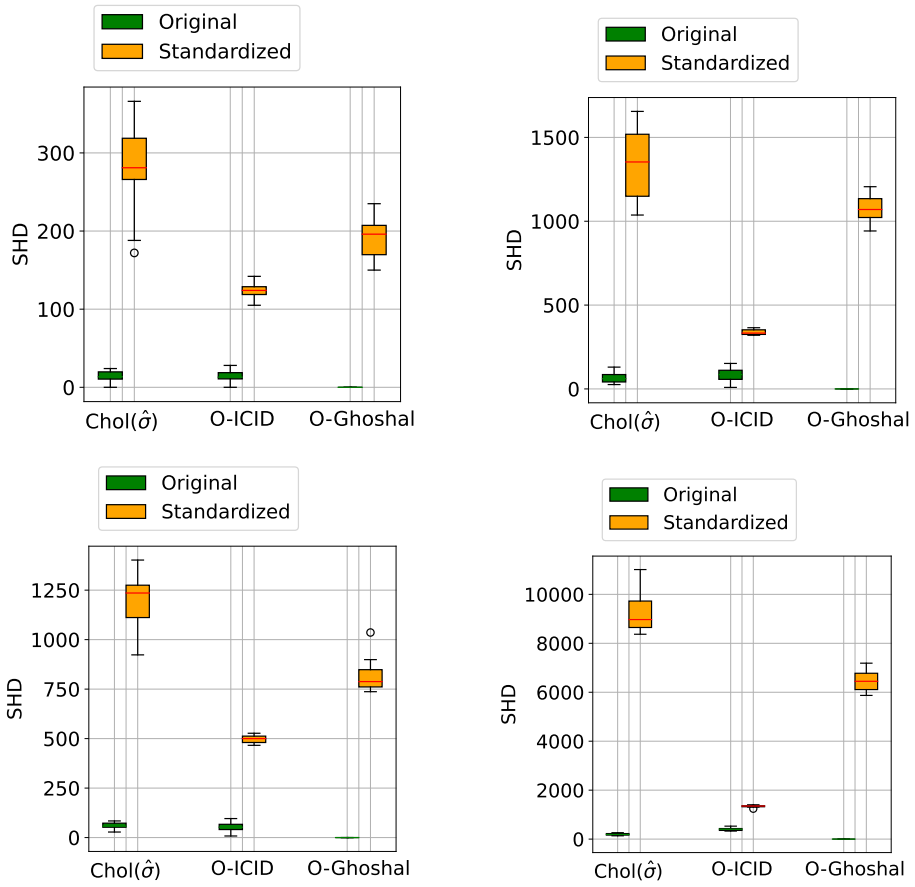
Figure 1: Learning curve of $\mathcal{O}$-ICID: Infeasibility measure (Left) and causal discovery error (Right) vs time. The true graph $\text{E}(B^\star)$ has $d = 50$ nodes and is generated from the ER1 or ER2 DAGs.

Impact of standardization. Based on the same linear SEM and data generation with the EV case in the last experiment, we produce a NV case through a standardization of the variable variances of $\mathbf{X} = (X_1, \ldots, X_d)$, which transforms $\mathbf{X}$ into $\tilde{X}_i = \frac{X_i}{\sqrt{\text{var}(X_i)}}$. Consequently, the inverse covariance matrix of $\tilde{\mathbf{X}}$ reads

$$\tilde{\Theta} := D\Theta D \quad \text{for } D = \text{diag}\left(\sqrt{\text{var}(X_1)}, \ldots, \sqrt{\text{var}(X_d)}\right).$$

The tests are conducted on graphs (with number of nodes d ranging from 100 to 400) with and without standardization.

From the results shown in Figure 2 and Table 3, we have the following observations: (i) In the original case with equal noise variances (EV), $\mathcal{O}$-Ghoshal outperforms the



(a) ER1, $d \in \{100, 400\}$ (b) ER2, $d \in \{100, 400\}$

Figure 2: Impact of var-sorting bias: comparative SHD performances on original vs standardized Θ by Chol($\hat{\sigma}$), $\mathcal{O}$-ICID and $\mathcal{O}$-Ghoshal for $d = 100$ (top) and $d = 400$ (bottom), on ER1 (left) and ER2 (right).

two other approaches in terms of accuracy (Table 3). The Cholesky decomposition of Θ via Chol($\hat{\sigma}$) also gives comparable results in both accuracy and running time when the variance order is close to the causal order. The accuracy of $\mathcal{O}$-ICID is comparable to the former two methods. (ii) In the standardized case, however, the accuracy of both $\mathcal{O}$-Ghoshal and Chol are significantly weakened, to the extent that their solutions become almost irrelevant to $B^\star$, while $\mathcal{O}$-ICID is much less affected. In particular, Table 3 shows that, in the standardized case, the true positive rates (TPR) of $\mathcal{O}$-ICID remain good enough (e.g., 78% on ER1 and 60% on ER2 for $d = 200$), which are much higher than the other two methods.

4.2 Impact of statistical estimation issues

The robustness of ICID and Ghoshal algorithms is assessed by varying two parameters for the estimation of the inverse covariance: i) the number n of samples; ii) a threshold τ that controls the hard threshold operation for edge selection; see Algorithm 4.

We test with 200 combinations of the ratio n/d, ranging from .5 to 5, and $\tau \in (0, 1)$. A global display of the results is presented in Figure 3, where each point corresponds to an instance of $\mathcal{O}$-ICID, where the input matrix $\hat{\Theta}$ is estimated for a given pair (n, τ). Each instance is represented in the 2D plane where the X-axis indicates the SHD error of $\hat{\Theta}$ (with respect to the true inverse covariance matrix Θ) and the Y-axis indicates the causal discovery error in SHD (SHD($\hat{B}$)). The general trend is that the causal error is bounded by the estimation error of the input $\hat{\Theta}$ (most instances are concentrated in the diagonal region), as could have been expected. However, a clear difference between ICID and Ghoshal is noted, as the causal discovery error SHD($\hat{B}$) is always bounded by and in many cases smaller than the statistical estimation error of $\hat{\Theta}$ while Ghoshal does not always have bounded errors (see instances above the diagonal) and tend to have more amplified errors than $\mathcal{O}$-ICID.

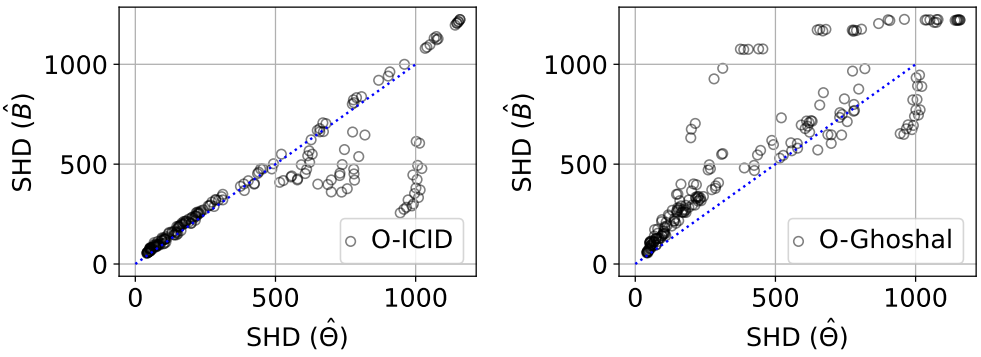
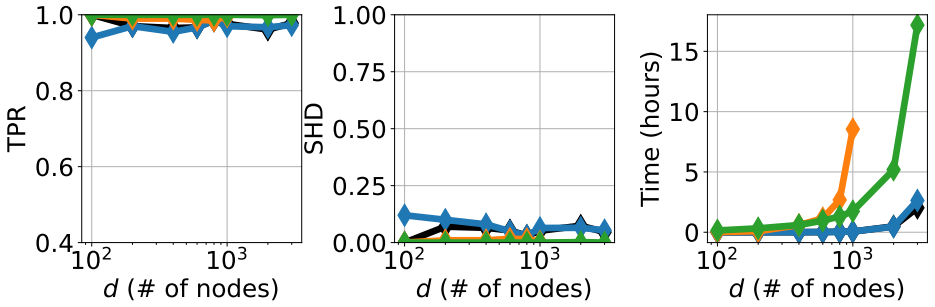


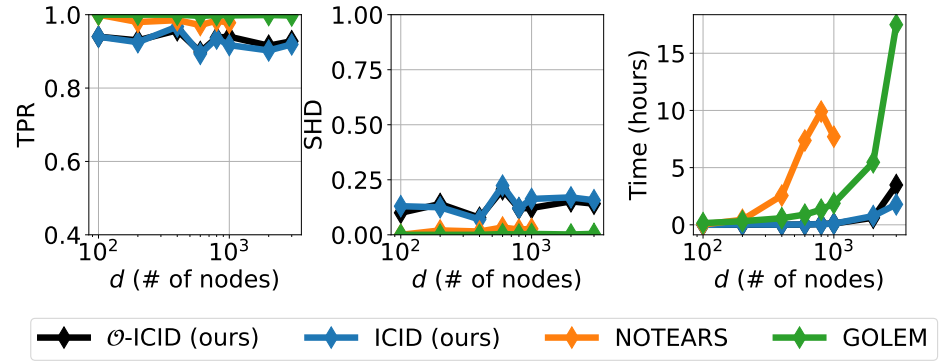
Figure 3: Impact of the statistical estimation of Θ: Causal discovery error (SHD($\hat{B}$)) vs estimation error in Θ (SHD($\hat{\Theta}$)). Left: $\mathcal{O}$-ICID. Right: $\mathcal{O}$-Ghoshal.

