

# CONVEX MIXED-INTEGER PROGRAMMING FOR CAUSAL ADDITIVE MODELS WITH OPTIMIZATION AND STATISTICAL GUARANTEES

XIAOZHU ZHANG <sup>1</sup>, NIR KERET <sup>2</sup>, ALI SHOJAIE <sup>1,2</sup>, AND ARMEEN TAEB <sup>1</sup>

**ABSTRACT.** We study the problem of learning a directed acyclic graph from data generated according to an additive, non-linear structural equation model with Gaussian noise. We express each non-linear function through a basis expansion, and derive a maximum likelihood estimator with a group  $\ell_0$ -regularization that penalizes the number of edges in the graph. The resulting estimator is formulated through a convex mixed-integer program, enabling the use of branch-and-bound methods to obtain a solution that is guaranteed to be accurate up to a pre-specified optimality gap. Our formulation can naturally encode background knowledge, such as the presence or absence of edges and partial order constraints among the variables. We establish consistency guarantees for our estimator in terms of graph recovery, even when the number of variables grows with the sample size. Additionally, by connecting the optimality guarantees with our statistical error bounds, we derive an early stopping criterion that allows terminating the branch-and-bound procedure while preserving consistency. Compared with existing approaches that either assume equal error variances, restrict to linear structural equation models, or rely on heuristic procedures, our method enjoys both optimization and statistical guarantees. Extensive simulations and real-data analysis show that the proposed method achieves markedly better graph recovery performance.

**Keywords:** causal discovery, directed acyclic graphs, discrete optimization, generalized additive models,  $\ell_0$ -penalization, non-linear structural equation models

## 1. INTRODUCTION

Understanding causal relationships is arguably the ultimate aim of science. It enables us to predict how a system will behave under external interventions—a crucial step toward both understanding and engineering that system. Directed acyclic graphs (DAGs) offer a convenient and powerful framework for modeling causal relationships. Although in some cases prior knowledge and expert intuition help suggest likely causal models, in many cases, the underlying causal structure is unknown and must be inferred from data. The task of inferring a DAG from observational or experimental data is known as *causal discovery*. The learned DAG then serves as a basis for making causal inferences.

Many methods in the causal discovery literature assume that the relationships among the variables are linear (Chen et al., 2019, Chickering, 2002, Ghoshal and Honorio, 2019, Shimizu et al., 2006, Wang and Drton, 2020, Xu et al., 2025a); see also the survey paper (Glymour et al., 2019). Such methods are naturally limited by their inability to capture non-linear effects. This paper focuses on causal discovery when relationships among variables may be non-linear.

Causal discovery in non-linear settings is commonly performed using a combinatorial search for a DAG structure, either via constraint-based or score-based methods. Constraint-based methods identify conditional independencies from the data. An example is the PC algorithm (Spirtes et al.,

---

<sup>1</sup> DEPARTMENT OF STATISTICS, UNIVERSITY OF WASHINGTON

<sup>2</sup> DEPARTMENT OF BIOSTATISTICS, UNIVERSITY OF WASHINGTON

*E-mail addresses:* xzzhang@uw.edu, nirkkeret@uw.edu, ashoojaie@uw.edu, ataeb@uw.edu.

2000) which initiates with a complete undirected graph and iteratively removes edges based on conditional independence assessments. However, conditional independence testing is known to be hard in non-linear settings (Shah and Peters, 2020); indeed, existing approaches require complicated tests (Chakraborty and Shojaie, 2022, Heinze-Deml et al., 2018, Strobl et al., 2019, Zhang et al., 2011), with low power and poor finite-sample performance. Additionally, constraint-based methods require a condition known as “strong-faithfulness”, which is known to be restrictive (Sondhi and Shojaie, 2019, Uhler et al., 2013). Score-based methods, which is the approach considered in this paper, often use penalized log-likelihood as a score function to seek the optimal graph within the entire space of DAGs. Unlike constraint-based methods, score-based approaches do not require the “strong-faithfulness” assumption.

A prominent example of a score-based method in non-linear settings assumes a *causal additive model* (CAM) where the non-linear functions as well as the noise terms are of additive form (Bühlmann et al., 2014). CAMs, which build on generalized additive models in standard regression (Hastie and Tibshirani, 1986), are quite expressive and can automatically model non-linear relationships that linear models miss. Furthermore, since the functional forms are non-linear and the noise is additive, the underlying DAG structure is identifiable from observational data (Hoyer et al., 2008, Peters et al., 2014). With this powerful model and under Gaussian noise, Bühlmann et al. (2014) adopt a multi-step approach to estimate a DAG: they obtain a topological ordering of the variables by maximizing a likelihood score over the space of permutations, and then perform variable selection with respect to the ordering. They prove that the first step of their algorithm is asymptotically consistent, provided that they can solve the optimization problem exactly. However, they deploy a heuristic greedy algorithm (called *IncEdge*) with no optimality guarantees.

Indeed, Gao et al. (2020) prove that the solution produced by the heuristic *IncEdge* algorithm in Bühlmann et al. (2014) can be inconsistent, highlighting a discrepancy between their optimization and statistical guarantees. To address this issue, for general non-linear models, Gao et al. (2020) propose a polynomial-time algorithm for estimating a topological ordering with statistical guarantees. Similar to Bühlmann et al. (2014), in a second step, they perform variable selection to obtain a DAG. Their approach relies critically on a condition on the residual variances. This assumption is closely related to assumptions made in prior work on linear models with equal noise variances (Chen et al., 2019, Ghoshal and Honorio, 2019, Peters and Bühlmann, 2013). As we illustrate next, when this condition is violated, the accuracy of the estimated graph deteriorates.

**Example 1.** Consider the following structural equation model (SEM):

$$(1) \quad \begin{cases} X_1 \sim \mathcal{N}(0, 0.5^2), \\ X_2 = X_1^2 - 0.25 + z_2, & z_2 \sim \mathcal{N}(0, \sigma_2^2), \\ X_3 = 2X_1^2 - 0.5 + h(X_2) - \mathbb{E}[h(X_2)] + z_3, & z_3 \sim \mathcal{N}(0, \sigma_3^2), \\ X_4 \sim \mathcal{N}(0, 0.5^2), & X_5 \sim \mathcal{N}(0, 0.5^2), \end{cases}$$

where  $X_1, z_2, z_3, X_4, X_5$  are mutually independent. We consider two scenarios: (i) equal noise variances, in which  $\sigma_2 = \sigma_3 = 0.5$ , ensuring that the conditions required for the approach in Gao et al. (2020) are satisfied; and (ii) unequal noise variances, in which  $\sigma_2 = 0.1$  and  $\sigma_3 = 0.3$ , thereby violating those conditions. In each setting, we generate  $n$  samples according to the SEM using five different functions for  $h(x)$ :  $h(x) = \sin(x)$ ,  $h(x) = 0.5x^3$ ,  $h(x) = \arctan(x^2)$ ,  $h(x) = |x|$ , and  $h(x) = \exp(x)$ . We apply the procedures in Bühlmann et al. (2014) (denoted CAM-*IncEdge*) and Gao et al. (2020) (denoted NPVAR) across all the ten settings, and evaluate the accuracy of their graph estimates, over 100 independent trials. We also apply the mixed-integer programming method in Xu et al. (2025a) (denoted MIP-linear), which was designed for optimal causal discovery in linear settings, to highlight

the benefits of deploying non-linear models. In Figure 1, we display the results, in terms of the probability of exact recovery and structural Hamming distance (SHD), for varying sample size  $n$ . We observe that CAM-IncEdge behaves unstably across different functions  $h(x)$  under the equal-variance setting. Moreover, regardless of the sample size, both CAM-IncEdge and NPVAR cannot estimate the correct graph in the unequal-variance setting. In fact, as the sample size increases, their performance does not seem to improve. Our mixed-integer programming approach (MIP-nonlinear), which provides both optimization and statistical guarantees and accounts for nonlinearities, consistently outperforms the other methods across different choices of  $h(x)$  when the sample size  $n$  is sufficiently large. In particular, it is the only approach that successfully recovers the correct graph under the unequal-variance setting. Furthermore, as we see in Section 6.2, a larger number of variables can result in greater deterioration in the performance of CAM and NPVAR (under unequal variances).

There are score-based approaches that do not perform a discrete search over DAGs. Rather, they relax the discrete search space to a continuous search space, allowing gradient descent and other techniques from continuous optimization to be applied to causal discovery (Yu et al., 2019, Zheng et al., 2018). However, the search space of these problems is highly non-convex, so that the optimization procedure may become stuck in a local minima (Wei et al., 2020). Indeed, there are indications that the performance relies on an artifact of the synthetic evaluation data, and can be matched by a simple algorithm that directly exploits it (Reisach et al., 2021). Furthermore, Seng et al. (2022) found this reliance to be a vulnerability to adversarial attacks. For additional discussion on non-linear causal discovery, we refer the reader to surveys by Peters et al. (2017) and Glymour et al. (2019).

**1.1. Our contributions.** We propose a mixed-integer programming (MIP) formulation for score-based causal discovery in causal additive models with Gaussian errors. Each non-linear function is parameterized as a linear combination of a fixed (possibly large) set of basis functions. The score function combines a convex negative log-likelihood with a sparsity-inducing regularizer that penalizes the number of edges in the DAG implied by the model parameters. This regularizer can be expressed as a group  $\ell_0$ -penalty on the model parameters, where each group corresponds to the coefficients of a non-linear basis expansion. Our estimator minimizes the score subject to DAG constraints, enforced through a layered-network formulation involving both continuous and integer variables (Manzour et al., 2021). In our layered-network formulation, the group  $\ell_0$ -penalty admits a linear representation in the integer variables and thus introduces no additional computational burden. Setting the basis functions to be linear recovers the estimator of Xu et al. (2025a), which was developed for linear causal discovery. Incorporating non-linear basis functions allows our method to capture non-linear dependencies that the previous approach cannot.

Our MIP framework is different from previous work (Bühlmann et al., 2014, Gao et al., 2020) in several key ways. First, it enables the use of branch-and-bound methods to obtain a solution that is guaranteed to be accurate up to a pre-specified optimality gap. Second, unlike the method in Gao et al. (2020), it does not assume homoscedastic noise (i.e., equal noise variance). Third, it involves a joint optimization over all model parameters, including a sparse DAG structure via a group  $\ell_0$ -regularization, as opposed to a two-step procedure that first estimates a topological ordering and then performs variable selection. Finally, unlike other methods, background knowledge—such as the presence or absence of certain edges and partial order constraints among variables—can be organically incorporated as linear constraints in our optimization framework.

We establish consistency guarantees for our estimator, allowing for the number of variables to grow with the sample size. Our analysis combines, and substantially extends, techniques from Bühlmann et al. (2014) and van de Geer and Bühlmann (2013). The approach in Bühlmann et al. (2014) addresses only the consistency of topological ordering and does not accommodate regularization, whereas the analysis in van de Geer and Bühlmann (2013) is restricted to linear models and cannot handle

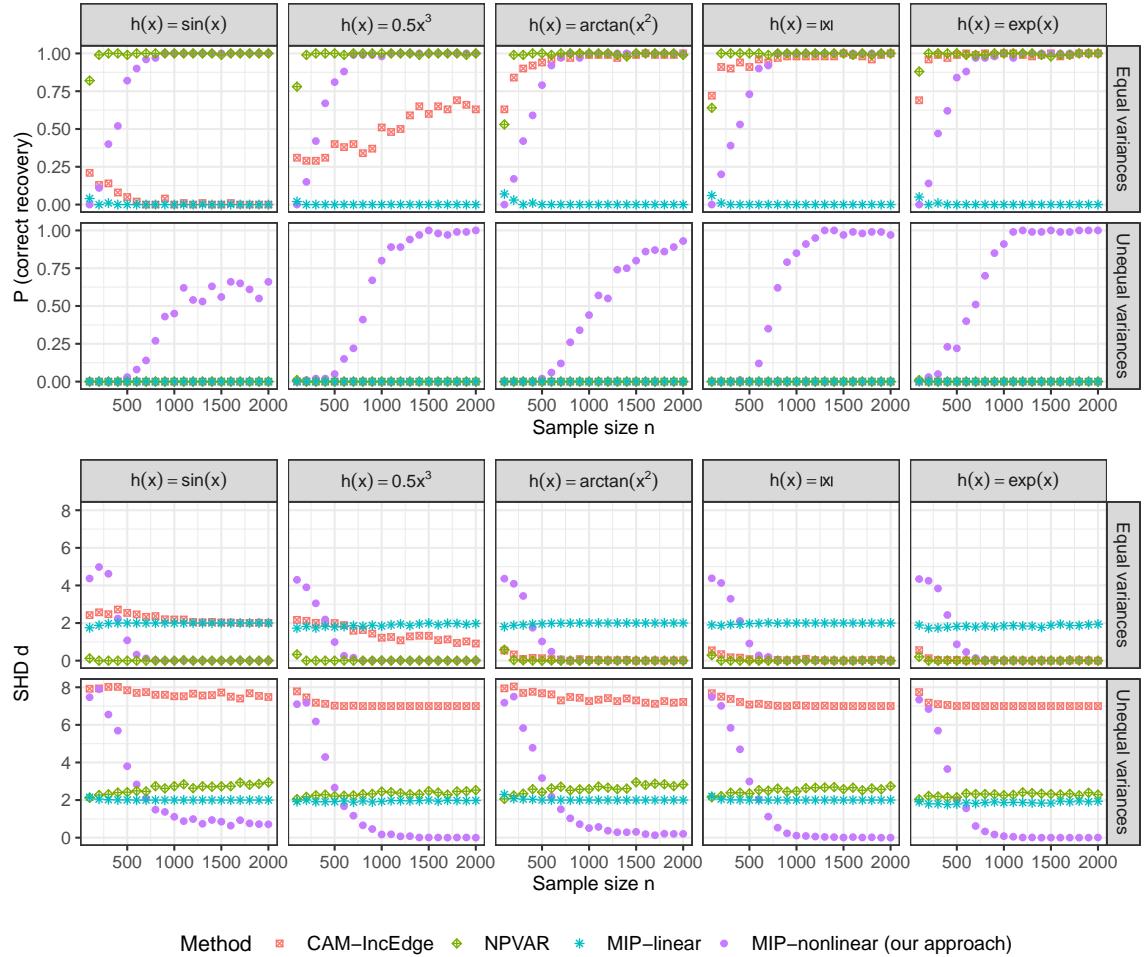


FIGURE 1. Graph recovery performance of existing methods and our proposed approach. Top: the probability of recovering the true graph. Bottom: the structural Hamming distance (SHD)  $d$  between the estimated graph and the true graph, defined as the number of incorrectly recovered or missed edges. The function  $h(x)$  corresponds to the one used in model (1). Implementation details are provided in Appendix H.1.

non-linear settings. Our work integrates these threads to handle both challenges simultaneously. Furthermore, by connecting the optimality and statistical guarantees of our estimator, we present an early stopping criterion under which we can terminate the branch-and-bound procedure while retaining both optimality and statistical consistency guarantees.

We develop several computational strategies to ensure that our approach scales to moderate-sized graphs (around 30 nodes) while retaining optimization guarantees. First, we re-parameterize the negative-log likelihood function to be convex. This ensures that the resulting MIP is convex once the integer constraints are relaxed, making it amenable to tools from the integer programming literature that provide optimality guarantees. Second, we identify a sufficient statistic and reformulate the

optimization problem in terms of it, eliminating the need to iterate over the full dataset each time the objective is evaluated. Third, as discussed above, we apply early stopping without sacrificing statistical guarantees. Finally, we restrict the search space using constraints computed prior to optimization: (i) a super-structure constraint, where we use efficient neighborhood selection methods to estimate an undirected graphical model (Meier et al., 2009, Ravikumar et al., 2009, Voorman et al., 2014) and restrict candidate edges to those in this graph; (ii) an edge-presence constraint, where we run an existing causal discovery method (e.g., Bühlmann et al., 2014, Gao et al., 2020) on many bootstrap samples and retain edges that appear with high stability; and (iii) a partial-order constraint, where we enforce topological ordering relations that are stable across the bootstrap estimates. All of these constraints can be imposed as linear constraints in our optimization program.

Finally, we demonstrate our method using both synthetic and real data from a tightly controlled physical system. First, we show that the computational strategies introduced earlier substantially improve efficiency while largely preserving accuracy. Second, in synthetic experiments for non-linear models, we observe that our approach consistently outperforms existing methods, almost always achieving the smallest structural Hamming distance under both homoscedastic and heteroscedastic noise. This advantage is observed across different true functions and choices of basis functions, and can be further enhanced by incorporating richer basis expansions when the underlying non-linear effects are more complex. Lastly, our method attains the most accurate recovery of an underlying DAG in a real physical system.

The paper is organized as follows. Section 2 describes the problem setup; Section 3 presents our MIP formulation; Section 4 outlines computational strategies for accelerating the method; Section 5 provides consistency guarantees; and Section 6 demonstrates the utility of our approach on synthetic and real data. Section 7 discusses future work.

**1.2. Notation.** Let  $V := \{1, \dots, p\}$  and  $\mathcal{E} := V \times V \setminus \{(j, j) : j \in V\}$ . A DAG  $\mathcal{G} = (V, E)$  over  $p$  nodes consists of a node set  $V$  and a directed edge set  $E \subseteq \mathcal{E}$ , such that the graph contains no directed cycles. Here, we use  $(k, j) \in E$  to indicate that  $k \rightarrow j$  in the DAG  $\mathcal{G}$ . The sparsity level of a DAG is defined as its number of edges:  $|\mathcal{G}| := |E|$ . The smallest eigenvalue of a matrix  $A$  is denoted as  $\lambda_{\min}(A)$ . For any deterministic vector  $x \in \mathbb{R}^n$ , we define its  $n$ -norm as  $\|x\|_n := \|x\|_2/\sqrt{n}$ . For a random variable  $X \in \mathbb{R}$  with  $n$  independent realizations  $\{x_1, \dots, x_n\}$ , we define its empirical  $n$ -norm as  $\|X\|_n := (n^{-1} \sum_{i=1}^n x_i^2)^{1/2}$ . We use  $X_j$ ,  $j = 1, \dots, p$ , to denote the random variable corresponding to the  $j$ -th node. Unless otherwise specified (e.g., in certain proofs), the notation  $X_j$ ,  $\epsilon_j$ , and their linear combinations refer to real-valued random variables, rather than vectors of their sample realizations. Lastly, for two sequences  $a_n$  and  $b_n$ , we write  $a_n \lesssim b_n$  if there exists a constant  $C > 0$  and  $n_0 \in \mathbb{N}$  such that  $a_n \leq Cb_n$  for all  $n \geq n_0$ . We write  $a_n \asymp b_n$  if both  $a_n \lesssim b_n$  and  $a_n \gtrsim b_n$  hold. We write  $a_n = o(b_n)$  if  $a_n/b_n \rightarrow 0$  as  $n \rightarrow \infty$ .

## 2. PROBLEM SETUP

**2.1. Model.** Consider an unknown oracle DAG  $\mathcal{G}^* := (V, E^*)$ , where  $V = \{1, \dots, p\}$  is the node set and  $E^* \subseteq \mathcal{E}$  is the true set of directed edges. The  $p$  nodes of  $\mathcal{G}^*$  correspond to observed random variables  $X \in \mathbb{R}^p$ . We assume for each node  $j \in V$ , the variable  $X_j \in \mathbb{R}$  satisfies a structural equation model (Lauritzen, 1996):

$$X_j = f_j(X_{\text{pa}(j)}, \epsilon_j), \quad \epsilon_1, \dots, \epsilon_p \text{ mutually independent},$$

where  $\text{pa}(j) := \{k : (k, j) \in E^*\}$  denotes the set of parents for node  $j$  in the graph  $\mathcal{G}^*$ . The independence of the error terms ensure that there is no latent confounding. We primarily focus on the setting where  $f_j(\cdot)$  is a non-linear function. Importantly, we assume that  $f_j(X_{\text{pa}(j)}, \epsilon_j)$  is additive in

each of its arguments, allowing the noise and the causal effects from each parental node to be decoupled from one another. This is known as a causal additive model (Bühlmann et al., 2014, Kertel and Klein, 2025, Maeda and Shimizu, 2021), which applies the principles of generalized additive models (Hastie and Tibshirani, 1986) to causal settings. We also model each noise variable as Gaussian  $\epsilon_j \sim \mathcal{N}(0, \sigma_j^{*2})$  with non-degenerate variances  $\sigma_j^* > 0$  ( $j = 1, \dots, p$ ). This specification enables us to formulate the distribution of  $X$ , which is crucial for establishing identifiability and a score-based estimator. We arrive at an additive structural equation model with Gaussian errors:

$$(2) \quad \begin{aligned} X_j &= \sum_{k \in \text{pa}(j)} f_{kj}^*(X_k) + \epsilon_j, \\ \epsilon_j &\sim \mathcal{N}(0, \sigma_j^{*2}), \quad \sigma_j^* > 0, \quad \mathbb{E}[f_{kj}^*(X_k)] = 0, \quad \forall j, k \in V, \end{aligned}$$

where  $f_{kj}^*(\cdot)$  is a function from  $\mathbb{R} \rightarrow \mathbb{R}$ . Note that  $f_{kj}^*(\cdot) \not\equiv 0$  if and only if  $k \in \text{pa}(j)$ . Without loss of generality, each additive component  $f_{kj}^*(X_k)$  has mean zero; see Appendix A for additional discussion. We denote by  $P$  the joint distribution induced by the structural equation model in (2):

$$(X_1, \dots, X_p) \sim P.$$

When the functions  $f_{kj}^*$  are linear, the model (2) reduces to a linear Gaussian structural equation model. In that setting, identifiability is a challenge: multiple structural equation models may induce the same distribution  $X$ , making it impossible to distinguish the true underlying structural equation model based solely on  $P$  (van de Geer and Bühlmann, 2013). However, when the functions  $f_{kj}^*$  are non-linear and, notably, smooth enough,  $\mathcal{G}^*$  is the only DAG that is compatible  $P$ :

**Proposition 2** (Theorem 1 in Hoyer et al. (2008), Corollary 30 in Peters et al. (2014), Lemma 1 in Bühlmann et al. (2014)). *Suppose all non-zero functions  $\{f_{kj}^*\}_{(k,j) \in \mathcal{E}}$  are non-linear and three times differentiable. Then any distribution  $Q$  that is generated by model (2) with a different DAG  $\mathcal{G}' \neq \mathcal{G}^*$  and non-constant, three times differentiable functions  $\{f'_{kj}\}$  differs from  $P$ .*

Proposition 2 reveals that the graph  $\mathcal{G}^*$  is guaranteed to be *identifiable* from the distribution  $P$  when non-zero functions  $f_{kj}^*$  are assumed to be non-linear and three-times differentiable. We adopt this assumption throughout the paper. This identifiability enables and underpins our goal: to recover the true DAG  $\mathcal{G}^*$ .

**2.2. The function class.** We formalize the class of functions to which the functions  $\{f_{kj}^*\}$  belong. For any  $(k, j) \in \mathcal{E}$ , we consider a set of basis functions  $\{b_{rkj}(\cdot)\}_{r=1}^\infty$ , where each function is three times differentiable and has infinite support. Let

$$\mathcal{F}_{kj} := \left\{ f_{kj} : f_{kj}(\cdot) = \beta_{0kj} + \sum_{r=1}^\infty \beta_{rkj} b_{rkj}(\cdot), \quad \beta_{0kj} = -\mathbb{E} \left[ \sum_{r=1}^\infty \beta_{rkj} b_{rkj}(X_k) \right] \right\},$$

where the choice of  $\beta_{0kj}$  ensures zero mean of  $f_{kj}(X_k)$ . We assume throughout that the non-linear true functions belong to these function classes:  $f_{kj}^* \in \mathcal{F}_{kj}$  for all  $k$  and  $j$ . Hence,

$$X_j = \sum_{k \in \text{pa}(j)} \beta_{0kj}^* + \sum_{r=1}^\infty \beta_{rkj}^* b_{rkj}(X_k) + \epsilon_j.$$

In practice, estimators are fitted using a finite collection of basis functions. With  $n$  samples, we further consider the space

$$\mathcal{F}_{kj,n} := \{f_{kj,n} \in \mathcal{F}_{kj} : \beta_{rkj} = 0, \quad \forall r \geq R_n + 1\} \subseteq \mathcal{F}_{kj},$$

where  $R_n$  is a truncation parameter that increases with  $n$ . In other words, estimation is carried out over a subset of the full function class, using only the first  $R_n$  basis functions, and increasing the

sample size expands the search space over which estimation is performed. In Section 5, in order to facilitate theoretical analysis, we focus on slightly more restricted function classes than  $\mathcal{F}_{kj}$  and  $\mathcal{F}_{kj,n}$  by imposing additional conditions on the basis functions and coefficients.

**2.3. An intractable group  $\ell_0$ -penalized maximum likelihood estimator.** We assume we have  $n$  independent and identically distributed samples of  $X$  generated according to model (2) with  $f_{kj}^* \in \mathcal{F}_{kj}$ . Denote by  $\theta$  the parameter with additive functions and error variances:  $\theta := (\{f_{kj}\}_{(k,j) \in \mathcal{E}}, \{\sigma_j\}_{j \in V})$ . Accordingly, we define  $\mathcal{G}(\theta) := (V, E(\theta))$  as the graph induced by  $\theta$ , where  $(k, j) \in E(\theta)$  if and only if  $f_{kj} \neq 0$ . Then the negative log-likelihood of model (2) parameterized by  $\theta$  is proportional to

$$(3) \quad \ell_n(\theta) := \sum_{j=1}^p \log \sigma_j^2 + \sum_{j=1}^p \frac{\left\|X_j - \sum_{k \neq j} f_{kj}(X_k)\right\|_n^2}{\sigma_j^2}.$$

The  $\|\cdot\|_n$  norm is defined in Section 1.2. Naturally, it is desirable to identify a model that not only fits the data well, as measured by a low negative log-likelihood, but also exhibits structural simplicity through a sparse DAG. To this end, we penalize the negative log-likelihood  $\ell_n(\theta)$  by the number of edges in the DAG, resulting in an  $\ell_0$ -penalized maximum likelihood estimator with the regularization parameter  $\lambda_n \geq 0$ :

$$(4) \quad \operatorname{argmin}_{\theta \in \Theta} \ell_n(\theta) + \lambda_n^2 \cdot |\mathcal{G}(\theta)|.$$

Here  $\Theta = \{\theta : f_{kj} \in \mathcal{F}_{kj,n}, \sigma_j > 0, \forall (k, j) \in \mathcal{E}, \mathcal{G}(\theta) \text{ is a DAG}\}$  is a non-convex constraint set. An equivalent formulation of the model is via the parameter  $\tilde{\theta} := (\{\beta_{rkj}\}_{(k,j) \in \mathcal{E}, r \in [R_n]}, \{\sigma_j\}_{j \in V})$ , where  $f_{kj}(\cdot) = \beta_{0kj} + \sum_{r=1}^{R_n} \beta_{rkj} b_{rkj}(\cdot)$ . Under this parameterization, the penalty term  $|\mathcal{G}(\theta)|$  is equivalent to a group  $\ell_0$ -penalty  $\sum_{(k,j) \in \mathcal{E}} \mathbb{1}\{\|\beta_{kj}\|_2 \neq 0\}$ , where  $\beta_{kj} = (\beta_{1kj}, \dots, \beta_{R_n kj}) \in \mathbb{R}^{R_n}$ . This leads to the following equivalent formulation of (4)

$$(5) \quad \hat{\theta}_n \in \operatorname{argmin}_{\tilde{\theta} \in \tilde{\Theta}} \ell_n(\tilde{\theta}) + \lambda_n^2 \sum_{(k,j) \in \mathcal{E}} \mathbb{1}\{\|\beta_{kj}\|_2 \neq 0\}.$$

Here,  $\tilde{\Theta} = \{\tilde{\theta} : \sigma_j > 0, \forall k \in V, \mathcal{G}(\tilde{\theta}) \text{ is a DAG}\}$ , where  $\mathcal{G}(\tilde{\theta}) = (V, E(\tilde{\theta}))$  with  $(k, j) \in E(\tilde{\theta})$  if  $\|\beta_{kj}\|_2 \neq 0$ .

The optimization problem (4) can be re-written as  $\operatorname{argmin}_{\theta \in \Theta} \sum_{j=1}^p \log \omega_j + \lambda_n^2 \cdot |\mathcal{G}(\theta)|$ , where  $\omega_j = \|X_j - \sum_{k \neq j} f_{kj}(X_k)\|_n^2$ . Disregarding the penalty term, this formulation closely resembles the maximum likelihood estimator proposed by Bühlmann et al. (2014), which seeks the topological ordering that minimizes  $\sum_{j=1}^p \log \omega_j$ . However, due to the enormous permutation space, the search procedure in *IncEdge* is carried out heuristically, greedily adding the edge that yields the largest increase in log-likelihood at each step. Consequently, it offers no optimality guarantees. Furthermore, this maximum likelihood estimation step alone does not produce a sparse graph; an additional pruning step is required to obtain sparsity given an estimated ordering. Such a multi-step procedure can accumulate errors. In contrast, our proposed  $\ell_0$ -penalized estimation process integrates variance estimation, permutation estimation, and graph recovery in a joint optimization procedure. We present methods for solving the optimal solution of (5), thereby improving the reliability of the recovered graph structure. The experiments in Section 6.2 empirically demonstrate the strong graph-recovery performance of our method compared with the method in Bühlmann et al. (2014).

The  $\ell_0$ -penalty has been widely used for causal discovery problem (Chickering, 2002, Hauser and Bühlmann, 2012, Silander and Myllymäki, 2006, van de Geer and Bühlmann, 2013, Xu et al., 2025a). When multiple basis functions are used to represent a single causal effect between two nodes, the group  $\ell_0$ -penalty arises as a natural extension. The group  $\ell_0$ -penalty directly counts the number of nonzero

edges in the graph and thus encourages the minimal edge set needed to characterize the distribution of  $X$ . This would not be true when choosing a group  $\ell_1$ -penalty:  $\sum_{(k,j) \in \mathcal{E}} \|\beta_{kj}\|_2$ . Although an  $\ell_1$ -penalty can be preferred in general for its convexity, in our setting, the constraint set  $\tilde{\Theta}$  still remains non-convex. Furthermore, as we will see in Section 3.2, when formulating the constraint set  $\tilde{\Theta}$  via integer programming, the group  $\ell_0$ -penalty admits a linear representation in the integer variables, and can thus be incorporated naturally. Figure 2 presents the regularization paths for when the regularization term is a group  $\ell_0$ -penalty—i.e., our formulation in (4)—and when the regularization term is swapped with a group  $\ell_1$ -penalty. The figure highlights that group  $\ell_0$ -penalty empirically achieves superior graph recovery.

