Biometrika (2024), **xx**, x, *pp*. 1–23
Printed in Great Britain

Integer Programming for Learning Directed Acyclic Graphs from Non-identifiable Gaussian Models

BY T. XU*
Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60208, U.S.A.
tongxu2027@u.northwestern.edu

A. TAEB*
Department of Statistics, University of Washington, Box 354322, Seattle, Washington 98195, U.S.A.
ataeb@uw.edu

S. KÜÇÜKYAVUZ
Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60208, U.S.A.
simge@northwestern.edu

AND A. SHOJAIE
Department of Biostatistics, University of Washington, Box 351617 Seattle, Washington 98195, U.S.A.
ashojaie@uw.edu

SUMMARY

We study the problem of learning directed acyclic graphs from continuous observational data, generated according to a linear Gaussian structural equation model. State-of-the-art structure learning methods for this setting have at least one of the following shortcomings: i) they cannot provide optimality guarantees and can suffer from learning sub-optimal models; ii) they rely on the stringent assumption that the noise is homoscedastic, and hence the underlying model is fully identifiable. We overcome these shortcomings and develop a computationally efficient mixed-integer programming framework for learning medium-sized problems that accounts for arbitrary heteroscedastic noise. We present an early stopping criterion under which we can terminate the branch-and-bound procedure to achieve an asymptotically optimal solution and establish the consistency of this approximate solution. In addition, we show via numerical experiments that our method outperforms three state-of-the-art algorithms and is robust to noise heteroscedasticity, whereas the performance of the competing methods deteriorates under strong violations of the identifiability assumption. The software implementation of our method is available as the Python package *micodag*.

Some key words: Directed acyclic graphs; Identifiability; Structural equation models; Mixed-integer programming.

© 2021 Biometrika Trust

1. INTRODUCTION

1.1. *Background and Related Works*

A Bayesian network is a probabilistic graphical model that represents conditional independencies among a set of random variables. A directed acyclic graph serves as the structure to encode these conditional independencies, where the random variables are represented as vertices (or nodes), and a pair of vertices that are not connected in a path are conditionally independent. A directed edge from node i to node j indicates that i causes j. The acyclic property of the graph prevents the occurrence of circular dependencies and allows for probabilistic inference and learning algorithms for Bayesian networks. At a high level, the focus of this paper is developing methods to efficiently learn provably optimal directed acyclic graphs from continuous observational data. Throughout, we assume that all relevant variables are observed.

There are various methods for learning directed acyclic graphs, with the majority falling into one of two categories: constraint-based methods and score-based methods. Constraint-based methods aim to identify conditional independencies from the data. An example is the PC algorithm (Spirtes et al., 1993) which initiates with a complete undirected graph and iteratively removes edges based on conditional independence assessments. Kalisch & Bühlmann (2007) shows that under a condition known as 'strong faithfulness', the PC algorithm is consistent in high-dimensional settings for learning sparse Gaussian directed acyclic graphs. Score-based methods, which is the approach considered in this paper, often employ penalized log-likelihood as a score function to seek the optimal graph within the entire space of directed acyclic graphs (van de Geer & Bühlmann, 2013). Penalized maximum likelihood estimation methods do not require the strong faithfulness assumption, which is known to be restrictive in high dimensions (Uhler et al., 2012). However, exactly solving the corresponding problem suffers from high computational complexity. As an example, learning an optimal graph using dynamic programming takes about 10 hours for a medium-size problem with 29 nodes (Silander & Myllymäki, 2006).

For faster computation with large graphs, one can resort to greedy search methods in a score-based setting. An example is the Greedy Equivalence Search algorithm (Chickering, 2002), which performs a greedy search on the space of completed partially directed acyclic graphs rather than directed acyclic graphs. Chickering (2002) established the asymptotic consistency of the resulting estimate. Furthermore, under a strong assumption that the graph has a fixed degree, Greedy Equivalence Search and its variants have polynomial-time complexity in the number of variables (Chickering, 2020). Despite their favorable properties, for any finite sample size, these algorithms do not generally recover the optimal scoring graph. Consequently, in moderate sample size regimes, the sub-optimal model that they learn can be rather different than the true model.

Several computationally efficient approaches rely on finding a topological ordering of the random variables under homoscedastic noise (also known as the equal-variance condition). When the nodes exhibit a natural ordering, the problem of estimating directed graphs reduces to the problem of estimating the network structure which can be efficiently solved using lasso-like methods (Shojaie & Michailidis, 2010). Assuming equal variances, Chen et al. (2019) proposed a top-down approach to obtain an ordering among conditional variances. Similarly, Ghoshal & Honorio (2018) introduced a bottom-up method that selects sinks by identifying the smallest diagonal element in conditional precision matrices obtained using the estimator of Cai et al. (2011), resulting in a topological ordering. These approaches are efficient and thus readily extendable to high-dimensional problems. They also provide better estimates than greedy methods, such as Greedy Equivalence Search. Consequently, they are currently two of the state-of-the-art approaches for causal structure learning. While being computationally efficient, these methods

heavily rely on the equal-variance assumption. With non-equal variances, the topological ordering of variances no longer exists, and their performances deteriorate (Küçükyavuz et al., 2023).

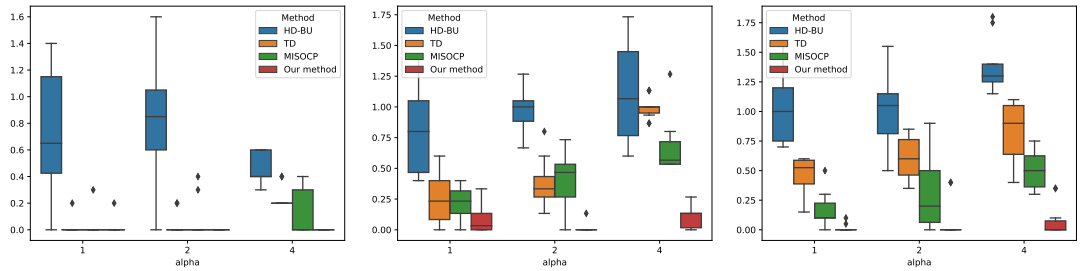
Integer and mixed integer programming formulations provide computationally efficient and rigorous optimization techniques for learning directed acyclic graphs. Typically, these methods consider a penalized maximum likelihood estimator over the space of directed acyclic graphs, and cast the acyclicity condition as a constraint over a collection of binary variables. Such a formulation enables the use of branch-and-bound algorithms for fast and accurate learning of graphs, both with discrete (Bartlett & Cussens, 2017; Cussens et al., 2017a,b) and continuous (Küçükyavuz et al., 2023) data. More specifically, in the context of continuous Gaussian data, Küçükyavuz et al. (2023) impose homoscedastic noise assumption to develop their procedure. The homoscedastic noise assumption results in both computational and statistical simplifications. From the computational perspective, the penalized maximum likelihood estimator can be expressed as a mixed integer programming formulation, with a convex quadratic loss function and a regularization penalty subject to linear constraints. From the statistical perspective, the homoscedastic noise assumption implies full identifiability of the underlying directed acyclic graph (Peters & Bühlmann, 2013) and avoids the typical challenge of non-identifiability (or identifiability of what is known as completed partially directed acyclic graph) from observational data. To the best of our knowledge, no integer programming formulation exists for continuous data without the stringent assumption of homoscedastic noise.

1.2. *Our Contributions*

We propose a mixed-integer programming formulation for general Gaussian structural equation models. Our estimator, which consists of a negative log-likelihood function that is a sum of a convex logarithmic term and a convex quadratic term, allows for heteroscedastic noise and reduces to the approach of Küçükyavuz et al. (2023) if the noise variances are constrained to be identical. Using a layered network formulation (Manzour et al., 2021), as well as tailored cutting plane methods and branch-and-bound techniques, we provide a computationally efficient approach to obtain a solution that is guaranteed to be accurate up to a pre-specified optimality gap. By connecting the optimality gap of our mixed-integer program to the statistical properties of the estimator, we establish an early stopping criterion under which we can terminate the branch-and-bound procedure and attain a solution that recovers the Markov equivalence class of the underlying directed acyclic graph. Compared to state-of-the-art benchmarks in non-identifiable instances, we demonstrate empirically that our method generally exhibits superior estimation performance, and displays greater robustness to variations in the level of non-identifiability compared to methods relying on the identifiability assumption. The improved performance is highlighted in the synthetic example of Figure 1, where our method is compared with the mixed-integer second-order conic program of Küçükyavuz et al. (2023), the high-dimensional bottom-up approach of Ghoshal & Honorio (2018), and the top-down approach of Chen et al. (2019); we omit Greedy Equivalence Search Chickering (2002) as it performed worse than all the other methods. Here, we randomly select the noise variances from the interval $[4-\alpha, 4+\alpha]$, where larger α leads to a greater amount of heteroscedasticity. Each method is implemented with a sample size of $n = 100$ over ten independent trials. The results show that in contrast to our method, the performance of competing approaches deteriorates under strong violations of the homoscedasticity assumption; see §6.5 for more details.

1.3. *Notations and Definitions*

A directed acyclic graph $\mathcal{G} = (V, E)$ among m nodes consists of vertices $V = \{1, 2, \ldots, m\}$ and directed edge set $E \subseteq V \times V$, where there are no directed cycles. We denote $\mathrm{MEC}(\mathcal{G})$ to



(a) $m = 10$ (b) $m = 15$ (c) $m = 20$

Fig. 1: Box plots of scaled d_{cpdag} for our methods and benchmarks with $\alpha = 1, 2, 4$ and number of nodes $m = 10, 15, 20$, respectively. Here, d_{cpdag} are scaled by the total number of edges in the true underlying directed acyclic graph. TD, top-down; HD-BU, high-dimensional bottom-up; MISOCP, mixed-integer second-order conic program; d_{cpdag}, differences between true and estimated completed partially directed acyclic graphs with smaller values being better; see §6.

be the Markov equivalence class of $\mathcal{G}$, consisting of directed acyclic graphs that have the same skeleton and same v-structures as $\mathcal{G}$. The skeleton of $\mathcal{G}$ is the undirected graph obtained from $\mathcal{G}$ by substituting directed edges with undirected ones. Furthermore, nodes i, j, and k form a v-structure if $(i, k) \in E$ and $(j, k) \in E$, and there is no edge between i, j. The Markov equivalence class can be compactly represented by a completed directed acyclic graph, which is a graph consisting of both directed and undirected edges. A completed directed acyclic graph has a directed edge from a node i to a node j if and only if this directed edge is present in every directed acyclic graph in the associated Markov equivalence class. A completed directed acyclic graph has an undirected edge between nodes i and j if the corresponding Markov equivalence class contains directed acyclic graphs with both directed edges from i to j and from j to i. For a matrix $B \in \mathbb{R}^{m\times m}$, we denote $\mathcal{G}(B)$ to be the directed graph on m nodes such that the directed edge from i to j appears in $\mathcal{G}(B)$ if and only if $B_{ij} \neq 0$. We denote the identity matrix by I with the size being clear from context. The collection of positive-definite diagonal matrices is denoted by $\mathbb{D}^m_{++}$. We use the Bachman-Landau symbols $\mathcal{O}$ to describe the limiting behavior of a function. Furthermore, we denote $z \asymp 1$ to express $z = \mathcal{O}(1)$ and $1/z = \mathcal{O}(1)$.

2. PROBLEM SETUP

2.1. *Modeling Framework*

Consider an unknown directed acyclic graph whose m nodes correspond to observed random variables $X \in \mathbb{R}^m$. We denote the directed acyclic graph by $\mathcal{G}^\star = (V, E)$ where $V = \{1, \ldots, m\}$ is the vertex set and $E \subseteq V \times V$ is the directed edge set.

We assume the random variables X satisfy a linear structural equation model:

$$X = B^{\star\mathrm{T}} X + \epsilon, \tag{1}$$

Here, the connectivity matrix $B^\star \in \mathbb{R}^{m\times m}$ is a matrix with zeros on the diagonal and $B^\star_{jk} \neq 0$ if $(j, k) \in E$. Further, ϵ is a mean-zero random vector with independent coordinates where the variances of individual coordinates can, in general, be different. Without loss of generality, we assume that all random variables are centered. Thus, each variable X_j in this model can be expressed as the linear combination of its parents—that is, the set of nodes with di-

rected edges pointing to j—plus independent centered noise. Our objective is to estimate the matrix $B^\star$, or an equivalence class if the underlying model is not identifiable, as its sparsity pattern encodes the structure of the graph $\mathcal{G}^\star$. To arrive at our procedure, we model the random variable $\epsilon \sim \mathcal{N}(0, \Omega^\star)$ to be Gaussian where noise variance matrices $\Omega^\star$ are positive definite and diagonal. Modeling the data to be Gaussian has multiple important implications. First, the structural equation model (1) is parameterized by the connectivity matrix $B^\star$ and the noise variances $\Omega^\star$. Second, the random variables X are distributed according to $\mathcal{P}^\star = \mathcal{N}(0, \Sigma^\star)$ where $\Sigma^\star = (I - B^\star)^{-\mathrm{T}}\Omega^\star(I - B^\star)^{-1}$. Throughout, we assume that the distribution $\mathcal{P}^\star$ is non-degenerate, or equivalently, $\Sigma^\star$ is positive definite.