4.3 Scalability

The comparison of ICID (Appendix B.2) with the other baselines on large-sized graphs is reported in Figure 4. The implementation and computation resources of NOTEARS and ICID (running on CPU) and GOLEM (running on GPU) are not the same as GOLEM leverages parallel computation. Nevertheless, all three methods rely on gradient-based optimization techniques and the comparison is to show a rough trend of complexity.



(a) ER0.5 (deg = 0.5)



(b) ER1 (deg = 1)

Figure 4: Comparative performances of structural causal learning vs number of nodes $d \in \{200, 400, \ldots, 2000, 3000\}$ on linear SEM problem instances with ER graph. Left: degree .5; Right: degree 1. The 'SHD' in the middle subplots denotes by abuse the normalized value: $\mathrm{SHD}(\hat{B})/\|B\|_0$.

Figure 4 shows that the computation time of both ICID methods increase with d with a much slower rate than GOLEM and NOTEARS: on ER1, the speedups over GOLEM are around 5 times at $d = 2000$ and 10 times at $d = 3000$; their speedups over NOTEARS are even greater when $d \geq 1000$. At the same time, the loss in learning accuracy of ICID compared to the GOLEM and NOTEARS are moderate, with a TPR above 85% and a normalized SHD below 20%.

The scalability of Ghoshal depends primarily on the running time of GraphicalLasso (used in the implementation of Ghoshal which takes up a majority of the total time

on large graphs), from our preliminary tests, which motivates us to evaluate the time efficiency of $\mathcal{O}$-Ghoshal instead in comparison with Chol($\hat{\sigma}$), for the similarity of these two methods; see results in Appendix C.3.

4.4 Real data

The benchmark protein signaling dataset [SPP$^+$05] consists of $n = 853$ observed expression signals of $d = 11$ proteins with an expert-provided ground-truth graph B, including 17 edges. $\mathcal{O}$-ICID considers the exact Θ computed from B; ICID considers the empirical $\hat{\Theta}_n$ estimated from $n = 100$ and $n = 853$ samples. Results are presented in Table 1. For $n = 100$, the average accuracy of ICID is on par with GOLEM (EV and NV version) and both are dominated by BCD-Nets (NV). For sanity check purposes, we test $\mathcal{O}$-ICID on the same dataset by using (as input matrix of Algorithm 1) the oracle inverse covariance matrix Θ obtained from the matrix equation (2) given $B^\star$. In this case, $\mathcal{O}$-ICID yields solutions with an average SHD of 9.0 which outperforms by far the best scores (13.9 by Gadget and 14.7 by BCD-Nets) reported in [CGE21].

Table 1: Causal structure learning on the protein dataset [SPP$^+$05]. The results of GOLEM and BCD-Nets are reported from [CGE21].

	n	# Edges	SHD	Time (sec)
BCD-Nets (EV)	100	11.3 ± 1.2	19.5 ± 0.3	–
BCD-Nets (NV)	100	9.2 ± 2.0	14.7 ± 0.9	–
GOLEM (EV)	100	1.5 ± 1.3	18.5 ± 1.3	–
GOLEM (NV)	100	1.5 ± 1.3	18.5 ± 1.3	–
ICID (ours)	100	11.6 ± 3.7	18.5 ± 2.5	3.92 ± 4.29
	853	5.7 ± 2.7	17.3 ± 1.4	1.74 ± 0.44
$\mathcal{O}$-ICID (ours)	NA	12.0 ± 0.0	9.0 ± 0.7	1.53 ± 0.01

5 Comparison with previous work

The results and algorithms proposed in this work are based on the matrix equation (2), which is studied and used previously in the work of [LB14, GH18]. [LB14] start from (2) to derive an identifiability result under the knowledge of the noise variances. This result ([LB14, Theorem 7]) underlies a score-based method that requires an enumeration of candidate DAGs in order to pick out a minimizer of the score function. The enumeration of DAGs can be efficient using dynamic programming, assuming that the support graph of Θ (which is the superset of the support of candidate DAGs) has a restricted tree-width. On the other hand, [GH18] recovers a topological order that is consistent with the causal structure iteratively: at each iteration a terminal variables (a variable with parent nodes only and no effects) is determined and removed from the linear SEM,

and at the same time, the corresponding row of the causal DAG is deduced using the matrix equation (2).

The proposed method $\mathcal{O}$-ICID uses the matrix equation (2) in a different way. By Algorithm 1, we use the inverse covariance matrix Θ in (2) as the target for the matrix decomposition model (6), and harness continuous optimization techniques (an augmented Lagrangian method) for the learning of this model. More precisely, the comparison with these two previous work is as follows.

- Compared to the score-based approach in [LB14], $\mathcal{O}$-ICID has verifiable optimal conditions while the former depends on a thorough enough enumeration of DAGs within the support of Θ. 3). Under the same assumptions as in [LB14, Theorem 7], the identification of the true causal structure can be realized by $\mathcal{O}$-ICID given the oracle inverse covariance matrix; see Theorem 6 and empirical results in Section 4.1 and Appendix C.5.

- Compared to the algorithm of [GH18], $\mathcal{O}$-ICID (Algorithm 1) differs in that the proposed algorithm seeks solution to the matrix equation (2) using gradient-based continuous optimization techniques instead of recovering a topological order sequentially while fitting the causal structure B to the matrix equation during the process. The computational cost of [GH18, Algorithm 1] is similar to the Cholesky decomposition of Θ (the special case algorithm $\mathrm{Chol}(\hat{\sigma})$), which is smaller than $\mathcal{O}$-ICID. However, the accuracy of [GH18, Algorithm 1] depends primarily on the consistency of its topological ordering method (using the variables' variances) with respect to the true causal structure, which holds under [GH18, Assumption 1]. In more general cases when such consistency does not hold, however, solutions by [GH18, Algorithm 1] become irrelevant. In such cases, the accuracy of $\mathcal{O}$-ICID is much less degraded than the former, as illustrated by the empirical results in Section 4.1.

The ICID framework in this work is also related to methods using maximum-likelihood estimation (MLE) [AAZ19, NGZ20]. Through the GraphicalLasso formulation [FHT07], we recover the M-estimator of GOLEM [NGZ20]. In fact, the log-likelihood function for inverse covariance estimation (via Graphical Lasso) rewritten in terms of the linear SEM parameters (B, Ω) (ignoring the constant terms),

$$\frac{1}{2} f(\Theta; X) := \frac{1}{2n} \operatorname{tr}(X^{\mathrm{T}} \Theta X) - \frac{1}{2} \log \det(\Theta) = -\log p(B, \Omega; X)$$

where (B, Ω) and Θ are related by the matrix equation (2), is the objective function of GOLEM for optimizing B.

The difference between ICID and GOLEM is that ICID divides the causal discovery problem in two consecutive parts that are lighter than the MLE problem of GOLEM computationally. Note that the MLE approach to inverse covariance estimation is a convex problem over the convex cone $\mathcal{S}^d_{++}$; and $\mathcal{O}$-ICID also has a lower per-iteration cost than GOLEM. The overall time complexity comparison between ICID and GOLEM (and NOTEARS) is shown in Section 4.3.

6 Discussion and perspectives

This paper addresses the causal discovery problem by separating it into two consecutive subproblems. The first subproblem in the presented ICID framework consists of estimating the inverse covariance matrix Θ, and the second subproblem, formulated as $\mathcal{O}$-ICID (6), consists of finding and optimizing a support-constrained decomposition of Θ.

One main contribution of the paper is the $\mathcal{O}$-ICID approach, which leverages continuous optimization techniques for finding the proposed specific type of matrix decomposition of Θ. The proposed method of $\mathcal{O}$-ICID is shown to have good scalability in learning large causal structures. By pairing $\mathcal{O}$-ICID with a basic, empirical inverse covariance estimation method, we observe significant speedups by the proposed method compared to state-of-the-art methods in learning causal graphs with a few thousand nodes.

