Multi-task Learning of Order-Consistent Causal Graphs

Xinshi Chen*
Georgia Institute of Technology
xinshi.chen@gatech.edu

Haoran Sun
Georgia Institute of Technology
haoransun@gatech.edu

Caleb Ellington
Carnegie Mellon University
cellingt@cs.cmu.edu

Eric Xing
Carnegie Mellon University
MBZUAI
eric.xing@mbzuai.ac.ae

Le Song
BioMap
MBZUAI
le.song@mbzuai.ac.ae

Abstract

We consider the problem of discovering K related Gaussian directed acyclic graphs (DAGs), where the involved graph structures share a consistent causal order and sparse unions of supports. Under the multi-task learning setting, we propose a l_1/l_2-regularized maximum likelihood estimator (MLE) for learning K linear structural equation models. We theoretically show that the joint estimator, by leveraging data across related tasks, can achieve a better sample complexity for recovering the causal order (or topological order) than separate estimations. Moreover, the joint estimator is able to recover non-identifiable DAGs, by estimating them together with some identifiable DAGs. Lastly, our analysis also shows the consistency of union support recovery of the structures. To allow practical implementation, we design a continuous optimization problem whose optimizer is the same as the joint estimator and can be approximated efficiently by an iterative algorithm. We validate the theoretical analysis and the effectiveness of the joint estimator in experiments.

1 Introduction

Estimating causal effects among a set of random variables is of fundamental importance in many disciplines such as genomics, epidemiology, health care and finance [1, 2, 3, 4, 5, 6]. Therefore, designing and understanding methods for causal discovery is of great interests in machine learning.

Causal discovery from finite observable data is often formulated as a directed acyclic graph (DAG) estimation problem in graphical models. A major class of DAG estimation methods are score-based, which search over the space of all DAGs for the best scoring one. However, DAG estimation remains a very challenging problem from both the *computational* and *statistical* aspects [7]. On the one hand, the number of possible DAG structures grows super-exponentially in the number of random variables, whereas the number of observational sample size is normally small. On the other hand, some DAGs are *non-identifiable* from observational data even with infinitely many samples.

Fortunately, very often multiple related DAG structures need to be estimated from data, which allows us to leverage their similarity to improve the estimator. For instance, in bioinformatics, gene expression levels are often measured over patients with different subtypes [8, 9] or under various experimental conditions [10]. In neuroinformatics, fMRI signals are often recorded for multiple subjects for studying the brain connectivity network [11, 12]. In these scenarios, multiple datasets will be collected, and their associated DAGs are likely to share similar characteristics. Intuitively, it may be beneficial to estimate these DAGs jointly.

*Work done partially during the visit at MBZUAI (Mohamed bin Zayed University of Artificial Intelligence)

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

In this paper, we focus on the analysis of the multi-task DAG estimation problem where the DAG structures can be related by a consistent causal order and a (partially) shared sparsity pattern, but allowed to have different connection strengths and edges, and differently distributed variables. In this setting, we propose a joint estimator for recovering multiple DAGs based on a group norm regularization. We prove that the joint l_1/l_2-penalized maximum likelihood estimator (MLE) can recover the causal order better than individual estimators.

Intuitively, it is not surprising that joint estimation is beneficial. However, our results provide a quantitative characterization on the improvement in sample complexity and the conditions under which such improvement can hold. We show that:

- For identifiable DAGs, if the shared sparsity pattern (union support) size s is of order $\mathcal{O}(1)$ in K where K is the number of tasks (DAGs), then the *effective* sample size for order recovery will be nK where n is the sample size in each problem. Furthermore, as long as s is of order $o(\sqrt{K})$ in K, the joint estimator with group norm regularization leads to an improvement in sample complexity.
- A non-identifiable DAG cannot be distinguished by single-task estimators even with indefinitely many observational data. However, non-identifiable DAGs can be recovered by our joint estimator if they are estimated together with other identifiable DAGs.

Apart from the theoretical guarantee, we design an efficient algorithm for approximating the joint estimator through a formulation of the combinatorial search problem to a continuous programming. This continuous formulation contains a novel design of a learnable masking matrix, which plays an important role in ensuring the acyclicity and shared order for estimations in all tasks. An interesting aspect of our design is that we can learn the masking matrix by differentiable search over a continuous space, but the optimum must be contained in a discrete space of cardinality $p!$ (reads p factorial, where p is the number of random variables).

We conduct a set of synthetic experiments to demonstrates the effectiveness of the algorithm and validates the theoretical results. Furthermore, we apply our algorithm to more realistic single-cell expression RNA sequencing data generated by SERGIO [13] based on real gene regulatory networks.