In the effort to estimate the parameters $(B^\star, \Omega^\star)$, one faces a challenge with identifiability: there may be multiple structural equation models that are compatible with $\mathcal{P}^\star$. Specifically, consider any connectivity matrix B such that the associated graph $\mathcal{G}(B)$ is a directed acyclic graph, and noise variance matrix $\Omega \in \mathbb{D}^m_{++}$. Then, the structural equation model entailed by (B, Ω), yields an equally representative model as the one entailed by the population parameters $(B^\star, \Omega^\star)$. Thus, the following question naturally arises: what is the set of equivalent directed acyclic graphs? The answer is that under an assumption called *faithfulness*, the sparsest directed acyclic graphs that are compatible with $\mathcal{P}^\star$ are precisely $\mathrm{MEC}(\mathcal{G}^\star)$, the Markov equivalence class of $\mathcal{G}^\star$ (van de Geer & Bühlmann, 2013). Since parsimonious models are generally more desirable, our objective is to estimate $\mathrm{MEC}(\mathcal{G}^\star)$ from data. In the next sections, we describe a maximum likelihood estimator, and subsequently an equivalent mixed-integer programming framework, and present its optimality and statistical guarantees.

2.2. *An Intractable Maximum-Likelihood Estimator*

We assume we have n samples of X which are independent and identically distributed. We denote $\hat{\Sigma}_n$ to be the sample covariance matrix of the data. Consider a Gaussian structural equation model parameterized by connectivity matrix B and noise variance Ω with $D = \Omega^{-1}$. The parameters (B, D) specify the following precision, or inverse covariance, matrix:

$$\Theta := \Theta(B, D) := (I - B)D(I - B)^{\mathrm{T}}.$$

The negative log-likelihood of this structural equation model is proportional to

$$\ell_n(\Theta) = \mathrm{trace}(\Theta\hat{\Sigma}_n) - \log\det(\Theta).$$

Naturally, we seek a model that not only has a small negative log-likelihood but is also specified by a sparse connectivity matrix, containing few nonzero elements. Thus, we deploy the following ℓ_0-penalized maximum likelihood estimator with regularization parameter $\lambda \geq 0$:

$$\min_{B \in \mathbb{R}^{m\times m}, D \in \mathbb{D}^m_{++}} \quad \ell_n((I - B)D(I - B)^{\mathrm{T}}) + \lambda^2\|B\|_{\ell_0} \tag{2a}$$

$$\text{s.t.} \quad B \in \mathcal{B}, \tag{2b}$$

where $\mathcal{B} = \{B \in \mathbb{R}^{m\times m} \mid \mathcal{G}(B) \text{ is a directed acyclic graph}\}$ and $\|B\|_{\ell_0}$ denotes the number of non-zeros in the connectivity matrix B. A few remarks are in order. First, the estimator (2) is equivalent to one proposed and analyzed in van de Geer & Bühlmann (2013). Second, when the diagonal term D is restricted to be a multiple of identity, the model will have homoscedastic noise. As desired, the resulting estimator reduces to the one considered in Küçükyavuz et al. (2023). Finally, ℓ_0 regularization is generally preferred over ℓ_1 regularization in the objective of (2). In particular, ℓ_0 regularization preserves the important property that equivalent directed acyclic graphs—those in the same Markov equivalence class—have the same penalized likelihood score, while this is not the case for ℓ_1 regularization (van de Geer & Bühlmann, 2013).

van de Geer & Bühlmann (2013) demonstrated that the solution of (2) has desirable statistical properties; however, solving it is in general intractable. As stated, the likelihood function $\ell_n(\cdot)$ is a non-linear and non-convex function of the parameters (B, D), and is thus not amenable to standard mixed-integer programming optimization techniques. In the following section, after a change of variables, we obtain a reformulation with a convex negative log-likelihood that enables a mixed-integer programming framework for our problem.

3. OUR CONVEX MIXED-INTEGER PROGRAMMING FRAMEWORK

3.1. *An Equivalent Reformulation*

Consider the following optimization problem:

$$\min_{\Gamma \in \mathbb{R}^{m \times m}} \sum_{i=1}^{m} -2\log(\Gamma_{ii}) + \operatorname{tr}(\Gamma\Gamma^{\mathrm{T}}\hat{\Sigma}_n) + \lambda^2 \|\Gamma - \operatorname{diag}(\Gamma)\|_{\ell_0}, \tag{3a}$$

$$\text{s.t. } \Gamma - \operatorname{diag}(\Gamma) \in \mathcal{B}, \tag{3b}$$

where $\operatorname{diag}(\Gamma)$ is a diagonal $m \times m$ matrix consisting of the diagonal entries of Γ. The following proposition establishes the equivalence between (2) and (3).

PROPOSITION 1. *Formulations* (2) *and* (3) *have the same minimum objective value. Further:*

(a) for any minimizer $(\hat{B}^{\mathrm{opt}}, \hat{D}^{\mathrm{opt}})$ *of* (2)*, the parameter* $\hat{\Gamma}^{\mathrm{opt}} = (I - \hat{B}^{\mathrm{opt}})(\hat{D}^{\mathrm{opt}})^{1/2}$ *is a minimizer of* (3)*;*

(b) for any minimizer $\hat{\Gamma}^{\mathrm{opt}}$ *of* (3)*, let* $\hat{D}^{\mathrm{opt}} = \operatorname{diag}(\hat{\Gamma}^{\mathrm{opt}})^2$ *and* $\hat{B}^{\mathrm{opt}} = I - \hat{\Gamma}^{\mathrm{opt}}(\hat{D}^{\mathrm{opt}})^{-1/2}$*. Then, the parameter set* $(\hat{B}^{\mathrm{opt}}, \hat{D}^{\mathrm{opt}})$ *is a minimizer of* (2)*.*

The proof of Proposition 1 is in Appendix A.1. This result states that instead of (2), one can equivalently solve the reformulated optimization problem (3). Furthermore, due to the equivalence of the minimizers, the nonzero sparsity pattern in the off-diagonal of an optimal solution $\hat{\Gamma}^{\mathrm{opt}}$ of (3) encodes the same directed acyclic graph structure as a minimizer $\hat{B}^{\mathrm{opt}}$ of (2). Note that the formulation (3) replaces the non-convex negative log-likelihood term of (2) with the term $\sum_{i=1}^{m} -2\log(\Gamma_{ii}) + \operatorname{tr}(\Gamma\Gamma^{\mathrm{T}}\hat{\Sigma}_n)$, which is a convex function of Γ consisting of a quadratic term plus sum of univariate logarithm terms. These appealing properties of the estimator (3) enable an efficient mixed-integer programming framework, which we discuss next.

3.2. *A Convex Mixed-Integer Program*

Various integer programming formulations have been proposed to encode acyclicity constraints, including the cutting plane method (Grötschel et al., 1985; Wolsey & Nemhauser, 1999), topological ordering (Park & Klabjan, 2017), and layered network formulation (Manzour et al., 2021). Here, we adopt the layered network formulation as it has been shown to perform the best with continuous data; see Manzour et al. (2021) for more details.

Before we present our formulation, we first introduce some notation. We denote the edge set E_{super} as a *super-structure* of E (true edge set) with the property that $E \subseteq E_{\text{super}}$. Here, the super-structure E_{super} may be bi-directional and consist of cycles with no self-loops. When E_{super} is sparse, we can significantly reduce the search space. A good candidate for a super-structure is the moral graph corresponding to the true directed acyclic graph, which can be readily estimated from data using the *graphical lasso*; see Drton & Maathuis (2017) for a review.

Following previous work on the layered network formulation, to efficiently encode the acyclicity constraint, we add two sets of decision variables to the optimization model. The first is the

set of binary variables $\{g_{jk} \in \{0,1\} : (j,k) \in E_{\text{super}}\}$. These variables are used to represent the presence or absence of edges in the estimated directed acyclic graph. The second is the set of continuous variables $\{\psi_k \in [1,m] : k \in \{1,\dots,m\}\}$ representing the *layer value* of each node in the estimated graph with the property that an ancestor of a node in the graph has a higher layer value. Formally, the layered network formulation of Problem (3) is given by:

$$\min_{\substack{\Gamma\in\mathbb{R}^{m\times m},\psi\in[1,m]\\ g\in\{0,1\}^{|E_{\text{super}}|}}} \sum_{i=1}^{m} -2\log(\Gamma_{ii}) + \operatorname{tr}(\Gamma\Gamma^{\mathrm{T}}\hat{\Sigma}_n) + \lambda^2 \sum_{(j,k)\in E_{\text{super}}} g_{jk} \tag{4a}$$

$$\text{s.t. } -Mg_{jk} \le \Gamma_{jk} \le Mg_{jk} \quad ((j,k)\in E_{\text{super}}), \tag{4b}$$

$$M \ge \Gamma_{ii} \quad (i = 1,\dots,m), \tag{4c}$$

$$1 - m + mg_{jk} \le \psi_k - \psi_j \quad ((j,k)\in E_{\text{super}}). \tag{4d}$$

The constraints (4b) are so-called *big-M constraints* (Park & Klabjan, 2017; Manzour et al., 2021) that bound the magnitude of the entries of Γ by a large M to ensure that when $g_{jk} = 0$, $\Gamma_{jk} = 0$ and when $g_{jk} = 1$ this constraint is redundant; we will explore choices for M below. While constraints (4c) are redundant for diagonal entries for large enough M, they are beneficial for computational efficiency. Furthermore, since more degrees of freedom yield better likelihood fit, $g_{jk} = 1$ generally yields $\Gamma_{jk} \neq 0$, and thus the regularization term $\|\Gamma - \operatorname{diag}(\Gamma)\|_{\ell_0}$ in (3) is identical to $\sum_{(j,k)\in E_{\text{super}}} g_{jk}$ in (4). Finally, the constraint (4d), together with (4b), ensures that Γ encodes a directed acyclic graph. To see why, suppose there is a directed path in Γ from node j to node k and a directed path from k to j. The constraint (4d) then ensures that $\psi_k \ge \psi_j + 1$ and $\psi_j \ge \psi_k + 1$, resulting in a contradiction. We have thus shown that for M large enough, e.g., M equaling the maximum nonzero entry, in magnitude, of the optimal Γ in (3), the mixed-integer program (4) is equivalent to (3). One can use heuristic approaches (Park & Klabjan, 2017; Küçükyavuz et al., 2023) to select the value of M. Specifically, we solve the problem without any cycle prevention constraints and obtain solution $\hat{\Gamma}$, then let $M = 2\max_{(i,j)\in E_{\text{super}}} |\hat{\Gamma}_{ij}|$.

3.3. *Perspective Strengthening for a Tighter Formulation*

Using a concept known as perspective strengthening (Cui et al., 2013; Küçükyavuz et al., 2023; Wei et al., 2020, 2022, 2023), we can tighten the constraint set of (4) without changing the optimal objective value. Such a tighter formulation can speed up computations by providing a better lower bound in the branch-and-bound procedure.