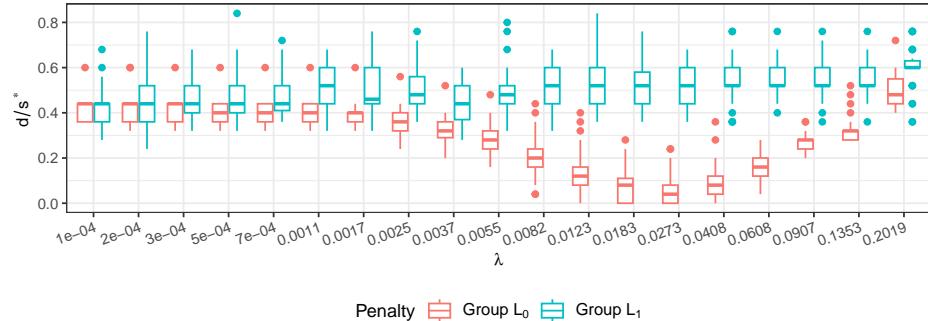


FIGURE 2. The performance of the group  $\ell_0$ -estimator and group  $\ell_1$ -estimator across a range of tuning parameter values  $\lambda_n$ . The performance is measured by  $d/s^*$ , where  $d$  indicates the number of incorrectly recovered and missed edges, and  $s^*$  indicates the number of true edges. A smaller  $d/s^*$  indicates better accuracy. The implementation details are provided in Appendix H.2.

Finally, as stated, both the objective function and the constraint set of (4) are non-convex, rendering this optimization problem generally intractable. In the next section, we reformulate our problem using a convex MIP framework to enable efficient computation.

### 3. CONVEX MIXED INTEGER PROGRAMMING FORMULATION

**3.1. An efficient log-likelihood using a sufficient statistic.** The negative log-likelihood function  $\ell_n(\theta)$ , as formulated in (3), can be cumbersome and computationally intensive to evaluate. We provide two main reasons. First,  $\ell_n(\theta)$  is not necessarily a convex function of  $\theta$ , which may hinder reaching the optimal solution and can substantially slow down standard optimization solvers, such as **Gurobi**. Second, the formulation (3) involves computing the empirical norm  $\|\cdot\|_n$ , which necessitates iterating over the entire dataset each time the objective function is evaluated.

We re-parameterize and reformulate the negative log-likelihood function as follows. Recall that for each  $f_{kj} \in \mathcal{F}_{k,j,n}$ , the function  $f_{kj}(\cdot) = \beta_{0kj} + \sum_{r=1}^{R_n} \beta_{rkj} b_{rkj}(\cdot)$  is a linear combination of  $R_n$  basis functions. Without loss of generality, we further assume that the same set of basis functions  $\{b_r(\cdot)\}_{r=1}^{R_n}$  is used across all edges  $(k, j) \in \mathcal{E}$ , so that each  $b_{rkj}(\cdot)$  reduces to  $b_r(\cdot)$ . Let  $Z \in \mathbb{R}^{pR_n+p+1}$  be an extended basis random vector:

$$Z := [X_1, \dots, X_p, \ 1, \ b_1(X_1), \dots, b_1(X_p), \ \dots, \ b_{R_n}(X_1), \dots, b_{R_n}(X_p)]^\top.$$

We then define a scaled coefficient matrix  $\Gamma \in \mathbb{R}^{p \times (pR_n + p + 1)}$  accordingly:

$$(6) \quad \Gamma := \text{diag}(\sigma_1, \dots, \sigma_p)^{-1} \cdot \begin{bmatrix} I_{p \times p} & -B_{p \times 1}^{(0)} & -B_{p \times p}^{(1)} & \cdots & -B_{p \times p}^{(R_n)} \end{bmatrix},$$

where  $I_{p \times p}$  is an identity matrix, the sub-matrix  $B_{p \times p}^{(r)}$ , ( $r = 1, \dots, R_n$ ) encapsulates the coefficients associated with the sub-vector  $[b_r(X_1), \dots, b_r(X_p)]$  in  $Z$ , and sub-matrix  $B_{p \times 1}^{(0)}$  encapsulates the intercept coefficients. Specifically, for  $r = 1, \dots, R_n$ , the  $j$ -th row and  $k$ -th column of  $B_{p \times p}^{(r)}$  is equal to  $\beta_{rkj}$ , with  $\beta_{rjj} = 0$  for all  $j \in V$ . The  $j$ -th row of  $B_{p \times 1}^{(0)}$  is equal to  $\sum_{k=1}^p \beta_{0kj}$ .

There exists a bijective mapping between the parameters  $\Gamma$  and  $\theta \in \Theta$ . Specifically, for any given  $\Gamma$ , the map to  $\theta$ ,  $\theta = \mu(\Gamma)$ , is defined as  $f_{kj}(\cdot) = [\sum_{r=1}^{R_n} \Gamma_{j,pr+1+k} \cdot (\mathbb{E}[b_r(X_k)] - b_r(\cdot))] \cdot \Gamma_{jj}^{-1}$  and  $\sigma_j = \Gamma_{jj}^{-1}$ . Particularly,  $\sum_{k=1}^p f_{kj}(\cdot) = -[\Gamma_{j,p+1} + \sum_{k=1}^p \sum_{r=1}^{R_n} \Gamma_{j,pr+1+k} \cdot b_r(\cdot)] \cdot \Gamma_{jj}^{-1}$ . Under this mapping, we establish the following equivalence.

**Proposition 3.** *Let  $\{Z^{(i)}\}_{i=1}^n$  be  $n$  independent and identically distributed samples from the distribution of  $Z$ . Let  $\widehat{\Sigma}_n := n^{-1} \sum_{i=1}^n Z^{(i)} Z^{(i)\top}$  be the Gram matrix. Then, the negative log-likelihood function can be equivalently formulated as:*

$$(7) \quad \ell_n(\theta) = \ell_n(\mu(\Gamma)) := \sum_{j=1}^p -2 \log \Gamma_{jj} + \text{tr}(\Gamma^\top \Gamma \widehat{\Sigma}_n).$$

The proof of Proposition 3 is given in Appendix B. First, we observe that the equivalent formulation in (7) is a convex function of  $\Gamma$ . Second, given the Gram matrix  $\widehat{\Sigma}_n$ , the trace term  $\text{tr}(\Gamma^\top \Gamma \widehat{\Sigma}_n)$  does not require iteration over all samples. It has computational complexity on the order of  $R_n^2 p^3$ , which is significantly smaller—when the sample size  $n$  is sufficiently large—than  $nR_n p^2$ , the complexity of evaluating  $\sum_{j=1}^p \|X_j - \sum_{k \neq j} f_{kj}(X_k)\|_n^2 / \sigma_j^2$ . The two observations above contribute to accelerating both the objective function evaluation and the optimization procedure itself. These observations are corroborated by Figure 3. Lastly, we point out that the result above can be straightforwardly extended to the case where basis functions vary across  $(k, j)$ .

**3.2. A convex mixed-integer constraint set.** Handling the acyclicity constraint, that  $\mathcal{G}(\theta)$  forms a DAG, is a key computational challenge in the  $\ell_0$ -penalized maximum likelihood estimator (4). Many integer programming approaches have been developed to encode this constraint (Grötschel et al., 1985, Manzour et al., 2021, Park and Klabjan, 2017). Among these formulations, we adopt the *layered network formulation* due to its demonstrated effectiveness in continuous-variable scenarios (Manzour et al., 2021). This approach has also been applied to a closely related problem under linear structural equation models (Xu et al., 2025a).

We introduce two sets of decision variables: (i) A set of binary variables  $\{g_{kj} \in \{0, 1\}\}_{(k,j) \in \mathcal{E}}$ , where  $g_{kj} = 1$  represents the presence of the edge from node  $k$  to  $j$  and  $g_{kj} = 0$  otherwise; (ii) A set of continuous variables  $\{\psi_j \in [1, p]\}_{j \in V}$ , where  $\psi_j$  represents the *layer value* of each node. The layer values impose a topological ordering on the nodes, such that any ancestor node must have a strictly

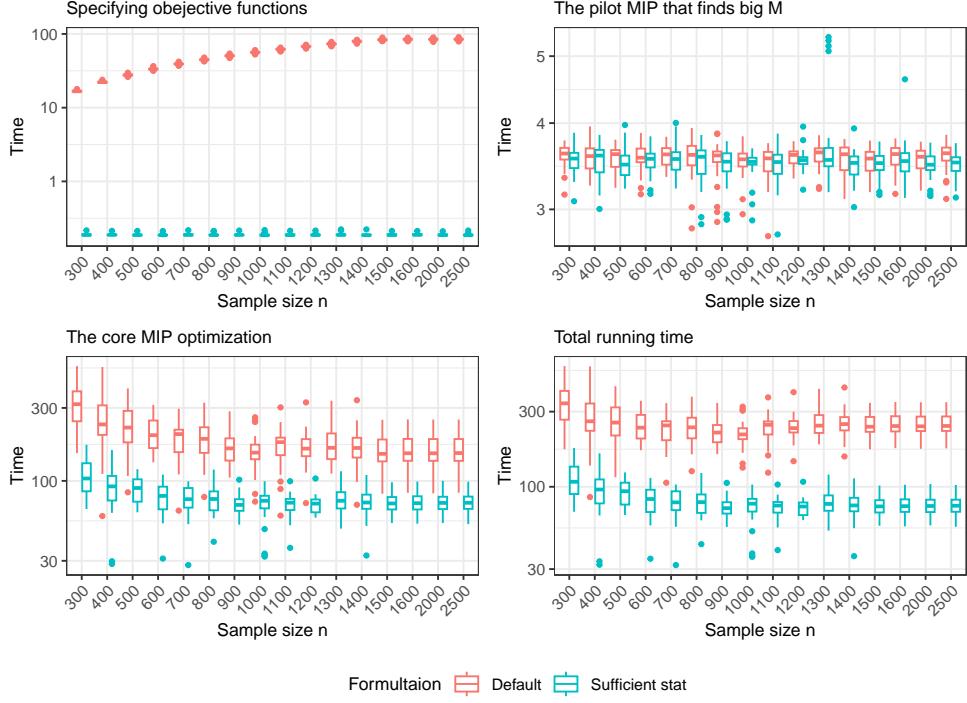


FIGURE 3. The running time of different components in the MIP (8) solved using **Gurobi**. All running times are reported in seconds, and the vertical axis is displayed on a log scale. Top left: running time for specifying the objective function involving decision variables before optimization. Top right: running time for the pilot optimization without any constraints to determine the big  $M$ . Bottom left: running time for the core mixed-integer optimization that finds the solution. Bottom right: total running time. Each box summarizes results from 30 independent trials. The implementation details are provided in Appendix H.3.

higher layer value than its descendants. Accordingly, we have:

$$(8a) \quad \min_{\substack{\Gamma \in \mathbb{R}^{p \times (pR_n + p + 1)} \\ g_{kj} \in \{0, 1\}, \forall (k, j) \in \mathcal{E} \\ \psi \in [1, p]^p}} \sum_{j=1}^p -2 \log \Gamma_{jj} + \text{tr}(\Gamma^\top \widehat{\Sigma}_n) + \lambda_n^2 \cdot \sum_{(k, j) \in \mathcal{E}} g_{kj}$$

$$(8b) \quad \text{s.t. } -Mg_{kj} \leq \Gamma_{j, pr+1+k} \leq Mg_{kj}, \quad \forall (k, j) \in \mathcal{E}, r \in \{1, \dots, R_n\},$$

$$(8c) \quad -M \sum_{k=1}^p g_{kj} \leq \Gamma_{j, p+1} \leq M \sum_{k=1}^p g_{kj}, \quad \forall j \in V,$$

$$(8d) \quad 0 < \Gamma_{jj} \leq M, \quad \forall j \in [p],$$

$$(8e) \quad 1 - p + pg_{kj} \leq \psi_j - \psi_k, \quad \forall (k, j) \in \mathcal{E}.$$

The constant  $M > 0$  is a pre-specified upper bound that exceeds all possible values of  $|\Gamma_{ij}|$  for  $i \in [p]$  and  $j \in [pR_n + p + 1]$ . It is used to linearize non-linear terms, resulting in the *big-M constraints* given in (8b)–(8d). A smaller value of  $M$  can improve computational efficiency. In practice, one may

first solve the optimization problem without the big- $M$  constraints to obtain a preliminary estimate  $\widehat{\Gamma}$ , and then set  $M = 2 \max_{i \in [p], j \in [pR_n+p+1]} |\widehat{\Gamma}_{ij}|$  (Küçükyavuz et al., 2023, Park and Klabjan, 2017, Xu et al., 2025a).

Any feasible solution to the optimization problem (8) necessarily forms a DAG. We establish this claim as follows. First, if there exists a path from node  $k$  to  $j$ , then the layer value assigned to node  $j$  must be higher than that of node  $k$ . To see this, consider a path  $k \rightarrow m \rightarrow \dots \rightarrow j$ ; by constraint (8e), the layer values satisfy  $\psi_j > \psi_m > \psi_k$ . This property automatically ensures that no cycle can exist between  $k$  and  $j$ , otherwise it would imply both  $\psi_k < \psi_j$  and  $\psi_k > \psi_j$  simultaneously. Second, if there is no edge between nodes  $k$  and  $j$ , which is encoded by  $g_{kj} = 0$ , the corresponding coefficients  $\beta_{rkj}$  must be zero for all  $r \in \{1, \dots, R_n\}$ ; this is guaranteed by constraint (8b). Moreover, if there are no parents for node  $j$ , which is encoded by  $\sum_{k=1}^p g_{kj} = 0$ , the corresponding intercept coefficient  $\sum_{k=1}^p \beta_{0kj}$  must also be zero; this is guaranteed by constraint (8c). Conversely, when  $g_{kj} = 1$ , encoding the presence of an edge between node  $k$  and  $j$ , the associated coefficients  $\beta_{rkj}$  are generally not exactly zero since additional degrees of freedom contribute to a better model fit. Finally, the penalty term, which is the number of edges in the DAG, is naturally encoded as  $\sum_{(k,j) \in \mathcal{E}} g_{kj}$ .

We recommend choosing the tuning parameter  $\lambda_n$  that results in the smallest the Bayesian information criterion (BIC) score (Schwarz, 1978). Specifically, if  $\widehat{\Gamma}$  and  $\widehat{g}$  are the estimates of (8), the BIC score is  $\sum_{j=1}^p -2 \log \widehat{\Gamma}_{jj} + \text{tr}(\widehat{\Gamma}^\top \widehat{\Gamma} \widehat{\Sigma}_n) + (p + \sum_{(k,j) \in \mathcal{E}} \widehat{g}_{kj}) \log(n)/n$ .

When the basis functions  $b_{jk}(\cdot)$  are linear, (8) reduces to the MIP proposed in Xu et al. (2025a). Thus, our proposed estimator is a generalization of the existing framework to non-linear settings.

**3.3. An equal-variance formulation.** When the noise variances are assumed to be equal, meaning that  $\sigma_j^* \equiv \sigma^* > 0$  for all  $j \in V$ , the optimization problem (8) can be substantially simplified. Let

$$B = \begin{bmatrix} I_{p \times p} & -B_{p \times 1}^{(0)} & -B_{p \times p}^{(1)} & \dots & -B_{p \times p}^{(R_n)} \end{bmatrix} \in \mathbb{R}^{p \times (pR_n+p+1)}.$$

Then, we have  $B = \sigma^* \Gamma$  under the equal-variance condition, and the objective function (8a) can be written as  $(\sigma^*)^{-2} \cdot [2p\sigma^{*2} \log \sigma^* + \text{tr}(B^\top B \widehat{\Sigma}_n) + \tilde{\lambda}_n^2 \cdot \sum_{(k,j) \in \mathcal{E}} g_{kj}]$ , where  $\tilde{\lambda}_n^2 = \lambda_n^2 \sigma^{*2}$ . Minimizing this resulting objective function is equivalent to

$$(9a) \quad \min_{\substack{B \in \mathbb{R}^{p \times (pR_n+p+1)} \\ g_{kj} \in \{0,1\}, \forall (k,j) \in \mathcal{E} \\ \psi \in [1,p]^p}} \text{tr}(B^\top B \widehat{\Sigma}_n) + \tilde{\lambda}_n^2 \cdot \sum_{(k,j) \in \mathcal{E}} g_{kj}$$

$$(9b) \quad \text{s.t.} \quad -Mg_{kj} \leq B_{j,pr+1+k} \leq Mg_{kj}, \quad \forall (k,j) \in \mathcal{E}, r \in \{1, \dots, R_n\},$$

$$(9c) \quad -M \sum_{k=1}^p g_{kj} \leq B_{j,p+1} \leq M \sum_{k=1}^p g_{kj}, \quad \forall j \in V,$$

$$(9d) \quad 1 - p + pg_{kj} \leq \psi_j - \psi_k, \quad \forall (k,j) \in \mathcal{E}.$$

Suppose that  $\widehat{B}$  is the estimated coefficient matrix obtained from (9), then the variance can be estimated as  $\widehat{\sigma}^2 = \|\widehat{B}Z\|_n^2$ . This formulation removes the logarithmic term and reduces the degrees of freedom by  $p$  compared with (8). Consequently, incorporating the equal-variance assumption can substantially improve computational efficiency. Similar to (8), in practice, we choose  $\tilde{\lambda}_n$  that yields the smallest the BIC score. Specifically, if  $\widehat{B}$  and  $\widehat{g}$  are the estimates of (9), the BIC score is  $\text{tr}(\widehat{B}^\top \widehat{B} \widehat{\Sigma}_n) + \sum_{(k,j) \in \mathcal{E}} \widehat{g}_{kj} \cdot \log(n)/n$ .

Since we do not usually know whether the data is generated according to SEM with equal variances or unequal variances, in practice, we implement both estimators (after choosing their regularization parameters via BIC scores), and choose the model with a smaller BIC score among the two.

#### 4. COMPUTATIONAL SPEEDUPS: EARLY STOPPING, AND INCORPORATING PARTIAL ORDER AND EDGE STRUCTURES

In this section, we introduce several strategies for improving the computational efficiency of the MIP (8).

**4.1. Super-structure.** For the DAG  $\mathcal{G}^* = (\mathcal{V}, \mathcal{E}^*)$ , a super-structure  $\mathcal{E}^\circ \subseteq \mathcal{E}$  is an edge set containing the true edge set  $\mathcal{E}^*$ :  $\mathcal{E}^* \subseteq \mathcal{E}^\circ$ . The super-structure  $\mathcal{E}^\circ$  may permit bidirectional edges and cycles, though self-loops are excluded. The idea of imposing a super-structure has been adopted by various related methods (Chickering, 2002, Küçükyavuz et al., 2023, Solus et al., 2021, Xu et al., 2025a). We impose this super-structure following the approach of Xu et al. (2025a): set  $g_{kj} = 0$  and  $\Gamma_{j,pr+1+k} = 0$  for all  $(k, j) \in \mathcal{E} \setminus \mathcal{E}^\circ$  and  $r \in \{1, \dots, R_n\}$ . That is, the search space of decision variables is restricted to edges within  $\mathcal{E}^\circ$ .

Enforcing a super-structure offers multiple advantages. First, in many applications, prior information about the presence or absence of specific edges is available, for example, from domain expertise or some widely accepted knowledge. Such information can be naturally incorporated into the super-structure. Second, as shown above, edges outside the super-structure  $\mathcal{E}^\circ$  are excluded from consideration. Consequently, when  $\mathcal{E}^\circ$  is sparse, the search space is greatly reduced, which in turn can substantially accelerate the optimization algorithm. Lastly, the existence of a sparse super-structure greatly facilitates the theoretical analysis of our proposed estimator (8). As we will see in Section 5, we assume access to a super-structure  $\mathcal{E}^\circ$  within which the number of parents for each node is uniformly bounded by a constant.

In addition to leveraging prior knowledge when available, we propose two approaches for estimating a super-structure. The goal here is to identify one that is as sparse as possible, while still capturing the true edge set  $\mathcal{E}^*$ . The first approach is neighborhood selection following the general idea in Meier et al. (2009), Voorman et al. (2014), where we perform variable selection in an additive model of  $X_j$  on the set of all other variables  $X_{-j} = \{X_k : k \neq j\}$ , via methods such as the group Lasso or the sparse additive model (Haris et al., 2022, Ravikumar et al., 2009). The tuning parameter may be selected using cross-validation, but alternative strategies that incorporate additional information are also possible. In particular, one may first apply a baseline method (e.g., Bühlmann et al., 2014, Gao et al., 2020, Peters et al., 2014, Zheng et al., 2018), and use the cardinality of the resulting graph as a reference. The tuning parameter for our neighborhood selection can then be gradually increased until the selected super-structure reaches a size, for instance, no more than twice of this reference. Note that the neighborhood selection approach obtains an undirected graph, serving as an estimate for the true moral graph of  $\mathcal{G}^*$ . (The true moral graph is the undirected graph obtained from  $\mathcal{G}^*$  by adding edges between any two nodes that share a common child and then making all edges undirected.) The second approach also leverages additional information from a related method. Specifically, we generate  $B$  bootstrap samples and apply the baseline method to each sample. An edge is included in the super-structure if its selection proportion across the  $B$  samples exceeds a prescribed threshold. We denote this threshold by  $\tau^\circ$ . Similar to the tuning parameter in neighborhood selection, one can create a reference for the cardinality of the graph, and then gradually increase the selection proportion threshold  $\tau^\circ$  until the size of the resulting super-structure is no larger than this reference. In practice, we use the union from both approaches as  $\mathcal{E}^\circ$ .

**4.2. Partial order and stable edges.** The topological ordering, or sometimes referred to as the permutation of nodes, is a key concept in DAG estimation. Several approaches focus on estimating the node ordering as the primary step, and then apply pruning procedures to obtain a fitted graph (Bühlmann et al., 2014, Chen et al., 2019, Gao et al., 2020, Peters et al., 2014). Although this is not the central idea pursued in our work, some prior information about the ordering of nodes in

$E^\circ$  can nevertheless be highly beneficial (Shojaie and Chen, 2024). If it is known a priori that node  $j$  cannot be a descendant of node  $k$ , then this relation can be encoded by imposing the constraint  $\psi_k < \psi_j$  in our MIP formulation (8). From here, we define the preliminary *partial order set*, as a collection of edges in  $E^\circ$  whose ordering is predetermined:  $E^p := \{(k, j) \in E^\circ : \psi_k < \psi_j\}$ .

The constraint  $\psi_k < \psi_j$  excludes any path from node  $j$  to  $k$ . However, the presence of edge from  $k$  to  $j$  remains undetermined: either  $g_{kj} = 0$  or  $g_{kj} = 1$ . If the presence of an edge from  $k$  to  $j$  is further known from prior knowledge, then the constraint  $g_{kj} = 1$  can be directly imposed. Accordingly, we define the preliminary *stable set* as the collection of edges in  $E^p$  that are predetermined to exist:  $E^s := \{(k, j) \in E^p : g_{kj} = 1\}$ . We view the stable set as a more “confident” set than the partial order set, since imposing it directly eliminates integer decision variables.

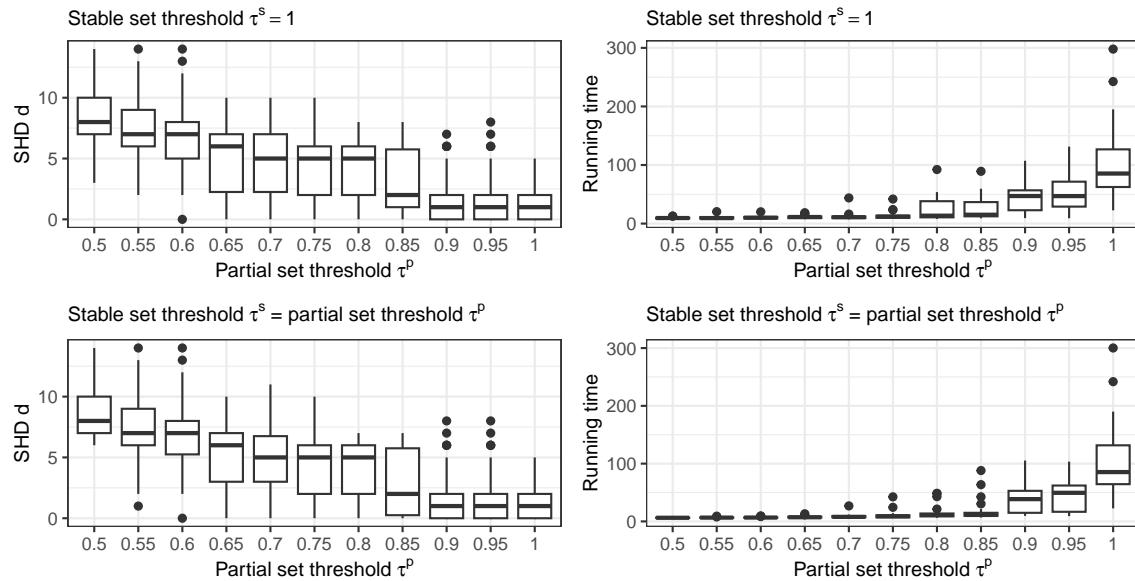


FIGURE 4. The performance of the estimator obtained from solving (8) across a range of threshold values. The metric  $d$  is the structural Hamming distance (SHD) which quantifies the number of incorrectly recovered and missed edges, with smaller values indicating better accuracy. All running times are reported in seconds. Top two panels: only  $\tau^p$  varies while  $\tau^s$  is fixed at 1. Bottom two panels: both thresholds vary simultaneously with  $\tau^p = \tau^s$ . Each box summarizes results from 30 independent trials. The implementation details are provided in Appendix H.4.

Enforcing the stable set and partial order set can substantially reduce the search space, thereby improving optimization efficiency. When prior information is unavailable or insufficient, a preliminary partial order set and stable set can be constructed using bootstrap procedures. Similar to the second approach of the super-structure estimation (see Section 4.1), we apply a baseline method to  $B$  bootstrap samples, and an edge is included in the partial order set or stable set if its selection proportion exceeds a prescribed threshold. We denote the thresholds for the partial order set and stable set by  $\tau^p$  and  $\tau^s$ , respectively. To ensure the nesting property  $E^s \subseteq E^p \subseteq E^\circ$ , the thresholds are chosen to satisfy  $\tau^\circ \leq \tau^p \leq \tau^s$ . Finally, note that the selected partial order set and stable set may contain

cycles. To prevent this, edges are added sequentially in decreasing order of their selection proportion; any edge that would introduce a cycle with previously added edges is omitted.

Figure 4 compares the performance of our proposed group  $\ell_0$ -estimator, where partial order and stable sets under different threshold levels are imposed. As the threshold increases, the partial order and stable set shrink in size. With fewer constraints imposed, the estimated graph selects fewer incorrect edges but requires more time to fit. Thus, a balance between accuracy and efficiency is needed, and in practice a good default can be  $\tau^p = 0.95$  and  $\tau^s = 1$ .

**4.3. Early stopping.** We employ a branch-and-bound procedure to solve the convex MIP (8). In this process, we iteratively update the lower and upper bounds of the objective function value. At the same time, we track the optimality gap of a solution, defined as the difference between these bounds for that solution. A solution is deemed optimal once its corresponding optimality gap reaches zero. At each iteration, we relax the integer constraints on  $g$  and solve the resulting, much simpler convex optimization problem. The obtained optimal objective value then serves as a lower bound. If any variable  $g$  takes a fractional value in this solution, the problem is then branched into two subproblems by enforcing either a smaller or larger integer value for  $g$ . A feasible solution obtained through this process provides an upper bound.

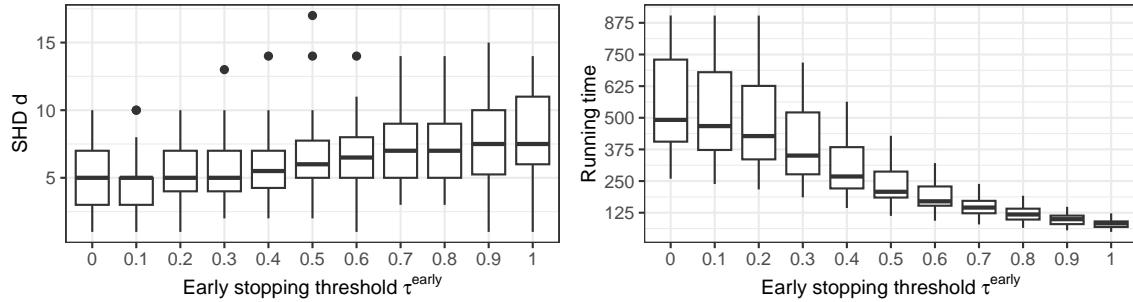


FIGURE 5. The performance of the estimator obtained from solving (8) across a range of  $\tau^{\text{early}}$  values. Left panel: structural Hamming distance (SHD)  $d$ , which quantifies the number of incorrectly recovered and missed edges, with smaller values indicating better accuracy. Right panel: total running time in seconds. Each box summarizes results from 30 independent trials. The implementation details are provided in Appendix H.5.

Rather than waiting for the lower and upper bounds to meet, the algorithm can be terminated early once the optimality gap falls below a specified threshold  $\tau^{\text{early}}$ . This practice is referred to as *early stopping*, and it is commonly used in practice to speed up the algorithm (Chen et al., 2023, Küçükayavuz et al., 2023, Xu et al., 2025a). One may wish that the sub-optimal solution obtained through early stopping, despite its non-zero optimality gap, still behaves similarly to the optimal solution. As shown in Figure 5, as  $\tau^{\text{early}}$  increases from zero, the performance of the estimator deteriorates only slightly, while the running time decreases substantially. Furthermore, it would be desirable if this sub-optimal solution achieved perfect recovery of the true DAG. In Section 5, we provide a theoretical justification for the early-stopping solution when the stopping threshold is chosen below a certain order.