The remainder of the paper is organized as follows. In Section 2, we introduce the linear structural equation model (SEM) interpretation of Gaussian DAGs and its properties. Section 3 is devoted to the statement of our main results, with some discussion on their consequences and implications. In Section 4, we present the efficient algorithm for approximating the joint estimator. Section 5 summarizes related theoretical and practical works. Experimental validations are provided in Section 6.

2 Background

A substantial body of work has focused on the linear SEM interpretation of Gaussian DAGs [14, 15, 16, 17]. Let $X = (X_1, \cdots, X_p)$ be a p-dimensional random variable. Then a linear SEM reads

$$X = \widetilde{G}_0^\top X + W, \quad W \sim \mathcal{N}(0, \Omega_0), \tag{1}$$

where Ω_0 is a $p \times p$ positive diagonal matrix which indicates the variances of the noise W. $\widetilde{G}_0 \in \mathbb{R}^{p\times p}$ is the connection strength matrix or adjacency matrix. Each nonzero entry $\widetilde{G}_{0ij}$ represents the direct causal effect of X_i on X_j. This model implies that X is Gaussian, $X \sim \mathcal{N}(0, \Sigma)$, where

$$\Sigma := (I - \widetilde{G}_0)^{-\top} \Omega_0 (I - \widetilde{G}_0)^{-1}. \tag{2}$$

Causal order π_0. The nonzero entries of $\widetilde{G}_0$ defines its causal order (also called topological order), which informs possible "parents" of each variable. A causal order can be represented by a permutation π_0 over $[p] := (1, 2, \cdots, p)$. We say $\widetilde{G}_0$ is **consistent** with π_0 if and only if

$$\widetilde{G}_{0ij} \neq 0 \Rightarrow \pi_0(i) < \pi_0(j). \tag{3}$$

There could exist more than one permutations that are consistent with a DAG structure $\widetilde{G}_0$, so we denote the set of permutations that satisfy Eq. 3 by Π_0. Once a causal order π_0 is identified, the connection strengths $\widetilde{G}_0$ can be estimated by ordinary least squares regression which is a comparatively easier problem.

Identifiability and equivalent class. However, estimating the true causal order π_0 is very challenging, largely due to the existence of the equivalent class described below.

Let $\mathbb{S}_p$ be the set of all permutations over $[p]$. For every $\pi \in \mathbb{S}_p$, there exists a connection strength matrix $\widetilde{G}(\pi)$, which is consistent with π, and a diagonal matrix $\Omega(\pi) = \text{diag}\left[\sigma_1(\pi)^2, \cdots, \sigma_p(\pi)^2\right]$ such that the variance Σ in Eq. 2 equals to

$$\Sigma = (I - \widetilde{G}(\pi))^{-\top}\Omega(\pi)(I - \widetilde{G}(\pi))^{-1}, \tag{4}$$

and therefore the random variable X in Eq. 1 can be equivalently described by the model [15]:

$$X = \widetilde{G}(\pi)^\top X + W(\pi), \quad W(\pi) \sim \mathcal{N}(0, \Omega(\pi)). \tag{5}$$

Without further assumption, the true underlying DAG, $\widetilde{G}_0 = \widetilde{G}(\pi_0)$, **cannot** be identified from the

$$\text{equivalent class:} \quad \{\widetilde{G}(\pi) : \pi \in \mathbb{S}_p\} \tag{6}$$

based on the distribution of X, even if infinitely many samples are observed.

3 Joint estimation of multiple DAGs

In the multi-task setting, we consider K linear SEMs:

$$X^{(k)} = \widetilde{G}_0^{(k)\top} X^{(k)} + W^{(k)}, \quad \text{for } k = 1, \cdots, K.$$

The superscript notation $^{(k)}$ indicates the k-th model. As mentioned in Sec 2, each model defines a random variable $X^{(k)} \sim \mathcal{N}(0, \Sigma^{(k)})$ with $\Sigma^{(k)} = (1 - \widetilde{G}_0^{(k)})^{-\top}\Omega_0^{(k)}(1 - \widetilde{G}_0^{(k)})^{-1}$.

3.1 Assumption

Similarity. In Condition 3.1 and Definition 3.1, we make a few assumptions about the similarity of these models. Condition 3.1 ensures the causal orders in the K DAGs do not have conflicts. Definition 3.1 measures the similarity of the sparsity patterns in the K DAGs. If the sparsity patterns are strongly overlapped, then the size of the support union s will be small. We do not enforce a strict constraint on the support union, but we will see in the later theorem that a smaller s can lead to a better recovery performance.

Condition 3.1 (Consistent causal orders). There exists a nonempty set of permutations $\Pi_0 \subseteq \mathbb{S}_p$ such that $\forall \pi_0 \in \Pi_0$, it holds for all $k \in [K]$ that $\widetilde{G}_{0\,ij}^{(k)} \neq 0 \Rightarrow \pi_0(i) < \pi_0(j)$.