Among the difficulties in causal discovery due to, e.g., the sparsity and node degree distribution of the sought DAG, limited amount of available observational data, many are directly reflected in the task of inverse covariance estimation, that is, the first subproblem of the ICID framework. Further work should focus on enhancing the two parts of ICID. One is about harnessing more advanced methods for inverse covariance estimation, and leveraging data augmentation techniques and bootstrapping. Another is concerned with more refined or novel techniques to enhance $\mathcal{O}$-ICID including supervised machine learning for identifying the true pairs from their estimates depending on the family of considered graphs, in the spirit of the Cause-Effect Pair Challenge [GSB19].

A Proofs

A.1 Example with chordal matrices

The following theorem illustrates the well-definedness of the support-constrained decomposition (Definition 4) in the special case with chordal matrices. This result builds upon the characterization of Cholesky decomposition of positive definite matrices supported on a chordal graph [FG65, Ros70, PPS89, VA15].

Proposition 8. *Let $\Theta \in \mathcal{S}^d_{++}$ be a positive definite matrix whose support graph $\mathrm{E}(\Theta)$ is chordal. Then Θ admits a support-constrained decomposition (Definition 4). Moreover, the set $\mathcal{S}(\Theta)$ (5) contains a pair (B, D) where $B \in \mathbb{R}^{d\times d}$ is supported on a DAG.*

Proof. For completeness, we state the following lemma on the connections between chordal graphs and positive definite matrices that can *factor without fill* [FG65, Ros70, PPS89]. Given an undirected graph $G = (V, E)$ endowed with a node ordering σ, the

following sets are considered:

$$\mathcal{S}_{G,\sigma} = \{\Theta \in \mathcal{S}_{++}^d : \Theta_{ij} = 0 \text{ for } (\sigma^{-1}(i), \sigma^{-1}(j)) \notin E\}, \tag{12}$$

$$\mathcal{L}_{G,\sigma} = \{L \in \mathbb{R}^{d\times d} : L_{ii} = 1, L_{ij} = 0 \text{ for } i < j \text{ or } (\sigma^{-1}(i), \sigma^{-1}(j)) \notin E\}. \tag{13}$$

Lemma 9 ([Ros70, PPS89])**.** *Let $G = (V, E)$ be a chordal graph, σ an ordering of V which corresponds to a perfect elimination ordering of G. Then it holds that $\Sigma \in \mathcal{S}_{G,\sigma}$ (12) if and only if $L \in \mathcal{L}_{G,\sigma}$ (13), where L is the Cholesky factor of Σ such that $\Sigma = LDL^{\mathrm{T}}$.*

Based on Lemma 9, G being chordal implies that, for a certain node permutation σ (corresponding to the permutation matrix P_σ), the positive definite matrix $\tilde{\Theta} := P_\sigma \Theta {P_\sigma}^{\mathrm{T}}$ belongs to $\mathcal{S}_{G,\sigma_0}$ (12) and that the (lower-triangular) Cholesky factor matrix $\tilde{L}$ of $\tilde{\Theta}$ (such that $\tilde{L}\tilde{D}\tilde{L}^{\mathrm{T}} = \tilde{\Theta}$ for a diagonal matrix $\tilde{D}$) satisfies $\tilde{L} \in \mathcal{L}_{G,\sigma_0}$ (13). As a consequence, the matrices (A, D) which are σ-similar to $(\tilde{L}, \tilde{D})$, i.e., $A := {P_\sigma}^{\mathrm{T}} \tilde{L} P_\sigma$ and $D := {P_\sigma}^{\mathrm{T}} \tilde{D} P_\sigma$ satisfy:

$$ADA^{\mathrm{T}} = {P_\sigma}^{\mathrm{T}} \tilde{L}\tilde{D}\tilde{L}^{\mathrm{T}} P_\sigma = {P_\sigma}^{\mathrm{T}} \tilde{\Theta} P_\sigma = \Theta,$$

which means that $A' = A\sqrt{D}$ satisfies $A'{A'}^{\mathrm{T}} = \Theta$. Moreover, it holds that $\mathrm{E}(A') \subset \mathrm{E}(\Theta)$ because (i) the two support graphs are identical to $\mathrm{E}(\tilde{L})$ and $\mathrm{E}(\tilde{\Theta})$, respectively, up to the node permutation σ and (ii) $\mathrm{E}(\tilde{L}) \subset \mathrm{E}(\tilde{\Theta})$ by Lemma 9. Therefore $A'{A'}^{\mathrm{T}} = \Theta$. Note that A' is σ-similar to $\tilde{L}D'$ (with diagonal $D' = P_\sigma \sqrt{D} {P_\sigma}^{\mathrm{T}}$), which is a strict triangular matrix. Hence A' represents a DAG. □

A.2 Proof of Proposition 5

Proof. For the first statement: $\varphi(B, \Omega) = \Theta$ by definition of the SEM given. It remains to prove that B also satisfies $\mathrm{E}(B) \subset \mathrm{E}(\Theta_{\mathrm{off}})$. Indeed, note that Assumption 3 is equivalent to $\Theta_{ij} \neq 0$ if $(i, j) \in \mathcal{M}(B)$ (moralization of B). Hence, in particular, $\Theta_{i,j} \neq 0$ if $B_{i,j} \neq 0$ or $B_{j,i} \neq 0$. □

A.3 Proof of Theorem 6

Proof. The pair satisfying $(B, {\Omega^\star}^{-1}) \in \mathcal{S}(\Theta^\star)$ entails that

$$(I - B){\Omega^\star}^{-1}(I - B)^{\mathrm{T}} = (I - B^\star){\Omega^\star}^{-1}(I - B^\star)^{\mathrm{T}}. \tag{14}$$

Now note that

$$(I - B){\Omega^\star}^{-\frac{1}{2}} = {\Omega^\star}^{-\frac{1}{2}}(I - {\Omega^\star}^{\frac{1}{2}} B {\Omega^\star}^{-\frac{1}{2}}) := {\Omega^\star}^{-\frac{1}{2}}(I - \tilde{B}),$$

and the same change of variable applies to $B^\star$ such that $(I - B^\star){\Omega^\star}^{-\frac{1}{2}} := {\Omega^\star}^{-\frac{1}{2}}(I - \tilde{B}^\star)$. Applying this change of variables to (14), we have ${\Omega^\star}^{-\frac{1}{2}}(I - \tilde{B})(I - \tilde{B})^{\mathrm{T}} {\Omega^\star}^{-\frac{1}{2}} = {\Omega^\star}^{-\frac{1}{2}}(I - \tilde{B}_0)(I - \tilde{B}_0)^{\mathrm{T}} {\Omega^\star}^{-\frac{1}{2}}$, i.e.,

$$(I - \tilde{B})(I - \tilde{B})^{\mathrm{T}} = (I - \tilde{B}^\star)(I - \tilde{B}^\star)^{\mathrm{T}}. \tag{15}$$

Since $B^\star$ and B (hence $\tilde{B}^\star$ and $\tilde{B}$) are supported on DAGs, both $(I-\tilde{B}^\star)$ and $(I-\tilde{B})$ are permutation similar to unit lower-triangular matrices (lower-triangular matrices with one's on the diagonal). The result then follows from [LB14, Lemma 25], which confirms that if two square matrices X and Y are permutation similar to unit lower-triangular matrices such that $XX^T = YY^T$, then $X = Y$. Through this lemma, $I - \tilde{B} = I - \tilde{B}^\star$, which concludes the proof. □

B Algorithms and computational details

Proposition 10. *Let $\ell_2(B, D) := \frac{1}{2}\|\Theta - \varphi(B, D)\|_{\mathrm{F}}^2$. The gradient of ℓ_2 at $(B, D) \in \mathbb{R}^{d\times d} \times \mathbb{R}^{d\times d}$ is*

$$\nabla \ell_2(B, D) = \Big(2\Delta(I - B)D, \quad -\Delta + (\Delta B + B^{\mathrm{T}}\Delta) - B^{\mathrm{T}}\Delta B\Big)$$

where $\Delta := \Theta - \varphi(B, D)$. The gradient of $\ell_2(B, D)$ restricted in the feasible space $\{B \in \mathbb{R}^{d\times d} : B_{\mathrm{off}} = 0\} \times \{D = \mathrm{diag}(d) : d \in \mathbb{R}^d_{++}\}$ is

$$\begin{aligned}\nabla_B \ell_2(B, D) &= 2(\Delta(I - B)D)_{\mathrm{off}},\\ \nabla_D \ell_2(B, D) &= -\mathrm{diag}(\Delta - (\Delta B + B^{\mathrm{T}}\Delta) + B^{\mathrm{T}}\Delta B).\end{aligned}$$