## 5. STATISTICAL GUARANTEES

We provide statistical guarantees for our estimator (8). Section 5.1 introduces the necessary notation as well as Assumptions 2–5, which are required for all subsequent theorems. In Section 5.2, we present Theorem 4, which establishes that our estimator achieves a correct permutation with high probability. Note that the additional Assumption 6 introduced in this section is also required for all theorems, although it is most naturally stated in the context of permutations. Section 5.3 presents Theorem 5, which shows that our estimated variances converge to the true variances with high probability. This theorem requires no further assumptions. Finally, Section 5.4 introduces Theorem 6, which states that our estimator perfectly recovers the true graph with high probability, under two additional Assumptions 7 and 8. For each result, we also provide an analogous guarantee for the early-stopping suboptimal solution in addition to the optimal solution.

The theoretical analysis in this section substantially extends those of Bühlmann et al. (2014) and van de Geer and Bühlmann (2013), as their results do not suffice for the guarantees we provide. Specifically, the analysis in Bühlmann et al. (2014) addresses only the consistency of the estimated topological ordering (and not the graph) and does not accommodate regularization; indeed, they view graph estimation as a second step after estimating an ordering among the variables, while we jointly estimate all the model parameters including the DAG structure. On the other hand, the analysis of van de Geer and Bühlmann (2013) is restricted to linear models and cannot handle non-linear settings. Our analysis unifies and extends these two lines of work, and in particular establishes consistency in terms of graph recovery in non-linear settings.

**5.1. Setup.** Let  $s^* := |\mathcal{G}^*|$  denote the sparsity level of the true graph  $\mathcal{G}^* = (V, E^*)$ . We consider the estimator (8) with two additional constraints. The first constraint is that the edges are restricted to be in a super-structure  $E^\circ$ , for which we assume that  $E^* \subseteq E^\circ$ . As described in Section 4.1, neighborhood selection methods estimate a moral graph, which can serve as the super-structure  $E^\circ$ . Existing theoretical results (e.g., Bühlmann et al. (2014, Lemma 4)) guarantee that the estimated moral graph is a supergraph of the true moral graph. By interpreting an undirected edge as bi-directed, the super-structure  $E^\circ$  would satisfy the desired assumption. As a result, for simplicity, we assume access to such a super-structure in the remainder. Second, we add a constraint that the number of edges does not exceed  $s_n$ , where conditions on  $s_n$  are described later.

Note that the parameter  $\Gamma$  in (8) is in bijective correspondence with the model parameterization  $\tilde{\theta} = (\{\beta_{rkj}\}_{(k,j) \in \mathcal{E}, r \in [R_n]}, \{\sigma_j\}_{j \in V}) \in \tilde{\Theta}$  through the formulation in (6). Given the basis functions, it is also equivalent to the model parameter  $\theta = (\{f_{kj}\}_{(k,j) \in \mathcal{E}}, \{\sigma_j\}_{j \in V}) \in \Theta$  via the mapping  $\mu$ . For notational convenience, we denote the estimator by  $\hat{\theta} = (\{\hat{\beta}_{rkj}\}_{(k,j) \in E^\circ, r \in [R_n]}, \{\hat{\sigma}_j\}_{j \in V})$ , while noting that all three parameterizations are mutually transformable. For the induced graph  $\mathcal{G}(\hat{\theta})$ , an edge from node  $k$  to node  $j$  is present if  $\sum_{r=1}^{R_n} |\beta_{rkj}|^2 \neq 0$ , and absent otherwise. The sparsity of the estimated graph  $\hat{s} := |\mathcal{G}(\hat{\theta})|$  is defined as the number of selected edges. We also denote their early-stopping counterparts as  $\hat{\theta}^{\text{early}} = (\{\hat{\beta}_{rkj}^{\text{early}}\}_{(k,j) \in E^\circ, r \in [R_n]}, \{\hat{\sigma}_j^{\text{early}}\}_{j \in V})$  and  $\hat{s}^{\text{early}}$ . Throughout, we allow the number of nodes  $p$  to vary with the sample size  $n$ . A good choice is  $n \gtrsim (\log p)^2$  for Theorem 4, and  $n \gtrsim p(\log p)^2$  for Theorems 5 and 6.

**Assumption 1** (super-structure sparsity). *The super-structure  $E^\circ$  is sparse:  $s_j := |\{k : (k,j) \in E^\circ\}| \leq K$  for each node  $j \in V$ .*

**Assumption 2** (model sparsity). *The quantity  $s_n$  that constrains the number of edges satisfies  $s^* \leq s_n$ .*

**Assumption 3** (Sobolev function class with bounded basis functions). *For all function classes  $\{\mathcal{F}_{kj}\}_{(k,j) \in \mathcal{E}}$ , the basis functions  $\{b_{rkj}(\cdot)\}_{r=1}^\infty$  are orthogonal, satisfying  $\mathbb{E}[b_{rkj}(X_k)b_{r'kj}(X_k)] = v\mathbf{1}(r =$*

$r')$  for some  $v \in (0, 1]$ , and uniformly bounded with  $\sup_r |b_{rkj}(\cdot)| \leq 1$ . Furthermore, their coefficients decay at a fast rate:  $\sum_{r=1}^{\infty} |\beta_{rkj}| \cdot r^{\eta} \leq C$  and  $\eta > 1$  for any  $(k, j) \in \mathcal{E}$ .

Assumptions 1 and 2 impose sparsity at both the node and graph levels. Specifically, for each node  $j$ , the super-structure  $\{k : (k, j) \in E^\circ\}$  contains all true parents while including no more than  $K$  candidate parents per node. As a result, both the true graph and the estimated graph have at most  $K$  parents per node. This assumption resembles Condition (B2) in Bühlmann et al. (2014). Moreover, the size of the true graph, upper bounded by  $s_n$ , may grow with the sample size. The added constraint  $\hat{s} \leq s_n$  ensures that the estimated graph adheres to the same sparsity level. Assumption 3 imposes additional structure on the function classes. In particular, it maps each function class  $\mathcal{F}_{kj}$  into a Sobolev ellipsoid, allowing control over both the complexity of  $\mathcal{F}_{kj}$  and the truncation error  $\|f_{kj} - \beta_{0kj} - \sum_{r=1}^{R_n} \beta_{rkj} b_{rkj}\|_\infty$  for all  $f_{kj} \in \mathcal{F}_{kj}$ . This assumption was also adopted in Ravikumar et al. (2009). In Appendix C.1, we provide a choice of basis functions that satisfies the conditions in Assumption 3.

*Additional notation for variances:* Let  $\vec{g}_j := \{k : g_{kj} = 1, (k, j) \in E^\circ\}$  be an index set of underlying parents of node  $j$ . Then we can define the additive function classes  $\mathcal{F}^{\oplus \vec{g}_j} := \bigoplus_{k \in \vec{g}_j} \mathcal{F}_{kj}$  and  $\mathcal{F}_n^{\oplus \vec{g}_j} := \bigoplus_{k \in \vec{g}_j} \mathcal{F}_{kj,n}$ . With the function  $\mathcal{R}_j(f; \vec{g}_j) := [X_j - \sum_{k \in \vec{g}_j} f_{kj}(X_k)]^2$ , the residual variances after projecting  $X_j$  onto  $\mathcal{F}^{\oplus \vec{g}_j}$  and  $\mathcal{F}_n^{\oplus \vec{g}_j}$  are given by  $\nu_j(\vec{g}_j) := \min_{f \in \mathcal{F}^{\oplus \vec{g}_j}} \mathbb{E}[\mathcal{R}_j(f; \vec{g}_j)]$  and  $\nu_j^n(\vec{g}_j) := \min_{f \in \mathcal{F}_n^{\oplus \vec{g}_j}} \mathbb{E}[\mathcal{R}_j(f; \vec{g}_j)]$ , respectively. Since each node has at most  $K$  parents, we can denote the maximal gap between these variances as  $d_{n,p} := \max_{j \in V} \max_{|\vec{g}_j| \in \{0, \dots, K\}} |\nu_j(\vec{g}_j) - \nu_j^n(\vec{g}_j)|$ , the maximal value of  $\nu_j(\vec{g}_j)$  as  $\bar{\nu}_p := \max_{j \in V} \mathbb{E}[X_j^2]$ , and the minimal value as  $\underline{\nu}_p := \min_{j \in V} \min_{|\vec{g}_j|=K} \nu_j(\vec{g}_j)$ .

**Assumption 4** (bounded noise variances). *There exist a constant  $\bar{\sigma}$  such that the noise variances are bounded:  $\sigma_j^{*2} \leq \bar{\sigma}^2 < \infty$  for all  $j \in V$ . In addition, the variance  $\nu_j(\vec{g}_j)$  is uniformly bounded below by a constant  $c > 0$ :  $\underline{\nu}_p \geq c$ .*

Assumption 4 ensures that  $\bar{\nu}_p \leq \bar{c}$  for some  $\bar{c} < \infty$ , thereby facilitating the boundedness of the estimated variances. This condition is similar to Condition (A3) in Bühlmann et al. (2014). Intuitively, the estimated variance  $\hat{\sigma}_j^2$  should concentrate around  $\nu_j^n(\hat{g}_j)$ , where  $\hat{g}_j := \{k : \hat{g}_{kj} = 1, (k, j) \in E^\circ\}$  denotes the estimated parent set on node  $j$ . Thus, the estimated variances are approximately bounded below and above by the smallest and largest possible values of  $\nu_j^n(\hat{g}_j)$ , respectively. One may observe that  $\underline{\nu}_p$  and  $\bar{\nu}_p$  are appropriate bounds if  $\nu_j^n(\vec{g}_j)$  does not deviate significantly from  $\nu_j(\vec{g}_j)$ , or equivalently  $d_{n,p} = o(1)$ . See Appendix C.2 for a detailed discussion on the control of  $\bar{\nu}_p$  and  $d_{n,p}$ .

**Assumption 5** (functional compatibility). *Under the distribution  $P$  generated from (2), there exists  $\phi^2 > 0$  such that for all  $\gamma \in \mathbb{R}^p$ ,  $\min_{j \in V} \|\sum_{k=1}^p \gamma_k f_{kj}(X_k)\|_{L_2}^2 \geq \phi^2 \|\gamma\|_2^2$  for all  $f_{kj} \in \mathcal{F}_{kj}$  with  $\|f_{kj}(X_k)\|_{L_2} = 1$ .*

Assumption 5 guarantees that the spaces  $\mathcal{F}^{\oplus \vec{g}_j}$  and  $\mathcal{F}_n^{\oplus \vec{g}_j}$  are closed under the  $L_2$  norm. This assumption is standard in high-dimensional additive models (Meier et al., 2009) and is directly adopted from Lemma 2 of Bühlmann et al. (2014).

**5.2. Correct permutation.** A permutation of the nodes is a mapping from an ordering to the nodes that respects the ancestor-descendant relationships in the graph. Specifically, for any permutation  $\pi$  of  $\{1, \dots, p\}$ , we say that  $\pi$  is consistent with a DAG if, whenever there is an edge from  $\pi(k)$  to  $\pi(j)$ , it holds that  $k < j$ . In particular, the node  $\pi(1)$  has no parents. A given DAG may admit multiple such permutations. For example, if the edge set is  $\{(1, 2), (2, 3), (2, 4)\}$ , then both  $\pi_1 = (1, 2, 3, 4)$  and  $\pi_2 = (1, 2, 4, 3)$  are valid permutations. Let  $\Pi^*$  denote the set of all permutations consistent

with the true graph  $\mathcal{G}^*$ , referred to as the set of true permutations. Similarly, let  $\widehat{\Pi}$  denote the set of permutations consistent with the estimated graph  $\mathcal{G}(\widehat{\theta})$ , referred to as the set of estimated permutations. We also denote its early-stopping counterpart as  $\widehat{\Pi}^{\text{early}}$ . We aim to show that with high probability, the estimated permutations are correct.

Let  $\pi^{-1}$  denote the inverse permutation that maps each node to its position in the ordering. For any permutation  $\pi$ , we define  $\vec{g}(\pi)_j := \{k : \pi^{-1}(k) < \pi^{-1}(j), (k, j) \in E^\circ\}$ , the set of parents of node  $j$  under permutation  $\pi$ . The residual variance after projecting  $X_j$  onto the function class  $\mathcal{F}^{\oplus \vec{g}(\pi)_j}$ , denoted by  $\nu_j(\vec{g}(\pi)_j)$ , quantifies the best possible fit of node  $j$  to a DAG with the permutation  $\pi$ . Particularly, when  $\pi \in \Pi^*$ , we have  $\nu_j(\vec{g}(\pi)_j) = \sigma_j^{*2}$ . When  $\pi \notin \Pi^*$ , we define the separation between these residual variances and the true variances as:

$$(10) \quad \xi_p := \min_{\pi \notin \Pi^*} p^{-1} \sum_{j=1}^p (\log \nu_j(\vec{g}(\pi)_j) - \log \sigma_j^{*2}).$$

Appealing to properties of the KL-divergence, one can show that  $\xi_p \geq 0$ . A more detailed discussion of these results is provided in Appendix C.3.

**Assumption 6** (non-negligible separation). *When the permutation is incorrect, the separation between the best fitted residual variances and the true variances is non-negligible:  $\lambda_n^2 s_n + p\sqrt{\log p/n} + s_n R_n^{-2\eta} = o(p\xi_p)$ .*

Assumption 6 ensures the convergence rate of  $\log \widehat{\sigma}_j$  towards  $\log \sigma_j$  dominates that of  $\xi_p$  towards zero, if the latter occurs. Some example choices for these quantities are:  $\lambda_n^2 \asymp p(\log p)^{1/2} n^{-1/2} s_n^{-1}$  and  $s_n \asymp R_n \asymp n^{1/5}$ . In the extreme case where the number of nodes  $p$  is fixed, the assumption holds trivially. Generally speaking, when  $p$  varies with  $n$ , the assumption simplifies to  $\sqrt{\log p/n} = o(\xi_p)$  in both the low-dimensional case where  $n \asymp p(\log p)^2$  and the high-dimensional case where  $n \asymp (\log p)^2$ . A similar condition appears in Theorem 3 of Bühlmann et al. (2014), where it is also assumed that  $\sqrt{\log p/n} = o(\xi_p)$ .

**Theorem 4.** *Under conditions in Proposition 2 and Assumptions 1–6, it holds, with probability at least  $1 - 4/p$ , that:*

- (1) *The estimated graph has correct permutations:  $\widehat{\Pi} \subseteq \Pi^*$ .*
- (2) *When the early-stopping threshold satisfies  $\tau^{\text{early}} = o(p\xi_p)$ , the resulting estimated graph has correct permutations:  $\widehat{\Pi}^{\text{early}} \subseteq \Pi^*$ .*

A proof is given in Appendix E. This proof generally follows the general idea of Theorem 1 in Bühlmann et al. (2014), but relies on the KL-divergence gap between any incorrect permutation distribution and the distribution induced by our estimator with regularization. Interestingly, even a sub-optimal solution obtained via early stopping can identify the correct permutations, provided that the optimality gap threshold is appropriately controlled. This result enables substantial computational savings without compromising estimation accuracy.

**5.3. Convergence of noise variances.** We next show convergence of estimated noise variances  $\{\widehat{\sigma}_j^2\}_{j \in V}$  to the true noise variances  $\{\sigma_j^{*2}\}_{j \in V}$ .

**Theorem 5.** *Under conditions in Proposition 2 and Assumptions 1–6, it holds, with probability at least  $1 - 5/p$ ,*

$$\sum_{j=1}^p (\sigma_j^{*2} - \widehat{\sigma}_j^2)^2 \lesssim (p + s_n R_n) \log p/n + s_n R_n^{-2\eta} + \lambda_n^2 s_n.$$

When  $s_n R_n^{-2\eta} + s_n R_n \log p/n \lesssim p \log p/n$  and with the choice  $\lambda_n^2 \asymp p \log p/(ns_n)$ ,

- (1)  $\sum_{j=1}^p (\sigma_j^{*2} - \hat{\sigma}_j^2)^2 \lesssim \lambda_n^2 s_n.$
- (2) When the early-stopping threshold satisfies  $\tau^{\text{early}} \lesssim \lambda_n^2 s_n + (p + s_n R_n) \log p/n + s_n R_n^{-2\eta}$  and  $\tau^{\text{early}} = o(p\xi_p)$ , we have  $\sum_{j=1}^p (\sigma_j^{*2} - (\hat{\sigma}_j^{\text{early}})^2)^2 \lesssim \lambda_n^2 s_n.$

Theorem 5 is proved in Appendix F. The proof generally follows the strategy of Lemma 7.1 in van de Geer and Bühlmann (2013), but requires adaptations to accommodate the non-linearity. Some possible choices for the scaling quantities are:  $n \asymp p(\log p)^2$ ,  $s_n \asymp R_n \asymp n^{1/5}$ . Since  $\lambda_n^2 s_n = o(1)$ , this result implies convergence of the estimated noise variances:  $|\hat{\sigma}_j^2 - \sigma_j^{*2}| = o(1)$  for all  $j \in V$  with high probability. This upper bound  $\lambda_n^2 s_n$  mirrors the result of linear structural equation models in van de Geer and Bühlmann (2013). Moreover, as noted in their Remark 3.1, a natural normalization of the quantity  $\sum_{j=1}^p (\sigma_j^{*2} - \hat{\sigma}_j^2)^2$  is to divide it by  $p$ . Under this perspective, one can instead choose  $n \asymp (\log p)^2$ ,  $s_n \asymp n$ , and  $R_n \asymp n$ , and obtain a pooled variance convergence:  $p^{-1} \sum_{j=1}^p (\sigma_j^{*2} - \hat{\sigma}_j^2)^2 \lesssim \log p/n$ . This version accommodates a high-dimensional setting where  $n \ll p$ . Finally, for variances convergence, the early-stopping threshold  $\tau^{\text{early}}$  may need to be much smaller than that required to ensure correct permutations in the worst case scenario. To see this, recall that we could choose  $o(p\xi_p) = p\sqrt{\log p/n}$  and  $\lambda_n^2 s_n + (p + s_n R_n) \log p/n + s_n R_n^{-2\eta} \asymp p \log p/n$ ; then in order to ensure the correct permutation  $\tau^{\text{early}}$  is at most  $p\sqrt{\log p/n}$ , whereas to ensure variance convergence it is at most  $p \log p/n$ .

**5.4. Graph recovery.** We next show consistency in terms of graph recovery.

**Assumption 7** (moderately increasing number of basis functions). *The number of basis functions does not grow too fast:  $R_n \lesssim \sqrt{n/\log p}$ .*

**Assumption 8** (beta-min condition). *For any  $(k, j) \in E^*$ , the true coefficients for the basis functions used in estimation are strong:  $\min_{(k,j) \in E^*} \sum_{r=1}^{R_n} |\beta_{rkj}^*|^2 \geq c_u \delta_{n,p} > 0$  for some  $c_u > 0$ , where  $(p + s_n R_n) \log p/n + s_n R_n^{-2\eta} = o(\delta_{n,p})$ .*

Assumption 7 regulates the growth rate of the number of basis functions. However,  $R_n$  cannot be too small either, as this may hinder Assumptions 6 and 8, due to poor approximation from an insufficient number of basis functions. Assumption 8 requires a lower bound on the strength of true causal effects so that they can be captured by the estimated graph. In an example setting we may choose  $n \asymp p(\log p)^3$ ,  $s_n \asymp n^{1/5}$ ,  $R_n \asymp n^{1/5} \log p$ , and  $\min_{(k,j) \in E^*} \sum_{r=1}^{R_n} |\beta_{rkj}^*|^2 \asymp \delta_{n,p} \asymp p(\log p)^2/n$ . A similar beta-min condition appears as Condition 3.5 in van de Geer and Bühlmann (2013) for linear structural equation models, which, without basis expansions, requires each nonzero  $\beta_{kj}^*$  to be at least of order  $\sqrt{p \log p / (s_n n)}$ . After translating their condition into our example setting, the two beta-min requirements align in order.

**Theorem 6.** *Suppose conditions in Proposition 2 and Assumptions 1–8 are satisfied. Take  $\lambda_n^2 \in [c_l \delta_{n,p}, c_u \delta_{n,p}]$  where  $0 < c_l < c_u$ . When  $n$  is sufficiently large, it holds, with probability at least  $1 - 6/p$ , that:*

- (1) *The estimated graph recovers the true structure:  $\mathcal{G}(\hat{\theta}) = \mathcal{G}^*$ .*
- (2) *By taking an early-stopping threshold satisfying  $\tau^{\text{early}} = o(p\xi_p \wedge \delta_{n,p})$ , the resulting estimated graph recovers the true structure:  $\mathcal{G}(\hat{\theta}^{\text{early}}) = \mathcal{G}^*$ .*

The proof of Theorem 6 is provided in Appendix G. This proof proceeds by showing that the optimal objective value attained under the true graph  $\mathcal{G}^*$  is strictly lower than that attained under any incorrect graphs. The tuning parameter  $\lambda_n^2$  cannot be too small so that the selection of redundant edges is prevented. The sample size should be sufficiently large in a sense that, there exists some  $N$  such that for all  $n > N$ ,  $\delta_{n,p} > (p + s_n R_n) \log p/n + s_n R_n^{-2\eta}$ . The existence of such an  $N$  is

guaranteed by Assumption 8. Furthermore, the early-stopping sub-optimal solution can also achieve perfect graph recovery although the corresponding threshold may need to be even smaller than that for noise convergence.

## 6. EXPERIMENTS

The code used to reproduce all experiments, along with the graph structures used in the analysis, is available in the GitHub repository<sup>1</sup>.

**6.1. Simulation setup.** We show the effectiveness of our method over ten graph structures. These structures are publicly available networks sourced from Manzour et al. (2021) and Xu et al. (2025a). They differ in scale, with the number of nodes  $p$  ranging from 6 to 25, and the number of true edges  $s^*$  varying between 6 and 68. For each graph structure, the node set  $V$  and true edge set  $E^*$  are pre-specified. One may freely choose the functional relationships  $\{f_{kj}^*\}_{(k,j) \in E}$  and noise variances  $\{\sigma_j^{*2}\}_{j \in V}$  to construct the corresponding structural equation model. For each dataset, we generate  $n = 500$  independent and identically distributed samples from the model (2).

We mainly focus on the ability of our method to correctly recover the true DAG. To measure the closeness of an estimated graph  $\hat{\mathcal{G}}$  to the true graph  $\mathcal{G}^*$ , we use the structural Hamming distance (SHD) (Tsamardinos et al., 2006). Specifically, let  $\hat{A} \in \mathbb{R}^{p \times p}$  and  $A^* \in \mathbb{R}^{p \times p}$  denote the estimated and true adjacency matrices, respectively, where the  $(k, j)$ -th entry equals 1 if there is an edge from node  $k$  to node  $j$ , and 0 otherwise. The SHD is then given by  $d := \sum_{k=1}^p \sum_{j=1}^p |\hat{A}_{kj} - A_{kj}^*|$ . We also record and compare the running time for each experiment. The MIP (8) is solved using the **Gurobi** 12.0.3 Optimizer via its Python interface. Notably, the **Gurobi** optimizer allows termination upon reaching a pre-specified time limit, while still returning a sub-optimal solution. This ensures that computation remains tractable without requiring indefinite runtimes.

**6.2. Simulations: comparison to existing methods.** We compare our proposed method with the following state-of-the-art approaches for learning DAG structures: the polynomial-time algorithm *NPVAR* for non-parametric DAG learning (Gao et al., 2020); the equal-variance algorithm *EqVar* using both top-down and bottom-up approaches to estimate the node ordering for linear models (Chen et al., 2019); the score-based method *NoTears* that learns non-linear SEMs by a continuous, and non-convex relaxation of acyclicity constraint (Zheng et al., 2020); the algorithm *RESET* using non-parametric regression and subsequent independence tests (Peters et al., 2014); the coordinate descent algorithm *CCDr* using sparse regularization for linear models (Aragam and Zhou, 2015); and the heuristic algorithm *CAM* for causal additive models (Bühlmann et al., 2014). Among these algorithms, *NPVAR* and *EqVar* rely specifically on the assumption of homoscedasticity. We refer to our proposed algorithm as *MIP*, and in particular, we implement two versions of *MIP*, each using a different super-structure. The first uses an oracle super-structure, the true moral graph, as adopted by Xu et al. (2025a). The second uses an estimated super-structure, obtained through the procedure described in Section 4.1.

In this simulation, the true model is specified as follows. For all  $(k, j) \in E^*$ , the true function is set to  $f_{kj}^*(x) = (\sin(x) - \mathbb{E}[\sin(X_k)] + \cos(x) - \mathbb{E}[\cos(X_k)])/2$ . Moreover, we consider both homoscedastic and heteroscedastic variance schemes. In the homoscedastic setting, we fix  $\sigma_j^* = 0.5$  for all  $j \in V$ . In the heteroscedastic setting, each  $\sigma_j^*$  is independently drawn from the interval  $[0.5, 1]$  using a transformed Beta distribution  $\text{Beta}(a_0 = 1, \mu_0) \cdot 0.5 + 0.5$ . Notably, this yields a uniform distribution

---

<sup>1</sup>[https://github.com/Xiaozhu-Zhang1998/FSSS\\_Reproducible\\_Codes](https://github.com/Xiaozhu-Zhang1998/FSSS_Reproducible_Codes).

over  $[0.5, 1]$  when  $\mu_0 = 1$ . As  $\mu_0$  increases, the sampled variances become increasingly concentrated towards 0.5, gradually approaching the homoscedastic regime.

The same set of basis functions is used across all edges. Specifically, we use degree-two splines with two internal knots to construct basis functions. The partial order and stable sets are estimated using the bootstrap procedure outlined in Section 4.2. The *CAM* algorithm is applied to each bootstrap sample under the heteroscedastic setting, while *NPVAR* is used under homoscedasticity. The tuning parameter  $\lambda_n^2$  in our estimator (8) is chosen from a grid of values  $\bar{c} \cdot p(\log p)^2/n$  via a BIC score (see Section 3); The parameter  $\tilde{\lambda}_n^2$  in the equal-variance version of our estimator (9) is chosen in the same manner. We apply both estimators, and choose the one with smaller BIC score. We set the early stopping threshold as  $\tau^{\text{early}} = 0.1\lambda_n^2/\log p$ , so that  $\tau^{\text{early}} = o(\delta_{n,p})$  when  $\lambda_n^2 \asymp \delta_{n,p} \asymp p(\log p)^2/n$ . To ensure computational efficiency, each MIP is terminated if the runtime exceeds  $60p$  seconds, in which case the sub-optimal solution at termination is returned.

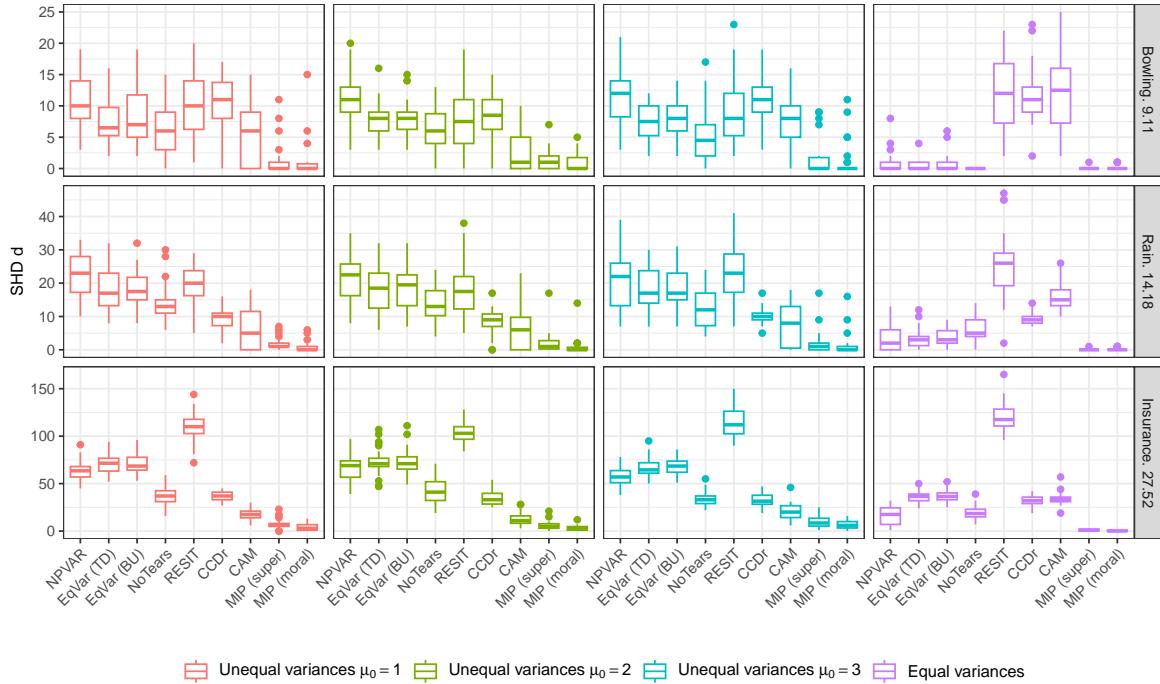


FIGURE 6. Performance comparison between our method *MIP* and baseline methods. The true functions are  $f_{kj}^*(x) = (\sin(x) - \mathbb{E}[\sin(X_k)] + \cos(x) - \mathbb{E}[\cos(X_k)])/2$ , and splines are used as the basis functions. Each row corresponds to one graph structure, indicated by name.*p.s\**. The EqVar (TD) and EqVar (BU) refer to the top-down and bottom-up version of *EqVar*, respectively. The MIP (super) and MIP (moral) refer to the *MIP* using the estimated superstructure and the true moral graph, respectively. Different colors correspond to different variance schemes. Each box summarizes results from 30 independent trials.

Figure 6 shows the performance of each method in terms of graph recovery on three representative graph structures, with the number of nodes  $p$  ranging from 9 to 27 and the number of true edges  $s^*$  ranging from 11 to 52. Implementation details for the baseline methods are provided in Appendix H.6.1. We observe that both versions of the *MIP* algorithm consistently achieve the lowest

SHD across all four variance settings. Notably, the *MIP* variant using the estimated super-structure, *MIP-super*, performs comparably to the version that relies on the oracle moral graph, *MIP-moral*, suggesting the robustness of the super-structure estimation procedure. Complete results for all ten graph structures, together with running times and further discussion, are presented in H.6.1 and Tables 1–4. We also provide additional experiments with another choice of  $f_{kj}^*$  and radial basis functions in Appendix H.6.2 and Tables 5–6. These yield similar observations, demonstrating the improved performance of our approach relative to the others.