Specifically, let $\delta \in \mathbb{R}^m_+$ be a non-negative vector with the property that $\hat{\Sigma}_n - D_\delta \succeq 0$, where $D_\delta = \operatorname{diag}(\delta_1, ..., \delta_m)$. By splitting the quadratic term $\Gamma\Gamma^{\mathrm{T}}\hat{\Sigma}_n = \Gamma\Gamma^{\mathrm{T}}(\hat{\Sigma}_n - D_\delta) + \Gamma\Gamma^{\mathrm{T}}D_\delta$ and the fact that $\operatorname{tr}(\Gamma\Gamma^{\mathrm{T}}D_\delta) = \sum_{j=1}^m \sum_{k=1}^m \delta_j \Gamma_{jk}^2$, we can express (4a) as $-2\sum_{i=1}^m \log(\Gamma_{ii}) + \operatorname{tr}(\Gamma\Gamma^{\mathrm{T}}Q) + \operatorname{tr}(\Gamma\Gamma^{\mathrm{T}}D_\delta) + \lambda^2 \sum_{(j,k)\in E_{\text{super}}} g_{jk}$, where $Q = \hat{\Sigma}_n - D_\delta$. We add a new set of non-negative continuous variables s_{jk} to represent Γ_{jk}^2, resulting in the following formulation:

$$\min_{\substack{\Gamma\in\mathbb{R}^{m\times m},\psi\in[1,m]\\ g\in\{0,1\}^{|E_{\text{super}}|}\\ s\in\mathbb{R}^{|E_{\text{super}}|}}} \sum_{i=1}^{m} -2\log(\Gamma_{ii}) + \operatorname{tr}(\Gamma\Gamma^{\mathrm{T}}Q) + \sum_{(j,k)\in E_{\text{super}}} \delta_j s_{jk} + \sum_{i=1}^{m} \delta_i s_{ii} + \lambda^2 \sum_{(j,k)\in E_{\text{super}}} g_{jk} \tag{5a}$$

$$\text{s.t. } -Mg_{jk} \le \Gamma_{jk} \le Mg_{jk} \quad ((j,k)\in E_{\text{super}}), \tag{5b}$$

$$M \ge \Gamma_{ii} \quad (i = 1,\dots,m), \tag{5c}$$

$$1 - m + mg_{jk} \le \psi_k - \psi_j \quad ((j,k)\in E_{\text{super}}), \tag{5d}$$

$$s_{jk} g_{jk} \geq \Gamma_{jk}^2 \quad ((j,k) \in E_{\text{super}}), \quad s_{ii} \geq \Gamma_{ii}^2 \quad (i \in 1, \ldots, m), \tag{5e}$$

$$s_{jk} \leq M^2 g_{jk} \quad ((j,k) \in E_{\text{super}}), \quad s_{ii} \leq M^2 \quad (i = 1, \ldots, m). \tag{5f}$$

To establish that the optimal objective values of (4) and (5) are identical, it suffices to show that the optimal set of variables $(\hat{s}^{\text{opt}}, \hat{\Gamma}^{\text{opt}})$ in (5) satisfy $\hat{s}_{jk}^{\text{opt}} = (\hat{\Gamma}_{jk}^{\text{opt}})^2$ for $(j,k) \in E_{\text{super}}$ and $\hat{s}_{ii}^{\text{opt}} = (\hat{\Gamma}_{ii}^{\text{opt}})^2$ for $i \in 1, \ldots, m$. Note that from the constraints (5b) and (5e), it follows that $s_{jk} \geq \Gamma_{jk}^2$ for $(j,k) \in E_{\text{super}}$. Since δ_j is non-negative in the objective (5a), we can get $\hat{s}_{jk}^{\text{opt}} = (\hat{\Gamma}_{jk}^{\text{opt}})^2$ for $(j,k) \in E_{\text{super}}$ and $\hat{s}_{ii}^{\text{opt}} = (\hat{\Gamma}_{ii}^{\text{opt}})^2$ for $i \in 1, \ldots, m$. The constraints $s_{jk} \leq M^2 g_{jk}$ and $s_{ii} \leq M^2$ are simply added to improve the computational efficiency.

The continuous relaxation of constraint set of (5)—replacing the integer constraints with $g \in [0,1]^{|E_{\text{super}}|}$—is contained in the continuous relaxation of the one from (4): for every set of feasible variables in (5), there exists a set of feasible variables in (4) yielding the same objective value. This fact, which follows from the analysis of Cui et al. (2013), leads to better lower bounds in the branch-and-bound process. Hence, throughout, we use the formulation (5).

In the above formulation, there is some flexibility in the choice of the vector δ. As larger values of δ lead to smaller continuous relaxation of the constraint set of (5) (Frangioni & Gentile, 2007), we choose δ by maximizing $\sum_{i=1}^m \delta_i$ subject to $\hat{\Sigma}_n - \text{diag}(\delta) \succeq 0$ for $\delta_i \geq 0, i = 1, \ldots, m$. This is a convex semi-definite program that can be solved efficiently.

4. PRACTICAL ASPECTS: OUTER APPROXIMATION AND EARLY STOPPING

4.1. *Outer Approximation to Handle the Logarithm*

Formulation (5) is a convex mixed-integer model that can be solved using the current optimization solvers, such as `Gurobi`. However, `Gurobi` handles the log terms in the objective by adding a new general constraint that uses a piecewise-linear approximation of the logarithmic function at predetermined breakpoints. Unless we use a large number of pieces in the piecewise-linear functions, which is computationally intensive, this approximation often leads to solutions that violate some constraints. To address this challenge, in this section, we describe an outer approximation method, based on ideas first proposed by Duran & Grossmann (1986).

Outer approximation is a cutting plane method that finds the optimal solution by constructing a sequence of piece-wise affine lower bounds for the logarithmic term in the objective function. We first replace $-2\log(\Gamma_{ii})$ with continuous variables $T_i \in \mathbb{R}, i = 1, \ldots, m$ so that the objective function (5a) becomes a quadratic function:

$$c(T, \Gamma) \coloneqq \sum_{i=1}^m T_i + \text{tr}(\Gamma\Gamma^{\text{T}} Q) + \sum_{(j,k) \in E_{\text{super}}} \delta_j s_{jk} + \sum_{i=1}^m \delta_i s_{ii} + \lambda^2 \sum_{(j,k) \in E_{\text{super}}} g_{jk}. \tag{6}$$

Without any constraint on T, the problem would be unbounded. Therefore, we iteratively add linear inequalities that are under-estimators of the logarithmic term to ensure that $T_i = -2\log(\Gamma_{ii})$ for $i = 1, \ldots, m$ at the optimal solution.

The linear approximation of function $f(x) = -2\log(x)$ at a point x_0 is $f(x_0) + \nabla f(x_0)(x - x_0) = -2\log(x_0) - 2(x - x_0)/x_0$. At iteration $t+1$, given the solution $\Gamma^{(t)}$, we add linear constraints $T_i \geq -2\log(\Gamma_{ii}^{(t)}) - 2(\Gamma_{ii} - \Gamma_{ii}^{(t)})/\Gamma_{ii}^{(t)}$ for $i = 1, \ldots, m$. As $f(x)$ is convex, these linear approximations are under-estimators and they cut off the current solution unless $T_i^{(t)}$ equals $-2\log(\Gamma_{ii}^{(t)})$ for $i = 1, \ldots, m$. The overall procedure is presented in Algorithm 1 and a simple illustrative example is given in Appendix A.4.

Algorithm 1. Pseudocode for outer approximation

Input: Sample covariance $\hat{\Sigma}_n$ and regularization $\lambda \in \mathbb{R}_+$
Output: $\hat{\Gamma}^{\text{opt}} \in \mathbb{R}^{m \times m}$
Initialize: $t \leftarrow 1$; $\Gamma^{(t)} \leftarrow$ starting point; $T^{(t)} \leftarrow -\infty$
while $T_j^{(t)} < -2\log(\Gamma_{jj}^{(t)})$ *for some* $j \in 1, \ldots, m$ **do**
  $\Gamma^{(t+1)}, T^{(t+1)} \leftarrow \arg\min_{T,\Gamma,g,\psi} c(T, \Gamma)$ s.t. $T_j \geq -2\log(\Gamma_{jj}^{(i)}) - \frac{2}{\Gamma_{jj}^{(i)}}(\Gamma_{jj} - \Gamma_{jj}^{(i)})$,
    $(i \in 1, \ldots, t; j \in 1, \ldots, m)$, (5b)-(5d)
  $t \leftarrow t + 1$
$\hat{\Gamma}^{\text{opt}} \leftarrow \Gamma^{(t)}$

4.2. *Early Stopping*

During the branch-and-bound algorithm, we maintain a lower bound and an upper bound on the objective value of the objective function (5a). Specifically, by relaxing the integer constraints to $g \in [0,1]^{|E_{\text{super}}|}$, one can solve the convex problem easily and the optimal objective value of the relaxed problem provides a lower bound. If any integer variable is fractional in this solution, we can split the problem into two sub-problems by considering smaller or larger integer values for that variable. This process, which is known as the *branch-and-bound* technique, creates a tree structure, where each node represents a problem and is connected to its potential sub-problems. If the relaxed solution of a node contains only integer values for all integer variables, then we have a feasible solution to the original problem, giving an upper bound of the objective, and the branching process terminates for that node. Throughout the optimization algorithm, we continuously update the upper and lower bounds. The optimality gap, GAP, of a solution is the difference between the upper and lower bounds of the objective at that solution and it should be 0 at an optimal solution. Alternatively, we can stop the algorithm early, before reaching the optimal solution, when the optimality gap reaches a specified threshold. Due to the intrinsic computational complexity of solving mixed-integer programming problems, early stopping is often used in practice but without statistical justification, because the solution may be suboptimal.

5. THEORETICAL RESULTS FOR EARLY STOPPING

In Section 4.2, we described how we can terminate the branch-and-bound algorithm to guarantee a desired optimality gap. In this section, we connect this optimality gap to the statistical properties of the terminated solution. This connection enables us to propose an optimality gap under which we can terminate the branch-and-bound procedure and attain a solution that is close to a member of true Markov equivalence class $\text{MEC}(\mathcal{G}^\star)$ and is asymptotically consistent. Following the analysis of van de Geer & Bühlmann (2013), we impose an additional constraint $\Gamma \in \mathcal{B}_\alpha$ in (3), where $\mathcal{B}_\alpha$ consists of matrices which have at most $\alpha\sqrt{n}/\log(m)$ nonzeros in any column for some constant α.

Recall from §2.1 that there may be multiple structural equation models that are compatible with the distributions $\mathcal{P}^\star$. Each equivalent structural equation model is specified by a directed acyclic graph; this directed acyclic graph defines a total ordering among the variables. Associated to each ordering π is a unique structural equation model that is compatible with the distribution $\mathcal{P}^\star$. We denote the set of parameters of this model as $(\tilde{B}^\star(\pi), \tilde{\Omega}^\star(\pi))$. For the tuple $(\tilde{B}^\star(\pi), \tilde{\Omega}^\star(\pi))$, we define $\tilde{\Gamma}^\star(\pi) := (I - \tilde{B}^\star(\pi))\tilde{\Omega}^\star(\pi)^{-1/2}$. Throughout, we will use the notation $s^\star = \|B^\star\|_{\ell_0}$ and $\tilde{s} := \tilde{s}^\star(\pi) = \|\tilde{B}^\star(\pi)\|_{\ell_0}$.

Assumption 1. (sparsity of every equivalent causal model) There exists some constant $\tilde{\alpha}$ such that for any π, $\|\tilde{B}^{\star}_{:j}(\pi)\|_{\ell_0} \leq \tilde{\alpha}\sqrt{n}/\log(m)$.

Assumption 2. (beta-min condition) There exist constants $0 \leq \eta_1 < 1$ and $0 < \eta_0^2 < 1 - \eta_1$, such that for any π, the matrix $\tilde{B}^{\star}(\pi)$ has at least $(1-\eta_1)\|\tilde{B}^{\star}(\pi)\|_{\ell_0}$ coordinates $k \neq j$ with $|\tilde{B}^{\star}_{kj}(\pi)| > \sqrt{\log(m)/n}(\sqrt{m/s^{\star}} \vee 1)/\eta_0$.

Assumption 3. (bounded minimum and maximum eigenvalues) The smallest and largest eigenvalues of $\Sigma^{\star}$, $\kappa_{\min}(\Sigma^{\star})$ and $\kappa_{\max}(\Sigma^{\star})$, satisfy $\underline{\kappa} \leq \kappa_{\min}(\Sigma^{\star})$ and $\kappa_{\max}(\Sigma^{\star}) \leq \bar{\kappa}$ for some nonzero constants $\underline{\kappa}$ and κ.

Assumption 4. (sufficiently large noise variances) For every permutation π, $\mathcal{O}(1) \geq \min_j[\tilde{\Omega}^{\star}(\pi)]_{jj} \geq \mathcal{O}(\sqrt{s^{\star}\log(m)/n})$.

Assumption 5. (faithfulness) Every conditional independence relationship entailed in the underlying distribution of the variables is encoded in the population directed acyclic graph $\mathcal{G}^{\star}$.

Here, Assumptions 1-3 are similar to those in van de Geer & Bühlmann (2013). Assumption 4 is used to characterize the behavior of the early stopped estimate and is thus new relative to van de Geer & Bühlmann (2013). Assumption 5 on faithfulness is used to connect our estimates to the population parameters in (1); we describe a relaxation of this condition shortly.