Proof. Applying the chain rule to $\ell_2(B, D)$ we have

$$\mathrm{D}\ell_2(B, D)[\xi, \eta] = \langle \phi(B) - \Theta, \mathrm{D}\varphi(B, D)[\xi, \eta]\rangle\,, \tag{16}$$

where $\mathrm{D}\varphi(B, D)[\xi, \eta]$ is calculated as follows:

$$\begin{aligned}\mathrm{D}\varphi(B, D)[\xi, \eta] &= \lim_{t\to 0} \frac{1}{t}\Big[- t(I - B)D\xi^{\mathrm{T}} - t\xi D(I - B)^{\mathrm{T}} + (I - (B + t\xi))(t\eta)(I - (B + t\xi))^{\mathrm{T}}\Big]\\ &= -(I - B)D\xi^{\mathrm{T}} - \xi D(I - B)^{\mathrm{T}} + \eta - (\eta B^{\mathrm{T}} + B\eta) + B\eta B^{\mathrm{T}}.\end{aligned}$$

By combining (16) with the identification $\mathrm{D}\ell_2(B, D)[\xi, \eta] := \langle \nabla_B \ell_2(B, D), \xi\rangle + \langle \nabla_D \ell_2(B, D), \eta\rangle$, it follows that

$$\begin{aligned}\nabla_B \ell_2(B, D) &= 2\Delta(I - B)D\\ \nabla_D \ell_2(B, D) &= -\Delta + (\Delta B + B^{\mathrm{T}}\Delta) - B^{\mathrm{T}}\Delta B\end{aligned}$$

where $\Delta := \Theta - \varphi(B, D)$. □

B.1 Primal descent solver of $\mathcal{O}$-ICID

The FISTA [BT09] is applied to solving (9) in view of the ℓ_1 norm penalty term.

Algorithm 2 FISTA for the primal descent problem (9) of $\mathcal{O}$-ICID

Input: Inverse covariance matrix $\Theta \in \mathbb{R}^{d\times d}$, objective function f of (9), λ_1', $\alpha_0 > 0$, $\gamma \in (0,1)$, $\beta = \frac{1}{2}$, tolerance ϵ, $\rho \in (0,1)$, low-rank parameter r.

1: Initialize: $W_0 = \mathbf{0}_{d\times d}$, set $Y_0 = W_0$.

2: **for** $s = 1, 2, \ldots$ **do**

3: Backtracking: find smallest integer $k_s \geq 0$ such that, for $\tilde{\eta}_s := \alpha_0 \beta^{k_s}$,

$$f(\tilde{W}) - f(Y_{s-1}) \leq -\gamma \tilde{\eta}_s \|\mathrm{grad}_{\mathcal{C}_\Theta} f(Y_{s-1})\|^2,$$

where

$$\tilde{W} = \mathrm{prox}_{\tilde{\eta}_s \lambda_1' \ell_1}\big(Y_{s-1} - \tilde{\eta}\mathrm{grad}_{\mathcal{C}_\Theta} f(Y_{s-1})\big).$$

 # see (18)-(19)

4: Update FISTA iterates:

$$W_s = \tilde{W} \quad \text{and} \quad Y_s = W_s + \frac{s-1}{s+2}(W_s - W_{s-1}).$$

5: Stop if $\|\Delta(W_s)\|_{\mathrm{F}} \leq \epsilon$: *# see* (20)
return W_s

6: **end for**

The operator for graph support projection in Algorithm 2, line 10 is as follows.

Definition 11. Given $S \in \mathbb{R}^{d\times d}$, the projection onto the support graph $\mathrm{E}(S)$ is denoted and defined as $P_S : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ such that

$$(P_S(Z))_{ij} = \begin{cases} Z_{ij} & \text{if } (i,j) \in \mathrm{E}(S) \\ 0 & \text{otherwise.} \end{cases}$$

The support constraint of (9), along with the prior knowledge that $\mathrm{diag}(B) = \mathbf{0}$ (since a candidate adjacency matrix B does not admit any self-cycle), imposes that the maximal search space of the problem is the following set

$$\mathcal{C}_\Theta := \{B \in \mathbb{R}^{d\times d} : B_{ij} = 0 \quad \forall i = j \text{ or } (i,j) \notin \mathrm{E}(\Theta)\}, \tag{17}$$

which is a (linear) subspace of $\mathbb{R}^{d\times d}$ with dimension $(\|\Theta\|_0 - d)$. This means that the constraint of (9) can be satisfied using subspace projection straightforwardly.

Denote the smooth part of the objective function of (9) by

$$f(B, \Delta) := \frac{\rho}{2}\|\Theta - \varphi(B)\|_{\mathrm{F}}^2 + \langle \Delta, \Theta - \varphi(B)\rangle \tag{9b}$$

where $\varphi(B) := \varphi(B, I)$ (for $\Omega = I$). The gradient $\nabla_B f(B, \Delta)$ is calculated using Proposition 10. From definition (17), it follows that the gradient of f restricted to subspace

$\mathcal{C}_\Theta$ (17), denoted as $\text{grad}_{\mathcal{C}_\Theta} f$, is

$$\text{grad}_{\mathcal{C}_\Theta} f(B) = P_{\mathcal{C}_\Theta}(\nabla_B f(B, \Delta)), \tag{18}$$

where $P_{\mathcal{C}_\Theta} : \mathbb{R}^{d\times d} \to \mathcal{C}_\Theta$ is the projection (Definition 11) onto the support of Θ_{off}.

On the other hand, the proximal operator associated with the ℓ_1 term of (9) is

$$\text{prox}_{\lambda\ell_1}(Z) = \begin{cases} \text{sign}(Z_{ij})(|Z_{ij}| - \eta) & \text{if } |Z_{ij}| \geq \lambda \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

In Algorithm 2, line 5, the stopping criterion is defined with respect to ℓ_1-subdifferential optimality. Hence $\Delta(B) \in \mathbb{R}^{d\times d}$ is as follows:

$$\begin{aligned} &(\Delta(B))_{ij} = -(\text{grad}_{\mathcal{C}_\Theta} f(B))_{ij} - \lambda_1' \text{sign}(B_{ij}) \text{ if } B_{ij} \neq 0, \\ &(\Delta(B))_{ij} = |(\text{grad}_{\mathcal{C}_\Theta} f(B))_{ij}| - \lambda_1' \text{ if } B_{ij} = 0 \text{ and} \\ &\qquad\qquad\qquad\qquad |(\text{grad}_{\mathcal{C}_\Theta} f(B))_{ij}| \geq \lambda_1', \\ &(\Delta(B))_{ij} = 0 \quad \text{otherwise.} \end{aligned} \tag{20}$$

Remark 12 (Alternatives). Decomposition (9) can be computed by standard solvers for matrix decomposition, with a light adaptation to the ℓ_1 penalty term. □

The line search parameters in Algorithm 2: $\alpha_0 = \lambda_{\max}(\Theta)$ (maximal eigenvalue), $\gamma = \frac{1}{2}$, and tolerance parameters $\epsilon = 10^{-4}$.

Remark 13 (Computational cost).

1. The computational cost of $\nabla_B f(B, D)$, similar to $\nabla_B \ell_2(B, D) = 2((\Theta - \varphi(B, D))(I - B)D)_{\text{off}}$ (see Proposition 10), costs $2k^2 d$ floating-point operations, where k is the maximal node degree of the moralization of B.

2. The computational cost of Algorithm 2 is dominated by the line 3, which requires the computation of (18)-(19). Given that the number of backtracking linesearchs is limited to $n_{ls} \leq 20$, the per-iteration cost of Algorithm 2 is dominated by $2n_{ls}dk^2$.

   Most often, the maximal node degree is limited ($k \leq K$ for a constant $K \geq 1$) which means there exists a constant $C > 0$ such that $Ck^2 \leq d$, and therefore the per-iteration cost of Algorithm 2 is bounded by $O(d^2)$.

B.2 ICID framework with an additional DAG constraint

Algorithm 3 gives an implementation of the ICID framework with an additional DAG constraint using the exponential trace function as in [ZARX18]. In this algorithm, a proximal mapping (22) is defined as a solution to

$$\min_{B\in\mathbb{R}^{d\times d}} \quad h(B) + \frac{1}{\gamma_2} \underbrace{\|B - c_0 B_{t+1}\|_{\text{F}}^2}_{g(B;B_{t+1})}, \tag{23}$$

Algorithm 3 An ICID framework with DAG constraint

Input: Observational data $X \in \mathbb{R}^{d\times n}$

1: Get Θ from Inverse covariance estimator (Algorithm 4)
2: Compute B_0 by $\mathcal{O}$-ICID (Algorithm 1)
3: **for** $t = 1, \ldots,$ **do**
4: **if** stopping criteria$(B_t, \tilde{B}_t)$ attained **then**
5: return B_t
6: **end if**
7: Compute proximal mappings:

$$B_{t+1} = (1-\rho)B_0 + \rho\tilde{B}_t \tag{21}$$

$$\tilde{B}_{t+1} = \mathrm{prox}_{\gamma_2 h}(c_0 B_{t+1}) \tag{22}$$

8: Increment γ_2
9: **end for**

where h is the exponential trace-based function

$$h(B) = \mathrm{tr}(\exp(|B|))$$

with the absolute value operation $|\cdot|$ applied to B element-wisely. Due to the exponential trace in h [ZARX18], problem (23) is nonconvex. We resort to the search of one proximal point (22) satisfying sufficient decrease in h. In view of reducing the cost for computing the gradients of the exponential trace function h in (22), the low-rank method LoRAM-AGD of [DS22] is used. The increment rule line 8 is an ad-hoc adaptation of the AMA (alternating minimization algorithm) for optimizing the Lagrangian of an equality constrained optimization.