**6.3. Simulations: role of number of basis functions  $R_n$ .** A larger number of basis functions  $R_n$  improves the approximation accuracy of each function class  $\mathcal{F}_{kj}$  by its truncated version  $\mathcal{F}_{kj,n}$ . Theoretically, a non-parametric model with  $R_n \asymp n^\zeta$  for some appropriately chosen  $\zeta > 0$  can substantially improve graph recovery and the convergence rate of variance estimators. We aim to demonstrate this through simulation.

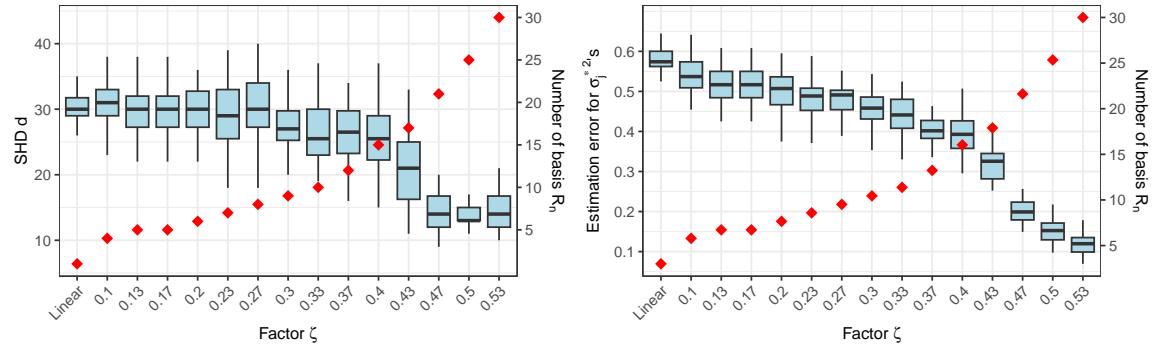


FIGURE 7. The performance of our method *MIP* with  $R_n = n^\zeta$  basis functions across a range of  $\zeta$  values. Left panel: the blue boxes represent the structural Hamming distance  $d$ , measured using its left vertical axis. Right panel: the blue boxes represent the values of  $p^{-1} \sum_j (\hat{\sigma}_j^2 - \sigma_j^{*2})^2$ , measured using its left vertical axis. Each box summarizes results from 30 independent trials. For both panels, the red dots indicate the corresponding  $R_n$  values, measured using their right vertical axis. The ‘‘linear’’ label refers to *MIP* applied directly to the design matrix  $X$  without incorporating any additional basis functions.

In this simulation, we use a graph structure with  $p = 15$  and  $s^* = 25$ . For all  $(k, j) \in E^*$ , the true function is set as  $f_{kj}^*(x) = \sin(20/x) - \mathbb{E}[\sin(20/X_k)]$ , which exhibits highly oscillatory behavior near the origin and is particularly challenging to approximate using basis functions. The noise variances  $\sigma_j^*$  alternate between 0.5 and 1 across nodes. We apply procedure (8) and use the same set of degree-three splines across all edges. For a given  $\zeta > 0$ , we place  $\lfloor n^\zeta \rfloor$  internal knots with evenly spaced percentiles, resulting in a total of  $R_n = 3 + \lfloor n^\zeta \rfloor$  basis functions. We adopt the true moral graph as the super-structure; however, we do not use any partial order set, stable set, or early stopping. The tuning parameter  $\lambda_n^2$  is fixed at 0.01, and the *MIP* algorithm is terminated after 30 minutes. Figure 7 compares the performance of *MIP* in terms of both graph recovery and variance estimation accuracy across different values of  $\zeta$ . As  $\zeta$  increases, the number of basis functions  $R_n$  grows exponentially with the sample size, enabling the model to capture increasingly complex patterns in the true functions. Consequently, both the SHD and the variance estimation error  $p^{-1} \sum_j (\hat{\sigma}_j^2 - \sigma_j^{*2})^2$  exhibit a decreasing trend. In contrast, when no basis functions are used, as indicated by the ‘‘linear’’ label, the model performs substantially worse in both aspects.

**6.4. Real data.** Causal chambers (Gamella et al., 2025) are computer-controlled devices that allow for manipulating and measuring variables from physical systems, providing a rich testbed for causal discovery algorithms. For the light tunnel chamber, we focus on the the ground truth involving eight variables, as shown in Figure 8a. Among all the causal edges, the only non-linear causal effect is  $\theta_1 - \theta_2 \rightarrow \tilde{I}_3$ , with the true function  $f_{(\theta_1 - \theta_2) \rightarrow \tilde{I}_3}^*(\theta_1 - \theta_2) = \beta_0 + \beta_1 \cos^2(\theta_1 - \theta_2)$  for some  $\beta_0, \beta_1 \in \mathbb{R}$ . We collect  $n = 3000$  observational samples of the eight variables and apply both *MIP* and the baseline methods. The implementation details follow those described in Section 6.2. Figure 8c presents a comparison of their performances, where *MIP* achieves the most accurate estimation. In particular, as shown in Figure 8b, *MIP* successfully recovers the skeleton of the true graph, including the non-linear effect. However, it incorrectly assigns the directions between  $(R, B, G)$  and  $\tilde{C}$ , which may stem from the non-identifiability of linear structural equation models.

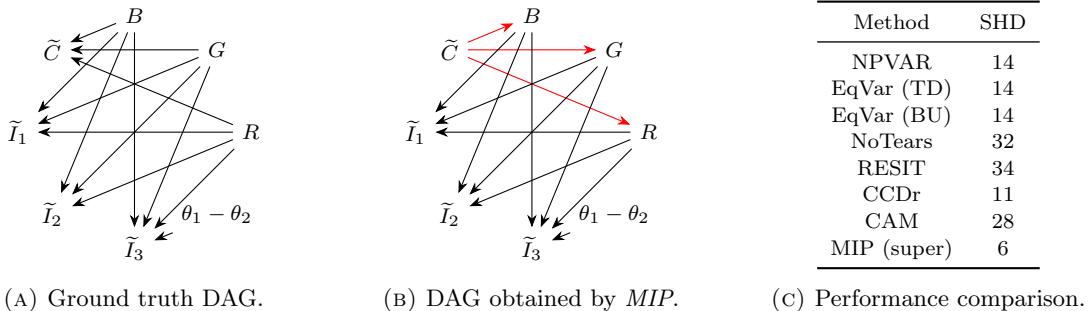


FIGURE 8. Learning causal models of the light tunnel chamber in Gamella et al. (2025). Left panel: the ground truth. Middle panel: the estimated graph by our method *MIP*, where red edges indicate incorrect discoveries. Right panel: performance comparison for all methods.

## 7. DISCUSSION

We propose and theoretically analyze a convex MIP formulation for causal discovery in causal additive models. Unlike existing methods, our method enjoys both optimization and statistical guarantees: it can obtain a solution up to a pre-specified optimality gap, and we prove that it consistently estimates the underlying DAG, even as the number of nodes grows with the sample size. Our experiments on synthetic and real data highlight improvements over other approaches.

Our work opens several avenues for future research. First, our causal additive model in (2) does not allow interactions among the variables; developing extensions of our framework to account for such interactions would be valuable. Second, our estimator relies on Gaussian errors; adapting it to non-Gaussian errors would broaden the utility of our framework. Third, in the linear causal discovery setting, Xu et al. (2025b) proposed a tailored coordinate descent method to approximately solve their MIP and proved its asymptotic optimality. Applying a similar approach to causal additive models could be of practical interest. Finally, we relied on modern commercial solvers to solve our optimization problem; designing specialized branch-and-bound methods that exploit the problem structure may lead to substantial computational gains (Hazimeh et al., 2020).

## ACKNOWLEDGMENTS

We acknowledge funding from NSF grant DMS-2413074 (Xiaozhu Zhang and Armeen Taeb).

## REFERENCES

- Aragam, B. and Zhou, Q. (2015), ‘Concave penalized estimation of sparse gaussian bayesian networks’, *Journal of Machine Learning Research* **16**(1), 2273–2328.
- Bennett, G. (1962), ‘Probability inequalities for the sum of independent random variables’, *Journal of the American Statistical Association* **57**(297), 33–45.
- Bühlmann, P., Peters, J. and Ernest, J. (2014), ‘CAM: Causal additive models, high-dimensional order search and penalized regression’, *The Annals of Statistics* **42**(6), 2526 – 2556.
- Chakraborty, S. and Shojaie, A. (2022), ‘Nonparametric causal structure learning in high dimensions’, *Entropy* **24**(3), 351.
- Chen, W., Drton, M. and Wang, Y. S. (2019), ‘On causal discovery with an equal-variance assumption’, *Biometrika* **106**(4), 973–980.
- Chen, Y., Ning, C. and Goulart, P. (2023), ‘A unified early termination technique for primal-dual algorithms in mixed integer conic programming’, *IEEE Control Systems Letters* **7**, 2803–2808.
- Chickering, D. M. (2002), ‘Optimal structure identification with greedy search’, *Journal of Machine Learning Research* **3**(Nov), 507–554.
- Efromovich, S. (1999), *Nonparametric Curve Estimation: Methods, Theory, and Applications*, Springer.
- Gamella, J. L., Peters, J. and Bühlmann, P. (2025), ‘Causal chambers as a real-world physical testbed for ai methodology’, *Nature Machine Intelligence* **7**(1), 107–118.
- Gao, M., Ding, Y. and Aragam, B. (2020), ‘A polynomial-time algorithm for learning nonparametric causal graphs’, *Advances in Neural Information Processing Systems* **33**, 11599–11611.
- Ghoshal, A. and Honorio, J. (2019), ‘Direct estimation of difference between structural equation models in high dimensions’, *arXiv preprint arXiv:1906.12024* .
- Glymour, C., Zhang, K. and Spirtes, P. (2019), ‘Review of causal discovery methods based on graphical models’, *Frontiers in Genetics* **10**, 524.
- Grötschel, M., Jünger, M. and Reinelt, G. (1985), ‘On the acyclic subgraph polytope’, *Mathematical Programming* **33**(1), 28–42.
- Haris, A., Simon, N. and Shojaie, A. (2022), ‘Generalized sparse additive models’, *Journal of machine learning research* **23**(70), 1–56.
- Hastie, T. and Tibshirani, R. (1986), ‘Generalized additive models’, *Statistical Science* **1**(3), 297–310.
- Hauser, A. and Bühlmann, P. (2012), ‘Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs’, *Journal of Machine Learning Research* **13**(1), 2409–2464.
- Hazimeh, H., Mazumder, R. and Saab, A. (2020), ‘Sparse regression at scale: branch-and-bound rooted in first-order optimization’, *Mathematical Programming* **196**, 347 – 388.
- Heinze-Deml, C., Peters, J. and Meinshausen, N. (2018), ‘Invariant causal prediction for nonlinear models’, *Journal of Causal Inference* **6**(2), 20170016.
- Hoyer, P., Janzing, D., Mooij, J. M., Peters, J. and Schölkopf, B. (2008), Nonlinear causal discovery with additive noise models, in ‘Advances in Neural Information Processing Systems’, Vol. 21.
- Kertel, M. and Klein, N. (2025), ‘Boosting causal additive models’, *Journal of Machine Learning Research* **26**(169), 1–49.
- Küçükyavuz, S., Shojaie, A., Manzour, H., Wei, L. and Wu, H.-H. (2023), ‘Consistent second-order conic integer programming for learning bayesian networks’, *Journal of Machine Learning Research* **24**(322), 1–38.

- Lauritzen, S. L. (1996), *Graphical Models*, Oxford University Press, Oxford, United Kingdom.
- Maeda, T. N. and Shimizu, S. (2021), Causal additive models with unobserved variables, in ‘Uncertainty in Artificial Intelligence’, PMLR, pp. 97–106.
- Manzour, H., Küçükyavuz, S., Wu, H.-H. and Shojaie, A. (2021), ‘Integer programming for learning directed acyclic graphs from continuous data’, *INFORMS Journal on Optimization* **3**(1), 46–73.
- Meier, L., van de Geer, S. and Bühlmann, P. (2009), ‘High-dimensional additive modeling’, *The Annals of Statistics* **37**(6B), 3779 – 3821.
- Park, Y. W. and Klabjan, D. (2017), ‘Bayesian network learning via topological order’, *Journal of Machine Learning Research* **18**(99), 1–32.
- Peters, J. and Bühlmann, P. (2013), ‘Identifiability of Gaussian structural equation models with equal error variances’, *Biometrika* **101**(1), 219–228.
- Peters, J., Janzing, D. and Schölkopf, B. (2017), *Elements of Causal Inference: Foundations and Learning Algorithms*, MIT press.
- Peters, J., Mooij, J. M., Janzing, D. and Schölkopf, B. (2014), ‘Causal discovery with continuous additive noise models’, *Journal of Machine Learning Research* **15**(1), 2009–2053.
- Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009), ‘Sparse additive models’, *Journal of the Royal Statistical Society Series B: Statistical Methodology* **71**(5), 1009–1030.
- Reisach, A., Seiler, C. and Weichwald, S. (2021), ‘Beware of the simulated DAG! causal discovery benchmarks may be easy to game’, *Advances in Neural Information Processing Systems* **34**, 27772–27784.
- Schwarz, G. (1978), ‘Estimating the dimension of a model’, *The annals of statistics* pp. 461–464.
- Seng, J., Zecević, M., Dhami, D. S. and Kersting, K. (2022), ‘Tearing apart NOTEARS: Controlling the graph prediction via variance manipulation’, *arXiv preprint arXiv:2206.07195*.
- Shah, R. D. and Peters, J. (2020), ‘The hardness of conditional independence testing and the generalised covariance measure’, *The Annals of Statistics* **48**(3), 1514 – 1538.
- Shimizu, S., Hoyer, P. O., Hyvärinen, A. and Kerminen, A. J. (2006), ‘A linear non-gaussian acyclic model for causal discovery’, *Journal of Machine Learning Research* **7**, 2003–2030.
- Shojaie, A. and Chen, W. (2024), ‘Learning directed acyclic graphs from partial orderings’, *arXiv preprint arXiv:2403.16031*.
- Silander, T. and 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, AUAI Press, Arlington, Virginia, USA, p. 445–452.
- Solus, L., Wang, Y. and Uhler, C. (2021), ‘Consistency guarantees for greedy permutation-based causal inference algorithms’, *Biometrika* **108**(4), 795–814.
- Sondhi, A. and Shojaie, A. (2019), ‘The reduced PC-algorithm: Improved causal structure learning in large random networks’, *Journal of Machine Learning Research* **20**(164), 1–31.
- Spirites, P., Glymour, C. N. and Scheines, R. (2000), *Causation, Prediction, and Search*, MIT press.
- Strobl, E. V., Zhang, K. and Visweswaran, S. (2019), ‘Approximate kernel-based conditional independence tests for fast non-parametric causal discovery’, *Journal of Causal Inference* **7**(1), 20180017.
- Tsamardinos, I., Brown, L. E. and Aliferis, C. F. (2006), ‘The max-min hill-climbing Bayesian network structure learning algorithm’, *Machine Learning* **65**(1), 31–78.
- Uhler, C., Raskutti, G., Bühlmann, P. and Yu, B. (2013), ‘Geometry of the faithfulness assumption in causal inference’, *The Annals of Statistics* **41**(2), 436–463.
- van de Geer, S. and Bühlmann, P. (2013), ‘ $\ell_0$ -penalized maximum likelihood for sparse directed acyclic graphs’, *The Annals of Statistics* **41**(2), 536 – 567.
- Van der Vaart, A. W. (2000), *Asymptotic Statistics*, Vol. 3, Cambridge university press.
- Vershynin, R. (2018), *High-dimensional Probability: An Introduction with Applications in Data Science*, Vol. 47, Cambridge university press.

- Voorman, A., Shojaie, A. and Witten, D. (2014), ‘Graph estimation with joint additive models’, *Biometrika* **101**(1), 85–101.
- Wang, Y. S. and Drton, M. (2020), ‘High-dimensional causal discovery under non-gaussianity’, *Biometrika* **107**(1), 41–59.
- Wei, D., Gao, T. and Yu, Y. (2020), ‘Dags with no fears: A closer look at continuous optimization for learning bayesian networks’, *Advances in Neural Information Processing Systems* **33**, 3895–3906.
- Xu, T., Küçükyavuz, S., Shojaie, A. and Taeb, A. (2025b), ‘An asymptotically optimal coordinate descent algorithm for learning bayesian networks from gaussian models’, *Accepted to Journal of Machine Learning Research* **abs/2408.11977**.
- Xu, T., Taeb, A., Küçükyavuz, S. and Shojaie, A. (2025a), ‘Integer programming for learning directed acyclic graphs from non-identifiable gaussian models’, *Biometrika* p. asaf032.
- Yu, Y., Chen, J., Gao, T. and Yu, M. (2019), DAG-GNN: DAG structure learning with graph neural networks, in ‘International Conference on Machine Learning’, PMLR, pp. 7154–7163.
- Zhang, K., Peters, J., Janzing, D. and Schölkopf, B. (2011), ‘Kernel-based conditional independence test and application in causal discovery’, *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence* p. 804–813.
- Zheng, X., Aragam, B., Ravikumar, P. K. and Xing, E. P. (2018), ‘DAGs with NO TEARS: continuous optimization for structure learning’, *Advances in Neural Information Processing Systems* **31**, 9492–9503.
- Zheng, X., Dan, C., Aragam, B., Ravikumar, P. and Xing, E. (2020), Learning sparse nonparametric dags, in ‘International conference on artificial intelligence and statistics’, Pmlr, pp. 3414–3425.

#### APPENDIX A. DISCUSSION OF THE ZERO-MEAN FORMULATION

Model (2) assumes

$$X_j = \sum_{k \in \text{pa}(j)} f_{kj}^*(X_k) + \epsilon_j, \quad \epsilon_j \sim N(0, \sigma_j^{*2}), \quad \mathbb{E}[f_{kj}^*(X_k)] = 0.$$

We show that the zero-mean requirement and the absence of intercepts are without loss of generality. Start from a general additive structural equation model on the same oracle DAG  $\mathcal{G}^* = (V, E^*)$ , allowing intercepts and non-zero mean components,

$$(11) \quad X_j = \mu_j + \sum_{k \in \text{pa}(j)} \tilde{f}_{kj}(X_k) + \tilde{\epsilon}_j, \quad j \in V,$$

where the  $\tilde{f}_{kj}$  are arbitrary measurable functions and  $\tilde{\epsilon}_1, \dots, \tilde{\epsilon}_p$  are mutually independent Gaussians with variances  $\sigma_j^{*2}$  but possibly non-zero means.

For each  $(k, j) \in E^*$  let  $m_{kj} := \mathbb{E}[\tilde{f}_{kj}(X_k)]$ , and for each  $j$  let  $m_j := \mathbb{E}[\tilde{\epsilon}_j]$ . Define

$$f_{kj}^*(x) := \tilde{f}_{kj}(x) - m_{kj}, \quad \epsilon_j := \tilde{\epsilon}_j - m_j, \quad \mu'_j := \mu_j + \sum_{k \in \text{pa}(j)} m_{kj} + m_j.$$

Then (11) can be equivalently formulated as

$$X_j = \mu'_j + \sum_{k \in \text{pa}(j)} f_{kj}^*(X_k) + \epsilon_j,$$

with  $\mathbb{E}[f_{kj}^*(X_k)] = 0$  and  $\epsilon_j \sim N(0, \sigma_j^{*2})$  by construction. Only constants were subtracted, so  $\text{pa}(j)$ ,  $\mathcal{G}^*$ , and the induced distribution of  $X$  are unchanged. Moreover, in this representation  $\mu'_j = \mathbb{E}[X_j]$ . Center the variables by  $X_j^c := X_j - \mu'_j$ . By defining the translated functions

$$g_{kj}^*(x) := f_{kj}^*(x + \mu'_k),$$

we obtain

$$X_j^c = \sum_{k \in \text{pa}(j)} g_{kj}^*(X_k^c) + \epsilon_j.$$

Each  $g_{kj}^*$  is a translation of  $f_{kj}^*$ , hence it is non-linear and three-times differentiable whenever  $f_{kj}^*$  is. Therefore, the centered model still satisfies the standing assumptions of Proposition 2. Finally, by relabeling the centered variables  $X_j^c$  again as  $X_j$ , and the translated functions  $g_{kj}^*$  again as  $f_{kj}^*$ , we recover exactly the normalized form (2). Thus the zero-mean convention is without loss of generality.

Note that when  $\text{pa}(j) = \emptyset$ , centering removes the intercept entirely, so the equation for  $X_j$  has no intercept term. This matches the constraint (8c) in our MIP that forces a zero intercept whenever node  $j$  has no parents.

#### APPENDIX B. PROOF OF PROPOSITION 3

*Proof.* Recall the definition

$$Z := [X_1, \dots, X_p, 1, b_1(X_1), \dots, b_1(X_p), \dots, b_{R_n}(X_1), \dots, b_{R_n}(X_p)] \in \mathbb{R}^{pR_n+p+1}.$$

Let  $Z^{(i)} \in \mathbb{R}^{pR_n+p+1}$  denote the  $i$ -th sample of the random vector  $Z$ . Moreover, we defined the mapping  $\theta = \mu(\Gamma)$ : for any  $\Gamma \in \mathbb{R}^{p \times (pR_n+p+1)}$ ,  $\sum_{k=1}^p f_{kj}(\cdot) = -[\Gamma_{j,p+1} + \sum_{k=1}^p \sum_{r=1}^{R_n} \Gamma_{j,pr+1+k}]$ .

$b_r(\cdot)] \cdot \Gamma_{jj}^{-1}$  and  $\sigma_j = \Gamma_{jj}^{-1}$ . Denote the  $j$ -th row of the matrix  $\Gamma$  as  $\gamma_j \in \mathbb{R}^{pR_n+p+1}$ . Now replacing  $\theta$  with  $\mu(\Gamma)$  in  $\ell_n(\theta)$ , we obtain

$$\begin{aligned}\ell_n(\mu(\Gamma)) &= \sum_{j=1}^p \log \Gamma_{jj}^{-2} + \sum_{j=1}^p \frac{\|X_j + \Gamma_{jj}^{-1}[\Gamma_{j,p+1} + \sum_{k=1}^p \sum_{r=1}^{R_n} \Gamma_{j,pr+1+k} b_r(\cdot)]\|_n^2}{\Gamma_{jj}^{-2}} \\ &= \sum_{j=1}^p -2 \log \Gamma_{jj} + \sum_{j=1}^p \left\| \Gamma_{jj} X_j + \Gamma_{j,p+1} + \sum_{k=1}^p \sum_{r=1}^{R_n} \Gamma_{j,pr+1+k} b_r(X_k) \right\|_n^2 \\ &= \sum_{j=1}^p -2 \log \Gamma_{jj} + \frac{1}{n} \sum_{j=1}^p \sum_{i=1}^n (\gamma_j^\top Z^{(i)})^2 = \sum_{j=1}^p -2 \log \Gamma_{jj} + \text{tr}(\Gamma^\top \Gamma \widehat{\Sigma}_n),\end{aligned}$$

where we defined  $\widehat{\Sigma}_n = n^{-1} \sum_{i=1}^n Z^{(i)\top} Z^{(i)}$ .  $\square$

## APPENDIX C. DISCUSSION AND IMPLICATIONS OF ASSUMPTIONS

**C.1. Valid basis functions and Sobolev ellipsoid.** Under Assumption 3, for each  $(k, j) \in \mathcal{E}$ , the function classes  $\mathcal{F}_{kj}$  can be rewritten as

$$(12) \quad \mathcal{F}_{kj} := \left\{ f_{kj} : f_{kj}(\cdot) = \beta_{0kj} + \sum_{r=1}^{\infty} \beta_{rkj} b_{rkj}(\cdot), \quad \sum_{r=1}^{\infty} |\beta_{rkj}| \cdot r^{\eta} \leq C, \quad \eta > 1, \right. \\ \left. \beta_{0kj} = -\mathbb{E} \left[ \sum_{r=1}^{\infty} \beta_{rkj} b_{rkj}(X_k) \right] \right\}.$$

Each basis function  $b_{rkj}(\cdot)$  for  $r \geq 1$  satisfies the following: (i) it has infinite support; (ii) it is uniformly bounded with  $|b_{rkj}(\cdot)| \leq 1$ ; (iii) it is orthogonal with  $\mathbb{E}[b_{rkj}(X_k) b_{r'kj}(X_k)] = v \mathbb{1}(r = r')$  for some  $v \in (0, 1]$ ; (iv) it is three times differentiable. One example construction of such basis functions, using the sine system (Efromovich, 1999), is given by

$$b_{rkj}(\cdot) = \sin(\pi r \Phi_k(\cdot)), \quad r \in \{1, \dots, R_n\},$$

where  $\Phi_k(\cdot)$  is the cumulative density function of  $X_k$ . Let  $U = \Phi_k(X_k) \sim \text{Unif}(0, 1)$ , then we see that

$$\mathbb{E}[b_{rkj}^2(X_k)] = \mathbb{E}[\sin^2(\pi r \Phi_k(X_k))] = \mathbb{E}[\sin^2(\pi r U)] = 1/2,$$

and for  $r \neq r'$ ,

$$\mathbb{E}[b_{rkj}(X_k) \cdot b_{r'kj}(X_k)] = \mathbb{E}[\sin(\pi r \Phi_k(X_k)) \cdot \sin(\pi r' \Phi_k(X_k))] = \mathbb{E}[\sin(\pi r U) \cdot \sin(\pi r' U)] = 0.$$

The structure on coefficients,  $\sum_{r=1}^{\infty} |\beta_{rkj}| \cdot r^{\eta} \leq C$  or equivalently  $\sum_{r=0}^{\infty} |\beta_{rkj}| \cdot r^{\eta} \leq C$ , corresponds to a Sobolev ellipsoid. This structure concentrates most of the “mass” in the early coefficients, resulting in a fast decay. The complexity of each function class  $\mathcal{F}_{kj}$ , as a result, can be effectively controlled.

**Lemma 7.** *Under Assumption 3, the bracketing entropy of  $\mathcal{F}_{kj}$  in (12) is bounded by:*

$$\log N_{[]}(\epsilon, \mathcal{F}_{kj}, \|\cdot\|_{\infty}) \lesssim \left( \frac{1}{\epsilon} \right)^{1/\eta} \log(1/\epsilon).$$

*Proof.* To simplify notation, we drop the subscripts  $k$  and  $j$ , and write  $\mathcal{F}$  for the function class  $\mathcal{F}_{kj}$ . Define the ellipsoid set  $\mathcal{B} = \{\beta : \sum_{r=0}^{\infty} |\beta_r| \cdot r^{\eta} \leq C\}$  with  $\eta > 1$ . The idea is to represent the complexity of  $\mathcal{F}$  via a mapping to the ellipsoid set  $\mathcal{B}$ .

First, we find the covering number of the set  $\mathcal{B}$ . We define a truncated version  $\tilde{\mathcal{B}} = \{\beta \in \mathcal{B} : \beta_r = 0 \text{ for all } r > t\}$ , where  $t$  is the smallest integer such that  $t^{-\eta} \leq \epsilon$ . We note that  $\tilde{\mathcal{B}}$  is contained in a  $(t+1)$ -dimensional  $\ell_1$ -ball with radius  $C$ , so

$$N(\epsilon, \tilde{\mathcal{B}}, \|\cdot\|_1) \leq \left(1 + \frac{2C}{\epsilon}\right)^{t+1}.$$

By selecting  $t = \lceil (1/\epsilon)^{1/\eta} \rceil - 1$ , we have

$$\log N(\epsilon, \tilde{\mathcal{B}}, \|\cdot\|_1) \lesssim \left(\frac{1}{\epsilon}\right)^{1/\eta} \log(1/\epsilon).$$

Now let  $\{\beta^1, \dots, \beta^N\}$  be an  $\epsilon$ -cover for  $\tilde{\mathcal{B}}$  w.r.t.  $\|\cdot\|_1$  norm. Then for any  $\beta \in \mathcal{B}$ , we take the  $\beta^i$  such that  $\sum_{j=0}^t |\beta_j - \beta_j^i| \leq \epsilon$ , and then

$$\|\beta - \beta^i\|_1 = \sum_{j=0}^t |\beta_j - \beta_j^i| + \sum_{j=t+1}^{\infty} |\beta_j| \leq \epsilon + t^{-\eta} \sum_{j=t+1}^{\infty} |\beta_j| \cdot t^{\eta} \leq (C+1)\epsilon.$$

This implies that  $\{\beta^1, \dots, \beta^N\}$  is a  $(C+1)\epsilon$ -cover of  $\mathcal{B}$ . Therefore,

$$\log N(\epsilon, \mathcal{B}, \|\cdot\|_1) = \log N(\epsilon/(C+1), \tilde{\mathcal{B}}, \|\cdot\|_1) \lesssim \left(\frac{1}{\epsilon}\right)^{1/\eta} \log(1/\epsilon).$$

Next, we find the covering number for the function class  $\mathcal{F}$ . Note that for any  $f \in \mathcal{F}$ , there exists some  $\beta \in \mathcal{B}$  we use to construct  $f$ :  $f = \beta_0 + \sum_{r=1}^{\infty} \beta_r b_r$ . Let  $\{\beta^1, \dots, \beta^N\}$  be an  $\epsilon$ -cover of  $\mathcal{B}$  w.r.t.  $\|\cdot\|_1$  norm. Then there must exist  $i \in \{1, \dots, N\}$  such that  $\|\beta^i - \beta\|_1 \leq \epsilon$ . Therefore,

$$\left\| f - \beta_0^i - \sum_{r=1}^{\infty} \beta_r^i b_r \right\|_{\infty} \leq |\beta_0 - \beta_0^i| + \left\| \sum_{r=1}^{\infty} (\beta_r - \beta_r^i) b_r \right\|_{\infty} \leq \sum_{r=0}^{\infty} |\beta_r - \beta_r^i| \leq \epsilon.$$

This implies that  $\{\beta_0^i + \sum_{r=1}^{\infty} \beta_r^i b_r\}_{i=1}^N$  is an  $\epsilon$ -cover of  $\mathcal{F}$  w.r.t  $\|\cdot\|_{\infty}$  norm. Hence,

$$\log N(\epsilon, \mathcal{F}, \|\cdot\|_{\infty}) \lesssim \left(\frac{1}{\epsilon}\right)^{1/\eta} \log(1/\epsilon).$$

Finally, we arrive at the bracketing entropy of  $\mathcal{F}$ :

$$\log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|_{\infty}) \leq \log N(\epsilon/2, \mathcal{F}, \|\cdot\|_{\infty}) \lesssim \left(\frac{1}{\epsilon}\right)^{1/\eta} \log(1/\epsilon).$$

□

Recall the definition of truncated function class  $\mathcal{F}_{kj,n} = \{f_{kj} \in \mathcal{F}_{kj} : f_{kj,n} = \beta_{0kj} + \sum_{r=1}^{R_n} \beta_{rkj} b_{rkj}\}$ . For any  $f \in \mathcal{F}_{kj}$  where  $f_{kj} = \beta_{0kj} + \sum_{r=1}^{\infty} \beta_{rkj} b_{rkj}$ , the residual after projecting it onto  $\mathcal{F}_{kj,n}$  is given by  $D_{kj} = f_{kj} - \beta_{0kj} - \sum_{r=1}^{R_n} \beta_{rkj} b_{rkj}$ . The Sobolev ellipsoid structure allows for effective control of the truncation error  $\|D_{kj}\|_{\infty}$ .