THEOREM 1. *Suppose Assumptions 1-5 are satisfied with constants $\alpha, \tilde{\alpha}, \eta_0$ sufficiently small. Further suppose that $m^2 \leq \mathcal{O}(n)$ and $s^{\star} \leq \mathcal{O}(m^2/\log m)$. Let $\alpha_0 := (4/m) \wedge 0.05$. Suppose we let the optimality gap criterion of our algorithm to be $\mathrm{GAP} = \mathcal{O}(m^2/n)$ and let $\hat{\Gamma}^{\mathrm{early}}$ be the early stopped estimate. Then, for $\lambda^2 \asymp \log(m)/n$, there exists a π such that with probability greater than $1 - 2\alpha_0$, $\|\hat{\Gamma}^{\mathrm{early}} - \tilde{\Gamma}^{\star}(\pi)\|_F^2 = \mathcal{O}(m^2/n)$, and $\|\tilde{\Gamma}^{\star}(\pi)\|_{\ell_0} \asymp s^{\star}$.*

The proof of Theorem 1 is in Appendix A.2. The result guarantees that our early stopping optimization procedure accurately estimates certain reordering of the population model. For accurately estimating the edges of the population model, we need a stronger version of the beta-min condition van de Geer & Bühlmann (2013), dubbed the strong beta-min condition.

Assumption 6. (strong beta-min condition) There exist constant $0 < \eta_0^2 < 1/s^{\star}$, such that for any π, the matrix $\tilde{B}^{\star}(\pi)$ has all of its nonzero coordinates (k, j) satisfy $|\tilde{B}^{\star}_{kj}(\pi)| > \sqrt{s^{\star}\log(m)/n}/\eta_0$.

With Assumption 6 we can guarantee that the estimated model is close to a member of the Markov equivalence class of the underlying directed acyclic graph. For a member of the population Markov equivalence class, let $(B^{\star}_{\mathrm{mec}}, \Omega^{\star}_{\mathrm{mec}})$ be the associated connectivity matrix and noise matrix that specify an equivalent model as (1). Furthermore, define $\Gamma^{\star}_{\mathrm{mec}} = (I - B^{\star}_{\mathrm{mec}}){\Omega^{\star}_{\mathrm{mec}}}^{-1/2}$.

THEOREM 2. *Suppose that $\lambda^2 \asymp s^{\star}\log(m)/n$, and assumptions of Theorem 1 as well as Assumption 6 hold. Then, with probability greater than $1 - 2\alpha_0$, there exists a member of the population Markov equivalence class with associated parameter $\Gamma^{\star}_{\mathrm{mec}}$ such that $\|\hat{\Gamma}^{early} - \Gamma^{\star}_{\mathrm{mec}}\|_F^2 \leq \mathcal{O}(m^2/n)$.*

The proof of Theorem 2 is given in Appendix A.3. We remark that without the faithfulness condition in Assumption 5, we can guarantee that the early stopping procedure is close to a member of what is known as the *minimal-edge I-MAP*. The minimal-edge I-MAP is the sparsest set of directed acyclic graphs that induce a structural equation model that is compatible with the

true data distribution. Under faithfulness, the minimal-edge I-MAP coincides with the population Markov equivalence class (van de Geer & Bühlmann, 2013).

6. EXPERIMENTS

6.1. *Setup*

In this section, we illustrate the utility of our method over competing methods on synthetic and real directed acyclic graphs. The state-of-the-art approaches include the high-dimensional top-down approach in Chen et al. (2019), the high-dimensional bottom-up approach in Ghoshal & Honorio (2018), and the mixed-integer second-order cone program in Küçükyavuz et al. (2023). We supply the true moral graph as the input superstructure for our and Küçükyavuz et al. (2023)'s methods (see §3.2). In Appendix A.5, we also present results where the moral graph is estimated from data using the graphical lasso (Friedman et al., 2007). All experiments are performed with a 3.2 GHz 8-Core AMD Ryzen 7 5800H CPU with 16 GB of RAM with `Gurobi` 10.0.0 Optimizer. Our method and the method by Küçükyavuz et al. (2023)'s method are implemented using `Python`. Chen et al. (2019)'s top-down methods and Ghoshal & Honorio (2018)'s bottom-up method are implemented in `R`.

As stated earlier, due to heteroscedastic noise, the underlying directed acyclic graph is generally identifiable up to the Markov equivalence class, represented by a completed partially directed acyclic graph. Thus, to evaluate the performance of the methods, we use the metric d_{cpdag}, which is the number of different entries between the unweighted adjacency matrices of the two completed partially directed acyclic graphs.

Unless otherwise specified, we set the desired optimality gap in our branch-and-bound algorithm to zero. If our branch-and-bound algorithm does not achieve the desired optimality gap within $50m$ seconds, we terminate the algorithm. We report the solution time (in seconds) and achieved relative optimality gap, $\text{RGAP} = (\text{upper bound} - \text{lower bound})/\text{lower bound}$, where upper bound is the objective value associated with the best feasible solution and lower bound represents the best obtained lower bound during the branch-and-bound process.

Unless stated otherwise, we use the Bayesian information criterion to choose the parameter λ. In our context, the Bayesian information criterion score is given by $-2n\sum_{i=1}^{m}\log(\hat{\Gamma}_{ii}) + n\text{tr}(\hat{\Gamma}\hat{\Gamma}^{\text{T}}\hat{\Sigma}_n) + k\log(n)$, where k is the number of nonzero entries in the estimated parameter $\hat{\Gamma}$. From our theoretical guarantees in §5, λ^2 should be on the order $\log(m)/n$. Hence, we choose λ with the smallest Bayesian information criterion score among $\lambda^2 = c^2\log(m)/n$, for $c = 1, 2, \ldots, 15$. The code to reproduce all the experiments is available at `https://github.com/AtomXT/MICP-NID`.

6.2. *Comparison to Other Benchmarks*

We compare the performance of our method with the benchmark methods on twelve publicly available networks sourced from Manzour et al. (2021) and the Bayesian Network Repository (bnlearn). These networks vary in size, ranging from $m = 6$ to $m = 70$ nodes. For a given network structure, we generate $n = 500$ independent and identically distributed samples from (1) where the nonzero entries of $B^\star$ are drawn uniformly at random from the set $\{-0.8, -0.6, 0.6, 0.8\}$ and diagonal entries of $\Omega^\star$ are chosen uniformly at random from the set $\{0.5, 1, 1.5\}$. In §6.5, we will explore a larger range of noise variances.

Table 1 summarizes the performance of all methods, averaged over 10 independent trials. Here, the symbol $*$ indicates that the achieved optimality gap is zero and so an optimal solution is found. The symbol $\geq$ in the time means that the time limit of $50m$ seconds was reached. We also report the structural hamming distances of the undirected skeleton of the true directed acyclic

graph and the corresponding skeleton of the estimated network, the true positive rate, and the false positive rate in Appendix A.6. Compared to the top-down approach of Chen et al. (2019) and the bottom-up approach by Ghoshal & Honorio (2018), our method produces more accurate estimates—giving smaller d_{cpdag}—and provides optimality guarantees, RGAP. We observe that when our algorithm terminates at proven optimality, which happens in 7 out of 12 instances, it also produces more accurate estimates than the method of Küçükyavuz et al. (2023). The improved statistical performance is because our method accounts for heteroscedastic noise, or non-identifiability, while the other methods do not. However, our optimization method can be slower, as it contains logarithmic terms in its objective—to handle heteroscedastic noise—that are not present in the second-order conic method of Küçükyavuz et al. (2023).

6.3. *Early Stopping*

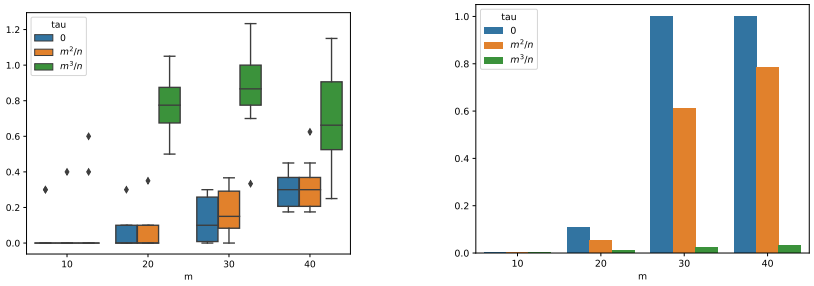
So far, we have set the desired level of optimality gap to be $\tau = 0$. As described in §4.2 and §5, one can set the optimality gap to $\tau = \mathcal{O}(m^2/n)$ to speed up computations while retaining good statistical properties. We next empirically explore how changing τ in the range $\tau \in \{0, m^2/n, m^3/n\}$ impacts both the computational and statistical performances of our algorithm. Concretely, we generate a directed acyclic graph among m nodes with m directed edges using `randomDAG` from `R` package `pcalg`. For this network structure, we generate $n = 400$ independent and identically distributed samples from (1) where the nonzero entries of $B^\star$ are drawn uniformly at random from the set $\{-0.8, -0.6, 0.6, 0.8\}$ and the diagonal entries of $\Omega^\star$ are chosen uniformly at random from the set $\{0.5, 1, 1.5\}$. We then evaluate the performance of our algorithm for $m \in \{10, 20, 30, 40\}$ and different values of τ over 10 independent datasets.

Figure 2 summarizes the results. Here, the metric d_{cpdag} is scaled by m, the total number of edges in the true underlying directed acyclic graph. Further, time is scaled by $50m$. We observe that applying an early stopping criterion—using $\tau \geq m^2/n$—-leads to notable computational benefits. For example, when $\tau = 0$ and $m = 40$, we cannot solve any of the ten instances within the time limit of $50m$ seconds, whereas early stopping with $\tau = m^2/n$ finishes within the time limit for five instances with a total average time of approximately 1000 seconds across the ten instances. Furthermore, as predicted by our theoretical results, the faster computation time does not come at a statistical cost when $\tau = m^2/n$. For example, when $m = 20$, all ten instances for

Table 1: Comparison of our method and benchmarks

	HD-BU		HU-TD		TD		MISOCP			OUR METHOD		
Network(m)	Time	d_{cpdag}	Time	d_{cpdag}	Time	d_{cpdag}	Time	RGAP	d_{cpdag}	Time	RGAP	d_{cpdag}
Dsep(6)	≤ 1	10.5	≤ 1	3.9	≤ 1	2.0	≤ 1	*	2.3	≤ 1	*	2.0
Asia(8)	≤ 1	18.0	≤ 1	14.5	≤ 1	12.1	≤ 1	*	10.8	≤ 1	*	2.2
Bowling(9)	≤ 1	5.6	≤ 1	8.5	≤ 1	6.0	≤ 1	*	5.7	3	*	2.0
InsSmall(15)	≤ 1	37.7	≤ 1	13.7	≤ 1	9.7	3	*	8.0	≥ 750	.065	6.9
Rain(14)	≤ 1	18.1	≤ 1	9.7	≤ 1	3.6	≤ 1	*	2.0	130	*	2.0
Cloud(16)	≤ 1	41.7	≤ 1	30.4	≤ 1	20.5	≤ 1	*	19.1	18	*	4.7
Funnel(18)	≤ 1	16.3	≤ 1	11.4	≤ 1	2.0	≤ 1	*	3.1	122	*	2.0
Galaxy(20)	≤ 1	45.7	≤ 1	36.1	≤ 1	29.1	≤ 1	*	8.9	150	*	1.0
Insurance(27)	1	45.2	2	46.1	≤ 1	34.6	590	.011	15.8	≥ 1350	.282	19.9
Factors(27)	1	25.1	2	49.0	≤ 1	48.1	759	.011	32.3	≥ 1350	.337	49.0
Hailfinder(56)	9	120.1	10	89.0	≤ 1	55.2	≥ 2800	.072	22.2	≥ 2800	.273	33.5
Hepar2(70)	20	126.6	22	82.4	2	68.4	≥ 3500	.067	43.9	≥ 3500	.316	55.2

Here, HD-BU, high-dimensional bottom-up; HD-TD, high-dimensional top-down; TD, top-down; MISOCP, mixed-integer second-order cone program; d_{cpdag}, differences between the true and estimated completed partially directed acyclic graphs; RGAP, relative optimality gap. All results are averaged across ten independent trials.



(a) Box plots of scaled d_{cpdag} (b) Bar plots of average scaled time

Fig. 2: Scaled d_{cpdag} and time in early stopping results with $\tau \in \{0, m^2/n, m^3/n\}$ across ten independent trials.