An empirical inverse covariance estimator. In the experiments, we use a basic empirical inverse covariance estimator for construction input matrices of $\mathcal{O}$-ICID and $\mathcal{O}$-Ghoshal.

Algorithm 4 Empirical inverse covariance estimator

Input: Data matrix $X \in \mathbb{R}^{n\times d}$, parameter $\lambda_1 \in (0,1)$

Output: $\widehat{\Theta}_{\lambda_1} \in \mathbb{R}^{d\times d}$

1: Compute empirical covariance and its inverse:

$$\widehat{C} = \frac{1}{n}(X - \bar{X})^{\mathrm{T}}(X - \bar{X}) \quad \text{and} \quad \widehat{\Theta} = \widehat{C}^{\dagger}, \tag{24}$$

where $\widehat{C}^{\dagger}$ denotes the pseudo-inverse of $\widehat{C}$.

2: Element-wise thresholding on off-diagonal entries:

$$\begin{aligned} \operatorname{diag}(\widehat{\Theta}_{\lambda_1}) &:= \operatorname{diag}(\widehat{\Theta}), \\ (\widehat{\Theta}_{\lambda_1})_{\text{off}} &:= \mathbb{H}(\widehat{\Theta}_{\text{off}}, \lambda_1 \|\widehat{\Theta}_{\text{off}}\|_{\max}), \end{aligned} \tag{25}$$

where $\mathbb{H}$ is defined as

$$\mathbb{H}(y,\tau) = \begin{cases} y & \text{if } |y| \geq \tau \\ 0 & \text{otherwise.} \end{cases}$$

In the computation of (24), the pseudo-inverse coincides with the inverse of $\widehat{C}$ when $\widehat{C}$ is positive definite (e.g., when the number n of samples is sufficiently large). In (25), the subscript 'off' indicates the following filtering operation

$$\Theta_{\text{off}} = \{\Theta_{ij} : i \neq j\}$$

where the indices of the remaining (off-diagonal) entries are preserved.

C Experiments

C.1 Random graphs and synthetic data

The experiments on synthetic data are conducted with DAGs generated from two sets of random graphs: (i) Erdős–Rényi (ER) graphs and (ii) Scale-free (SF) [BA99] graphs, as characterized in Table 2.

Table 2: Features of different graph models.

	Parameter	Degree distribution
Erdős-Rényi	$p \in (0,1)$	Binomial $\mathcal{B}(d,p)$
Scale-free	γ	$P(k) \propto k^{-\gamma}$

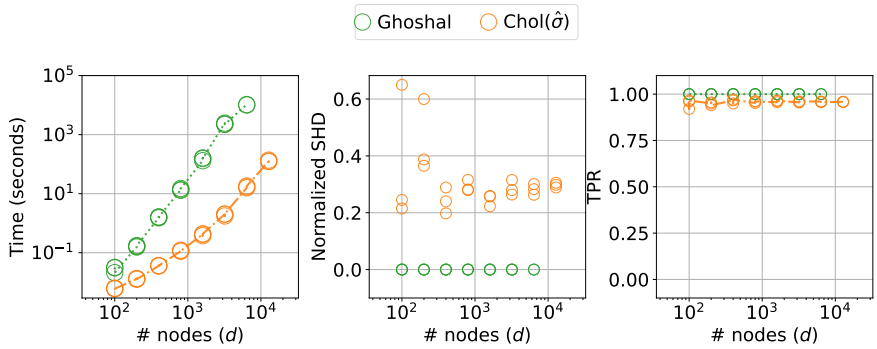
The generation of random DAGs from the two sets above is the same as in [ZARX18, NGZ20]. The naming of these graphs has a node degree specification, such as 'ER1', where the number indicates the average node degree of the graph. Specifically, for a

given DAG $G^\star$, its weighted adjacency matrix $B^\star$ is generated by assigning weights to the nonzeros of $\mathbb{B}(G^\star) \in \{0,1\}^{d\times d}$ independently from the uniform distribution: $B^\star_{ij} \sim \text{Unif}([-2,-0.5] \cup [0.5,2])$, for $(i,j) \in \text{E}(\mathbb{B}(G^\star))$. We generate observational data according to the linear SEM model (1), and store them in dataset $X \in \mathbb{R}^{n\times d}$ where n is the number of samples. The additive noises of the linear SEM such that $X = B^{\star\text{T}}X+E$, belong to either of the following models: (i) Gaussian noise (Gaussian): $E \sim \mathcal{N}(0,\Omega)$ and (ii) Exponential noise (Exponential): $E \sim \text{Exp}(\Omega)$, where Ω is a diagonal matrix of noise variances. Therefore, the dataset X belongs to one of the following categories $\{\text{ER deg}, \text{SF deg}\} \times \{\text{Gaussian}, \text{Exponential}\}$ (where deg is the aforementioned average node degree).

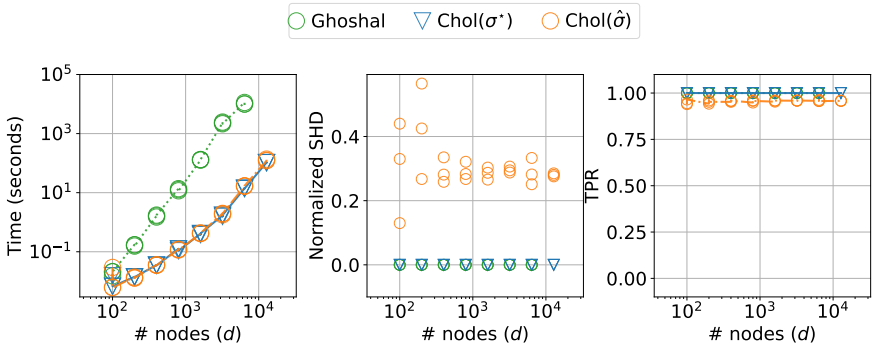
C.2 Evaluation Metrics

The graph metrics for the comparison of graph edge sets are the commonly used (e.g., by the aforementioned baseline methods) ones as follows:
(1) TPR = TP/T (higher is better),
(2) FDR = (R + FP)/P (lower is better),
(3) FPR = (R + FP)/F (lower is better),
(4) SHD = E +M + R (lower is better).
More precisely, SHD is the (minimal) total number of edge additions (E), deletions (M), and reversals (R) needed to convert an estimated DAG into a true DAG. Since a pair of directed graphs are compared, a distinction between True Positives (TP) and Reversed edges (R) is needed: the former is estimated with correct direction whereas the latter is not. Likewise, a False Positive (FP) is an edge that is not in the undirected skeleton of the true graph. In addition, Positive (P) is the set of estimated edges, True (T) is the set of true edges, False (F) is the set of non-edges in the ground truth graph. Finally, let (E) be the extra edges from the skeleton, (M) be the missing edges from the skeleton.

C.3 Cholesky decomposition as a special case of $\mathcal{O}$-ICID

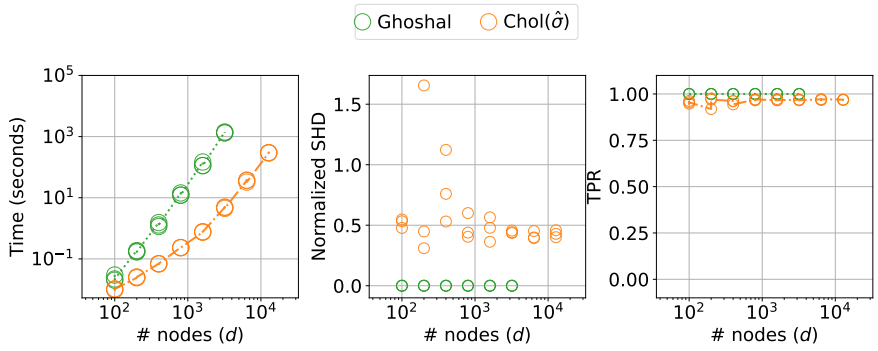


(a) $\sigma^\star$ unknown, ER2

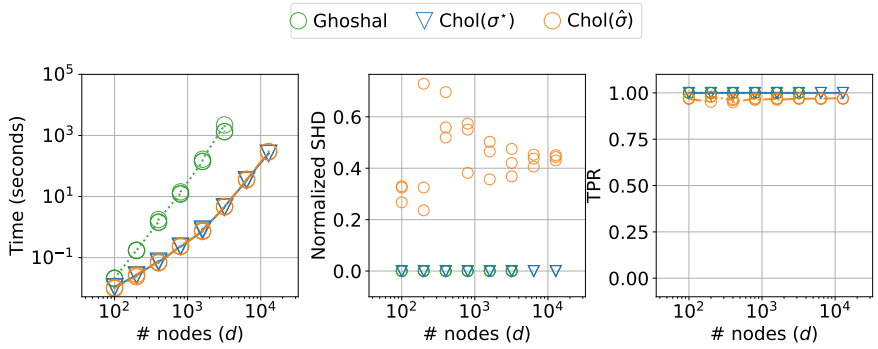


(b) $\sigma^\star$ known, ER2

Figure 5: Causal discovery from inverse covariance matrices: from mid-sized to large graphs.