**Lemma 8.** *Under Assumption 3, the truncation error decreases with  $R_n$ :  $\|D_{kj}\|_{\infty} \lesssim R_n^{-\eta}$ .*

*Proof.* We note that  $D_{kj} = f_{kj} - \beta_{0kj} - \sum_{r=1}^{R_n} \beta_{rkj} b_{rkj} = \sum_{r=R_n+1}^{\infty} \beta_{rkj} (b_{rkj} - \mu_{rkj})$ , where  $\mu_{rkj} = \mathbb{E}[b_{rkj}(X_k)]$ . Therefore,

$$\|D_{kj}\|_{\infty} = \left\| \sum_{r=R_n+1}^{\infty} \beta_{rkj} (b_{rkj} - \mu_{rkj}) \right\|_{\infty} \leq 2 \sum_{r=R_n+1}^{\infty} |\beta_{rkj}| = 2R_n^{-\eta} \sum_{r=R_n+1}^{\infty} |\beta_{rkj}| \cdot R_n^{\eta} \leq 2C R_n^{-\eta}.$$

□

Finally, we show that the function classes are uniformly bounded.

**Lemma 9.** *Under Assumption 3, the functions in  $\mathcal{F}_{kj}$  are uniformly bounded:*

$$\sup_{(k,j) \in \mathcal{E}} \sup_{f_{kj} \in \mathcal{F}_{kj}} \|f_{kj}(\cdot)\|_{\infty} \leq 2C,$$

where  $C$  is the constant appeared in Assumption 3.

*Proof.* For any  $f_{kj} \in \mathcal{F}_{kj}$ ,  $\forall (k,j) \in \mathcal{E}$ , we have

$$\|f_{kj}(x)\|_{\infty} = \left\| \beta_{0kj} + \sum_{r=1}^{\infty} \beta_{rkj} b_{rkj}(x) \right\|_{\infty} \leq \sum_{r=0}^{\infty} |\beta_{rkj}| \leq |\beta_{0kj}| + \sum_{r=1}^{\infty} |\beta_{rkj}| \cdot r^{\eta} \leq |\beta_{0kj}| + C,$$

where

$$\beta_{0kj} = -\mathbb{E} \left[ \sum_{r=1}^{\infty} \beta_{rkj} b_{rkj}(X_k) \right] \leq \sum_{r=1}^{\infty} |\beta_{rkj}| \leq \sum_{r=1}^{\infty} |\beta_{rkj}| \cdot r^{\eta} \leq C.$$

The result hence follows. □

**C.2. More on boundedness of variances.** Recall that in Section 5, we defined the maximal variances  $\bar{\nu}_p := \max_{j \in V} \mathbb{E}[X_j^2]$ , and the minimal variance  $\underline{\nu}_p := \min_{j \in V} \min_{|\vec{g}_j|=K} \nu_j(\vec{g}_j)$ . First, we point out that for any noise variance  $\sigma_j^{*2}$ , there exists  $\vec{g}_j$  such that  $\nu_j(\vec{g}_j) = \sigma_j^{*2}$ ; see Lemma 12 for a proof. Next, we see that for any  $\nu_j(\vec{g}_j)$ ,

$$\underline{\nu}_p = \min_{j \in V} \min_{|\vec{g}_j| \in \{0, \dots, K\}} \nu_j(\vec{g}_j) \leq \nu_j(\vec{g}_j) \leq \max_{j \in V} \max_{|\vec{g}_j| \in \{0, \dots, K\}} = \bar{\nu}_p.$$

The conditions in Assumption 4 ensure uniform boundedness of these quantities.

**Lemma 10.** *Denote  $\underline{\sigma} = \min_{j \in V} \sigma_j^*$ . Under Assumptions 3 and 4, the variances are uniformly bounded:  $\underline{c} \leq \underline{\nu}_p \leq \underline{\sigma}^2 \leq \bar{\sigma}^2 \leq \bar{\nu}_p \leq \bar{c}$  for some  $\underline{c} > 0$  and  $\bar{c} < \infty$ .*

*Proof.* It suffices to show that there exists  $\bar{c} < \infty$  such that  $\bar{\nu}_p \leq \bar{c}$ . Note that

$$\bar{\nu}_p = \max_{j \in V} \mathbb{E}[X_j^2] = \max_{j \in V} \mathbb{E} \left[ 2 \left( \sum_{k \in \text{pa}(j)} f_{kj}^*(X_k) \right)^2 + 2\epsilon_j^2 \right] \leq 8C^2 K^2 + 2\bar{\sigma}^2 < \infty.$$

The last inequality follows from the fact that  $|\text{pa}(j)| \leq K$ , and Lemma 9. □

We also defined the gap  $d_{n,p} := \max_{j \in V} \max_{|\vec{g}_j| \in \{0, \dots, K\}} |\nu_j(\vec{g}_j) - \nu_j^n(\vec{g}_j)|$ . As  $n$  is sufficiently large, the function class  $\mathcal{F}_n^{\oplus \vec{g}_j}$  is expected to closely approximate its non-truncated counterpart  $\mathcal{F}^{\oplus \vec{g}_j}$ . Particularly, when  $n = \infty$ , the function class  $\mathcal{F}_n^{\oplus \vec{g}_j}$  and  $\mathcal{F}^{\oplus \vec{g}_j}$  coincide exactly for any  $\vec{g}_j$ , and thus  $\nu_j^n(\vec{g}_j) = \nu_j(\vec{g}_j)$ . This observation motivates the following result:  $d_{n,p} = o(1)$ .

**Lemma 11.** *Under Assumptions 3 and 4, the gap between  $\nu_j(\vec{g}_j)$  and  $\nu_j^n(\vec{g}_j)$  is sufficiently small for any  $\vec{g}_j$ :  $d_{n,p} = o(1)$ .*

*Proof.* We denote

$$\bar{f} \in \operatorname{argmin}_{f \in \mathcal{F}^{\oplus \vec{g}_j}} \mathbb{E}[X_j - \sum_{k \in \vec{g}_j} f_{kj}(X_k)]^2,$$

and let  $\bar{f}^{(n)} \in \mathcal{F}_n^{\oplus \vec{g}_j}$  be the truncation of  $\bar{f}$  to the first  $R_n$  basis functions. Note that for any  $\vec{g}_j$  and any  $j \in V$ ,

$$\begin{aligned} \nu_j^n(\vec{g}_j) - \nu_j(\vec{g}_j) &\leq \mathbb{E} \left[ X_j - \sum_{k \in \vec{g}_j} \bar{f}_{kj}^{(n)}(X_j) \right]^2 - \mathbb{E} \left[ X_j - \sum_{k \in \vec{g}_j} \bar{f}_{kj}(X_j) \right]^2 \\ &= \mathbb{E} \left[ X_j - \sum_{k \in \vec{g}_j} \bar{f}_{kj}(X_j) + \sum_{k \in \vec{g}_j} \bar{f}_{kj}(X_j) - \sum_{k \in \vec{g}_j} \bar{f}_{kj}^{(n)}(X_j) \right]^2 - \mathbb{E} \left[ X_j - \sum_{k \in \vec{g}_j} \bar{f}_{kj}(X_j) \right]^2 \\ &= 2 \cdot \mathbb{E} \left[ \left( X_j - \sum_{k \in \vec{g}_j} \bar{f}_{kj}(X_j) \right) \left( \sum_{k \in \vec{g}_j} \bar{f}_{kj}(X_j) - \sum_{k \in \vec{g}_j} \bar{f}_{kj}^{(n)}(X_j) \right) \right] + \mathbb{E} \left[ \sum_{k \in \vec{g}_j} \bar{f}_{kj}(X_j) - \sum_{k \in \vec{g}_j} \bar{f}_{kj}^{(n)}(X_j) \right]^2 \\ &\leq 2 \cdot (\nu_j(\vec{g}_j))^{1/2} \cdot \Delta_f^{1/2} + \Delta_f, \end{aligned}$$

where  $\Delta_f = \mathbb{E} \left[ \sum_{k \in \vec{g}_j} \bar{f}_{kj}(X_j) - \sum_{k \in \vec{g}_j} \bar{f}_{kj}^{(n)}(X_j) \right]^2$ . By Lemma 10, we have  $(\nu_j(\vec{g}_j))^{1/2} \leq \bar{c}^{1/2}$ . Moreover, letting  $\mu_{rkj} = \mathbb{E}[b_{rkj}(X_k)] \leq 1$ , we have

$$\begin{aligned} \Delta_f &\leq |\vec{g}_j| \sum_{k \in \vec{g}_j} \mathbb{E} \left[ \bar{f}_{kj}(X_j) - \bar{f}_{kj}^{(n)}(X_j) \right]^2 \\ &\leq K \sum_{k \in \vec{g}_j} \sum_{r=R_n+1}^{\infty} \sum_{r'=R_n+1}^{\infty} \beta_{rkj} \beta_{r'kj} \mathbb{E} [(b_{rkj}(X_k) - \mu_{rkj})(b_{r'kj}(X_k) - \mu_{r'kj})] \\ &\leq K^2 v \sum_{r=R_n+1}^{\infty} \beta_{rkj}^2 + K^2 \left[ \sum_{r=R_n+1}^{\infty} |\beta_{rkj} \mu_{rkj}| \right]^2 \\ &\leq K^2 v R_n^{-2\eta} \left[ \sum_{r=R_n+1}^{\infty} \beta_{rkj}^2 r^{2\eta} \right] + K^2 R_n^{-2\eta} \left[ \sum_{r=R_n+1}^{\infty} \beta_{rkj}^2 r^{2\eta} \right] \leq (1+v) K^2 C^2 R_n^{-2\eta}, \end{aligned}$$

where the last inequality follows from Assumption 3 where  $\sum_{r=R_n+1}^{\infty} \beta_{rkj}^2 \cdot r^{2\eta} \leq [\sum_{r=1}^{\infty} |\beta_{rkj}| \cdot r^{\eta}]^2 \leq C^2$ . As a result, we have

$$0 \leq \nu_j^n(\vec{g}_j) - \nu_j(\vec{g}_j) \leq 2\bar{c}^{1/2}(1+v)^{1/2} K C R_n^{-\eta} + (1+v) K^2 C^2 R_n^{-2\eta}.$$

The result hence follows.  $\square$

**C.3. More on permutations and the likelihood.** Recall that for any permutation  $\pi$ , we define the parents of node  $j$  as  $\vec{g}(\pi)_j = \{k : \pi^{-1}(k) < \pi^{-1}(j), (k, j) \in E^\circ\}$ , and the corresponding log-likelihood  $\log p(X; \theta, \pi) := \sum_{j=1}^p \log p_j(X; \theta, \pi)$ , where

$$\log p_j(X; \theta, \pi) := -\frac{1}{2} \log(2\pi) - \log \sigma_j - \frac{\left[ X_j - \sum_{k \in \vec{g}(\pi)_j} f_{kj}(X_k) \right]^2}{2\sigma_j^2}.$$

Let  $\theta^* := (\{f_{kj}^*\}_{(k,j) \in E^*}, \{\sigma_j^*\}_{j \in V})$ , and consider the expected negative log-likelihood:

$$\min_{\theta} \mathbb{E}_{\theta^*} [-\log p(X; \theta, \pi)] = \sum_{j=1}^p \min_{\theta} \mathbb{E}_{\theta^*} [-\log p_j(X; \theta, \pi)].$$

In order to find  $\theta^\pi := \operatorname{argmin}_{\theta} \mathbb{E}_{\theta^*} [-\log p(X; \theta, \pi)]$ , we analyze the minimum of each component individually. For each  $j$ , we observe that

$$\operatorname{argmin}_{\sigma_j > 0} \mathbb{E}_{\theta^*} [-\log p_j(X; \theta, \pi)] = \mathbb{E}_{\theta^*} \left[ \left( X_j - \sum_{k \in \vec{g}(\pi)_j} f_{kj}(X_k) \right)^2 \right] = \mathbb{E}_{\theta^*} [\mathcal{R}_j(f; \vec{g}(\pi)_j)].$$

Hence,

$$\min_{\theta} \mathbb{E}_{\theta^*} [-\log p_j(X; \theta, \pi)] = C + \frac{1}{2} \min_{f \in \mathcal{F}^{\oplus \vec{g}(\pi)_j}} \log \mathbb{E}_{\theta^*} [\mathcal{R}_j(f; \vec{g}(\pi)_j)] = C + \frac{1}{2} \log \nu_j(\vec{g}(\pi)_j),$$

where  $C = \log(2\pi)/2 + 1/2$ .

**Lemma 12.** *For any permutation  $\pi$ , we have  $\sum_{j=1}^p \log \nu_j(\vec{g}(\pi)_j) \geq \sum_{j=1}^p \log \sigma_j^{*2}$ . Particularly, when  $\pi \in \Pi^*$ ,  $\nu_j(\vec{g}(\pi)_j) = \sigma_j^{*2}$  for any  $j \in V$ .*

*Proof.* First, we show that when  $\pi \in \Pi^*$ ,  $\{f_{kj}^*\}_{k \in V} \in \operatorname{argmin}_f \mathbb{E}_{\theta^*} [\mathcal{R}_j(f; \vec{g}(\pi)_j)]$  and  $\nu_j(\vec{g}(\pi)_j) = \sigma_j^{*2}$ . To see this, for any  $\{f_{kj}^0\}_{k \in V}$ , we have

$$\mathbb{E}_{\theta^*} [\mathcal{R}_j(f^0; \vec{g}(\pi)_j)] = \mathbb{E}_{\theta^*} \left[ \epsilon_j^2 + \sum_{k \in \vec{g}(\pi)_j} (f_{kj}^*(X_k) - f_{kj}^0(X_k))^2 \right] \geq \sigma_j^{*2},$$

and the equality holds when  $f_{kj}^* = f_{kj}^0$ . The desired result thus follows. In addition, the result implies that, when  $\pi \in \Pi^*$ : (i)  $\theta^\pi = \theta^*$ ; (ii)  $\mathbb{E}_{\theta^*} [-\log p_j(X; \theta^*, \pi)] = C + \log(\sigma_j^*)$ .

Next, we show that  $\sum_{j=1}^p \log \nu_j(\vec{g}(\pi)_j) \geq \sum_{j=1}^p \log \sigma_j^{*2}$  for any permutation  $\pi$ . Note that

$$\frac{1}{2} \sum_{j=1}^p \log \nu_j(\vec{g}(\pi)_j) - \frac{1}{2} \sum_{j=1}^p \log(\sigma_j^{*2}) = \mathbb{E}_{\theta^*} [-\log p(X; \theta^\pi, \pi)] - \mathbb{E}_{\theta^*} [-\log p(X; \theta^*, \pi)] \geq 0.$$

The last inequality holds since the KL-divergence is non-negative. The result hence follows.  $\square$

Lemma 12 shows that when  $\pi \in \Pi^*$ , the residual variance  $\nu_j(\vec{g}(\pi)_j)$  arrives at its best: the irreducible variance. Moreover, the separation quantity  $\xi_p$  defined in (10) is non-negative. We now investigate into a more stringent argument such that  $\xi_p > 0$ .

**Lemma 13.** *Under Assumption 5, the separation is strictly positive:  $\xi_p > 0$ .*

*Proof.* For any  $\pi$ , define  $\vec{g}(\pi^\circ)_j := \{k : \pi^{-1}(k) < \pi^{-1}(j)\}$ . We have  $\vec{g}(\pi)_j \subseteq \vec{g}(\pi^\circ)_j$ . Then when  $\pi \notin \Pi^*$ ,

$$\xi_p = \min_{\pi \notin \Pi^*} p^{-1} \sum_{j=1}^p (\log \nu_j(\vec{g}(\pi)_j) - \log \sigma_j^{*2}) \geq \min_{\pi \notin \Pi^*} p^{-1} \sum_{j=1}^p (\log \nu_j(\vec{g}(\pi^\circ)_j) - \log \sigma_j^{*2}) > 0.$$

The last strict inequality follows from Lemma 3 in Bühlmann et al. (2014).  $\square$

#### APPENDIX D. PRELIMINARILY RESULTS

**D.1. Events that hold with high probability.** In this section,  $\epsilon_j \in \mathbb{R}^n$ ,  $X_j \in \mathbb{R}^n$  and  $D_{kj}^* \in \mathbb{R}^n$  represent vectors containing  $n$  independent samples, and  $X^{(i)} \in \mathbb{R}^p$  represents the  $i$ -th sample containing  $p$  variables.

**Lemma 14.** Suppose  $\epsilon_j \sim \mathcal{N}(0, \sigma_j^2)$ ,  $j = 1, \dots, p$  are independently distributed. Define the event

$$\mathcal{T}_1 := \left\{ \left| \|\epsilon_j\|_n^2 - \sigma_j^{*2} \right| \lesssim \sqrt{\log p/n}, \quad \forall j \in [p] \right\}.$$

Then  $\mathbb{P}(\mathcal{T}_1) \geq 1 - 2/p$ .

*Proof.* By the Bernstein's inequality (Bennett, 1962) we have  $\mathbb{P}(\mathcal{T}_{1,j}) \geq 1 - 2e^{-t}/p$ , where

$$\mathcal{T}_{1,j} := \left\{ \frac{\left| \|\epsilon_j\|_n^2 - \sigma_j^{*2} \right|}{\sigma_j^{*2}} \leq 2\sqrt{(t + \log p)/n} + 2(t + \log p)/n \right\}$$

By letting  $t = \log p$  and using the fact  $\mathcal{T}_1 = \cap_{j=1}^p \mathcal{T}_{1,j}$ , the result follows.  $\square$

Let  $P$  be the true distributions induced by model (2), and  $P_n$  the empirical distribution. We denote  $dx := (dx_1, \dots, dx_p)$ .

**Lemma 15.** Define  $\mu_{kk'j}(f)(x_1, \dots, x_p) := f_{kj}(x_k)f_{k'j}(x_{k'})$ . Let the combined function class for node  $j$  be  $\mathcal{F}_j^\cup := \bigcup_{k=1}^p \bigcup_{k'=1}^p \{\mu_{kk'j}(f_{kj}, f_{k'j}) : f_{kj} \in \mathcal{F}_{kj}, f_{k'j} \in \mathcal{F}_{k'j}\}$ . Define the function  $g_j(x^{(1)}, \dots, x^{(n)}) := \sup_{f \in \mathcal{F}_j^\cup} |\int f(x)(P_n - P)(dx)|$ , and the event

$$\mathcal{T}_2 := \left\{ g_j(X^{(1)}, \dots, X^{(n)}) \lesssim \sqrt{\frac{\log p}{n}}, \quad \forall j \in [p] \right\}.$$

Then  $\mathbb{P}(\mathcal{T}_2) \geq 1 - 1/p$ .

*Proof.* Note that each function class  $\mathcal{F}_{kj}$  has finite bracketing entropy by Lemma 7. Moreover, each  $f_{kj} \in \mathcal{F}_{kj}$  is uniformly bounded with  $|f_{kj}| \leq 2C$  by Lemma 9. Now for the function class  $\mathcal{F}_j^\cup$ , its bracketing entropy is given by

$$\log N_\square(\delta, \mathcal{F}_j^\cup, \|\cdot\|_\infty) \leq \log(p^2 \cdot \max_{k,j} N_\square^2(\delta/(4C), \mathcal{F}_{kj}, \|\cdot\|_\infty)) \lesssim \log p + (1/\delta)^{1/\eta} \log(1/\delta).$$

When  $\delta \asymp 1$ , the corresponding bracketing integral is given by

$$J_\square(\delta, \mathcal{F}_j^\cup, \|\cdot\|_\infty) = \int_0^\delta \sqrt{\log N_\square(u, \mathcal{F}_j^\cup, \|\cdot\|_\infty)} du \lesssim \sqrt{\log p}.$$

Hence, from Corollary 19.35 of Van der Vaart (2000), with  $F(x) = \sup_{f \in \mathcal{F}_j^\cup} |f(x)| \asymp 1$ ,

$$\mathbb{E} [g_j(X^{(1)}, \dots, X^{(n)})] \lesssim \frac{J_\square(\|F\|_\infty, \mathcal{F}_j^\cup, \|\cdot\|_\infty)}{\sqrt{n}} \lesssim \sqrt{\frac{\log p}{n}}.$$

Note that the function  $g_j(x^{(1)}, \dots, x^{(n)})$  satisfies the bounded difference property with bound  $8C^2/n$  for all  $n$  samples. To see this, note that for any  $i = 1, \dots, n$ ,

$$\begin{aligned} & \left| g_j(x^{(1)}, \dots, x^{(i)}, \dots, x^{(n)}) - g_j(x^{(1)}, \dots, x^{(i)'}, \dots, x^{(n)}) \right| \\ & \leq \sup_{f \in \mathcal{F}_j^\cup} \left| \frac{1}{n} f(x^{(i)}) - \frac{1}{n} \mathbb{E}[f(X^{(i)})] - \frac{1}{n} f(x^{(i)'}) + \frac{1}{n} \mathbb{E}[f(X^{(i)})] \right| \leq \frac{1}{n} \sup_{f \in \mathcal{F}_j^\cup} |f(x_i) - f(x^{(i)'})| \leq \frac{8C^2}{n}. \end{aligned}$$

Then by the McDiarmid's inequality, we have  $\mathbb{P}(\mathcal{T}_{2,j}) \geq 1 - e^{-t}/p$ , where

$$\mathcal{T}_{2,j} := \left\{ g_j(X^{(1)}, \dots, X^{(n)}) \leq \mathbb{E} [g_j(X^{(1)}, \dots, X^{(n)})] + C^2 \sqrt{32(t + \log p)/n} \right\}.$$

By letting  $t = \log p$  and using the fact that  $\mathcal{T}_2 = \cap_{j=1}^p \mathcal{T}_{2,j}$ , the result follows.  $\square$

**Lemma 16.** Let  $D_{kj}^* = \sum_{r=R_n+1}^\infty \beta_{rkj}^* b_{rkj}(X_k)$ . Define the event

$$\mathcal{T}_3 := \left\{ n^{-1} \sum_{k \in \text{pa}(j)} |\epsilon_j^\top D_{kj}^*| \lesssim s_j R_n^{-2\eta} + s_j \log p/n, \quad \forall j \in [p] \right\}.$$

Then  $\mathbb{P}(\mathcal{T}_3) \geq 1 - 1/p$ .

*Proof.* First consider the set

$$\mathcal{T}_{3,j} := \left\{ n^{-1} \sum_{k \in \text{pa}(j)} |\epsilon_j^\top D_{kj}^*| \lesssim s_j R_n^{-2\eta} + s_j (t + \log p)/n \right\}.$$

Since  $\epsilon_j$  and  $D_{kj}^*$  are independent for all  $k \in \text{pa}(j)$ , we focus on the conditional event  $\mathcal{T}_{3,j} \mid D_{kj}^*$ . When  $D_{kj}^*$  is fixed, by Lemma 7.4 of van de Geer and Bühlmann (2013), with probability at least  $1 - e^{-t}/p$ ,

$$(n\sigma_j^*)^{-1} \sup_{\|D_{kj}^*\|_n \leq 1} |\epsilon_j^\top D_{kj}^*| \leq \sqrt{2/n} + \sqrt{2(t + \log p)/n}.$$

Consequently, for some  $\delta_1 \in (0, 1)$ .

$$\begin{aligned} 2(n\sigma_j^*)^{-1} |\epsilon_j^\top D_{kj}^*| &= 2(n\sigma_j^*)^{-1} |\epsilon_j^\top D_{kj}^*| \cdot \frac{\|D_{kj}^*\|_n}{\|D_{kj}^*\|_n} \leq 2 \left( \sqrt{2/n} + \sqrt{2(t + \log p)/n} \right) \cdot \|D_{kj}^*\|_n \\ &\leq \delta_1 \|D_{kj}\|_n^2 + 4/(n\delta_1) + 4(t + \log p)/(n\delta_1) \lesssim R_n^{-2\eta} + (t + \log p)/n. \end{aligned}$$

The last inequality follows from Lemma 8. This implies that  $\mathbb{P}(\mathcal{T}_{3,j} \mid D_{kj}^*) \geq 1 - e^{-t}/p$ , and consequently  $\mathbb{P}(\mathcal{T}_{3,j}) \geq 1 - e^{-t}/p$ . By letting  $t = \log p$  and using the fact that  $\mathcal{T}_3 = \cap_{j=1}^p \mathcal{T}_{3,j}$ , the result follows.  $\square$

We define the following notations used in Lemma 17 and 18. Let

$$Z_j := [1_n, b_{11j}(X_1), \dots, b_{1pj}(X_p), \dots, b_{R_n 1j}(X_1), \dots, b_{R_n pj}(X_p)] \in \mathbb{R}^{n \times (pR_n + 1)},$$

and

$$\beta_j = [\beta_{0*j}, \beta_{11j}, \dots, \beta_{1pj}, \dots, \beta_{R_n1j}, \dots, \beta_{R_npj}]^\top \in \mathbb{R}^{pR_n+1}, \quad \beta_{0*j} = \sum_{k=1}^p \beta_{0kj}.$$

Moreover, denote all permutations corresponding to  $\mathcal{G}(\beta)$  as  $\Pi_\beta$ , where node  $k$  is a parent of node  $j$  if  $\sum_{r=1}^{R_n} |\beta_{rkj}|^2 \neq 0$ , and  $\beta_j = 0$  if node  $j$  has no parents.

**Lemma 17.** *We define the event*

$$\mathcal{T}_4 := \left\{ 2n^{-1} \left| \epsilon_j^\top Z_j(\beta_j^* - \tilde{\beta}_j) \right| \leq \delta_2 \|Z_j(\beta_j^* - \tilde{\beta}_j)\|_n^2 + \Delta_{n,p}, \quad \forall \tilde{\beta}_j \text{ with } \Pi_{\tilde{\beta}} \subseteq \Pi^*, \quad \forall j \in [p] \right\},$$

where  $\delta_2 \in (0, 1)$  and  $\Delta_{n,p} \lesssim s_j R_n \log p/n$ . Then  $\mathbb{P}(\mathcal{T}_4) \geq 1 - 1/p$ .

*Proof.* First we consider the set

$$\mathcal{T}_{4,j} := \left\{ 2n^{-1} \left| \epsilon_j^\top Z_j(\beta_j^* - \tilde{\beta}_j) \right| \leq \delta_2 \|Z_j(\beta_j^* - \tilde{\beta}_j)\|_n^2 + \Delta(t)_{n,p}, \quad \Pi_{\tilde{\beta}} \subseteq \Pi^* \right\},$$

where  $\Delta(t)_{n,p} \lesssim s_j(R_n + t + \log p)/n$ . When the permutations  $\Pi_{\tilde{\beta}}$  are correct,  $\epsilon_j$  and  $Z_j(\beta_j^* - \tilde{\beta}_j)$  are independent. Therefore, we focus on the conditional event  $\mathcal{T}_{4,j} \mid Z_j(\beta_j^* - \tilde{\beta}_j)$ . When  $Z_j(\beta_j^* - \tilde{\beta}_j)$  is fixed, by Lemma 7.4 of van de Geer and Bühlmann (2013), with probability at least  $1 - e^{-t}/p$ ,

$$\begin{aligned} (n\sigma_j^*)^{-1} \sup_{\|Z_j(\beta_j^* - \tilde{\beta}_j)\|_n \leq 1} \left| \epsilon_j^\top Z_j(\beta_j^* - \tilde{\beta}_j) \right| &\leq \sqrt{2(2R_n s_j + 1)/n} + \sqrt{2(t + \log p)/n} \\ &\leq \sqrt{2(2R_n s_j + 1)/n} + \sqrt{2s_j(t + \log p)/n}. \end{aligned}$$

The last inequality can be justified by considering two cases: (1) when  $s_j \geq 1$ , the inequality is trivially satisfied; (2) when  $s_j = 0$ , we have  $\beta_j^* = \tilde{\beta}_j = 0$  due to  $\{k : (k, j) \in E^\circ\} = \emptyset$ , and thus the inequality again holds trivially. Consequently, for some  $\delta_2 \in (0, 1)$ ,

$$\begin{aligned} 2(n\sigma_j^*)^{-1} \left| \epsilon_j^\top Z_j(\beta_j^* - \tilde{\beta}_j) \right| &= 2(n\sigma_j^*)^{-1} \left| \epsilon_j^\top Z_j(\beta_j^* - \tilde{\beta}_j) \right| \cdot \frac{\|Z_j(\beta_j^* - \tilde{\beta}_j)\|_n}{\|Z_j(\beta_j^* - \tilde{\beta}_j)\|_n} \\ &\leq 2 \left( \sqrt{2(2R_n s_j + 1)/n} + \sqrt{2s_j(t + \log p)/n} \right) \|Z_j(\beta_j^* - \tilde{\beta}_j)\|_n \\ &\leq \delta_2 \|Z_j(\beta_j^* - \tilde{\beta}_j)\|_n^2 + (8R_n s_j + 4)/(n\delta_2) + 4s_j(t + \log p)/(n\delta_2). \end{aligned}$$

This implies that  $\mathbb{P}(\mathcal{T}_{4,j} \mid Z_j(\beta_j^* - \tilde{\beta}_j)) \geq 1 - e^{-t}/p$ , and consequently  $\mathbb{P}(\mathcal{T}_{4,j}) \geq 1 - e^{-t}/p$ . By letting  $t = \log p$  and using the fact that  $\mathcal{T}_4 = \cap_{j=1}^p \mathcal{T}_{4,p}$ , the result follows.  $\square$

**Lemma 18.** *Define the event*

$$\mathcal{T}_5 := \left\{ \|Z_j(\beta_j^* - \beta_j)\|_n^2 \gtrsim \|\beta_j^* - \beta_j\|_2^2, \quad \forall \beta_j \text{ feasible in (8)}, \quad \forall j \in [p] \right\}.$$

Then  $\mathbb{P}(\mathcal{T}_5) \geq 1 - 1/p$ .