$\tau = \{0, m^2/n\}$ are completed and the d_{cpdag} metrics are similar. We also observe that when $\tau = m^3/n$, the scaled d_{cpdag} values are substantially larger than those of $\tau = m^2/n$, further supporting our theoretical results.

6.4. *Comparison of General Constraint Attributes Versus Outer Approximation*

In §4.1, we proposed an outer approximation technique for handling the logarithm term in our mixed integer program (5). We next numerically illustrate the computational and statistical benefits of using our outer approximation approach over `Gurobi`'s general constraint attribute. Specifically, we generate 10 independent datasets according to the setup in §6.2. Table 2 compares the performance of solving (5) using the outer approximation technique versus using the general constraint attribute of `Gurobi`. For small networks that are solved to optimality, before reaching the time limit, our method with outer approximation has similar statistical performance, with similar values of d_{cpdag}, while offering significant computational advantages. For larger networks that cannot be solved within the time limit, our proposed method has comparable performance to the approach without outer approximation. For the "Galaxy" network, our method yields smaller d_{cpdag} while completing the task in 150 seconds whereas the approach with `Gurobi`'s general constraint attribute does not finish within the allotted time limit.

Table 2: Comparison of general constraint attribute with outer approximation for our method

	General constraint attribute			Outer approximation		
Network(m)	Time	RGAP	d_{cpdag}	Time	RGAP	d_{cpdag}
Dsep(6)	6	*	2.0	≤ 1	*	2.0
Asia(8)	8	*	2.0	≤ 1	*	2.2
Bowling(9)	33	*	2.0	3	*	2.0
InsSmall(15)	≥ 750	.093	4.0	≥ 750	.065	6.9
Rain(14)	≥ 700	.018	2.0	130	*	2.0
Cloud(16)	228	*	5.0	18	*	4.7
Funnel(18)	240	*	2.0	122	*	2.0
Galaxy(20)	≥ 1000	.078	12.0	150	*	1.0
Insurance(27)	≥ 1350	.273	22.0	≥ 1350	.282	19.9
Factors(27)	≥ 1350	.337	54.4	≥ 1350	.337	49.0
Hailfinder(56)	≥ 2800	.223	15.0	≥ 2800	.273	33.5
Hepar2(70)	≥ 3500	.321	49.0	≥ 3500	.316	55.2

Here, d_{cpdag}, differences between the true and estimated completed partially directed acyclic graphs; RGAP, relative optimality gap. All results are averaged across ten independent trials.

6.5. *Robustness to Different Amounts of Noise Heteroscedasticity*

We provide more details of the experiment presented in §1.2. We consider the empirical setup in §6.3, with the exception that the diagonal entries of $\Omega^\star$ are sampled uniformly at random from the interval $[4-\alpha, 4+\alpha]$, where $\alpha \in \{1, 2, 4\}$. Larger values of α thus indicate a larger degree of noise heteroscedasticity. We empirically evaluate the performance of our method and competing methods across ten independent trials.

Table 3 and Figure 1 show that our method is robust to different levels of noise heteroscedasticity, producing small values of d_{cpdag} for different α. This is in contrast to the competing methods that rely on homoscedastic noise: they yield large values of d_{cpdag} under strong violation of the homoscedasticity assumption (e.g., $\alpha = 4$).

Table 3: Comparison to benchmarks for different amounts of noise heteroscedasticity

		TD		HD-BU		MISOCP			OUR METHOD		
m	α	Time	d_{cpdag}	Time	d_{cpdag}	Time	RGAP	d_{cpdag}	Time	RGAP	d_{cpdag}
10	1	0.01	0.2	0.23	7.0	0.07	*	0.3	0.76	*	0.2
10	2	0.01	0.4	0.21	8.3	0.08	*	0.7	0.76	*	0
10	4	0.01	2.4	0.21	5.4	0.07	*	1.2	0.66	*	0
15	1	0.01	3.7	0.35	12.1	0.13	*	3.2	7.41	*	1.4
15	2	0.01	5.6	0.36	14.6	0.14	*	6.2	5.17	*	0.2
15	4	0.01	14.9	0.36	16.7	0.12	*	10.1	4.60	*	1.5
20	1	0.03	9.1	0.58	20.4	0.22	*	3.2	177.06	*	0.3
20	2	0.03	12.0	0.57	20.1	0.27	*	6.3	68.09	*	1.6
20	4	0.03	16.9	0.57	27.6	0.27	*	9.9	116.20	*	1.6

Here, TD, top-down; HD-BU, high-dimensional bottom-up; MISOCP, mixed-integer second-order cone program; d_{cpdag}, differences between the true and estimated completed partially directed acyclic graphs; RGAP, relative optimality gap. All results are averaged across ten independent trials.

7. DISCUSSION

We discuss some future research questions that arise from our work. The mixed-integer framework allows for more than one optimal solution to be identified. It would be of interest to explore this feature to identify multiple solutions in the Markov equivalence class. In addition, the outer approximation algorithm can be further enhanced by developing stronger cutting planes that exploit the constraint structure. Finally, an open question is whether, in the context of our statistical guarantees with early stopping, the sample-size requirement of $n \geq \mathcal{O}(m^2)$ is fundamental.

ACKNOWLEDGEMENT

TX and SK are supported, in part, by a National Science Foundation Grant. AT is supported by the Royalty Research Fund at the University of Washington. AS was supported by grants from the National Science Foundation and the National Institutes of Health.

REFERENCES

BARTLETT, M. & CUSSENS, J. (2017). Integer linear programming for the bayesian network structure learning problem. *Artificial Intelligence* **244**, 258–271. Combining Constraint Solving with Mining and Learning.

CAI, T., LIU, W. & LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106**, 594–607.

CHEN, W., DRTON, M. & WANG, Y. S. (2019). On causal discovery with an equal-variance assumption. *Biometrika* **106**, 973–980.

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

CHICKERING, D. M. (2020). Statistically efficient greedy equivalence search. In *Uncertainty in Artificial Intelligence*.

CUI, X. T., ZHENG, X. J., ZHU, S. S. & SUN, X. L. (2013). Convex relaxations and MIQCQP reformulations for a class of cardinality-constrained portfolio selection problems. *J. of Global Optimization* **56**, 1409–1423.

CUSSENS, J., HAWS, D. & STUDENỲ, M. (2017a). Polyhedral aspects of score equivalence in Bayesian network structure learning. *Mathematical Programming* **164**, 285–324.

CUSSENS, J., JÄRVISALO, M., KORHONEN, J. H. & BARTLETT, M. (2017b). Bayesian network structure learning with integer programming: Polytopes, facets and complexity. *Journal of Artificial Intelligence Research* **58**, 185–229.

DRTON, M. & MAATHUIS, M. H. (2017). Structure learning in graphical modeling. *Annual Review of Statistics and Its Application* **4**, 365–393.

DURAN, M. A. & GROSSMANN, I. E. (1986). An outer-approximation algorithm for a class of mixed-integer nonlinear programs. *Mathematical Programming* **36**, 307–339.

FRANGIONI, A. & GENTILE, C. (2007). SDP diagonalizations and perspective cuts for a class of nonseparable miqp. *Operations Research Letters* **35**, 181–185.

FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441.

GHOSHAL, A. & HONORIO, J. (2018). Learning linear structural equation models in polynomial time and sample complexity. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, A. Storkey & F. Perez-Cruz, eds., vol. 84 of *Proceedings of Machine Learning Research*. PMLR.

GRÖTSCHEL, M., JÜNGER, M. & REINELT, G. (1985). On the acyclic subgraph polytope. *Mathematical Programming* **33**, 28–42.

KALISCH, M. & BÜHLMANN, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research* **8**, 613–636.

KÜÇÜKYAVUZ, S., SHOJAIE, A., MANZOUR, H., WEI, L. & WU, H.-H. (2023). Consistent second-order conic integer programming for learning Bayesian networks. *Journal of Mchine Learning Research* (in press) .

MANZOUR, H., KÜÇÜKYAVUZ, S., WU, H.-H. & SHOJAIE, A. (2021). Integer programming for learning directed acyclic graphs from continuous data. *INFORMS Journal on Optimization* **3**, 46–73.

PARK, Y. W. & KLABJAN, D. (2017). Bayesian network learning via topological order. *Journal of Machine Learning Research* **18**, 1–32.

PETERS, J. & BÜHLMANN, P. (2013). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika* **101**, 219–228.

SHOJAIE, A. & MICHAILIDIS, G. (2010). Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika* **97**, 519–538.

SILANDER, T. & MYLLYMÄKI, P. (2006). A simple approach for finding the globally optimal Bayesian network structure. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'06. Arlington, Virginia, USA: AUAI Press.

SPIRTES, P., GLYMOUR, C. & SCHEINES, R. (1993). *Causation, Prediction, and Search*. The MIT Press.

UHLER, C., RASKUTTI, G., BUHLMANN, P. & YU, B. (2012). Geometry of the faithfulness assumption in causal inference. *Annals of Statistics* **41**, 436–463.

VAN DE GEER, S. & BÜHLMANN, P. (2013). ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *Annals of Statistics* **41**, 536 – 567.

WEI, L., ATAMTÜRK, A., GÓMEZ, A. & KÜÇÜKYAVUZ, S. (2023). On the convex hull of convex quadratic optimization problems with indicators. *Mathematical Programming* Article in Advance.

WEI, L., GÓMEZ, A. & KÜÇÜKYAVUZ, S. (2022). Ideal formulations for constrained convex optimization problems with indicator variables. *Mathematical Programming* **192**, 57–88.

WEI, L., GÓMEZ, A. & KÜÇÜKYAVUZ, S. (2020). On the convexification of constrained quadratic optimization problems with indicator variables. In *International Conference on Integer Programming and Combinatorial Optimization*. Springer.

WOLSEY, L. & NEMHAUSER, G. (1999). *Integer and Combinatorial Optimization*. Wiley Series in Discrete Mathematics and Optimization. Wiley.

APPENDIX

A.1. *Proof of Proposition 1*

Proof of Proposition 1. By replacing Θ with $\Gamma\Gamma^{\mathrm{T}}$ for Γ such that $\Gamma = (I - B)D^{1/2}$, problem (2) is equivalent to the following problem:

$$\begin{aligned} \min_{\Gamma, B, D} \; & -2\log\det(\Gamma) + \mathrm{tr}(\Gamma\Gamma^{\mathrm{T}}\hat{\Sigma}_n) + \lambda^2\|\Gamma - \mathrm{diag}(\Gamma)\|_{\ell_0}, \\ \text{s.t. } & \Gamma = (I - B)D^{1/2}, \\ & B \in \mathcal{B}, D \succ 0. \end{aligned}$$

Since B has zero diagonal elements and D is diagonal with strict positive diagonal entries, we have $\|\Gamma - \mathrm{diag}(\Gamma)\|_{\ell_0} = \|B\|_{\ell_0}$. We do not require Γ to be a lower-triangular matrix at this time.

These observations lead to the next claim that allows us to replace the non-convex constraint $\Gamma = (I - B)D^{1/2}$, with convex (linear) constraints.

LEMMA A1. $\{\Gamma : \textit{There exists } D \succ 0, B \in \mathcal{B} \textit{ s.t. } \Gamma = (I - B)D^{1/2}\} = \{\Gamma : \Gamma_{ii} > 0, i = 1, \ldots, m; \Gamma - \mathrm{diag}(\Gamma) \in \mathcal{B}\}$.

Proof of Lemma A1. " $\subseteq$ ": One can write $\Gamma = (I - B)D^{1/2} = D^{1/2} - BD^{1/2}$. Since B is a directed acyclic graph and D is a positive definite diagonal matrix, $BD^{1/2}$ is also a directed acyclic graph. Therefore,

$$\Gamma - \mathrm{diag}(\Gamma) = D^{1/2} - BD^{1/2} - D^{1/2} = -BD^{1/2}$$

is a directed acyclic graph, and we have $\Gamma_{ii} = D_{ii}^{1/2} > 0$, for any i.

" $\supseteq$ ": If $\Gamma - \mathrm{diag}(\Gamma) \in \mathcal{B}$ and $\Gamma_{ii} > 0$, for any i, let $D^{1/2} = \mathrm{diag}(\Gamma)$, then $D \succ 0$.