(a) $\sigma^\star$ unknown, ER4



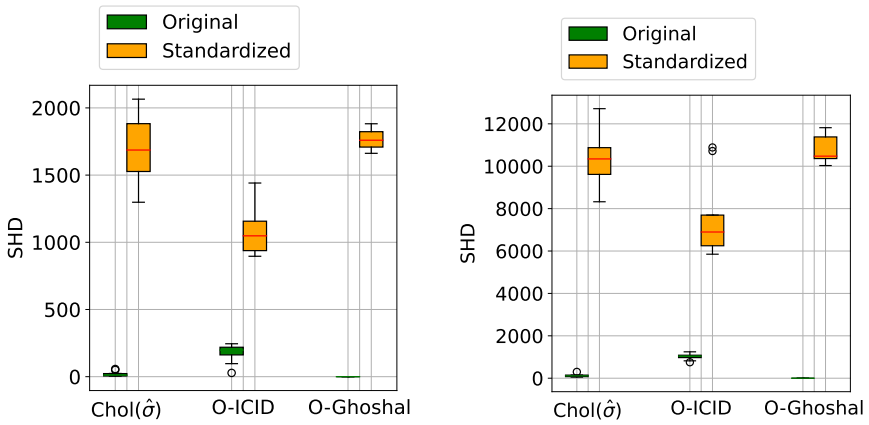
(b) $\sigma^\star$ known, ER4

Figure 6: Causal discovery from inverse covariance matrices: from mid-sized to large graphs.

C.4 Tests of var-sortability

Table 3: Results of causal discovery from inverse covariance matrices with and without standardization.

G	d	St'd	Algorithm	TPR	SHD	Median SHD	time (sec)
ER1	50	False	$\mathcal{O}$-ICID	0.982 ± 0.035	3.800 ± 6.844	0.000	6.935 ± 3.372
			Chol($\hat{\sigma}$)	0.942 ± 0.022	9.900 ± 3.957	10.000	0.003 ± 0.001
			$\mathcal{O}$-Ghoshal	1.000 ± 0.000	0.000 ± 0.000	0.000	0.002 ± 0.000
		True	$\mathcal{O}$-ICID	0.780 ± 0.065	60.200 ± 7.983	60.000	2.917 ± 0.131
			Chol($\hat{\sigma}$)	0.506 ± 0.114	104.600 ± 38.902	96.500	0.002 ± 0.000
			$\mathcal{O}$-Ghoshal	0.414 ± 0.087	82.600 ± 17.424	81.500	0.002 ± 0.000
ER2	50	False	$\mathcal{O}$-ICID	0.928 ± 0.078	26.500 ± 22.530	20.500	10.871 ± 2.527
			Chol($\hat{\sigma}$)	0.959 ± 0.027	21.200 ± 19.971	13.000	0.003 ± 0.000
			$\mathcal{O}$-Ghoshal	1.000 ± 0.000	0.000 ± 0.000	0.000	0.002 ± 0.000
		True	$\mathcal{O}$-ICID	0.717 ± 0.061	173.500 ± 14.676	172.500	3.205 ± 0.988
			Chol($\hat{\sigma}$)	0.504 ± 0.077	439.300 ± 107.440	413.500	0.003 ± 0.000
			$\mathcal{O}$-Ghoshal	0.262 ± 0.035	412.500 ± 33.124	412.500	0.002 ± 0.000
ER1	100	False	$\mathcal{O}$-ICID	0.959 ± 0.024	13.700 ± 8.138	13.000	21.285 ± 7.704
			Chol($\hat{\sigma}$)	0.963 ± 0.023	13.400 ± 8.422	13.500	0.004 ± 0.000
			$\mathcal{O}$-Ghoshal	1.000 ± 0.000	0.000 ± 0.000	0.000	0.017 ± 0.000
		True	$\mathcal{O}$-ICID	0.764 ± 0.063	123.300 ± 11.795	124.000	8.470 ± 0.198
			Chol($\hat{\sigma}$)	0.448 ± 0.064	278.500 ± 62.038	281.000	0.004 ± 0.000
			$\mathcal{O}$-Ghoshal	0.385 ± 0.042	192.900 ± 28.579	196.000	0.016 ± 0.003
ER2	100	False	$\mathcal{O}$-ICID	0.876 ± 0.083	81.900 ± 47.431	85.500	34.961 ± 1.049
			Chol($\hat{\sigma}$)	0.952 ± 0.020	68.900 ± 35.763	61.000	0.007 ± 0.001
			$\mathcal{O}$-Ghoshal	1.000 ± 0.000	0.000 ± 0.000	0.000	0.017 ± 0.000
		True	$\mathcal{O}$-ICID	0.651 ± 0.033	338.500 ± 16.828	333.000	10.108 ± 3.120
			Chol($\hat{\sigma}$)	0.447 ± 0.060	1346.100 ± 232.460	1353.500	0.006 ± 0.000
			$\mathcal{O}$-Ghoshal	0.269 ± 0.039	1073.200 ± 89.057	1070.000	0.017 ± 0.000
ER1	200	False	$\mathcal{O}$-ICID	0.964 ± 0.026	20.800 ± 16.665	19.000	99.687 ± 21.220
			Chol($\hat{\sigma}$)	0.959 ± 0.012	32.000 ± 15.804	28.000	0.009 ± 0.001
			$\mathcal{O}$-Ghoshal	1.000 ± 0.000	0.000 ± 0.000	0.000	0.109 ± 0.005
		True	$\mathcal{O}$-ICID	0.787 ± 0.049	236.600 ± 10.844	237.500	31.491 ± 0.386
			Chol($\hat{\sigma}$)	0.489 ± 0.039	504.900 ± 74.312	501.500	0.009 ± 0.001
			$\mathcal{O}$-Ghoshal	0.401 ± 0.036	382.300 ± 31.002	378.000	0.108 ± 0.005
ER2	200	False	$\mathcal{O}$-ICID	0.858 ± 0.044	163.800 ± 48.563	168.500	128.268 ± 1.281
			Chol($\hat{\sigma}$)	0.950 ± 0.010	152.200 ± 57.401	149.500	0.014 ± 0.001
			$\mathcal{O}$-Ghoshal	1.000 ± 0.000	0.000 ± 0.000	0.000	0.098 ± 0.024
		True	$\mathcal{O}$-ICID	0.603 ± 0.035	670.500 ± 23.978	668.000	33.582 ± 4.494
			Chol($\hat{\sigma}$)	0.466 ± 0.052	3983.400 ± 696.295	4156.000	0.014 ± 0.001
			$\mathcal{O}$-Ghoshal	0.261 ± 0.009	2912.400 ± 310.722	2836.000	0.089 ± 0.025
ER1	400	False	$\mathcal{O}$-ICID	0.954 ± 0.022	53.500 ± 24.437	52.500	447.128 ± 80.374
			Chol($\hat{\sigma}$)	0.958 ± 0.009	60.900 ± 17.091	57.500	0.021 ± 0.001
			$\mathcal{O}$-Ghoshal	1.000 ± 0.000	0.000 ± 0.000	0.000	0.739 ± 0.051
		True	$\mathcal{O}$-ICID	0.753 ± 0.035	498.400 ± 21.083	500.000	133.187 ± 0.130
			Chol($\hat{\sigma}$)	0.465 ± 0.019	1195.900 ± 136.425	1235.500	0.021 ± 0.001
			$\mathcal{O}$-Ghoshal	0.392 ± 0.027	822.500 ± 90.792	788.000	0.718 ± 0.038
ER2	400	False	$\mathcal{O}$-ICID	0.835 ± 0.021	409.200 ± 65.207	406.500	386.745 ± 82.128
			Chol($\hat{\sigma}$)	0.963 ± 0.008	195.500 ± 41.032	191.000	0.037 ± 0.003
			$\mathcal{O}$-Ghoshal	1.000 ± 0.000	0.000 ± 0.000	0.000	0.721 ± 0.054
		True	$\mathcal{O}$-ICID	0.567 ± 0.026	1340.900 ± 45.642	1351.500	129.332 ± 4.140
			Chol($\hat{\sigma}$)	0.475 ± 0.015	9263.700 ± 831.401	8970.000	0.038 ± 0.003
			$\mathcal{O}$-Ghoshal	0.271 ± 0.015	6474.300 ± 439.669	6448.000	0.712 ± 0.044



(a) SF2, $d \in \{100, 400\}$ (b) SF2, $d \in \{100, 400\}$

Figure 7: The var-sorting bias in structural causal discovery: comparative SHD performances on non-standardized vs standardized Θ of Cholesky, $\mathcal{O}$-ICID and $\mathcal{O}$-Ghoshal for $d = 100$ (top) and $d = 400$ (bottom), on SF2 (left) and SF2 (right).