*Proof.* Denote the  $i$ -th row of  $Z_j$  by  $Z_j^{(i)} \in \mathbb{R}^{pR_n+1}$ . Note that all entries of  $Z_j$  are bounded in  $[-1, 1]$ , and thus they are sub-Gaussian with the same parameter. Without loss of generality, we assume that  $\|\beta_j^* - \beta_j\|_2 = 1$ . Since  $\|\beta_j^* - \beta_j\|_0 \leq 2KR_n + 1$ , we have  $(Z_j^{(i)\top} (\beta_j^* - \beta_j))^2$  being sub-exponential and

$|Z_j^{(i)\top}(\beta_j^* - \beta_j)| \leq \|\beta_j^* - \beta_j\|_1 \leq \sqrt{2KR_n + 1}$ . By the sub-exponential Bernstein inequality (Vershynin, 2018), with probability at least  $1 - e^{-t}/p$ , we have

$$\frac{1}{n} \sum_{i=1}^n (\beta_j^* - \beta_j)^\top Z_j^{(i)\top} Z_j^{(i)\top} (\beta_j^* - \beta_j) \gtrsim \mu_{Z,\beta} - \bar{\kappa}^2 \left( \sqrt{(t + \log p)/n} + (t + \log p)/n \right),$$

where  $\bar{\kappa} = \max_i \|\beta_j^\top Z_j^{(i)\top} \beta_j\|_{\psi_1} \lesssim R_n^{1/2}$ , and

$$\mu_{Z,\beta} = \mathbb{E} \left[ (\beta_{0*}^* - \beta_{0*}) + \sum_{k=1}^p \sum_{r=1}^{R_n} (\beta_{rkj}^* - \beta_{rkj}) \cdot b_{rkj}(X_k^{(1)}) \right]^2.$$

This concentration result implies that  $\mathbb{P}(\mathcal{T}_{5,j}) \geq 1 - e^{-t}/p$ , where

$$\mathcal{T}_{5,j} := \left\{ \|Z_j(\beta_j^* - \beta_j)\|_n^2 \gtrsim \mu_{Z,\beta} - \bar{\kappa}^2 \sqrt{(t + \log p)/n} - (t + \log p)/n, \quad \forall \beta_j \text{ feasible with } \|\beta_j^* - \beta_j\|_2 = 1 \right\}.$$

By Assumption 7, we have  $\bar{\kappa}^2 \lesssim \sqrt{n/\log p}$ . If  $\mu_{Z,\beta} \geq c$  for some  $c > 0$ , then by letting  $t = \log p$  and using the fact that  $\mathcal{T}_5 = \cap_{j=1}^p \mathcal{T}_{5,j}$ , the result follows.

Finally, we show that  $\mu_{Z,\beta} \geq c$  for some  $c > 0$ . Note that the function  $\bar{f}_{kj}(x_k) = \frac{1}{2} \sum_{r=1}^{R_n} (\beta_{rkj}^* - \beta_{rkj})(b_{rkj}(x_k) - \mu_{rkj}) \in \mathcal{F}_{kj}$ , where  $\mu_{rkj} = \mathbb{E}[b_{rkj}(X_k)]$ . Let  $\bar{\beta}_{rkj} = \frac{1}{2}(\beta_{rkj}^* - \beta_{rkj})$ , and we have  $\sum_{r=0}^{R_n} |\bar{\beta}_{rkj}|^2 = 1/4$ . Then

$$\begin{aligned} \frac{\mu_{Z,\beta}}{4} &= \mathbb{E} \left[ \sum_{k=1}^p \bar{f}_{kj}(X_k) \right]^2 \geq \phi^2 \sum_{k=1}^p \mathbb{E} [\bar{f}_{kj}(X_k)]^2 = \phi^2 \sum_{k=1}^p \mathbb{E} \left[ \bar{\beta}_{0kj} + \sum_{r=1}^{R_n} \bar{\beta}_{rkj} b_{rkj}(X_k) \right]^2 \\ &= \phi^2 \left( \bar{\beta}_{0kj}^2 + v \sum_{r=1}^{R_n} \bar{\beta}_{rkj}^2 \right) \geq \frac{\phi^2 v}{4} > 0. \end{aligned}$$

The first inequality follows from Assumption 5. The last equality follows from  $\mathbb{E}[b_{rkj}(X_k)b_{r'kj}(X_k)] = v\mathbb{1}(r = r')$  with  $v \in (0, 1]$ , as well as  $\bar{\beta}_{0kj} = -\sum_{r=1}^{R_n} \bar{\beta}_{rkj} \cdot \mathbb{E}[b_{rkj}(X_k)]$ .

□

**D.2. Lemmas for the main results.** For any  $\vec{g}_j$ , we introduce the empirical counterpart of the residual variance  $\nu_j^n(\vec{g}_j)$ , which is denote by  $\widehat{\nu}_j^n(\vec{g}_j) := \min_{f \in \mathcal{F}_n^{\oplus \vec{g}_j}} \int \mathcal{R}_j(f; \vec{g}_j) P_n(dx)$ .

**Lemma 19.** *Under event  $\mathcal{T}_1 \cap \mathcal{T}_2$ , we have*

$$|\nu_j^n(\vec{g}_j) - \widehat{\nu}_j^n(\vec{g}_j)| \lesssim s_j \sqrt{\log p/n}.$$

*Proof.* Note that we can rewrite

$$\nu_j^n(\vec{g}_j) = \min_{f \in \mathcal{F}_n^{\oplus \vec{g}_j}} \int \mathcal{R}_j(f; \vec{g}_j) P(dx), \quad \widehat{\nu}_j^n(\vec{g}_j) = \min_{f \in \mathcal{F}_n^{\oplus \vec{g}_j}} \int \mathcal{R}_j(f; \vec{g}_j) P_n(dx).$$

For any  $f \in \mathcal{F}^{\oplus \vec{g}_j}$ ,

$$\begin{aligned} \int \mathcal{R}_j(f; \vec{g}_j)(P - P_n)(dx) &= \int \left[ x_j - \sum_{k \in \vec{g}_j} f_{kj}(x_k) \right]^2 (P - P_n)(dx) \\ &\leq 2 \int \left[ \sum_{k \in \text{pa}(j)} f_{kj}^*(x_k) - \sum_{k \in \vec{g}_j} f_{kj}(x_k) \right]^2 (P - P_n)(dx) + 2 \int \epsilon_j^2(P - P_n)(dx) \\ &\leq 8 \cdot (|\vec{g}_j| \vee |\text{pa}(j)|)^2 \cdot \max_{k, k'} \sup_{f_{kj} \in \mathcal{F}_{kj}, f_{k'j} \in \mathcal{F}_{k'j}} \left| \int f_{kj}(x_k) f_{k'j}(x_{k'})(P - P_n)(dx) \right| + 2 \int \epsilon_j^2(P - P_n)(dx). \end{aligned}$$

Under the event  $\mathcal{T}_1 \cap \mathcal{T}_2$ , since  $(|\vec{g}_j| \vee |\text{pa}(j)|)^2 \leq K s_j \lesssim s_j$ , this implies that

$$\int \mathcal{R}_j(f; \vec{g}_j)(P - P_n)(dx) \lesssim s_j \sqrt{\log p/n}.$$

Now let

$$\tilde{f}^j \in \operatorname{argmin}_{f \in \mathcal{F}_n^{\oplus \vec{g}_j}} \int \mathcal{R}_j(f; \vec{g}_j) P(dx), \quad \bar{f}^j \in \operatorname{argmin}_{f \in \mathcal{F}_n^{\oplus \vec{g}_j}} \int \mathcal{R}_j(f; \vec{g}_j) P_n(dx),$$

and then

$$\nu_j^n(\vec{g}_j) = \int \mathcal{R}_j(\tilde{f}^j; \vec{g}_j) P(dx), \quad \hat{\nu}_j^n(\vec{g}_j) = \int \mathcal{R}_j(\bar{f}^j; \vec{g}_j) P_n(dx).$$

We conclude that

$$\int \mathcal{R}_j(\bar{f}^j; \vec{g}_j) P_n(dx) \leq \int \mathcal{R}_j(\tilde{f}^j; \vec{g}_j) P_n(dx) \lesssim \int \mathcal{R}_j(\tilde{f}^j; \vec{g}_j) P(dx) + s_j \sqrt{\log p/n},$$

and

$$\int \mathcal{R}_j(\tilde{f}^j; \vec{g}_j) P(dx) \leq \int \mathcal{R}_j(\bar{f}^j; \vec{g}_j) P(dx) \lesssim \int \mathcal{R}_j(\bar{f}^j; \vec{g}_j) P_n(dx) + s_j \sqrt{\log p/n}.$$

In summary,

$$|\nu_j^n(\vec{g}_j) - \hat{\nu}_j^n(\vec{g}_j)| \lesssim s_j \sqrt{\log p/n}.$$

□

**Lemma 20.** Suppose conditions of Lemma 10, 11 and 19 are satisfied. Then  $\underline{c} \leq \nu_j^n(\vec{g}_j) \leq \bar{c} + o(1)$  and  $\underline{c} - o(1) \leq \hat{\nu}_j^n(\vec{g}_j) \leq \bar{c} + o(1)$  for any  $\vec{g}_j$ .

*Proof.* First, by Lemma 10, we have  $\nu_j^n(\vec{g}_j) \geq \nu_j(\vec{g}_j) \geq \nu_j(\vec{g}(\pi)_j) \geq \underline{\nu}_p \geq \underline{c}$ , where  $\vec{g}_j \subseteq \vec{g}(\pi)_j$ . By Lemma 19, we further conclude that  $\hat{\nu}_j^n(\vec{g}_j) \geq \nu_j^n(\vec{g}_j) - o(1) \geq \underline{c} - o(1)$ .

Next, recall that  $d_{n,p} := \max_{j \in V} \max_{|\vec{g}_j| \in \{0, \dots, K\}} |\nu_j(\vec{g}_j) - \nu_j^n(\vec{g}_j)|$ . Then by Lemma 10 and 11, we have  $\nu_j^n(\vec{g}_j) \leq \nu_j(\vec{g}_j) + d_{n,p} \leq \bar{\nu}_p + d_{n,p} \leq \bar{c} + o(1)$ . By Lemma 19, we further conclude that  $\hat{\nu}_j^n(\vec{g}_j) \leq \nu_j^n(\vec{g}_j) + o(1) \leq \bar{c} + o(1)$ .

□

**Lemma 21.** Suppose the conditions of Lemma 19 and Lemma 20 are satisfied. For any  $\pi \notin \Pi^*$  and any  $\vec{g}_j \subseteq \vec{g}(\pi)_j$ , we have

$$p^{-1} \sum_{j=1}^p [\log \hat{\nu}_j^n(\vec{g}_j) - \log \sigma_j^{*\star}] \gtrsim \xi_p - s_n/p \sqrt{\log p/n}.$$

*Proof.* First, note that  $\nu_j^n(\vec{g}_j) \geq c$  by Lemma 20. Then

$$\begin{aligned} \sum_{j=1}^p [\log \hat{\nu}_j^n(\vec{g}_j) - \log \sigma_j^{*2}] &\gtrsim \sum_{j=1}^p [\log \nu_j^n(\vec{g}_j) - \log \sigma_j^{*2}] - s_n \sqrt{\log p/n} \\ &\geq \sum_{j=1}^p [\log \nu_j(\vec{g}(\pi)_j) - \log \sigma_j^{*2}] - s_n \sqrt{\log p/n} \geq p\xi_p - s_n \sqrt{\log p/n}. \end{aligned}$$

The first inequality follows from Lemma 19, and  $\log(a-x) \geq \log(a) - x/a$  for any  $a > 0$  and  $x < a$ . The second inequality follows from the fact that  $\nu_j^n(\vec{g}_j) \geq \nu_j(\vec{g}_j) \geq \nu_j(\vec{g}(\pi)_j)$  for any  $\vec{g}_j$ . The last inequality follows from the definition of  $\xi_p$ .  $\square$

## APPENDIX E. PROOF OF THEOREM 4

### E.1. Proof of Theorem 4 Part 1.

*Proof.* First, note that for any feasible  $f_{kj}$  and  $\vec{g}_j$ , the optimization

$$\min_{\sigma_j \geq 0} \sum_{j=1}^p \left[ \log \sigma_j^2 + \frac{\|X_j - \sum_{k \in \vec{g}_j} f_{kj}(X_k)\|_n^2}{\sigma_j^2} \right] + \lambda_n^2 \sum_{j=1}^p |\vec{g}_j|$$

is solved at  $\sigma_j^2 = \|X_j - \sum_{k \in \vec{g}_j} f_{kj}(X_k)\|_n^2$ . As a result, we can re-write the objective function as

$$\min_{\{f_{kj}\}, \{\vec{g}_j\}} \sum_{j=1}^p \log \left\| X_j - \sum_{k \in \vec{g}_j} f_{kj}(X_k) \right\|_n^2 + p + \lambda_n^2 \sum_{j=1}^p |\vec{g}_j|$$

with the constraint that each  $f_{kj} \in \mathcal{F}_{kj,n}$  and the selected edges form a DAG. Given the solution  $\hat{f}_{kj}$  and  $\hat{g}_j$ , the estimated variance is thus  $\hat{\sigma}_j^2 = \|X_j - \sum_{k \in \hat{g}_j} \hat{f}_{kj}(X_k)\|_n^2$ . Recall that  $D_{kj}^* = \sum_{r=R_n+1}^{\infty} \beta_{rkj}^* b_{rkj}$ . Appealing to the basic inequality,

$$\sum_{j=1}^p \log \hat{\sigma}_j^2 + p + \lambda_n^2 \hat{s} \leq \sum_{j=1}^p \log \sigma_j^{*2} + \sum_{j=1}^p \frac{\|\epsilon_j + \sum_{k \in \text{pa}(j)} D_{kj}^*\|_n^2}{\sigma_j^{*2}} + \lambda_n^2 s^*$$

or equivalently

$$(13) \quad \sum_{j=1}^p \log \frac{\hat{\sigma}_j^2}{\sigma_j^{*2}} \leq \sum_{j=1}^p \frac{\|\epsilon_j\|_n^2 - \sigma_j^{*2}}{\sigma_j^{*2}} + \sum_{j=1}^p \frac{\|\sum_{k \in \text{pa}(j)} D_{kj}^*\|_n^2}{\sigma_j^{*2}} + \frac{2}{n} \sum_{j=1}^p \frac{|\epsilon_j^\top \sum_{k \in \text{pa}(j)} D_{kj}^*|}{\sigma_j^{*2}} + \lambda_n^2 (s^* - \hat{s}).$$

Under event  $\mathcal{T}_1 \cap \mathcal{T}_3$ , as well as Lemma 8, we have

$$(14) \quad p^{-1} \sum_{j=1}^p [\log \hat{\sigma}_j^2 - \log \sigma_j^{*2}] \lesssim p^{-1} (\lambda_n^2 s_n + p \sqrt{\log p/n} + s_n R_n^{-2\eta}).$$

Under Assumption 6, the upper bound is  $\xi_p o(1)$ . This implies that  $\widehat{\Pi} \subseteq \Pi^*$ . Otherwise, there must exist  $\widehat{\pi} \in \widehat{\Pi}$  such that  $\widehat{\pi} \notin \Pi^*$ . Appealing to Lemma 21, where  $\widehat{\sigma}_j^2 = \widehat{\nu}_j^n(\widehat{g}_j)$ , we have

$$p^{-1} \sum_{j=1}^p [\log \widehat{\sigma}_j^2 - \log \sigma_j^{*2}] \gtrsim \xi_p - s_n/p\sqrt{\log p/n}.$$

This lower bound, however, is  $\xi_p(1 + o(1))$ , contradicting with the upper bound  $\xi_p o(1)$  when  $\xi_p > 0$ .

Finally, note that this proof relies on the event  $\mathcal{T}_1 \cap \mathcal{T}_2 \cap \mathcal{T}_3$ , which holds with probability at least  $1 - 4/p$ .  $\square$

## E.2. Proof of Theorem 4 Part 2.

*Proof.* The basic inequality for the early stopping solution  $\widehat{\theta}^{\text{early}}$  can be written as:

$$\ell_n(\widehat{\theta}^{\text{early}}) + \lambda_n^2 \widehat{s}^{\text{early}} \leq \ell_n(\widehat{\theta}) + \lambda_n^2 \widehat{s} + \tau^{\text{early}} \leq \ell_n(\{\beta_{rkj}^*\}_{(k,j) \in E^*, r \in [R_n]}, \{\sigma_j^*\}_{j \in V}) + \lambda^2 s^* + \tau^{\text{early}},$$

We use the same proof technique as in Part 1. When the early stopping threshold satisfies  $\tau^{\text{early}} = o(p\xi_p)$ , the upper bound of  $p^{-1} \sum_{j=1}^p [\log(\widehat{\sigma}_j^{\text{early}})^2 - \log \sigma_j^{*2}]$ , similar as in (14), remains to be  $\xi_p o(1)$ . Moreover, the estimated variance from the early-stopping solution satisfies  $(\widehat{\sigma}_j^{\text{early}})^2 \geq \widehat{\nu}_j^n(g_j^{\text{early}})$ , so when  $\widehat{\pi}^{\text{early}} \notin \Pi^*$ , the lower bound of  $p^{-1} \sum_{j=1}^p [\log(\widehat{\sigma}_j^{\text{early}})^2 - \log \sigma_j^{*2}]$  remains to be  $o(\xi_p)$ . With these two key facts in place, the early-stopping permutations are guaranteed to be correct.  $\square$

## APPENDIX F. PROOF OF THEOREM 5

### F.1. Proof of Theorem 5 Part 1.

*Proof.* We first note that, by Theorem 4, the estimated permutations are correct:  $\widehat{\Pi} \subseteq \Pi^*$ . Applying the inequality  $\log(1 + x) \leq x - x^2/(2(1 + c)^2)$ ,  $-1 < x \leq c$ , we obtain

$$(15) \quad \log \left( \frac{\widehat{\sigma}_j^2}{\sigma_j^{*2}} \right) = -\log \left( \frac{\sigma_j^{*2}}{\widehat{\sigma}_j^2} \right) \geq \frac{\widehat{\sigma}_j^2 - \sigma_j^{*2}}{\widehat{\sigma}_j^2} + \frac{(\min_{j \in V} \widehat{\sigma}_j^2)^2}{2\bar{\sigma}^4} \left( \frac{\sigma_j^{*2} - \widehat{\sigma}_j^2}{\widehat{\sigma}_j^2} \right)^2.$$

Moreover, by (13), we have

$$(16) \quad K_0 \sum_{j=1}^p \left( \frac{\sigma_j^{*2} - \widehat{\sigma}_j^2}{\widehat{\sigma}_j^2} \right)^2 \leq \sum_{j=1}^p \left( \frac{\|\epsilon_j\|_n^2}{\sigma_j^{*2}} - 1 \right) - \sum_{j=1}^p \frac{\widehat{\sigma}_j^2 - \sigma_j^{*2}}{\widehat{\sigma}_j^2} + \sum_{j=1}^p (\mathcal{D}_{1j} + \mathcal{D}_{2j}) + \lambda_n^2 (s^* - \widehat{s}),$$

where

$$\mathcal{D}_{1j} = \frac{\|\sum_{k \in \text{pa}(j)} D_{kj}^* \|_n^2}{\sigma_j^{*2}}, \quad \mathcal{D}_{2j} = \frac{2 \sum_{k \in \text{pa}(j)} |\epsilon_j^\top D_{kj}^*|}{n \sigma_j^{*2}},$$

and  $K_0 > 0$  is a constant such that  $0 < K_0 \leq (\min_{j \in V} \widehat{\sigma}_j^2)^2 / (2\bar{\sigma}^4)$ . Such a constant  $K_0$  exists since by Lemma 20, where  $\widehat{\sigma}_j^2 = \widehat{\nu}_j^n(\widehat{g}_j)$ , we have  $\widehat{\sigma}_j^2 \geq c - o(1)$  for all  $j \in V$ . In the following, we use the parameterization with  $f_{kj}(\cdot) = \beta_{0kj} + \sum_{r=1}^{R_n} \beta_{rkj} b_{rkj}(\cdot)$ . Recall the definition

$$Z_j = [1, b_{11j}(X_1), \dots, b_{1pj}(X_p), \dots, b_{R_n 1j}(X_1), \dots, b_{R_n pj}(X_p)] \in \mathbb{R}^{pR_n+1}, \quad (\text{as a random vector})$$

and

$$\beta_j = [\beta_{0*j}, \beta_{11j}, \dots, \beta_{1pj}, \dots, \beta_{R_n1j}, \dots, \beta_{R_npj}]^\top \in \mathbb{R}^{pR_n+1}, \quad \beta_{0\star j} = \sum_{k=1}^p \beta_{0kj}.$$

Then we have  $\sum_{k=1}^p f_{kj}(X_k) = \beta_j^\top Z_j$ . For each  $j \in V$ ,

$$\begin{aligned} \widehat{\sigma}_j^2 - \sigma_j^{*2} &= \left\| X_j - Z_j \widehat{\beta}_j \right\|_n^2 - \sigma_j^{*2} = \left\| Z_j(\beta_j^* - \widehat{\beta}_j) + \epsilon_j + \sum_{k \in \text{pa}(j)} D_{kj}^* \right\|_n^2 - \sigma_j^{*2} \\ &= \left\| Z_j(\beta_j^* - \widehat{\beta}_j) + \sum_{k \in \text{pa}(j)} D_{kj}^* \right\|_n^2 + (\|\epsilon_j\|_n^2 - \sigma_j^{*2}) + \frac{2}{n} \epsilon_j^\top (Z_j(\beta_j^* - \widehat{\beta}_j)) + 2 \frac{\epsilon_j^\top}{n} \sum_{k \in \text{pa}(j)} D_{kj}^* \end{aligned}$$

Under the event  $\mathcal{T}_4$ , we obtain that for some  $\delta_3 > 0$  with  $0 < \delta_2 + \delta_3 < 1$ ,

$$\begin{aligned} (17) \quad & (\|\epsilon_j\|_n^2 - \sigma_j^{*2}) - (\widehat{\sigma}_j^2 - \sigma_j^{*2}) \\ & \leq - \left\| Z_j(\beta_j^* - \widehat{\beta}_j) + \sum_{k \in \text{pa}(j)} D_{kj}^* \right\|_n^2 + \delta_2 \left\| Z_j(\beta_j^* - \widehat{\beta}_j) \right\|_n^2 + \Delta_{n,p} + \sigma_j^{*2} \mathcal{D}_{2j} \\ & \leq -(1 - \delta_2 - \delta_3) \left\| Z_j(\beta_j^* - \widehat{\beta}_j) \right\|_n^2 - (1 - 1/\delta_3) \left\| \sum_{k \in \text{pa}(j)} D_{kj}^* \right\|_n^2 + \Delta_{n,p} + \sigma_j^{*2} \mathcal{D}_{2j} \\ & \leq (1/\delta_3 - 1) \sigma_j^{*2} \mathcal{D}_{1j} + \sigma_j^{*2} \mathcal{D}_{2j} + \Delta_{n,p}. \end{aligned}$$

Furthermore, for some  $0 < \delta_4 < K_0$ ,

$$(18) \quad \frac{\|\epsilon_j\|_n^2 - \sigma_j^{*2}}{\sigma_j^{*2}} - \frac{\|\epsilon_j\|_n^2 - \sigma_j^{*2}}{\widehat{\sigma}_j^2} = \frac{\|\epsilon_j\|_n^2 - \sigma_j^{*2}}{\sigma_j^{*2}} \cdot \frac{\widehat{\sigma}_j^2 - \sigma_j^{*2}}{\widehat{\sigma}_j^2} \leq \frac{1}{\delta_4} \left( \frac{\|\epsilon_j\|_n^2 - \sigma_j^{*2}}{\sigma_j^{*2}} \right)^2 + \delta_4 \left( \frac{\widehat{\sigma}_j^2 - \sigma_j^{*2}}{\widehat{\sigma}_j^2} \right)^2.$$

Now we plug (17) and (18) back into (16). Under event  $\mathcal{T}_1 \cap \mathcal{T}_2 \cap \mathcal{T}_3 \cap \mathcal{T}_4$  and Lemma 8, we obtain

$$\begin{aligned} & (K_0 - \delta_4) \sum_{j=1}^p \left( \frac{\sigma_j^{*2} - \widehat{\sigma}_j^2}{\widehat{\sigma}_j^2} \right)^2 \\ & \leq \sum_{j=1}^p \left[ \frac{\|\epsilon_j\|_n^2 - \sigma_j^{*2}}{\sigma_j^{*2}} - \frac{\|\epsilon_j\|_n^2 - \sigma_j^{*2}}{\widehat{\sigma}_j^2} \right] + \sum_{j=1}^p \left[ \frac{\|\epsilon_j\|_n^2 - \sigma_j^{*2}}{\widehat{\sigma}_j^2} - \frac{\widehat{\sigma}_j^2 - \sigma_j^{*2}}{\widehat{\sigma}_j^2} \right] - \delta_4 \sum_{j=1}^p \left( \frac{\sigma_j^{*2} - \widehat{\sigma}_j^2}{\widehat{\sigma}_j^2} \right)^2 \\ & \quad + \sum_{j=1}^p (\mathcal{D}_{1j} + \mathcal{D}_{2j}) + \lambda_n^2 (s^* - \widehat{s}) \\ & \lesssim (p + s_n R_n) \log p/n + s_n R_n^{-2\eta} + \lambda_n^2 s_n. \end{aligned}$$

The result hence follows. Note that this proof relies on the event  $\mathcal{T}_1 \cap \mathcal{T}_2 \cap \mathcal{T}_3 \cap \mathcal{T}_4$ , which holds with probability at least  $1 - 5/p$ .  $\square$

### F.2. Proof of Theorem 5 Part 2.

*Proof.* Note that the estimated permutations are correct:  $\widehat{\Pi}^{\text{early}} \subseteq \Pi^*$  by Theorem 4. Then using the same proof technique as in Part 1, this is a direct result from the basic inequality

$$\ell_n(\widehat{\theta}^{\text{early}}) + \lambda_n^2 \widehat{s}^{\text{early}} \leq \ell_n(\widehat{\theta}) + \lambda_n^2 \widehat{s} + \tau^{\text{early}} \leq \ell_n(\{\beta_{rkj}^*\}_{(k,j) \in E^*, r \in [R_n]}, \{\sigma_j^*\}) + \lambda^2 s^* + \tau^{\text{early}}.$$

□

## APPENDIX G. PROOF OF THEOREMS 6

### G.1. Proof of Theorem 6 Part 1.

*Proof.* In this proof, we keep using the parameterization with  $Z_j \in \mathbb{R}^{pR_n+1}$  and  $\beta_j \in \mathbb{R}^{pR_n+1}$ . Let  $\mathcal{G}(\beta)$  be the graph induced by  $\beta$ , where an edge between node  $k$  and node  $j$  is present if  $\sum_{r=1}^{R_n} |\beta_{rkj}|^2 \neq 0$ , and absent otherwise. We aim to show that the solution to the following optimization problem, which is equivalent to the original one, recovers the true graph  $\mathcal{G}^*$ :

$$(19) \quad \min_{\beta \in \mathfrak{B}} l(\beta), \quad l(\beta) := \sum_{j=1}^p \log \frac{\|X_j - Z_j \beta_j\|_n^2}{\sigma_j^{*2}} - \sum_{j=1}^p \frac{\|\epsilon_j\|_n^2 - \sigma_j^{*2}}{\sigma_j^{*2}} + \lambda_n^2 (\mathbf{1}\{\|\beta_{kj}\|_2 \neq 0\} - s^*),$$

where  $\mathfrak{B}$  corresponds to the feasible set in (8). The proof proceeds in three steps. In each step, we solve an optimization problem that is more constrained than the original one.

- In step I, we restrict the feasible region to  $\mathcal{R}_1 := \{\beta \in \mathfrak{B} : \mathcal{G}(\beta) = \mathcal{G}^*\}$ , and derive an upper bound for the minimized objective function over  $\beta \in \mathcal{R}_1$ . We denote this upper bound by  $l_1$ .
- In step II, we restrict the feasible region to

$$\mathcal{R}_2 := \{\beta \in \mathfrak{B} : |\mathcal{G}(\beta)| \leq s^*, \mathcal{G}(\beta) \neq \mathcal{G}^*, \mathcal{G}(\beta) \text{ has correct permutations}\},$$

and derive a lower bound for the minimized objective function over  $\beta \in \mathcal{R}_2$ . We denote this lower bound by  $l_2$ , and verify that  $l_1 < l_2$  when  $n$  is sufficiently large.

- In step III, we restrict the feasible region to

$$\mathcal{R}_3 := \{\beta \in \mathfrak{B} : |\mathcal{G}(\beta)| > s^*, \mathcal{G}(\beta) \text{ has correct permutations}\},$$

and derive a lower bound for the minimized objective function over  $\beta \in \mathcal{R}_3$ . We denote this lower bound by  $l_3$ , and verify that  $l_1 < l_3$  when  $n$  is sufficiently large.

The three steps above guarantee that the solution to the original problem falls inside  $\mathcal{R}_1$  when  $n$  is sufficiently large. Note that we restrict the search only to the space of correct permutations, since, by Theorem 4, any feasible value associated with an incorrect permutation cannot be optimal. This proof relies on the event  $\mathcal{T}_1 \cap \mathcal{T}_2 \cap \mathcal{T}_3 \cap \mathcal{T}_4 \cap \mathcal{T}_5$ , which holds with probability at least  $1 - 6/p$ .