Then, $D^{-1/2}(\Gamma - \mathrm{diag}(\Gamma)) = D^{-1/2}\Gamma - I$ is also a directed acyclic graph. Let $B = I - \Gamma D^{-1/2}$, then we have $(I - B)D^{1/2} = \Gamma$. □

Therefore, Problem (2) is equivalent to the following optimization problem:

$$\begin{aligned} \min_{\Gamma} \; & -2\log\det(\Gamma) + \mathrm{tr}(\Gamma\Gamma^{\mathrm{T}}\hat{\Sigma}_n) + \lambda^2\|\Gamma - \mathrm{diag}(\Gamma)\|_{\ell_0}, \\ \text{s.t. } & \Gamma_{ii} > 0 \quad (i = 1, \ldots, m), \\ & \Gamma - \mathrm{diag}(\Gamma) \in \mathcal{B}. \end{aligned}$$

Furthermore, for any permutation matrix P, we have $\det(\Gamma) = \det(P\Gamma P^{\mathrm{T}})$. Since $\Gamma - \mathrm{diag}(\Gamma) \in \mathcal{B}$, if $\Gamma_{ij} \neq 0$ for $i \neq j$, we deduce that Γ_{ji} must be 0. Therefore, there exists a permutation matrix $\bar{P}$ such that $\bar{P}\Gamma\bar{P}^{\mathrm{T}}$ is a lower-triangular matrix. As a result, $\log\det(\Gamma) = \sum_{i=1}^{m}\log(\Gamma_{ii})$. Further, the constraint $\Gamma_{ii} > 0$ can be removed due to the logarithm barrier function.

If $(\hat{B}^{\mathrm{opt}}, \hat{D}^{\mathrm{opt}})$ is a minimizer of (2), as shown in the proof of Lemma A1, we have the corresponding feasible point $\hat{\Gamma}^{\mathrm{opt}} = (I - \hat{B}^{\mathrm{opt}})(\hat{D}^{\mathrm{opt}})^{1/2}$ of (3). Since (2) and (3) have the same objective function after changing variables, $\hat{\Gamma}^{\mathrm{opt}}$ is also the minimizer of (3). Similarly, if $\hat{\Gamma}^{\mathrm{opt}}$ is a minimizer of (3), then $\hat{D}^{\mathrm{opt}} = \mathrm{diag}(\hat{\Gamma}^{\mathrm{opt}})^2$ and $\hat{B}^{\mathrm{opt}} = I - \hat{\Gamma}^{\mathrm{opt}}(\hat{D}^{\mathrm{opt}})^{-1/2}$ is the minimizer of (2). □

A.2. *Proof of Theorem 1*

In the following proofs, for any vector $x \in \mathbb{R}^m$ and matrix $X \in \mathbb{R}^{m\times m}$, we use $\|X\|_F, \|X\|_2, \|X\|_\star$ to denote the Frobenius norm, spectral norm, and the nuclear norm of X respectively. In addition, $\|x\|_{\ell_2}$ is the ℓ_2 norm of a vector x. Our proof relies on the following result from van de Geer & Bühlmann (2013).

PROPOSITION A1. *(Theorem 3.1 of van de Geer & Bühlmann (2013)) Assume Assumptions 1-3 hold with constants $\alpha, \tilde{\alpha}, \eta_0$ sufficiently small. Let $\hat{\Gamma}^{\mathrm{opt}}$ be any optimum of* (3) *with associated ordering π. Let $(\hat{B}^{\mathrm{opt}}, \hat{\Omega}^{\mathrm{opt}})$ be the associated connectivity and noise variance matrix satisfying $\hat{\Gamma}^{\mathrm{opt}} = (I - \hat{B}^{\mathrm{opt}})\hat{\Omega}^{\mathrm{opt}\,-1/2}$. Let $\alpha_0 := (4/m) \wedge 0.05$. Then for*

$$\lambda^2 \asymp \frac{\log(m)}{n}$$

we have, with probability greater than $1-\alpha_0$*,*

$$\|\hat{B}^{\text{opt}} - \tilde{B}^\star(\pi)\|_F^2 + \|\hat{\Omega}^{\text{opt}} - \tilde{\Omega}^\star(\pi^{\text{opt}})\|_F^2 = \mathcal{O}(\lambda^2 s^\star),$$

and $\|\tilde{B}^\star\|_{\ell_0} \asymp s^\star$*.*

COROLLARY A1. *Consider the setup in Proposition* A*1. Then,*

$$\left\|\hat{\Gamma}^{\text{opt}} - \tilde{\Gamma}^\star(\pi)\right\|_F^2 \leq \frac{16\max\{1, \|\tilde{B}^\star(\pi)\|_F^2, \|\tilde{\Omega}^\star(\pi)^{-1/2}\|_F^2\}\lambda^2 s^\star}{\min\{1, \min_j(\tilde{\Omega}^\star(\pi)_{jj})^3\}}.$$

We will prove Corollary A1 after we complete the proof of Theorem 1. For notational simplicity, we let $\Gamma^\star = \tilde{\Gamma}^\star(\pi)$.

Proof of Theorem 1. For a matrix $\Gamma \in \mathbb{R}^{m\times m}$, let $\ell(\Gamma) := \sum_{i=1}^m -2\log(\Gamma_{ii}) + \text{tr}(\Gamma\Gamma^{\text{T}}\hat{\Sigma}_n)$. Here, $\ell(\Gamma)$ represents the negative log-likelihood of Γ, where we have, for simplicity, abused the notation relative to the main paper. Note that the attained objective (after early stopping) and global optimal objective are given by:

$$\mathcal{L}(\hat{\Gamma}^{\text{early}}) = \ell(\hat{\Gamma}^{\text{early}}) + \lambda^2\hat{s}^{\text{early}}$$
$$\mathcal{L}(\hat{\Gamma}^{\text{opt}}) = \ell(\hat{\Gamma}^{\text{opt}}) + \lambda^2\hat{s}^{\text{opt}}.$$

Suppose first $\hat{s}^{\text{early}} \geq \hat{s}^{\text{opt}}$. Then, since $\mathcal{L}(\hat{\Gamma}^{\text{early}}) \geq \mathcal{L}(\hat{\Gamma}^{\text{opt}})$, we have that $\ell(\hat{\Gamma}^{\text{early}}) - \ell(\hat{\Gamma}^{\text{opt}}) \leq \text{GAP}$. On the other hand, if we suppose $\hat{s}^{\text{opt}} \geq \hat{s}^{\text{early}}$, $\ell(\hat{\Gamma}^{\text{early}}) - \ell(\hat{\Gamma}^{\text{opt}}) \leq \text{GAP} + \lambda^2\hat{s}^{\text{opt}} = \mathcal{O}(m^2/n)$, where we have used $\hat{s}^{\text{opt}} \asymp s^\star$ from Proposition A1 and that $s^\star \leq \mathcal{O}(m^2/\log m)$ from Theorem 1 conditions.

Thus, based on the stopping criterion, we conclude that:

$$\ell(\hat{\Gamma}^{\text{early}}) - \ell(\hat{\Gamma}^{\text{opt}}) = \mathcal{O}\left(\frac{m^2}{n}\right). \tag{A3}$$

For notational simplicity, we will consider a vectorized objective. Let $T \subseteq \{1,\ldots,m^2\}$ be indices corresponding to diagonal elements of an $m\times m$ matrix being vectorized. With abuse of notation, let $\hat{\Gamma}^{\text{early}}$, $\hat{\Gamma}^{\text{opt}}$, and $\Gamma^\star$ be the vectorized form of their corresponding matrices.

Then,

$$\begin{aligned}
\ell(\hat{\Gamma}^{\text{early}}) - \ell(\hat{\Gamma}^{\text{opt}}) &= \nabla\ell(\hat{\Gamma}^{\text{opt}})^T(\hat{\Gamma}^{\text{early}} - \hat{\Gamma}^{\text{opt}}) + 1/2(\hat{\Gamma}^{\text{early}} - \hat{\Gamma}^{\text{opt}})^T\nabla^2\ell(\tilde{\Gamma})(\hat{\Gamma}^{\text{early}} - \hat{\Gamma}^{\text{opt}}) \\
&= [\nabla\ell(\hat{\Gamma}^{\text{opt}}) - \nabla\ell(\Gamma^\star)]^T(\hat{\Gamma}^{\text{early}} - \hat{\Gamma}^{\text{opt}}) + \nabla\ell(\Gamma^\star)^T(\hat{\Gamma}^{\text{early}} - \hat{\Gamma}^{\text{opt}}) + \\
&\quad 1/2(\hat{\Gamma}^{\text{early}} - \hat{\Gamma}^{\text{opt}})^T\nabla^2\ell(\tilde{\Gamma})(\hat{\Gamma}^{\text{early}} - \hat{\Gamma}^{\text{opt}}) \\
&= (\Gamma^\star - \hat{\Gamma}^{\text{opt}})^T\nabla^2\ell(\bar{\Gamma})(\hat{\Gamma}^{\text{early}} - \hat{\Gamma}^{\text{opt}}) + \nabla\ell(\Gamma^\star)^T(\hat{\Gamma}^{\text{early}} - \hat{\Gamma}^{\text{opt}}) + \\
&\quad 1/2(\hat{\Gamma}^{\text{early}} - \hat{\Gamma}^{\text{opt}})^T\nabla^2\ell(\tilde{\Gamma})(\hat{\Gamma}^{\text{early}} - \hat{\Gamma}^{\text{opt}}).
\end{aligned}$$

Here, entries of $\tilde{\Gamma}$ lie between $\hat{\Gamma}^{\text{early}}$ and $\hat{\Gamma}^{\text{opt}}$, and entries of $\bar{\Gamma}$ lie between $\hat{\Gamma}^{\text{opt}}$ and $\Gamma^\star$. Some algebra then gives:

$$\begin{aligned}
1/2(\hat{\Gamma}^{\text{early}} - \hat{\Gamma}^{\text{opt}})^T\nabla^2\ell(\tilde{\Gamma})(\hat{\Gamma}^{\text{early}} - \hat{\Gamma}^{\text{opt}}) \leq &[\ell(\hat{\Gamma}^{\text{early}}) - \ell(\hat{\Gamma}^{\text{opt}})] \\
&+ \|\nabla\ell(\Gamma^\star)\|_{\ell_2}\|\hat{\Gamma}^{\text{early}} - \hat{\Gamma}^{\text{opt}}\|_{\ell_2} \\
&+ \|\hat{\Gamma}^{\text{early}} - \hat{\Gamma}^{\text{opt}}\|_{\ell_2}\|\hat{\Gamma}^{\text{opt}} - \Gamma^\star\|_{\ell_2}\kappa_{\max}(\nabla^2\ell(\bar{\Gamma})).
\end{aligned} \tag{A4}$$

Due to the convexity of the negative log-likelihood function, for any Γ,

$$\nabla^2\ell(\Gamma) \succeq \hat{\Sigma}_n \otimes I_m.$$

Standard Gaussian concentration results state that for every $0 < \xi < 1$ and $n \geq \mathcal{O}(m)$, with probability greater than $1 - \xi$,

$$\|\hat{\Sigma}_n - \Sigma^\star\|_2 \leq \mathcal{O}\left(\sqrt{\frac{m}{n}}\right). \tag{A5}$$

Letting $\xi = \alpha_0$, under Assumption 3, we have that with, probability greater than $1 - \alpha_0$

$$\kappa_{\min}(\nabla^2 \ell(\Gamma)) \geq \kappa_{\min}(\hat{\Sigma}_n \otimes I_m) = \kappa_{\min}(\hat{\Sigma}_n) > \underline{\kappa} - \mathcal{O}\left(\sqrt{\frac{m}{n}}\right) \geq \underline{\kappa}/2. \tag{A6}$$

Using the inequality $(\hat{\Gamma}^{\text{early}} - \hat{\Gamma}^{\text{opt}})^T \nabla^2 \ell(\tilde{\Gamma})(\hat{\Gamma}^{\text{early}} - \hat{\Gamma}^{\text{opt}}) \geq \kappa_{\min}(\nabla^2 \ell(\tilde{\Gamma}))\|\hat{\Gamma}^{\text{early}} - \hat{\Gamma}^{\text{opt}}\|_{\ell_2}^2$, from (A4), with probability greater than $1 - 2\alpha_0$,

$$\begin{aligned}\|\hat{\Gamma}^{\text{early}} - \hat{\Gamma}^{\text{opt}}\|_{\ell_2}^2 \leq & 4\underline{\kappa}^{-1}\left\{\ell(\hat{\Gamma}^{\text{early}}) - \ell(\hat{\Gamma}^{\text{opt}})\right\} \\ & + 4\underline{\kappa}^{-1}\left\{\|\hat{\Gamma}^{\text{early}} - \hat{\Gamma}^{\text{opt}}\|_{\ell_2}\left(\|\hat{\Gamma}^{\text{opt}} - \Gamma^\star\|_2 \kappa_{\max}(\nabla^2 \ell(\bar{\Gamma})) + \|\nabla \ell(\Gamma^\star)\|_{\ell_2}\right)\right\}.\end{aligned} \tag{A7}$$