Table 4: Results of causal discovery from inverse covariance matrices with and without standardization. Graph type: SF.

G	d	St'd	Algorithm	TPR	SHD	Median SHD	time (sec)
SF2	50	False	$\mathcal{O}$-ICID	0.794 ± 0.107	84.667 ± 46.276	98.000	30.756 ± 9.536
			Chol($\hat{\sigma}$)	0.991 ± 0.010	10.100 ± 11.561	6.000	0.015 ± 0.006
			$\mathcal{O}$-Ghoshal	1.000 ± 0.000	0.000 ± 0.000	0.000	0.008 ± 0.002
SF2	50	True	$\mathcal{O}$-ICID	0.638 ± 0.049	355.222 ± 71.372	359.000	25.134 ± 11.208
			Chol($\hat{\sigma}$)	0.533 ± 0.162	511.200 ± 172.535	502.500	0.013 ± 0.004
			$\mathcal{O}$-Ghoshal	0.105 ± 0.033	641.300 ± 31.812	635.500	0.005 ± 0.004
SF2	100	False	$\mathcal{O}$-ICID	0.762 ± 0.098	172.556 ± 69.712	198.000	93.124 ± 40.584
			Chol($\hat{\sigma}$)	0.989 ± 0.005	19.400 ± 19.265	11.500	0.040 ± 0.008
			$\mathcal{O}$-Ghoshal	1.000 ± 0.000	0.000 ± 0.000	0.000	0.025 ± 0.004
SF2	100	True	$\mathcal{O}$-ICID	0.585 ± 0.067	1071.889 ± 176.337	1048.000	86.412 ± 40.048
			Chol($\hat{\sigma}$)	0.403 ± 0.116	1679.500 ± 267.260	1686.500	0.042 ± 0.009
			$\mathcal{O}$-Ghoshal	0.087 ± 0.020	1767.500 ± 72.809	1759.500	0.025 ± 0.003
SF2	200	False	$\mathcal{O}$-ICID	0.768 ± 0.099	360.556 ± 97.381	343.000	359.662 ± 158.652
			Chol($\hat{\sigma}$)	0.988 ± 0.007	61.100 ± 53.530	48.500	0.129 ± 0.022
			$\mathcal{O}$-Ghoshal	1.000 ± 0.000	0.000 ± 0.000	0.000	0.165 ± 0.007
SF2	200	True	$\mathcal{O}$-ICID	0.529 ± 0.053	2809.111 ± 459.324	2890.000	324.586 ± 171.684
			Chol($\hat{\sigma}$)	0.412 ± 0.068	4168.400 ± 492.283	4237.000	0.124 ± 0.022
			$\mathcal{O}$-Ghoshal	0.097 ± 0.020	4386.800 ± 251.035	4395.500	0.162 ± 0.006
SF2	400	False	$\mathcal{O}$-ICID	0.634 ± 0.071	1020.667 ± 155.093	1045.000	1450.952 ± 607.777
			Chol($\hat{\sigma}$)	0.989 ± 0.002	121.900 ± 72.209	110.000	0.350 ± 0.070
			$\mathcal{O}$-Ghoshal	1.000 ± 0.000	0.000 ± 0.000	0.000	0.877 ± 0.016
SF2	400	True	$\mathcal{O}$-ICID	0.622 ± 0.217	7586.333 ± 1921.822	6897.000	1342.516 ± 635.851
			Chol($\hat{\sigma}$)	0.423 ± 0.057	10264.800 ± 1238.370	10344.500	0.346 ± 0.053
			$\mathcal{O}$-Ghoshal	0.086 ± 0.017	10774.600 ± 642.635	10475.500	0.860 ± 0.015

C.5 Convergence properties of $\mathcal{O}$-ICID

Table 5: Results of $\mathcal{O}$-ICID: each row shows the result in a random test. The true random graph of $d = 50$ nodes are drawn from ERk for $k(\text{deg}) = 2$.

deg	Convergence indicators			Scores					
	$\|P_{S_{\widehat{\Theta}}^c}(R)\|$	$\gamma(\tilde{B})$	(Infeasi., P-Opt, Conv)	**TPR**	**FDR**	**SHD**	nnz	$\|B - B^\star\|_{\text{F}}$	time (sec.)
2.0			(, , 0)	1.000	0.774	221	442		2.57
2.0	2.92e-2	0.490	(5.96e-2, 2.08e-1, 0)	1.000	0.735	193	378	6.60e+0	6.64
2.0	6.03e-2	0.460	(1.32e-1, 1.47e-1, 0)	0.980	0.716	174	345	1.03e+1	6.38
2.0	3.24e-2	0.490	(6.56e-2, 3.17e-1, 0)	1.000	0.714	178	350	7.35e+0	5.43
2.0	5.53e-2	0.570	(9.66e-2, 3.12e+0, 0)	0.990	0.748	198	393	8.71e+0	5.99
2.0	6.06e-2	0.560	(1.08e-1, 1.71e+0, 0)	1.000	0.774	228	443	6.71e+0	5.58
2.0	1.20e-7	0.040	(3.11e-6, 9.64e-5, 1)	1.000	0.038	4	104	2.25e-6	3.13
2.0	0.00e+0	0.000	(9.97e-7, 4.62e-5, 2)	1.000	0.000	0	100	6.34e-7	3.46
2.0	2.74e-9	0.020	(1.29e-7, 9.84e-5, 2)	1.000	0.010	1	101	1.46e-7	3.08
2.0	0.00e+0	0.000	(3.94e-9, 9.84e-5, 2)	1.000	0.000	0	100	3.21e-9	2.34

Table 6: Results of $\mathcal{O}$-ICID: each row shows the result in a random test. The true random graph of $d = 50$ nodes are drawn from ERk for $k(\text{deg}) \in \{0.5, 1\}$.

deg	Convergence indicators			Scores					
	$\|P_{S_{\widehat{\Theta}}^c}(R)\|$	$\gamma(\tilde{B})$	(Infeasi., P-Opt, Conv)	**TPR**	**FDR**	**SHD**	nnz	$\|B - B^\star\|_{\text{F}}$	time (sec.)
0.5	0.00e+0	0.000	(3.67e-1, 7.49e-5, 0)	1.000	0.000	0	25	2.55e-1	0.45
0.5	1.14e-2	0.020	(6.08e-1, 7.13e-5, 0)	1.000	0.074	2	27	4.00e-1	1.07
0.5	0.00e+0	0.000	(2.00e-5, 9.78e-5, 0)	1.000	0.000	0	25	1.42e-5	0.84
0.5	0.00e+0	0.000	(2.35e-5, 9.38e-5, 0)	1.000	0.000	0	25	1.66e-5	0.76
0.5	0.00e+0	0.000	(2.54e-5, 9.89e-5, 0)	1.000	0.000	0	25	1.73e-5	0.50
0.5	0.00e+0	0.000	(6.71e-3, 8.38e-5, 1)	1.000	0.000	0	25	4.68e-3	1.57
0.5	0.00e+0	0.000	(6.26e-3, 3.33e-5, 1)	1.000	0.000	0	25	4.22e-3	1.53
0.5	0.00e+0	0.000	(4.84e-5, 9.57e-5, 0)	1.000	0.000	0	25	3.39e-5	0.68
0.5	0.00e+0	0.000	(2.01e-5, 9.84e-5, 0)	1.000	0.000	0	25	1.42e-5	0.97
0.5	0.00e+0	0.000	(6.20e-6, 9.20e-5, 1)	1.000	0.000	0	25	4.38e-6	0.79
1.0	4.54e-2	0.250	(1.80e-1, 8.11e-1, 0)	1.000	0.537	40	108	4.92e+0	1.47
1.0	0.00e+0	0.000	(4.20e-6, 9.78e-5, 1)	1.000	0.000	0	50	3.26e-6	0.85
1.0	0.00e+0	0.000	(6.83e-9, 9.45e-5, 2)	1.000	0.000	0	50	3.76e-9	2.85
1.0	1.34e-8	0.050	(2.64e-7, 9.63e-5, 2)	1.000	0.074	4	54	1.42e-7	2.31
1.0	0.00e+0	0.000	(3.75e-7, 9.43e-5, 2)	1.000	0.000	0	50	3.35e-7	0.94
1.0	0.00e+0	0.000	(8.96e-7, 9.75e-5, 2)	1.000	0.000	0	50	1.22e-6	0.91
1.0	3.86e-8	0.350	(1.10e-7, 9.64e-5, 2)	1.000	0.020	1	51	7.98e-8	1.30
1.0	0.00e+0	0.000	(2.97e-8, 9.86e-5, 2)	1.000	0.000	0	50	2.22e-8	0.71
1.0	0.00e+0	0.000	(4.08e-6, 9.72e-5, 1)	1.000	0.000	0	50	3.00e-6	0.65
1.0	0.00e+0	0.000	(1.68e-7, 8.65e-5, 2)	1.000	0.000	0	50	1.38e-7	2.50