**Step I.** In this step, we use  $\bar{\beta}$  to denote the solution to optimization problem (19) over the additional constraint set  $\mathcal{R}_1$ . The estimated variance is hence  $\bar{\sigma}_j^2 := \|X_j - Z_j \bar{\beta}_j\|_n^2$ . Appealing to the basic inequality similar as (13), we can arrive at an upper bound for the objective function:

$$(20) \quad \begin{aligned} l(\bar{\beta}) &= \sum_{j=1}^p \log \frac{\bar{\sigma}_j^2}{\sigma_j^{*2}} - \sum_{j=1}^p \frac{\|\epsilon_j\|_n^2 - \sigma_j^{*2}}{\sigma_j^{*2}} \leq \sum_{j=1}^p \frac{\|\sum_{k \in \text{pa}(j)} D_{kj}^*\|_n^2}{\sigma_j^{*2}} + \frac{2}{n} \sum_{j=1}^p \frac{|\epsilon_j^\top \sum_{k \in \text{pa}(j)} D_{kj}^*|}{\sigma_j^{*2}} \\ &\lesssim s_n R_n^{-2\eta} + s_n \log p/n. \end{aligned}$$

We call the upper bound  $l_1 := s_n R_n^{-2\eta} + s_n \log p/n$ . By Assumption 8, we have  $l_1 = o(\delta_{n,p})$ .

**Step II.** In this step, we overload the notation  $\bar{\beta}$  to represent the solution to optimization problem (19) over the additional constraint set  $\mathcal{R}_2$ . The estimated variance is  $\bar{\sigma}_j^2 := \|X_j - Z_j \bar{\beta}_j\|_n^2$ , and the estimated sparsity is  $\bar{s} := |\mathcal{G}(\bar{\beta})|$ . Then by (15), (17) and (18),

$$\begin{aligned} l(\bar{\beta}) &= \sum_{j=1}^p \log \frac{\bar{\sigma}_j^2}{\sigma_j^{*2}} - \sum_{j=1}^p \frac{\|\epsilon_j\|_n^2 - \sigma_j^{*2}}{\sigma_j^{*2}} + \lambda_n^2(\bar{s} - s^*) \\ &\geq \sum_{j=1}^p \left[ \frac{\bar{\sigma}_j^2 - \sigma_j^{*2}}{\bar{\sigma}_j^2} - \frac{\|\epsilon_j\|_n^2 - \sigma_j^{*2}}{\bar{\sigma}_j^2} \right] + \sum_{j=1}^p \left[ \frac{\|\epsilon_j\|_n^2 - \sigma_j^{*2}}{\bar{\sigma}_j^2} - \frac{\|\epsilon_j\|_n^2 - \sigma_j^{*2}}{\sigma_j^{*2}} \right] + K_0 \sum_{j=1}^p \left( \frac{\sigma_j^{*2} - \bar{\sigma}_j^2}{\bar{\sigma}_j^2} \right)^2 + \lambda_n^2(\bar{s} - s^*) \\ &\gtrsim \sum_{j=1}^p \left\| Z_j(\beta_j^* - \bar{\beta}_j) \right\|_n^2 - (p + s_n R_n) \log p/n - s_n R_n^{-2\eta} + \lambda_n^2(\bar{s} - s^*). \end{aligned}$$

Now we attempt to connect this lower bound with the number of missed edges  $\gamma_1 := |\mathcal{G}^* \setminus \mathcal{G}(\bar{\beta})|$ . We also define  $\gamma_2 := |\mathcal{G}(\bar{\beta}) \setminus \mathcal{G}^*|$ . For any  $\bar{\beta} \in \mathcal{R}_2$ , we consider two cases.

Case 1: If  $|\mathcal{G}(\bar{\beta})| = s^*$ , then  $\gamma_1 = \gamma_2 \geq 1$ . Under the event  $\mathcal{T}_5$ , we have

$$\sum_{j=1}^p \left\| Z_j(\beta_j^* - \bar{\beta}_j) \right\|_n^2 \gtrsim \sum_{j=1}^p \|\beta_j^* - \bar{\beta}_j\|_2^2 \geq \gamma_1 \min_{(k,j) \in E^*} \sum_{r=1}^{R_n} |\beta_{rkj}^*|^2 \geq \min_{(k,j) \in E^*} \sum_{r=1}^{R_n} |\beta_{rkj}^*|^2.$$

Then

$$l(\bar{\beta}) \gtrsim \min_{(k,j) \in E^*} \sum_{r=1}^{R_n} |\beta_{rkj}^*|^2 - (p + s_n R_n) \log p/n - s_n R_n^{-2\eta}.$$

We call this lower bound  $l_2^{(1)}$ . By Assumption 8, we have  $l_2^{(1)} \gtrsim \delta_{n,p} > o(\delta_{n,p}) = l_1$  when  $n$  is sufficiently large.

Case 2: If  $|\mathcal{G}(\bar{\beta})| < s^*$ , then  $\gamma_2 \geq 0$  and  $\gamma_1 = s^* - \bar{s} + \gamma_2$ . Under the event  $\mathcal{T}_5$ , we have

$$\sum_{j=1}^p \left\| Z_j(\beta_j^* - \bar{\beta}_j) \right\|_n^2 \gtrsim \sum_{j=1}^p \|\beta_j^* - \bar{\beta}_j\|_2^2 \geq \gamma_1 \min_{(k,j) \in E^*} \sum_{r=1}^{R_n} |\beta_{rkj}^*|^2 \geq (s^* - \bar{s}) \min_{(k,j) \in E^*} \sum_{r=1}^{R_n} |\beta_{rkj}^*|^2.$$

Then

$$l(\bar{\beta}) \gtrsim (s^* - \bar{s}) \min_{(k,j) \in E^*} \sum_{r=1}^{R_n} |\beta_{rkj}^*|^2 + \lambda_n^2(\bar{s} - s^*) - (p + s_n R_n) \log p/n - s_n R_n^{-2\eta}.$$

We call this lower bound  $l_2^{(2)}$ . By Assumption 8 and the choice of  $\lambda_n^2$ , we see when  $n$  is sufficiently large,

$$l_2^{(2)} \geq (s^* - \bar{s}) \left[ \min_{(k,j) \in E^*} \sum_{r=1}^{R_n} |\beta_{rkj}^*|^2 - \lambda_n^2 - (p + s_n R_n) \log p/n - s_n R_n^{-2\eta} \right] \gtrsim \delta_{n,p} > o(\delta_{n,p}) = l_1.$$

Both cases lead to the result that  $l_2 = \min\{l_2^{(1)}, l_2^{(2)}\} > l_1$  when  $n$  is sufficiently large.

**Step III.** In this step, we overload the notation  $\bar{\beta}$  to represent the solution to optimization problem (19) over the additional constraint set  $\mathcal{R}_3$ . The estimated sparsity is also overloaded as  $\bar{s} := |\mathcal{G}(\bar{\beta})|$ .

Similarly as in Step II, we obtain a lower bound for  $l(\bar{\beta})$ :

$$\begin{aligned} l(\bar{\beta}) &\gtrsim \sum_{j=1}^p \|Z_j(\beta_j^* - \bar{\beta}_j)\|_n^2 + \lambda_n^2(\bar{s} - s^*) - (p + s_n R_n) \log p/n - s_n R_n^{-2\eta} \\ &\geq \lambda_n^2 - (p + s_n R_n) \log p/n - s_n R_n^{-2\eta} \end{aligned}$$

We call this lower bound  $l_3$ . By the choice of  $\lambda_n^2 \gtrsim \delta_{n,p}$ , we verify that  $l_3 \gtrsim \delta_{n,p} > o(\delta_{n,p}) = l_1$  when  $n$  is sufficiently large.  $\square$

## G.2. Proof of Theorem 6 Part 2.

*Proof.* Note that the estimated permutations are correct:  $\widehat{\Pi}^{\text{early}} \subseteq \Pi^*$  by Theorem 4. Then the basic inequality for the early stopping solution  $\widehat{\theta}^{\text{early}}$  can be written as:

$$l(\widehat{\beta}^{\text{early}}) \leq l(\widehat{\beta}) + \tau^{\text{early}} \leq l(\{\beta_1^*, \dots, \beta_p^*\}) + \tau^{\text{early}},$$

We use the same proof technique as in Part 1. When the early stopping threshold satisfies  $\tau^{\text{early}} = o(p\xi_p \wedge \delta_{n,p})$ , the upper bound  $l_1$  remains to be  $o(\delta_{n,p})$ . Moreover, the lower bounds  $l_2$  and  $l_3$  remain at least on the order of  $\delta_{n,p}$ . The result hence follows.  $\square$

## APPENDIX H. MORE ON EXPERIMENTS

**H.1. Simulation Setup for Figure 1.** The true model is given in (1), and each reported result is averaged over 100 independent trials. The *CAM-IncEdge* procedure is implemented using the R package *CAM* with default settings, except that we set `pruning = TRUE`. The *NPVAR* procedure is implemented using the publicly available package at <https://github.com/MingGao97/NPVAR> with default parameters. For *MIP (linear)*, we use whichever of procedures (8) or (9) yields the smaller BIC score, with basis functions  $\{b_r(X_k)\}_{k=1}^{R_n}$  taken simply as  $X_k$  itself. Our proposed nonlinear MIP approach is implemented in the same manner: selecting between (8) and (9) based on the BIC score; however, in order to handle non-linearity, we use degree-three splines with five internal knots for all edges. We set  $\lambda_n = 0.01$  for both procedures. No super-structure, partial order sets, stable sets, or early stopping is used in any of the MIP-based methods.

**H.2. Simulation setup for Figure 2.** We use the graph structure *Insurance Small* with  $p = 15$  nodes and  $s^* = 25$  true edges in this experiment. For all  $(k, j) \in E^*$ , the true function is set as  $f_{kj}^*(x) = (\sin(x) - \mathbb{E}[\sin(X_k)] + \cos(x) - \mathbb{E}[\cos(X_k)])/2$ . The noise variances  $\sigma_j^*$  alternate between 0.5 and 1 across nodes. For each trial, we generate  $n = 500$  independent and identically distributed samples. We apply procedure (8) in this simulation. To construct basis functions, the same set of degree-two splines with two internal knots is used across all edges. We adopt the true moral graph as the super-structure. In this experiment, we do not use any partial order sets, stable sets, or early stopping. Each run of the MIP is terminated after 15 minutes.

**H.3. Simulation setup for Figure 3.** The “default” formulation in Figure 3 refers to the following MIP:

(21a)

$$\min_{\begin{array}{l} \nu_j \in \mathbb{R}, \vartheta_{0*j} \in \mathbb{R}, \forall j \in V \\ \vartheta_{rkj} \in \mathbb{R}, \forall r \in [R_n], \forall (k,j) \in \mathcal{E} \\ g_{kj} \in \{0,1\}, \forall (k,j) \in \mathcal{E} \\ \psi \in [1,p]^p \end{array}} \sum_{j=1}^p -2 \log \nu_j + \sum_{j=1}^p \left\| \nu_j X_j - \vartheta_{0*j} - \sum_{k=1}^p \sum_{r=1}^{R_n} \vartheta_{rkj} b_r(X_k) \right\|_n^2 + \lambda_n^2 \cdot \sum_{(k,j) \in \mathcal{E}} g_{kj}$$

$$(21b) \quad \text{s.t.} \quad -Mg_{kj} \leq \vartheta_{rkj} \leq Mg_{kj}, \quad \forall (k,j) \in \mathcal{E}, r \in \{1, \dots, R_n\},$$

$$(21c) \quad -M \sum_{k=1}^p g_{kj} \leq \vartheta_{0*j} \leq M \sum_{k=1}^p g_{kj}, \quad \forall j \in V,$$

$$(21d) \quad 0 < \nu_j \leq M, \quad \forall j \in V,$$

$$(21e) \quad 1 - p + pg_{kj} \leq \psi_j - \psi_k, \quad \forall (k,j) \in \mathcal{E}.$$

Suppose  $\sigma_j$  denotes the standard deviation of the noise term  $\epsilon_j$ . Then, the decision variable  $\nu_j$  corresponds to  $1/\sigma_j$ ,  $\vartheta_{rkj}$  represents the scaled coefficient  $\beta_{rkj}/\sigma_j$ , and  $\vartheta_{0*j}$  represents the scaled intercepts  $\sum_{k=1}^p \beta_{0kj}/\sigma_j$ .

Regardless of the specific formulation, the implementation of the MIP procedure using **Gurobi** can be broadly decomposed into three computationally intensive components. First, the objective function must be specified by operating on the decision variables. When this step requires iterating over the entire dataset, it can become a computational bottleneck. Second, an unconstrained version of the optimization problem is solved to obtain a suitable value for the big- $M$  constant. Third, the core optimization process is executed to solve the full MIP problem. The computational cost at this stage is influenced by both the complexity of evaluating the objective function and the structural properties of the optimization problem, such as convexity.

We now present the simulation setup, which is very similar to that of Figure 2, but we provide a full description here for completeness. The true model is specified using the graph structure *Insurance Small* with  $p = 15$  nodes and  $s^* = 25$  true edges. For all  $(k,j) \in E^*$ , the true function is set as  $f_{kj}^*(x) = (\sin(x) - \mathbb{E}[\sin(X_k)] + \cos(x) - \mathbb{E}[\cos(X_k)])/2$ . The noise variances  $\sigma_j^*$  alternate between 0.5 and 1 across nodes. We apply procedure (8) in this simulation. For graph estimation, the same set of degree-two splines with two internal knots is used across all edges. We adopt the true moral graph as the super-structure. The tuning parameter is set as  $\lambda_n^2 = 0.5$ . In this experiment, we do not use any partial order sets, stable sets, or early stopping. Each run of the MIP is terminated after 50 minutes.

**H.4. Simulation setup for Figure 4.** The simulation setup is very similar to that of Figure 3, but we provide a full description here for completeness. We use the graph structure *Insurance Small* with  $p = 15$  nodes and  $s^* = 25$  true edges in this experiment. For all  $(k,j) \in E^*$ , the true function is set as  $f_{kj}^*(x) = (\sin(x) - \mathbb{E}[\sin(X_k)] + \cos(x) - \mathbb{E}[\cos(X_k)])/2$ . The noise variances  $\sigma_j^*$  alternate between 0.5 and 1 across nodes. For each trial, we generate  $n = 500$  independent and identically distributed samples. We apply procedure (8) in this simulation. To construct basis functions, the same set of degree-two splines with two internal knots is used across all edges. We adopt the true moral graph as the super-structure. The partial order and stable sets are estimated via the bootstrap procedure described in Section 4.2, where the *CAM* algorithm (Bühlmann et al., 2014) is applied to each bootstrap sample. We set  $\lambda_n^2 = 0.01$  and fix the early stopping threshold at  $\tau^{\text{early}} = 0$ . Each run of *MIP* is terminated after 15 minutes.

**H.5. Simulation setup for Figure 5.** The simulation setup is very similar to that of Figure 4, but we provide a full description here for completeness. We use the graph structure *Insurance Small* with  $p = 15$  nodes and  $s^* = 25$  true edges in this experiment. For all  $(k, j) \in E^*$ , the true function is set as  $f_{kj}^*(x) = (\sin(x) - \mathbb{E}[\sin(X_k)] + \cos(x) - \mathbb{E}[\cos(X_k)])/2$ . The noise variances  $\sigma_j^*$  alternate between 0.5 and 1 across nodes. For each trial, we generate  $n = 500$  independent and identically distributed samples. We apply procedure (8) in this simulation. To construct basis functions, the same set of degree-two splines with two internal knots is used across all edges. We adopt the true moral graph as the super-structure. In this experiment, we do not use any partial order sets or stable sets. We set  $\lambda_n^2 = 0.01$ . Each run of the MIP is terminated after 15 minutes.

## H.6. More on comparison to existing methods.

**H.6.1. More on Section 6.2.** We describe the implementation details of six baseline approaches here:

- *NPVAR*: The implementation is available at <https://github.com/MingGao97/NPVAR>. We first used the function `NPVAR()` with `layer.select = TRUE` to find the topological ordering of nodes. The `eta` parameter was chosen such that the each resulting layer contains only one node. We then used the function `prune()` to obtain the estimated adjacency matrix, where `cutoff` was chosen based on the Bayesian information criterion score.
- *EqVar*: The implementation is available at <https://github.com/WY-Chen/EqVarDAG>. For the top-down and bottom-up versions, we used the functions `EqVarDAG_TD()` and `EqVarDAG_BU()`, respectively, with the parameter set to `mtd = "cvlasso"`. The `cutoff` parameter was chosen based on the Bayesian information criterion score.
- *NoTears*: The implementation is available at <https://github.com/xunzheng/notears>. We used neural networks for the non-linearities with a single hidden layer of 5 neurons. We used the function `notears_non-linear()` with the training parameters `lambda1` and `lambda2` chosen based on the Bayesian information criterion score.
- *RESIT*: The implementation is available at <http://people.tuebingen.mpg.de/jpeters/onlineCodeANM.zip>. We used the function `ICML()` with parameter `model = train_gam`, and its parameter `alpha` was chosen based on the Bayesian information criterion score.
- *CCDr*: The implementation is available at the Python Causal Discovery Toolbox (`cdt`). We used the function `CCDr()` to fit the model.
- *CAM*: The implementation is available at the R library `CAM` and the Python library `cdt`. We used the function `CAM()` in Python interface, with parameters `score = "non-linear"`, `variablesel = True`, `pruning = True`. The parameter `cutoff` was chosen based on the Bayesian information criterion score.

Tables 1 – 4 show complete results for all ten graph structures. We observe that *NPVAR* and *EqVar*, which rely on the homoscedasticity assumption, perform particularly well when the noise variances are equal, but their performance deteriorates significantly under heteroscedastic settings. A similar pattern is observed for *NoTears*. In contrast, the method *CAM* tends to perform substantially better under heteroscedastic conditions, while its performance worsens when the variances are equal. Overall, among the baseline methods, *CAM* yields the best performance in heteroscedastic settings, whereas *NPVAR* performs best in the homoscedastic case. These observations support our choice of base procedures applied to the bootstrap samples in each scenario. Finally, we note that the total running time for *MIP* should account not only for the values reported in each table, but also for the time spent on the bootstrap procedure. If the bootstrap samples are processed in parallel, this additional

cost is simply equal to the running time of a single execution of *CAM* (or *NPVAR*). In summary, the *MIP* algorithm achieves the most accurate graph recovery within the allotted optimization time budget of  $60p$  seconds.

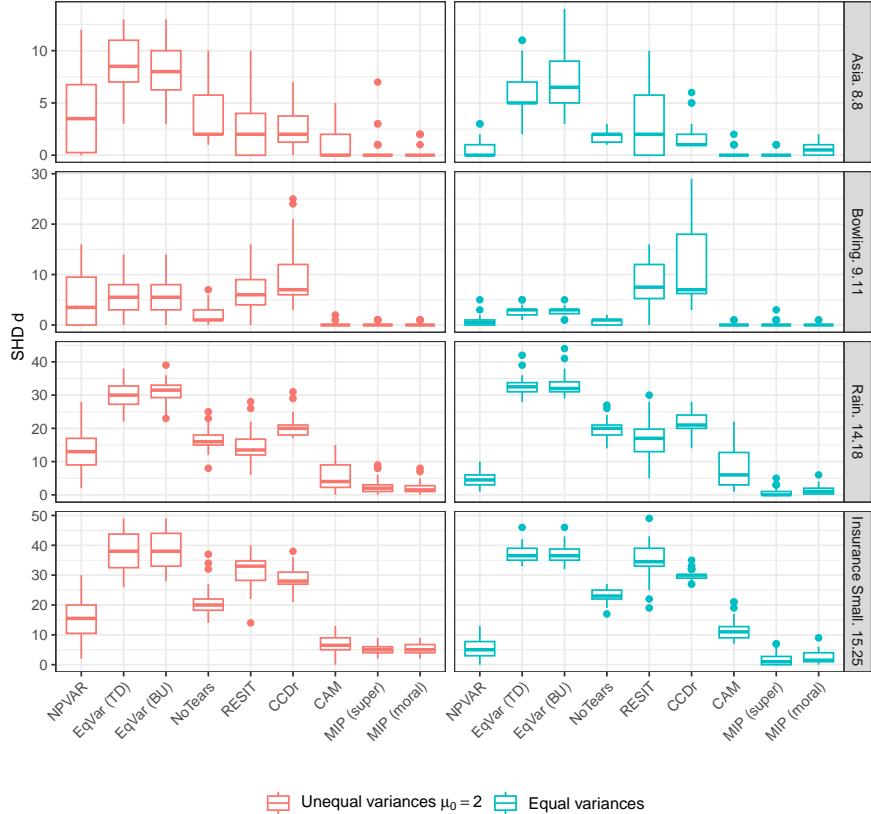


FIGURE 9. The performance comparison between *MIP* and baseline methods. The true functions are  $f_{kj}^*(x) = \sin(x + \sin(x)) - \mathbb{E}[\sin(X_k + \sin(X_k))]$ , and radial basis functions are used as basis functions. Each row corresponds to one graph structure, indicated by name. $p.s^*$ . The EqVar (TD) and EqVar (BU) refer to the top-down and bottom-up version of *EqVar*, respectively. The MIP (super) and MIP (moral) refer to the *MIP* using the estimated super-structure and the true moral graph, respectively. Different colors correspond to different variance schemes. Each box summarizes results from 30 independent trials.

**H.6.2. Another choice of true and basis functions.** We use the same setup as in Section 6.2, except for the choice of the true functions and basis functions. Specifically, for all  $(k, j) \in E^*$ , we set the true function as  $f_{kj}^*(x) = \sin(x + \sin(x)) - \mathbb{E}[\sin(X_k + \sin(X_k))]$ . Across all edges, we use the same set of radial basis functions defined as  $b_r(x; k_r) = \exp(-(x - k_r)^2)$ , where  $k_r$  is the  $r$ -th knot,  $r = 1, \dots, R_n$ . These  $R_n$  knots are placed at evenly spaced percentiles. We note that for one-dimensional features, radial basis functions typically have weaker approximation power compared to spline basis functions when using a small number of basis terms. To achieve sufficient approximation accuracy in this experiment, we set  $R_n = 30$ , which, however, imposes a heavier computational burden. Figure 9

compares the graph recovery performance of *MIP* with baseline methods. The comparison focuses on graphs with up to  $p = 15$  nodes, under both the homoscedastic and the heteroscedastic  $\mu_0 = 2$  settings. Our proposed algorithm *MIP* achieves the best overall performance across all setups. Complete results are presented in Tables 5 and 6.

|                      | NPVAR       |              | EqVar (TD)   |               | EqVar (BU)  |               |
|----------------------|-------------|--------------|--------------|---------------|-------------|---------------|
|                      | $d$         | Time         | $d$          | Time          | $d$         | Time          |
| Network. $p.s^*$     |             |              |              |               |             |               |
| Dsep.6.6             | 5.9 (2.5)   | 0.3 (0.0)    | 5.3 (2.6)    | 0.2 (0.0)     | 5.4 (2.7)   | 0.2 (0.0)     |
| Asia.8.8             | 9.6 (3.6)   | 0.6 (0.0)    | 7.9 (3.0)    | 0.3 (0.0)     | 8.0 (3.0)   | 0.3 (0.0)     |
| Bowling.9.11         | 10.7 (4.1)  | 0.8 (0.1)    | 8.0 (2.7)    | 0.4 (0.0)     | 7.9 (2.8)   | 0.4 (0.0)     |
| InsuranceSmall.15.25 | 27.1 (7.6)  | 4.1 (0.2)    | 27.5 (6.7)   | 1.0 (0.1)     | 27.7 (6.9)  | 0.9 (0.1)     |
| Rain.14.18           | 21.5 (7.1)  | 3.5 (0.2)    | 18.4 (6.6)   | 0.9 (0.1)     | 18.4 (6.4)  | 0.8 (0.1)     |
| Cloud.16.19          | 24.2 (6.0)  | 5.1 (0.2)    | 19.9 (5.0)   | 1.0 (0.1)     | 20.2 (5.0)  | 1.0 (0.1)     |
| Funnel.18.18         | 24.9 (9.9)  | 5.8 (0.2)    | 17.1 (5.9)   | 0.9 (0.1)     | 17.5 (6.0)  | 0.9 (0.0)     |
| Galaxy.20.22         | 29.1 (10.0) | 7.8 (0.2)    | 20.6 (5.6)   | 1.0 (0.0)     | 21.2 (5.2)  | 1.0 (0.1)     |
| Insurance.27.52      | 67.9 (15.1) | 19.6 (0.5)   | 73.2 (14.0)  | 1.4 (0.1)     | 72.7 (14.2) | 1.4 (0.1)     |
| Factors.27.68        | 54.3 (12.2) | 19.6 (0.5)   | 82.8 (21.2)  | 1.4 (0.0)     | 83.3 (20.6) | 1.4 (0.1)     |
|                      | NoTears     |              | RESIT        |               | CCDr        |               |
| Network. $p.s^*$     | $d$         | Time         | $d$          | Time          | $d$         | Time          |
| Dsep.6.6             | 4.9 (3.1)   | 3.5 (1.3)    | 3.1 (2.5)    | 3.4 (0.2)     | 3.4 (1.4)   | 3.5 (0.6)     |
| Asia.8.8             | 7.5 (3.2)   | 5.1 (3.2)    | 6.1 (3.4)    | 7.1 (0.4)     | 2.1 (2.7)   | 4.5 (0.6)     |
| Bowling.9.11         | 6.0 (3.6)   | 5.5 (1.8)    | 7.7 (4.7)    | 10.0 (0.3)    | 8.0 (3.7)   | 4.5 (0.8)     |
| InsuranceSmall.15.25 | 18.1 (6.2)  | 16.7 (7.4)   | 24.1 (6.2)   | 67.3 (1.4)    | 18.1 (4.0)  | 4.3 (1.0)     |
| Rain.14.18           | 13.8 (5.5)  | 13.2 (4.8)   | 18.0 (7.9)   | 51.7 (1.4)    | 8.5 (3.5)   | 4.6 (2.0)     |
| Cloud.16.19          | 17.2 (4.9)  | 15.0 (6.5)   | 20.8 (6.0)   | 86.2 (4.4)    | 5.5 (3.3)   | 3.9 (0.7)     |
| Funnel.18.18         | 12.7 (5.6)  | 11.1 (3.3)   | 17.6 (6.1)   | 110.3 (2.1)   | 4.9 (3.0)   | 4.0 (0.6)     |
| Galaxy.20.22         | 18.8 (7.1)  | 19.8 (5.3)   | 23.6 (7.5)   | 169.0 (4.7)   | 6.4 (2.6)   | 3.9 (0.6)     |
| Insurance.27.52      | 42.1 (12.6) | 61.1 (16.1)  | 103.4 (10.2) | 497.3 (9.8)   | 34.4 (7.1)  | 4.1 (0.6)     |
| Factors.27.68        | 46.8 (14.1) | 140.9 (51.7) | 119.4 (9.3)  | 489.2 (11.1)  | 72.1 (7.7)  | 4.3 (0.6)     |
|                      | CAM         |              | MIP (super)  |               | MIP (moral) |               |
| Network. $p.s^*$     | $d$         | Time         | $d$          | Time          | $d$         | Time          |
| Dsep.6.6             | 2.8 (1.9)   | 10.8 (0.7)   | 1.2 (1.6)    | 4.9 (0.7)     | 1.1 (1.6)   | 4.7 (0.5)     |
| Asia.8.8             | 0.3 (0.7)   | 16.1 (1.3)   | 0.5 (0.9)    | 5.7 (1.2)     | 0.3 (0.8)   | 4.5 (0.8)     |
| Bowling.9.11         | 2.9 (3.5)   | 16.5 (1.8)   | 1.1 (1.5)    | 15.0 (19.4)   | 0.9 (1.3)   | 5.3 (1.1)     |
| InsuranceSmall.15.25 | 11.3 (5.1)  | 46.3 (4.4)   | 4.0 (2.8)    | 681.8 (326.0) | 3.2 (3.0)   | 31.2 (40.3)   |
| Rain.14.18           | 6.1 (6.0)   | 45.8 (3.7)   | 1.9 (3.1)    | 297.7 (324.0) | 0.8 (2.6)   | 11.6 (7.9)    |
| Cloud.16.19          | 3.0 (2.0)   | 57.0 (5.8)   | 2.4 (2.1)    | 133.6 (246.1) | 0.9 (1.5)   | 6.8 (0.9)     |
| Funnel.18.18         | 0.2 (0.6)   | 51.2 (4.8)   | 1.4 (1.1)    | 511.5 (452.7) | 0.5 (0.8)   | 11.5 (6.9)    |
| Galaxy.20.22         | 1.1 (1.6)   | 67.4 (5.4)   | 2.0 (1.8)    | 594.5 (500.6) | 0.7 (1.3)   | 8.0 (0.9)     |
| Insurance.27.52      | 12.8 (6.6)  | 92.2 (6.9)   | 5.9 (4.3)    | 1631.8 (1.4)  | 3.1 (2.7)   | 718.0 (593.5) |
| Factors.27.68        | 55.4 (11.4) | 112.9 (6.4)  | 25.8 (10.7)  | 1633.6 (1.9)  | 11.2 (9.3)  | 1627.8 (0.5)  |

TABLE 1. The full performance of *MIP* and baseline methods for the heteroscedastic scheme with  $\mu_0 = 1$ . The true functions are  $f_{kj}^*(x) = (\sin(x) - \mathbb{E}[\sin(X_k)] + \cos(x) - \mathbb{E}[\cos(X_k)])/2$ , and splines are used as basis functions. Each entry gives the mean value and standard deviation in parenthesis over 30 independent trials. The EqVar (TD) and EqVar (BU) refer to the top-down and bottom-up version of *EqVar*, respectively. The MIP (super) and MIP (moral) refer to the *MIP* using the estimated super-structure and the true moral graph, respectively.