Let $\tau := 4(\|\hat{\Gamma}^{\text{opt}} - \Gamma^\star\|_{\ell_2} \kappa_{\max}(\nabla^2 \ell(\bar{\Gamma})) + \|\nabla \ell(\Gamma^\star)\|_{\ell_2})/\underline{\kappa}$. Note that for non-negative Z, W, Π, the inequality $Z^2 \leq \Pi Z + W$ implies $Z \leq (\Pi + \sqrt{\Pi^2 + 4W})/2$. Using this fact, in conjunction with (A7), we obtain with probability greater than $1 - 2\alpha_0$ the bound

$$\|\hat{\Gamma}^{\text{early}} - \hat{\Gamma}^{\text{opt}}\|_{\ell_2} \leq \frac{\tau}{2} + \frac{1}{2}\sqrt{\tau^2 + 16\underline{\kappa}^{-1}\left\{\ell(\hat{\Gamma}^{\text{early}}) - \ell(\hat{\Gamma}^{\text{opt}})\right\}}. \tag{A8}$$

We next bound the quantity τ. From Corollary A1 and (A6), we have control over the terms $\|\hat{\Gamma}^{\text{opt}} - \Gamma^\star\|_{\ell_2}$ in τ and $\kappa_{\min}(\nabla^2 \ell(\tilde{\Gamma}))$. It remains to control $\kappa_{\max}(\nabla^2 \ell(\bar{\Gamma}))$ and $\|\nabla \ell(\Gamma^\star)\|_{\ell_2}$. Let $\Gamma \in \mathbb{R}^{m^2}$. Suppose that for every $j \in T$, $\Gamma_j \geq \nu$. Then, some calculations yield the bound

$$\nabla^2 \ell(\Gamma) \preceq \hat{\Sigma}_n \otimes I_m + \frac{2}{\nu^2} I_{m^2} = \hat{\Sigma}_n \otimes I_m + \frac{2}{\nu^2} I_{m^2}.$$

We have that for every $j \in T$, $\hat{\Gamma}_j^{\text{opt}} \geq \Gamma_j^\star - \|\hat{\Gamma}^{\text{opt}} - \Gamma^\star\|_{\ell_2}$. From Corollary A1, Assumption 4, and that $\lambda\sqrt{s^\star} \leq 1$, we then have $\hat{\Gamma}_j^{\text{opt}} \geq \Gamma_j^\star/2 \geq 1/2(\Omega_j^\star)^{-1/2}$. Since the entries of $\bar{\Gamma}$ are between those of $\Gamma^\star$ and $\hat{\Gamma}^{\text{opt}}$,

$$\kappa_{\max}(\nabla^2 \ell(\bar{\Gamma})) \leq \kappa_{\max}(\hat{\Sigma}_n) + 8 \min_j \Omega_j^\star \leq \bar{\kappa} + 8 \min_j \Omega_j^\star + \mathcal{O}\left(\sqrt{\frac{m}{n}}\right) = \mathcal{O}(1). \tag{A9}$$

To control $\nabla \ell(\Gamma^\star)$, we first note that $\mathbb{E}[\nabla \ell(\Gamma^\star)] = 0$. Therefore, $\|\nabla \ell(\Gamma^\star)\|_{\ell_2} = \|\nabla \ell(\Gamma^\star) - \mathbb{E}[\nabla \ell(\Gamma^\star)]\|_{\ell_2}$. Since

$$\nabla \ell(\Gamma^\star) - \mathbb{E}[\nabla \ell(\Gamma^\star)] = \left((\hat{\Sigma}_n - \Sigma^\star) \otimes I_m\right)\Gamma^\star,$$

letting $K^\star = (\Sigma^\star)^{-1}$ we get

$$\begin{aligned}\|\nabla \ell(\Gamma^\star) - \mathbb{E}[\nabla \ell(\Gamma^\star)]\|_{\ell_2}^2 &= \operatorname{tr}((\hat{\Sigma}_n - \Sigma^\star)(\hat{\Sigma}_n - \Sigma^\star)^T K^\star) \\ &\leq \|\hat{\Sigma}_n - \Sigma^\star\|_2^2 \|K^\star\|_\star \leq m\|\hat{\Sigma}_n - \Sigma^\star\|_2^2 \|K^\star\|_2 \leq \mathcal{O}\left(\frac{m^2}{n}\right).\end{aligned}$$

Thus,

$$\|\nabla \ell(\Gamma^\star) - \mathbb{E}[\nabla \ell(\Gamma^\star)]\|_{\ell_2} \leq \mathcal{O}\left(\frac{m}{\sqrt{n}}\right). \tag{A10}$$

□

Putting $\hat{\Gamma}^{\text{early}}$, $\hat{\Gamma}^{\text{opt}}$ and $\Gamma^\star$ into matrix form, and plugging the bounds (A3), (A6), and (A10) into (A8), we get

$$\|\hat{\Gamma}^{\text{early}} - \hat{\Gamma}^{\text{opt}}\|_F^2 = \mathcal{O}\left(\frac{m^2}{n}\right).$$

Combining this bound with Proposition A1, we get the desired result:

$$\|\hat{\Gamma}^{\text{early}} - \Gamma^\star\|_F^2 \leq 2\|\hat{\Gamma}^{\text{early}} - \hat{\Gamma}^{\text{opt}}\|_F^2 + 2\|\hat{\Gamma}^{\text{opt}} - \Gamma^\star\|_F^2.$$

Proof of Corollary A*1.* For notational simplicity, we let $\Omega = \tilde{\Omega}^\star(\pi)$ and $B = \tilde{B}^\star(\pi)$. First note by mean-value theorem that for any $j \in \{1, \ldots, m\}$, there exists some ω in the interval between $\hat{\Omega}_{jj}^{\text{opt}}$ and Ω_{jj} such that:

$$(\Omega_{jj}^{-1/2} - (\hat{\Omega}^{\text{opt}})_{jj}^{-1/2})^2 \leq \frac{1}{4}(\Omega_{jj} - \hat{\Omega}_{jj}^{\text{opt}})^2\omega^{-3}.$$

Noticing that $\omega \geq \min_j \Omega_{jj} - \|\Omega - \hat{\Omega}^{\text{opt}}\|_F \geq \min_j \Omega_{jj} - \lambda\sqrt{s^\star} \geq \min_j \Omega_{jj}/2$, where the final equality follows from Assumption 4. Putting things together, we have that:

$$\|\Omega^{-1/2} - (\hat{\Omega}^{\text{opt}})^{-1/2}\|_F^2 \leq \frac{2\|\Omega - \hat{\Omega}^{\text{opt}}\|_F^2}{\min \Omega_{jj}^3}.$$

Then,

$$\begin{aligned}
\|(I - \hat{B}^{\text{opt}})(\hat{\Omega}^{\text{opt}})^{-1/2} - (I - B)\Omega^{-1/2}\|_F^2 &\leq \|(I - \hat{B}^{\text{opt}})(\hat{\Omega}^{\text{opt}})^{-1/2} - (I - B)(\hat{\Omega}^{\text{opt}})^{-1/2} \\
&\quad + (I - B)(\hat{\Omega}^{\text{opt}})^{-1/2} - (I - B)\Omega^{-1/2}\|_F^2 \\
&= \|(B - \hat{B}^{\text{opt}})((\hat{\Omega}^{\text{opt}})^{-1/2} - \Omega^{-1/2}) + (B - \hat{B}^{\text{opt}})\Omega^{-1/2} \\
&\quad + (I - B)((\hat{\Omega}^{\text{opt}})^{-1/2} - \Omega^{-1/2})\|_F^2 \\
&\leq \frac{4\lambda^2(s^\star)^2 + 4\lambda^2\|\Omega^{-1/2}\|_F^2 s^\star + 4\|B\|_F^2\lambda^2 s^\star + 4\lambda^2 s^\star}{\min\{1, \min_j(\Omega_{jj})^3\}} \\
&\leq \frac{16\max\{1, \|B\|_F^2, \|\Omega^{-1/2}\|_F^2\}\lambda^2 s^\star}{\min\{1, \min_j(\Omega_{jj})^3\}}.
\end{aligned}$$

A.3. *Proof of Theorem 2*

Recall that van de Geer & Bühlmann (2013) consider the equivalent optimization problem to (3) where Γ is parameterized by the connectivity matrix B and noise variance matrix Ω with the transformation $\Gamma = (I - B)\Omega^{-1/2}$. Appealing to Remark 3.2 of van de Geer & Bühlmann (2013), we have that under assumptions of Theorem 1 as well as Assumption 6, the graph encoded by any optimal connectivity matrix $\hat{B}^{\text{opt}}$ of this optimization problem encodes with probability $1 - \alpha_0$ a member of the Markov equivalence class of the population directed acyclic graph. Let $(B_{\text{mec}}^\star, \Omega_{\text{mec}}^\star)$ be the associated connectivity matrix and noise matrix of this population model. Furthermore, define $\Gamma_{\text{mec}}^\star = (I - B_{\text{mec}}^\star){\Omega_{\text{mec}}^\star}^{-1/2}$.

The proof of the theorem relies on the following lemma.

LEMMA A2. *Under the conditions of Theorem 2, we have with probability greater than* $1 - 2\alpha_0$*,*

$$\|\hat{\Gamma}^{\text{opt}} - \Gamma_{\text{mec}}^\star\|_F^2 = \mathcal{O}\left(\frac{m^2}{n}\right).$$

Proof of Lemma A*2.* We have by Remark 3.2 of van de Geer & Bühlmann (2013) that under the setting of Theorem 2, with probability greater than $1 - \alpha_0$, $\hat{\Gamma}^{\text{opt}}$ has the same support as $\Gamma_{\text{mec}}^\star$. Note that by optimality,

$$\ell(\hat{\Gamma}^{\text{opt}}) + \lambda^2\hat{s}^{\text{opt}} \leq \ell(\Gamma_{\text{mec}}^\star) + \lambda^2 s^\star.$$

Since $s^\star = \hat{s}^{\text{opt}}$, we have

$$\ell(\hat{\Gamma}^{\text{opt}}) - \ell(\Gamma^\star_{\text{mec}}) \leq 0.$$

For notational simplicity, we will consider a vectorized objective. With abuse of notation, let $\hat{\Gamma}^{\text{opt}}$ and $\Gamma^\star_{\text{mec}}$ be the vectorized form of their corresponding matrices. Then,

$$\ell(\hat{\Gamma}^{\text{opt}}) - \ell(\Gamma^\star_{\text{mec}}) = \nabla\ell(\Gamma^\star_{\text{mec}})^T(\hat{\Gamma}^{\text{opt}} - \Gamma^\star_{\text{mec}}) + \frac{1}{2}(\hat{\Gamma}^{\text{opt}} - \Gamma^\star_{\text{mec}})^T\nabla^2\ell(\tilde{\Gamma})(\hat{\Gamma}^{\text{opt}} - \Gamma^\star_{\text{mec}}).$$

Here, entries of $\tilde{\Gamma}$ lie between $\hat{\Gamma}^{\text{opt}}$ and $\Gamma^\star_{\text{mec}}$. Some algebra then gives:

$$(\hat{\Gamma}^{\text{opt}} - \Gamma^\star_{\text{mec}})^T\nabla^2\ell(\tilde{\Gamma})(\hat{\Gamma}^{\text{opt}} - \Gamma^\star_{\text{mec}}) \leq 2\|\nabla\ell(\Gamma^\star_{\text{mec}})\|_{\ell_2}\|\hat{\Gamma}^{\text{opt}} - \Gamma^\star_{\text{mec}}\|_{\ell_2}. \tag{A11}$$

Due to the convexity of the negative log-likelihood function, we have that, for any Γ,

$$\nabla^2\ell(\Gamma) \succeq \hat{\Sigma}_n \otimes I_m.$$

Recall from (A5) that $\|\hat{\Sigma}_n - \Sigma^\star\|_2 \leq \mathcal{O}\left(\sqrt{m/n}\right)$ with probability greater than $1-\alpha_0$. Then, with a probability greater than $1-\alpha_0$

$$\kappa_{\min}(\nabla^2\ell(\tilde{\Gamma})) \geq \kappa_{\min}(\hat{\Sigma}_n \otimes I_m) = \kappa_{\min}(\hat{\Sigma}_n) > \underline{\kappa} - \mathcal{O}\left(\sqrt{\frac{m}{n}}\right) \geq \underline{\kappa}/2. \tag{A12}$$