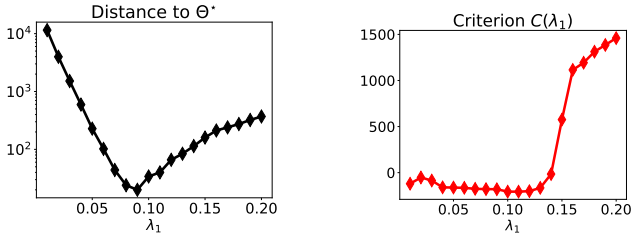
C.6 Selection of λ_1 for Algorithm 4

In the experiment for Figure 4, a parameter selection is needed. We use grid search for selecting values of λ_1 for the empirical inverse covariance estimator (Algorithm 4).

Note that the total time for selecting the value of λ_1 using Algorithm 4 is counted as the computation time of ICID in the benchmark of Figure 4.

We start by estimating the grid search area of λ_1, based on observational data on ER graphs with $200 \leq d \leq 2.10^3$ nodes. The same methodology applies to SF graphs.

Given that most desired causal structures have an average degree $1 \leq \deg \leq 4$, the target sparsity of $\widehat{\Theta}_{\lambda_1}$ by Algorithm 4 is bounded by $\bar{\rho}_{\deg} = \max(\frac{\deg}{d}) \approx 2.0\%$ for graphs with $d \geq 200$ nodes. This gives us an approximate target percentile of around 98%, i.e., top 2% edges in terms of absolute weight of $\widehat{\Theta}_{\text{off}}$. In other words, the maximal value $\lambda_1^{\max}$ of the grid search area is set as $\lambda_1^{\max} := \frac{|\widehat{\Theta}_{\text{off}}(\tau_{98})|}{\|\widehat{\Theta}_{\text{off}}\|_{\max}}$, where τ_{98} refers to the index of the 98-th percentile in $\{|\widehat{\Theta}_{\text{off}}|\}$. For the experiments with ER2 graphs in Section 4, the estimated $\lambda_1^{\max}$ is 6.10^{-1}. Hence, the search grid of λ_1 is set up as $n_{I_1} = 20$ equidistant values on $I_1 = [10^{-2}, 6.10^{-1}]$.



(a) Distance($\widehat{\Theta}_{\lambda_1}, \Theta^\star$) (b) Criterion used

Figure 8: Grid search of λ_1 with Algorithm 4 based on criterion $C(\lambda_1)$ (26). Data X is from linear SEM with Gaussian noise, on ER2 graph with $d = 200$ nodes.

The selection criterion, similar to GraphicalLasso, is defined as

$$C(\lambda_1) := \operatorname{tr}(\widehat{C}\widehat{\Theta}_{\lambda_1}) - \log\det(\tilde{\Theta}_{\lambda_1}), \tag{26}$$

where $\tilde{\Theta}_{\lambda_1} = \widehat{\Theta}_{\lambda_1} + \frac{9}{10}\operatorname{diag}(\widehat{\Theta}_{\lambda_1})$ is used in the $\log\det$-evaluation for an enhanced positive definiteness in all cases.

Figure 8 shows the criterion values compared to the Hamming distances with the oracle precision matrix $\Theta^\star := \phi(B^\star)$. We observe that the selection criterion with $\arg\min_{I_1} C(\lambda_1)$ gives an answer that is rather close to the optimal value in terms of distance of $\widehat{\Theta}_{\lambda_1}$ to the oracle precision matrix $\Theta^\star$.

References

[AAZ19] Bryon Aragam, Arash Amini, and Qing Zhou. Globally optimal score-based learning of directed acyclic graphs in high-dimensions. *Advances in Neural Information Processing Systems*, 32, 2019.

[ABGLP19] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[BA99] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[Ber99] D. P. Bertsekas. *Nonlinear programming*, volume 2nd Editio. 1999. URL: http://www.citeulike.org/group/4340/article/1859441.

[BT09] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[CGE21] Chris Cundy, Aditya Grover, and Stefano Ermon. BCD nets: Scalable variational approaches for bayesian causal discovery. In *Advances in Neural Information Processing Systems*, 2021. URL: https://openreview.net/forum?id=gbtDcLzwKUb.

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

[Chi02] David Maxwell Chickering. Learning equivalence classes of bayesian-network structures. *The Journal of Machine Learning Research*, 2:445–498, 2002.

[CT07] Emmanuel Candes and Terence Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351, 2007.

[DS22] Shuyu Dong and Michèle Sebag. From graphs to DAGs: a low-complexity model and a scalable algorithm, 2022. URL: https://arxiv.org/abs/2204.04644.

[FG65] Delbert Fulkerson and Oliver Gross. Incidence matrices and interval graphs. *Pacific journal of mathematics*, 15(3):835–855, 1965.

[FHT07] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. 2007. URL: http://statweb.stanford.edu/{~}tibs/ftp/graph.pdf.

[GH18] Asish Ghoshal and Jean Honorio. Learning linear structural equation models in polynomial time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pages 1466–1475. PMLR, 2018.

[GSB19] Isabelle Guyon, Alexander R. Statnikov, and Berna Bakir Batu, editors. *Cause Effect Pairs in Machine Learning*. Springer, 2019. URL: https://doi.org/10.1007/978-3-030-21810-2, doi:10.1007/978-3-030-21810-2.

[GYKZ20] AmirEmad Ghassami, Alan Yang, Negar Kiyavash, and Kun Zhang. Characterizing distribution equivalence and structure learning for cyclic and acyclic directed graphs. In *International Conference on Machine Learning*, pages 3494–3504. PMLR, 2020.

[KF09] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[LB14] Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research*, 15(1):3065–3105, 2014.

[Mee95] Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, page 403–410. Morgan Kaufmann Publishers Inc., 1995.

[NGZ20] Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and DAG constraints for learning linear DAGs. *Advances in Neural Information Processing Systems*, 33:17943–17954, 2020.

[NZZZ21] Ignavier Ng, Yujia Zheng, Jiji Zhang, and Kun Zhang. Reliable causal discovery with improved exact search and weaker assumptions. *Advances in Neural Information Processing Systems*, 34:20308–20320, 2021.

[PBM16] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016.

[Pea00] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

[PJS17] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[PPS89] Vern I Paulsen, Stephen C Power, and Roger R Smith. Schur products and matrix completions. *Journal of functional analysis*, 85(1):151–178, 1989.

[Ros70] Donald J Rose. Triangulated graphs and the elimination process. *Journal of Mathematical Analysis and Applications*, 32(3):597–609, 1970.

[RSW21] Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated DAG! Causal discovery benchmarks may be easy to

game. *Advances in Neural Information Processing Systems*, 34:27772–27784, 2021.

[RU18] Garvesh Raskutti and Caroline Uhler. Learning directed acyclic graph models based on sparsest permutations. *Stat*, 7(1):e183, 2018.

[SAU20] Chandler Squires, Joshua Amaniampong, and Caroline Uhler. Efficient permutation discovery in causal DAGs. *arXiv preprint arXiv:2011.03610*, 2020.

[SG21] Axel Sauer and Andreas Geiger. Counterfactual generative networks. In *International Conference on Learning Representations (ICLR)*, 2021.

[SGSH00] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[SIS+11] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *The Journal of Machine Learning Research*, 12:1225–1248, 2011.

[SPP+05] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

[SWU21] Liam Solus, Yuhao Wang, and Caroline Uhler. Consistency guarantees for greedy permutation-based causal inference algorithms. *Biometrika*, 108(4):795–814, 2021.

[TKL+21] Stratis Tsirtsis, Amir-Hossein Karimi, Ana Lucic, Manuel Gomez-Rodriguez, Isabel Valera, and Hima Lakkaraju. ICML workshop on algorithmic recourse. 2021.

[VA15] Lieven Vandenberghe and Martin S. Andersen. Chordal graphs and semidefinite optimization. *Foundations and Trends® in Optimization*, 1(4):241–433, 2015. URL: `http://dx.doi.org/10.1561/2400000006`, `doi:10.1561/2400000006`.

[ZARX18] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL: `https://proceedings.neurips.cc/paper/2018/file/e347c51419ffb23ca3fd5050202f9c3d-Paper.pdf`.