|                       | NPVAR       |               | EqVar (TD)   |               | EqVar (BU)  |                |
|-----------------------|-------------|---------------|--------------|---------------|-------------|----------------|
|                       | <i>d</i>    | Time          | <i>d</i>     | Time          | <i>d</i>    | Time           |
| Network. <i>p.s</i> * |             |               |              |               |             |                |
| Dsep.6.6              | 6.7 (3.5)   | 0.3 (0.0)     | 6.0 (3.3)    | 0.2 (0.0)     | 6.2 (3.5)   | 0.2 (0.0)      |
| Asia.8.8              | 8.4 (3.8)   | 0.6 (0.0)     | 6.7 (3.3)    | 0.3 (0.0)     | 6.8 (3.2)   | 0.3 (0.0)      |
| Bowling.9.11          | 11.0 (4.3)  | 0.8 (0.1)     | 7.7 (4.0)    | 0.4 (0.0)     | 8.5 (4.6)   | 0.3 (0.0)      |
| InsuranceSmall.15.25  | 27.7 (8.0)  | 3.4 (0.1)     | 26.5 (8.0)   | 0.7 (0.0)     | 27.4 (7.6)  | 0.7 (0.0)      |
| Rain.14.18            | 22.3 (6.3)  | 2.8 (0.2)     | 18.3 (6.1)   | 0.7 (0.1)     | 18.0 (5.5)  | 0.6 (0.1)      |
| Cloud.16.19           | 26.0 (8.4)  | 4.2 (0.2)     | 22.1 (5.3)   | 0.7 (0.0)     | 21.9 (5.8)  | 0.8 (0.1)      |
| Funnel.18.18          | 23.3 (7.8)  | 6.2 (0.5)     | 15.3 (5.0)   | 0.9 (0.1)     | 15.8 (5.9)  | 0.9 (0.1)      |
| Galaxy.20.22          | 30.0 (7.5)  | 7.8 (0.2)     | 20.1 (5.1)   | 1.0 (0.1)     | 20.9 (4.7)  | 1.0 (0.1)      |
| Insurance.27.52       | 63.9 (10.1) | 18.8 (0.8)    | 69.8 (10.2)  | 1.3 (0.1)     | 70.2 (11.5) | 1.4 (0.1)      |
| Factors.27.68         | 54.8 (10.7) | 21.5 (0.8)    | 81.7 (22.9)  | 1.5 (0.1)     | 82.4 (22.3) | 1.6 (0.1)      |
|                       | NoTears     |               | RESIT        |               | CCDr        |                |
| Network. <i>p.s</i> * | <i>d</i>    | Time          | <i>d</i>     | Time          | <i>d</i>    | Time           |
| Dsep.6.6              | 5.5 (4.0)   | 4.3 (2.1)     | 3.4 (2.9)    | 3.4 (0.2)     | 3.7 (2.1)   | 4.1 (1.6)      |
| Asia.8.8              | 5.9 (3.7)   | 5.5 (3.2)     | 6.1 (4.1)    | 7.3 (0.4)     | 2.6 (4.1)   | 4.1 (0.8)      |
| Bowling.9.11          | 5.9 (4.1)   | 6.4 (8.0)     | 10.0 (4.8)   | 9.0 (0.3)     | 10.6 (3.9)  | 3.6 (0.5)      |
| InsuranceSmall.15.25  | 20.5 (6.8)  | 18.9 (6.8)    | 29.0 (7.4)   | 51.7 (2.0)    | 18.4 (3.9)  | 3.7 (0.5)      |
| Rain.14.18            | 14.1 (5.7)  | 12.6 (5.6)    | 19.1 (6.0)   | 39.0 (2.7)    | 9.5 (3.3)   | 3.5 (0.6)      |
| Cloud.16.19           | 19.5 (4.5)  | 15.5 (7.4)    | 21.2 (7.5)   | 69.0 (1.5)    | 4.3 (3.1)   | 3.9 (0.6)      |
| Funnel.18.18          | 11.5 (5.0)  | 10.0 (3.8)    | 21.3 (8.7)   | 114.7 (9.0)   | 4.7 (2.7)   | 5.0 (2.6)      |
| Galaxy.20.22          | 17.0 (5.8)  | 14.9 (6.0)    | 25.3 (7.6)   | 170.4 (4.7)   | 5.4 (3.1)   | 4.1 (0.5)      |
| Insurance.27.52       | 38.1 (9.7)  | 72.1 (29.7)   | 110.6 (15.4) | 470.7 (29.4)  | 36.6 (5.0)  | 4.1 (0.5)      |
| Factors.27.68         | 46.5 (12.9) | 196.3 (100.8) | 116.0 (11.0) | 535.1 (20.7)  | 71.3 (6.4)  | 4.6 (1.0)      |
|                       | CAM         |               | MIP (super)  |               | MIP (moral) |                |
| Network. <i>p.s</i> * | <i>d</i>    | Time          | <i>d</i>     | Time          | <i>d</i>    | Time           |
| Dsep.6.6              | 3.0 (2.1)   | 11.5 (1.4)    | 0.8 (1.6)    | 5.2 (1.0)     | 0.7 (1.6)   | 4.8 (0.6)      |
| Asia.8.8              | 0.0 (0.0)   | 15.5 (1.4)    | 0.6 (0.8)    | 5.2 (0.9)     | 0.3 (0.6)   | 4.1 (0.3)      |
| Bowling.9.11          | 5.7 (5.2)   | 15.4 (1.5)    | 1.1 (2.6)    | 47.3 (91.3)   | 1.0 (2.9)   | 6.4 (2.1)      |
| InsuranceSmall.15.25  | 14.3 (4.1)  | 41.6 (4.7)    | 6.4 (4.7)    | 861.6 (160.5) | 4.2 (4.1)   | 46.2 (37.6)    |
| Rain.14.18            | 6.3 (6.4)   | 40.4 (4.4)    | 1.7 (1.8)    | 460.1 (344.3) | 0.9 (1.5)   | 17.0 (11.1)    |
| Cloud.16.19           | 3.7 (2.3)   | 53.0 (4.9)    | 2.6 (2.0)    | 196.5 (319.6) | 1.2 (1.5)   | 7.1 (1.6)      |
| Funnel.18.18          | 0.2 (0.7)   | 50.7 (6.0)    | 1.4 (1.3)    | 609.6 (488.7) | 0.5 (1.0)   | 12.9 (7.3)     |
| Galaxy.20.22          | 1.6 (2.1)   | 66.9 (4.3)    | 2.4 (1.8)    | 683.4 (542.9) | 0.7 (1.0)   | 9.1 (1.7)      |
| Insurance.27.52       | 17.4 (5.6)  | 94.4 (6.5)    | 7.6 (4.9)    | 1631.6 (1.1)  | 4.1 (3.7)   | 1173.0 (573.2) |
| Factors.27.68         | 63.4 (11.1) | 117.0 (12.8)  | 30.4 (10.7)  | 1633.2 (2.0)  | 16.2 (7.7)  | 1627.6 (0.7)   |

TABLE 2. The full performance of *MIP* and baseline methods for the heteroscedastic scheme with  $\mu_0 = 2$ . The true functions are  $f_{kj}^*(x) = (\sin(x) - \mathbb{E}[\sin(X_k)] + \cos(x) - \mathbb{E}[\cos(X_k)])/2$ , and splines are used as basis functions. Each entry gives the mean value and standard deviation in parenthesis over 30 independent trials. The EqVar (TD) and EqVar (BU) refer to the top-down and bottom-up version of *EqVar*, respectively. The MIP (super) and MIP (moral) refer to the *MIP* using the estimated super-structure and the true moral graph, respectively.

|                       | NPVAR       |              | EqVar (TD)   |               | EqVar (BU)  |                |
|-----------------------|-------------|--------------|--------------|---------------|-------------|----------------|
|                       | <i>d</i>    | Time         | <i>d</i>     | Time          | <i>d</i>    | Time           |
| Network. <i>p.s</i> * |             |              |              |               |             |                |
| Dsep.6.6              | 4.8 (3.4)   | 0.3 (0.0)    | 4.0 (3.1)    | 0.2 (0.0)     | 3.9 (3.1)   | 0.2 (0.0)      |
| Asia.8.8              | 8.3 (3.7)   | 0.6 (0.0)    | 6.8 (2.9)    | 0.3 (0.0)     | 6.8 (3.2)   | 0.3 (0.0)      |
| Bowling.9.11          | 11.3 (4.6)  | 0.8 (0.1)    | 7.8 (3.0)    | 0.4 (0.0)     | 8.0 (2.8)   | 0.4 (0.0)      |
| InsuranceSmall.15.25  | 24.6 (6.9)  | 4.1 (0.1)    | 24.2 (5.8)   | 0.9 (0.1)     | 24.2 (5.3)  | 0.9 (0.1)      |
| Rain.14.18            | 21.1 (8.5)  | 3.5 (0.1)    | 18.2 (6.2)   | 0.9 (0.1)     | 18.3 (6.3)  | 0.8 (0.1)      |
| Cloud.16.19           | 21.7 (7.8)  | 4.5 (0.2)    | 20.6 (6.3)   | 0.8 (0.1)     | 21.0 (6.6)  | 0.9 (0.1)      |
| Funnel.18.18          | 25.0 (5.8)  | 5.8 (0.2)    | 17.3 (5.5)   | 0.9 (0.1)     | 17.0 (5.8)  | 0.9 (0.1)      |
| Galaxy.20.22          | 33.1 (7.9)  | 7.8 (0.3)    | 21.6 (5.0)   | 1.0 (0.1)     | 22.3 (5.3)  | 1.0 (0.1)      |
| Insurance.27.52       | 58.0 (10.4) | 19.7 (1.1)   | 66.8 (10.0)  | 1.4 (0.1)     | 68.0 (8.7)  | 1.4 (0.1)      |
| Factors.27.68         | 46.1 (8.4)  | 19.6 (0.5)   | 67.5 (12.6)  | 1.4 (0.1)     | 67.0 (13.3) | 1.4 (0.1)      |
|                       | NoTears     |              | RESIT        |               | CCDr        |                |
| Network. <i>p.s</i> * | <i>d</i>    | Time         | <i>d</i>     | Time          | <i>d</i>    | Time           |
| Dsep.6.6              | 4.2 (2.9)   | 4.2 (3.1)    | 5.0 (3.0)    | 3.4 (0.3)     | 3.6 (1.5)   | 3.9 (0.9)      |
| Asia.8.8              | 5.6 (3.4)   | 5.9 (2.5)    | 7.0 (2.8)    | 7.2 (0.4)     | 2.2 (2.7)   | 3.8 (0.7)      |
| Bowling.9.11          | 5.0 (4.3)   | 6.1 (3.0)    | 9.3 (4.8)    | 9.8 (0.4)     | 10.9 (4.0)  | 4.0 (0.7)      |
| InsuranceSmall.15.25  | 16.3 (4.6)  | 20.2 (6.5)   | 28.5 (7.7)   | 66.3 (1.2)    | 18.2 (3.6)  | 3.9 (0.5)      |
| Rain.14.18            | 12.6 (5.8)  | 11.8 (3.9)   | 22.3 (8.4)   | 50.4 (1.1)    | 10.1 (2.5)  | 3.6 (0.4)      |
| Cloud.16.19           | 16.0 (5.5)  | 13.5 (4.4)   | 21.4 (7.7)   | 73.4 (3.2)    | 4.1 (2.5)   | 7.3 (1.2)      |
| Funnel.18.18          | 11.2 (5.3)  | 9.6 (4.1)    | 20.7 (7.6)   | 109.7 (2.7)   | 4.0 (2.4)   | 3.9 (0.5)      |
| Galaxy.20.22          | 18.4 (7.4)  | 19.3 (9.0)   | 28.4 (8.7)   | 167.8 (3.7)   | 7.5 (4.0)   | 4.5 (0.9)      |
| Insurance.27.52       | 33.7 (7.7)  | 91.7 (49.2)  | 113.7 (15.8) | 511.1 (24.8)  | 32.7 (7.3)  | 8.8 (2.1)      |
| Factors.27.68         | 33.1 (8.2)  | 182.0 (79.3) | 118.5 (10.2) | 487.5 (6.7)   | 69.7 (5.5)  | 4.0 (0.6)      |
|                       | CAM         |              | MIP (super)  |               | MIP (moral) |                |
| Network. <i>p.s</i> * | <i>d</i>    | Time         | <i>d</i>     | Time          | <i>d</i>    | Time           |
| Dsep.6.6              | 2.9 (1.9)   | 11.0 (0.8)   | 1.4 (2.0)    | 4.4 (0.7)     | 1.5 (2.2)   | 4.1 (0.5)      |
| Asia.8.8              | 0.1 (0.4)   | 14.8 (0.9)   | 0.4 (0.8)    | 5.2 (0.9)     | 0.2 (0.6)   | 4.1 (0.4)      |
| Bowling.9.11          | 7.4 (3.9)   | 17.2 (1.7)   | 1.5 (2.8)    | 46.6 (80.4)   | 1.1 (2.8)   | 7.3 (4.4)      |
| InsuranceSmall.15.25  | 15.8 (5.2)  | 48.7 (5.1)   | 6.6 (5.9)    | 858.9 (176.1) | 5.3 (5.7)   | 69.5 (48.7)    |
| Rain.14.18            | 8.0 (6.2)   | 45.6 (4.7)   | 1.8 (3.4)    | 569.0 (356.0) | 1.3 (3.4)   | 22.4 (17.5)    |
| Cloud.16.19           | 4.4 (1.7)   | 55.3 (4.7)   | 2.4 (2.0)    | 172.1 (279.7) | 1.1 (1.5)   | 8.0 (1.5)      |
| Funnel.18.18          | 0.2 (0.6)   | 51.2 (5.5)   | 1.9 (1.4)    | 675.3 (447.5) | 0.8 (1.1)   | 10.9 (1.5)     |
| Galaxy.20.22          | 3.2 (3.0)   | 68.9 (4.5)   | 1.6 (1.8)    | 696.1 (557.5) | 0.4 (0.9)   | 11.3 (7.9)     |
| Insurance.27.52       | 20.9 (8.6)  | 97.0 (7.6)   | 10.2 (6.4)   | 1632.4 (1.4)  | 6.9 (4.8)   | 1382.3 (456.5) |
| Factors.27.68         | 59.8 (7.1)  | 115.2 (6.5)  | 34.1 (9.1)   | 1633.2 (2.0)  | 16.4 (8.0)  | 1627.8 (0.8)   |

TABLE 3. The full performance of *MIP* and baseline methods for the heteroscedastic scheme with  $\mu_0 = 3$ . The true functions are  $f_{kj}^*(x) = (\sin(x) - \mathbb{E}[\sin(X_k)] + \cos(x) - \mathbb{E}[\cos(X_k)])/2$ , and splines are used as basis functions. Each entry gives the mean value and standard deviation in parenthesis over 30 independent trials. The EqVar (TD) and EqVar (BU) refer to the top-down and bottom-up version of *EqVar*, respectively. The MIP (super) and MIP (moral) refer to the *MIP* using the estimated super-structure and the true moral graph, respectively.

|                       | NPVAR      |              | EqVar (TD)   |               | EqVar (BU)  |                |
|-----------------------|------------|--------------|--------------|---------------|-------------|----------------|
|                       | <i>d</i>   | Time         | <i>d</i>     | Time          | <i>d</i>    | Time           |
| Network. <i>p.s</i> * |            |              |              |               |             |                |
| Dsep.6.6              | 0.7 (1.3)  | 0.3 (0.0)    | 0.7 (1.3)    | 0.3 (0.1)     | 0.7 (1.4)   | 0.3 (0.1)      |
| Asia.8.8              | 0.6 (1.1)  | 0.7 (0.0)    | 0.5 (1.0)    | 0.4 (0.0)     | 0.5 (0.7)   | 0.4 (0.0)      |
| Bowling.9.11          | 0.8 (1.7)  | 1.0 (0.1)    | 0.4 (0.8)    | 0.5 (0.1)     | 0.7 (1.4)   | 0.5 (0.1)      |
| InsuranceSmall.15.25  | 2.5 (3.4)  | 3.7 (0.2)    | 9.4 (2.9)    | 0.8 (0.0)     | 9.8 (2.8)   | 0.8 (0.1)      |
| Rain.14.18            | 4.0 (4.5)  | 3.0 (0.2)    | 3.2 (2.8)    | 0.7 (0.0)     | 3.9 (2.5)   | 0.7 (0.1)      |
| Cloud.16.19           | 2.2 (3.0)  | 4.4 (0.3)    | 3.7 (2.4)    | 0.8 (0.1)     | 5.5 (4.5)   | 0.9 (0.1)      |
| Funnel.18.18          | 2.9 (5.4)  | 7.0 (0.3)    | 2.1 (1.3)    | 1.1 (0.1)     | 2.2 (1.4)   | 1.1 (0.1)      |
| Galaxy.20.22          | 3.0 (4.3)  | 7.9 (0.6)    | 4.2 (4.3)    | 1.0 (0.1)     | 4.5 (3.8)   | 1.0 (0.1)      |
| Insurance.27.52       | 16.5 (9.4) | 21.2 (1.6)   | 36.0 (5.9)   | 1.6 (0.1)     | 37.3 (5.9)  | 1.6 (0.2)      |
| Factors.27.68         | 22.9 (5.2) | 21.0 (1.0)   | 55.0 (6.1)   | 1.6 (0.1)     | 53.7 (6.9)  | 1.6 (0.2)      |
|                       | NoTears    |              | RESIT        |               | CCDr        |                |
| Network. <i>p.s</i> * | <i>d</i>   | Time         | <i>d</i>     | Time          | <i>d</i>    | Time           |
| Dsep.6.6              | 2.0 (2.0)  | 3.8 (2.1)    | 5.5 (3.6)    | 4.0 (0.3)     | 3.7 (1.1)   | 4.7 (1.2)      |
| Asia.8.8              | 1.0 (1.8)  | 4.7 (1.9)    | 8.4 (3.4)    | 8.7 (0.3)     | 2.6 (2.8)   | 3.4 (0.5)      |
| Bowling.9.11          | 0.0 (0.0)  | 5.0 (1.7)    | 12.3 (5.5)   | 11.7 (0.6)    | 11.5 (4.2)  | 4.7 (1.0)      |
| InsuranceSmall.15.25  | 11.2 (3.8) | 34.2 (17.5)  | 36.3 (7.7)   | 56.9 (2.8)    | 19.1 (3.5)  | 3.7 (0.6)      |
| Rain.14.18            | 5.8 (3.5)  | 19.0 (11.4)  | 25.1 (10.1)  | 44.1 (2.6)    | 9.3 (1.8)   | 5.8 (2.5)      |
| Cloud.16.19           | 5.8 (3.7)  | 16.3 (8.4)   | 26.3 (7.9)   | 75.1 (1.7)    | 3.9 (1.9)   | 4.1 (0.6)      |
| Funnel.18.18          | 1.7 (1.7)  | 11.0 (5.1)   | 25.2 (7.4)   | 134.7 (4.8)   | 3.1 (2.3)   | 3.7 (0.5)      |
| Galaxy.20.22          | 5.0 (3.8)  | 22.0 (9.5)   | 31.3 (6.2)   | 170.0 (14.5)  | 8.3 (3.9)   | 5.1 (0.8)      |
| Insurance.27.52       | 19.3 (6.7) | 107.8 (51.1) | 120.5 (14.7) | 533.8 (15.2)  | 32.1 (5.0)  | 4.4 (0.6)      |
| Factors.27.68         | 17.1 (9.9) | 182.4 (41.0) | 118.7 (9.9)  | 522.6 (23.8)  | 63.2 (3.4)  | 5.6 (1.3)      |
|                       | CAM        |              | MIP (super)  |               | MIP (moral) |                |
| Network. <i>p.s</i> * | <i>d</i>   | Time         | <i>d</i>     | Time          | <i>d</i>    | Time           |
| Dsep.6.6              | 4.9 (0.6)  | 11.4 (0.8)   | 0.0 (0.0)    | 0.9 (0.6)     | 0.0 (0.0)   | 0.4 (0.5)      |
| Asia.8.8              | 0.3 (0.8)  | 15.9 (0.9)   | 0.0 (0.0)    | 1.1 (0.7)     | 0.1 (0.3)   | 0.7 (0.6)      |
| Bowling.9.11          | 12.1 (6.6) | 18.7 (1.3)   | 0.0 (0.2)    | 2.3 (1.5)     | 0.1 (0.3)   | 1.6 (0.5)      |
| InsuranceSmall.15.25  | 23.6 (4.4) | 47.6 (3.6)   | 0.2 (0.5)    | 180.3 (280.8) | 0.2 (0.5)   | 4.2 (3.2)      |
| Rain.14.18            | 16.3 (4.4) | 41.6 (2.1)   | 0.0 (0.2)    | 40.5 (114.3)  | 0.1 (0.3)   | 3.7 (0.7)      |
| Cloud.16.19           | 4.2 (1.6)  | 53.7 (2.7)   | 0.3 (0.8)    | 58.4 (111.8)  | 0.2 (0.7)   | 3.0 (0.8)      |
| Funnel.18.18          | 0.4 (1.1)  | 56.0 (4.2)   | 0.1 (0.4)    | 57.2 (195.8)  | 0.2 (0.6)   | 3.3 (1.2)      |
| Galaxy.20.22          | 9.9 (6.0)  | 76.3 (5.0)   | 0.0 (0.2)    | 271.2 (325.2) | 0.1 (0.3)   | 4.7 (2.0)      |
| Insurance.27.52       | 33.9 (6.4) | 103.8 (9.2)  | 1.0 (1.0)    | 1625.4 (1.5)  | 0.3 (0.5)   | 606.9 (598.2)  |
| Factors.27.68         | 61.3 (5.8) | 131.0 (7.2)  | 2.8 (2.3)    | 1626.1 (0.8)  | 0.7 (0.9)   | 1130.9 (633.6) |

TABLE 4. The full performance of *MIP* and baseline methods for the homoscedastic scheme. The true functions are  $f_{kj}^*(x) = (\sin(x) - \mathbb{E}[\sin(X_k)] + \cos(x) - \mathbb{E}[\cos(X_k)])/2$ , and splines are used as basis functions. Each entry gives the mean value and standard deviation in parenthesis over 30 independent trials. The EqVar (TD) and EqVar (BU) refer to the top-down and bottom-up version of *EqVar*, respectively. The MIP (super) and MIP (moral) refer to the *MIP* using the estimated super-structure and the true moral graph, respectively.

|                       | NPVAR      |             | EqVar (TD)  |               | EqVar (BU)  |              |
|-----------------------|------------|-------------|-------------|---------------|-------------|--------------|
|                       | <i>d</i>   | Time        | <i>d</i>    | Time          | <i>d</i>    | Time         |
| Network. <i>p.s</i> * |            |             |             |               |             |              |
| Dsep.6.6              | 0.6 (0.8)  | 0.3 (0.0)   | 4.3 (1.1)   | 0.2 (0.0)     | 4.3 (1.1)   | 0.2 (0.0)    |
| Asia.8.8              | 0.7 (0.9)  | 0.6 (0.0)   | 6.2 (2.5)   | 0.3 (0.0)     | 7.0 (2.9)   | 0.3 (0.0)    |
| Bowling.9.11          | 0.8 (1.1)  | 0.8 (0.0)   | 2.9 (1.0)   | 0.4 (0.0)     | 2.9 (0.9)   | 0.4 (0.0)    |
| InsuranceSmall.15.25  | 5.2 (3.1)  | 3.3 (0.1)   | 37.2 (2.9)  | 0.7 (0.1)     | 37.0 (3.0)  | 0.7 (0.0)    |
| Rain.14.18            | 4.7 (2.3)  | 2.8 (0.1)   | 32.9 (2.9)  | 0.6 (0.1)     | 33.2 (3.4)  | 0.6 (0.0)    |
|                       | NoTears    |             | RESIT       |               | CCDr        |              |
|                       | <i>d</i>   | Time        | <i>d</i>    | Time          | <i>d</i>    | Time         |
| Network. <i>p.s</i> * |            |             |             |               |             |              |
| Dsep.6.6              | 1.3 (0.6)  | 4.5 (1.8)   | 1.9 (2.4)   | 3.5 (0.2)     | 4.4 (0.7)   | 3.6 (0.5)    |
| Asia.8.8              | 1.8 (0.6)  | 7.2 (5.2)   | 3.4 (3.4)   | 7.3 (0.3)     | 1.7 (1.4)   | 4.2 (0.5)    |
| Bowling.9.11          | 0.8 (0.7)  | 11.7 (5.4)  | 8.2 (4.5)   | 10.2 (0.4)    | 12.0 (7.9)  | 6.5 (3.3)    |
| InsuranceSmall.15.25  | 23.0 (2.4) | 33.9 (16.8) | 34.6 (6.3)  | 56.9 (1.3)    | 29.9 (1.9)  | 3.6 (0.4)    |
| Rain.14.18            | 19.9 (2.8) | 51.7 (46.9) | 16.8 (6.1)  | 43.7 (1.9)    | 21.4 (3.3)  | 3.6 (0.5)    |
|                       | CAM        |             | MIP (super) |               | MIP (moral) |              |
|                       | <i>d</i>   | Time        | <i>d</i>    | Time          | <i>d</i>    | Time         |
| Network. <i>p.s</i> * |            |             |             |               |             |              |
| Dsep.6.6              | 4.7 (2.2)  | 7.9 (0.4)   | 0.1 (0.4)   | 1.5 (1.2)     | 0.1 (0.3)   | 0.8 (0.8)    |
| Asia.8.8              | 0.2 (0.5)  | 9.7 (0.5)   | 0.1 (0.3)   | 35.8 (123.5)  | 0.7 (0.8)   | 34.1 (121.3) |
| Bowling.9.11          | 0.1 (0.3)  | 10.9 (0.4)  | 0.3 (0.6)   | 0.7 (0.8)     | 0.1 (0.3)   | 1.8 (1.5)    |
| InsuranceSmall.15.25  | 11.8 (3.7) | 24.9 (1.7)  | 1.8 (2.2)   | 438.8 (419.8) | 2.4 (2.1)   | 64.6 (227.9) |
| Rain.14.18            | 7.9 (5.6)  | 21.1 (1.3)  | 0.7 (1.3)   | 279.5 (363.3) | 1.5 (1.5)   | 61.7 (212.2) |

TABLE 5. The full performance of *MIP* and baseline methods for the homoscedastic scheme. The true functions are  $f_{kj}^*(x) = \sin(x + \sin(x)) - \mathbb{E}[\sin(X_k + \sin(X_k))]$ , and radial basis functions are used as basis functions. Each entry gives the mean value and standard deviation in parenthesis over 30 independent trials. The EqVar (TD) and EqVar (BU) refer to the top-down and bottom-up version of *EqVar*, respectively. The MIP (super) and MIP (moral) refer to the *MIP* using the estimated super-structure and the true moral graph, respectively.

|                      | NPVAR      |            | EqVar (TD)  |               | EqVar (BU)  |               |
|----------------------|------------|------------|-------------|---------------|-------------|---------------|
|                      | $d$        | Time       | $d$         | Time          | $d$         | Time          |
| Network. $p.s^*$     |            |            |             |               |             |               |
| Dsep.6.6             | 1.8 (2.4)  | 0.3 (0.0)  | 4.1 (2.3)   | 0.2 (0.0)     | 4.0 (2.1)   | 0.2 (0.0)     |
| Asia.8.8             | 4.1 (3.9)  | 0.6 (0.0)  | 8.6 (2.7)   | 0.3 (0.0)     | 8.2 (2.7)   | 0.3 (0.0)     |
| Bowling.9.11         | 4.8 (5.0)  | 0.9 (0.1)  | 5.7 (3.4)   | 0.4 (0.0)     | 5.7 (3.4)   | 0.4 (0.0)     |
| InsuranceSmall.15.25 | 15.2 (7.3) | 3.4 (0.2)  | 38.0 (6.3)  | 0.7 (0.1)     | 38.7 (6.4)  | 0.7 (0.0)     |
| Rain.14.18           | 13.2 (6.7) | 2.9 (0.1)  | 30.2 (3.9)  | 0.6 (0.0)     | 30.9 (3.5)  | 0.6 (0.0)     |
|                      | NoTears    |            | RESIT       |               | CCDr        |               |
| Network. $p.s^*$     | $d$        | Time       | $d$         | Time          | $d$         | Time          |
| Dsep.6.6             | 1.1 (0.9)  | 4.3 (3.0)  | 2.2 (2.8)   | 3.5 (0.3)     | 4.3 (2.4)   | 4.4 (1.3)     |
| Asia.8.8             | 4.0 (2.8)  | 5.8 (4.5)  | 2.7 (2.7)   | 7.2 (0.4)     | 2.7 (1.8)   | 3.6 (0.7)     |
| Bowling.9.11         | 2.3 (2.1)  | 10.5 (9.0) | 6.8 (4.2)   | 10.2 (0.4)    | 9.5 (6.3)   | 3.5 (0.5)     |
| InsuranceSmall.15.25 | 21.2 (5.4) | 21.7 (8.8) | 30.9 (5.7)  | 56.2 (1.7)    | 28.6 (3.7)  | 3.8 (0.6)     |
| Rain.14.18           | 16.6 (3.6) | 14.7 (5.9) | 14.7 (5.2)  | 43.0 (1.2)    | 20.2 (3.3)  | 4.3 (0.7)     |
|                      | CAM        |            | MIP (super) |               | MIP (moral) |               |
| Network. $p.s^*$     | $d$        | Time       | $d$         | Time          | $d$         | Time          |
| Dsep.6.6             | 2.6 (2.8)  | 13.7 (3.1) | 2.2 (2.0)   | 252.3 (230.2) | 1.9 (1.9)   | 232.8 (218.0) |
| Asia.8.8             | 0.8 (1.3)  | 9.6 (0.6)  | 0.6 (1.5)   | 420.8 (347.5) | 0.3 (0.7)   | 340.5 (334.9) |
| Bowling.9.11         | 0.1 (0.4)  | 10.5 (0.4) | 0.2 (0.4)   | 663.3 (318.7) | 0.2 (0.4)   | 714.8 (317.5) |
| InsuranceSmall.15.25 | 6.8 (3.1)  | 22.3 (2.2) | 4.8 (1.7)   | 756.1 (654.2) | 5.2 (2.0)   | 828.3 (575.3) |
| Rain.14.18           | 5.7 (4.6)  | 21.7 (2.2) | 2.4 (2.4)   | 827.8 (619.7) | 2.0 (2.0)   | 782.9 (598.3) |

TABLE 6. The full performance of *MIP* and baseline methods for the heteroscedastic scheme with  $\mu_0 = 2$ . The true functions are  $f_{kj}^*(x) = \sin(x + \sin(x)) - \mathbb{E}[\sin(X_k + \sin(X_k))]$ , and radial basis functions are used as basis functions. Each entry gives the mean value and standard deviation in parenthesis over 30 independent trials. The EqVar (TD) and EqVar (BU) refer to the top-down and bottom-up version of *EqVar*, respectively. The MIP (super) and MIP (moral) refer to the *MIP* using the estimated super-structure and the true moral graph, respectively.