Based on Assumption 3, we have $\kappa_{\min}(\nabla^2\ell(\tilde{\Gamma})) > 0$. Using the inequality $(\hat{\Gamma}^{\text{opt}} - \Gamma^\star_{\text{mec}})^T\nabla^2\ell(\tilde{\Gamma})(\hat{\Gamma}^{\text{opt}} - \Gamma^\star_{\text{mec}}) \geq \kappa_{\min}(\nabla^2\ell(\tilde{\Gamma}))\|\hat{\Gamma}^{\text{opt}} - \Gamma^\star_{\text{mec}}\|^2_{\ell_2}$, from (A11), with probability greater than $1-2\alpha_0$

$$\|\hat{\Gamma}^{\text{opt}} - \Gamma^\star_{\text{mec}}\|_{\ell_2} \leq \frac{4\|\nabla\ell(\Gamma^\star_{\text{mec}})\|_{\ell_2}}{\underline{\kappa}} \leq \frac{4\|\nabla\ell(\Gamma^\star_{\text{mec}}) - \mathbb{E}[\nabla\ell(\Gamma^\star_{\text{mec}})]\|_{\ell_2}}{\underline{\kappa}} + \frac{4\mathbb{E}[\nabla\ell(\Gamma^\star_{\text{mec}})]\|_{\ell_2}}{\underline{\kappa}}. \tag{A13}$$

By (A10), $\|\nabla\ell(\Gamma^\star_{\text{mec}}) - \mathbb{E}[\nabla\ell(\Gamma^\star_{\text{mec}})]\|_{\ell_2} \leq \mathcal{O}(m/\sqrt{n})$. Furthermore, $\mathbb{E}[\nabla\ell(\Gamma^\star_{\text{mec}})] = 0$. Combining the results with (A13), we obtain $\|\hat{\Gamma}^{\text{opt}} - \Gamma^\star_{\text{mec}}\|^2_{\ell_2} = \mathcal{O}\left(m^2/n\right)$. Putting things back into matrix form, we have the desired result. □

We now complete the proof of Theorem 2.

Proof of Theorem 2. First, by Lemma A2, with probability greater than $1-2\alpha_0$

$$\|\hat{\Gamma}^{\text{early}} - \Gamma^\star_{\text{mec}}\|^2_F \leq 2\|\hat{\Gamma}^{\text{early}} - \hat{\Gamma}^{\text{opt}}\|^2_F + 2\|\hat{\Gamma}^{\text{opt}} - \Gamma^\star_{\text{mec}}\|^2_F = \text{GAP} + \mathcal{O}\left(\frac{m^2}{n}\right).$$

Since the GAP is on the order $\mathcal{O}\left(m^2/n\right)$, we get $\|\hat{\Gamma}^{\text{early}} - \Gamma^\star_{\text{mec}}\|^2_F = \mathcal{O}\left(m^2/n\right)$.

A.4. *Example of outer approximation*

In this section, we give a simple example to illustrate outer approximation. Consider the following integer programming problem:

$$\min_{x\in\mathbb{Z}_+} -2\log x + x,$$

which we know the optimal solution is $x = 2$. Replacing $-2\log x$ with y, the outer approximation works as follows. Starting from any feasible point, say $x^{(1)} = 4, y^{(1)} = -\infty$, we have $y^{(1)} < -2\log(x^{(1)})$, so we need to add the first cutting plane: $y \geq -2\log(4) - 0.5(x-4) = -0.5x + 2 - 2\log(4)$. Then, by solving

$$\min_{x\in\mathbb{Z}_+, y\in\mathbb{R}} y + x \text{ s.t. } y \geq -0.5x + 2 - 2\log(4),$$

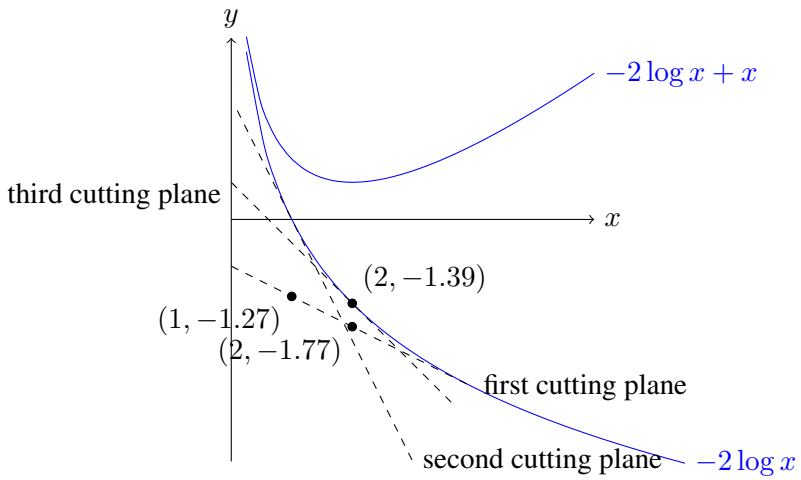


Fig. 3: Illustration of the outer approximation algorithm.

we obtain the solution $y^{(2)} = -1.27, x^{(2)} = 1$. Since $y^{(2)} < -2\log(x^{(2)})$, we need to add another cutting plane to get a better lower bound. By solving

$$\min_{x\in\mathbb{Z}_+, y\in\mathbb{R}} y + x \text{ s.t. } y \geq -0.5x + 2 - 2\log(4), y \geq -2x + 2,$$

we obtain the solution $y^{(3)} = -1.77, x^{(3)} = 2$. Since $y^{(3)} < -2\log(x^{(3)})$, we add the third cutting plane: $y \geq -x + 2 - 2\log(2)$, which leads to the solution $y^{(2)} = -2\log(2) = -1.39, x^{(2)} = 2$. Then the outer approximation algorithm stops.

As shown in Figure 3, the cutting planes are dynamically added to the problem, which offers two distinct advantages compared to the piece-wise linear approximation at predetermined breakpoints. First, we solve the problem exactly. Second, it is computationally more efficient by obviating the need for the a priori creation of a piece-wise linear approximation function with a large number of breakpoints to reduce the approximation error as is done by `Gurobi`'s general constraint attribute.

A.5. *Comparison with other benchmarks using estimated super-structure*

In Table 4, we present results where the moral graph is estimated from data using the graphical lasso (Friedman et al., 2007).

A.6. *Other comparison metrics*

In Tables 5 and 6 we report the structural hamming distances (SHD) of the undirected skeleton of the true directed acyclic graph and the corresponding skeleton of the estimated network, the true positive rate (TPR), and the false positive rate (FPR).

Table 4: Comparison of mixed-integer convex program and other benchmarks using estimated super-structure

	HD-BU		HU-TD		TD		MISOCP			OUR METHOD		
Network(m)	Time	d_{cpdag}	Time	d_{cpdag}	Time	d_{cpdag}	Time	RGAP	d_{cpdag}	Time	RGAP	d_{cpdag}
Dsep(6)	≤ 1	10.5	≤ 1	3.9	≤ 1	2.0	≤ 1	*	2.5	≤ 1	*	2.0
Asia(8)	≤ 1	18.0	≤ 1	14.5	≤ 1	12.1	≤ 1	*	11.4	≤ 1	*	2.2
Bowling(9)	≤ 1	5.6	≤ 1	8.5	≤ 1	6.0	≤ 1	*	5.3	3	*	2.0
InsSmall(15)	≤ 1	37.7	≤ 1	13.7	≤ 1	9.7	4	*	8.2	≥ 750	.080	7.0
Rain(14)	≤ 1	18.1	≤ 1	9.7	≤ 1	3.6	1	*	2.0	151	*	2.0
Cloud(16)	≤ 1	41.7	≤ 1	30.4	≤ 1	20.5	≤ 1	*	19.4	93	*	5.2
Funnel(18)	≤ 1	16.3	≤ 1	11.4	≤ 1	2.0	≤ 1	*	2.7	70	*	2.0
Galaxy(20)	≤ 1	45.7	≤ 1	36.1	≤ 1	29.1	≤ 1	*	11.6	237	*	1.0
Insurance(27)	1	45.2	2	46.1	≤ 1	34.6	≥ 1350	.083	18.1	≥ 1350	.340	22.8
Factors(27)	1	25.1	2	49.0	≤ 1	48.1	1073	.018	30.1	≥ 1350	.311	56.1
Hailfinder(56)	9	120.1	10	89.0	≤ 1	55.2	≥ 2800	.097	44.6	≥ 2800	.245	41.4
Hepar2(70)	20	126.6	22	82.4	2	68.4	≥ 3500	.128	57.0	≥ 3500	5.415	76.9

HD-BU, high-dimensional bottom-up; HD-TD, high-dimensional top-down; TD, top-down; MISOCP, mixed-integer second-order cone program; d_{cpdag}, differences between the true and estimated completed partially directed acyclic graphs; RGAP, relative optimality gap;

Table 5: SHDs/TPR/FPR results part I

	HD-BU			HU-TD			TD		
Network(m)	SHDs	TPR	FPR	SHDs	TPR	FPR	SHDs	TPR	FPR
Dsep(6)	4.2	.983	.137	1.5	.900	.030	1.0	.833	0
Asia(8)	10.5	.900	.173	6.7	.900	.105	4.9	.875	.070
Bowling(9)	3.6	.909	.037	4.4	.909	.049	2.0	.909	.014
InsSmall(15)	20.8	.956	.099	5.9	.960	.025	2.1	.952	.005
Rain(14)	9.6	.950	.049	5.0	.956	.024	1.7	.944	.004
Cloud(16)	20.5	.926	.081	14.9	.921	.057	7.1	.916	.023
Funnel(18)	10.3	.950	.031	7.1	.950	.020	1.0	.944	0
Galaxy(20)	25.1	.955	.064	16.9	.959	.042	10.1	.955	.024
Insurance(27)	26.1	.990	.038	27.1	.967	.038	17.7	.950	.022
Factors(27)	9.6	.968	.011	22.4	.869	.020	22.2	.768	.010
Hailfinder(56)	78.8	.973	.025	49.7	.983	.016	29.3	.921	.008
Hepar2(70)	85.9	.985	.018	53.6	.982	.011	38.5	.856	.004

HD-BU, high-dimensional bottom-up; HD-TD, high-dimensional top-down; TD, top-down; SHDs, structural hamming distance of undirected graph skeletons; TPR, true positive rate; FPR, false positive rate;

Table 6: SHDs/TPR/FPR results part II

	MISOCP-True			OUR METHOD-True			MISOCP-Est			OUR METHOD-Est		
Network(m)	SHDs	TPR	FPR	SHDs	TPR	FPR	SHDs	TPR	FPR	SHDs	TPR	FPR
Dsep(6)	0.9	.833	.022	1.0	.833	0	0.8	.833	.033	1.0	.833	0
Asia(8)	2.5	.875	.225	1.0	.875	0	2.9	.875	.255	1.0	.875	0
Bowling(9)	2.0	.909	.116	1.0	.909	0	1.9	.909	.108	1.0	.909	0
InsSmall(15)	1.0	.960	.005	2.5	.900	0	1.2	.960	.013	2.6	.908	.004
Rain(14)	1.0	.944	.041	1.0	.944	0	1.0	.944	.041	1.0	.944	.014
Cloud(16)	2.0	.947	.069	1.9	.900	0	3.3	.916	.079	2.1	.895	.001
Funnel(18)	1.3	.944	.004	1.0	.944	0	1.5	.944	.004	1.0	.944	0
Galaxy(20)	2.9	.955	.032	1.0	.955	0	3.4	.945	.039	1.0	.955	0
Insurance(27)	3.8	.973	.017	5.5	.954	.010	4.8	.962	.021	6.7	.937	.011
Factors(27)	12.4	.906	.041	19.1	.776	.014	12.3	.887	.034	22.8	.725	.014
Hailfinder(56)	8.6	.980	.010	10.0	.955	.005	14.8	.962	.018	12.6	.942	.006
Hepar2(70)	12.3	.985	.010	12.2	.954	.003	17.9	.977	.012	20.4	.936	.005

SHDs, the structural hamming distance of undirected graph skeletons; TPR, true positive rate; FPR, false positive rate; MISOCP, mixed-integer second-order conic program; True, using the true moral graph as superstructure; Est, using estimated moral graph;