Definition 3.1 (Support union). Recall $\widetilde{G}(\pi)$ in Eq. 6. The support union of the K DAGs associated with permutation π is denoted by $S(\pi) := \left\{(i,j) : \exists k \in [K]\ s.t.\ \widetilde{G}_{ij}^{(k)}(\pi) \neq 0\right\}$. The support union of the K true DAGs $\widetilde{G}_0^{(k)}$ is $S_0 := S(\pi_0)$. We further denote $s_0 := |S_0|$ and $s := \sup_{\pi \in \mathbb{S}_p} |S(\pi)|$.

Identifiability. To ensure the consistency of the estimator based on least squared loss, we first assume that the DAGs to be recovered are minimum-trace DAGs.

Condition 3.2 (Minimum-trace). Recall the equivalent class defined in Eq. 4 to Eq. 4. Assume for all $k = 1, \cdots, K$, $\text{trace}(\Omega_0^{(k)}) = \min_{\pi \in \mathbb{S}_p} \text{trace}(\Omega(\pi)^{(k)})$.

However, the minimum-trace DAG may not be unique without further assumptions, making the true DAG indistinguishable from other minimum-trace DAGs. Therefore, we consider the equal variance condition in Condition 3.3 which ensures the uniqueness of the minimum-trace DAG. In this paper, we assume the first $K' \leq K$ models satisfy this condition, so they are identifiable. We do not make such an assumption on the other $K - K'$ models, so the $K - K'$ models may not be identifiable.

Condition 3.3 (Equal variance). For all $k = 1, \cdots, K'$ with $K' \leq K$, the noise $W^{(k)} \sim \mathcal{N}(0, \Omega_0^{(k)})$ has equal variance with $\Omega_0^{(k)} = \sigma_0^{(k)} I_p$.

3.2 l_1/l_2-penalized joint estimator

Denote the sample matrix by $\boldsymbol{X}^{(k)}$ whose row vectors are n i.i.d. samples from $\mathcal{N}(0, \Sigma^{(k)})$. Based on the task similarity assumptions, we propose the following l_1/l_2-penalized joint maximum likelihood estimator (MLE) for jointly estimating the connection strength matrices $\{\widetilde{G}_0^{(k)}\}_{k=1}^K$:

$$\hat{\pi}, \{\widehat{G}^{(k)}\} = \underset{\pi \in \mathbb{S}_p,\, \left\{G^{(k)} \in \mathbb{D}(\pi)\right\}}{\arg\min} \sum_{k=1}^{K} \frac{1}{2n} \|\boldsymbol{X}^{(k)} - \boldsymbol{X}^{(k)} G^{(k)}(\pi)\|_F^2 + \lambda \|G^{(1:K)}(\pi)\|_{l_1/l_2}. \tag{7}$$

Similar to the notation in Eq. 4, $G^{(k)}(\pi)$ indicates its consistency with π. $\mathbb{D}(\pi)$ denote the space of all DAGs that are consistent with π. It is notable that a single π shared across all K tasks is optimized in Eq. 7, which respects Condition 3.1. The group norm over the set of K matrices is defined as

$$\|G^{(1:K)}(\pi)\|_{l_1/l_2} := \sum_{i=1}^{p}\sum_{j=1}^{p}\|G_{ij}^{(1:K)}\|_2, \quad \text{where } G_{ij}^{(1:K)} := [G_{ij}^{(1)}(\pi), \cdots, G_{ij}^{(K)}(\pi)].$$

It will penalize the size of union support in a soft way. When $K = 1$, this joint estimator will be reduced to the l_1-penalized maximum likelihood estimation.

Remark 3.1. The optimization in Eq. 7 is used for analysis only. A continuous program with the same optimizer will be discussed in Section 4 for practical implementation.

3.3 Main result: causal order recovery

We start with a few pieces of notations and definitions. Then the theorem statement follows.

Definition 3.2. Let $\widetilde{g}_j^{(k)}(\pi)$ denote the j-th column of $\widetilde{G}^{(k)}(\pi)$. Let

$$d := \sup_{j\in[p],\pi\in\mathbb{S}_p}\left|\cup_{k\in[K]}\operatorname{supp}(\widetilde{g}_j^{(k)}(\pi))\right|, \quad g_{\max} := \sup_{\pi\in\mathbb{S}_p,(i,j)\in S(\pi)}\|\widetilde{G}_{ij}^{(1:K)}(\pi)\|_2/\sqrt{K}.$$

In Definition 3.2, d is the maximal number of parents (in union) in the DAGs $\widetilde{G}^{(k)}(\pi)$, which is also a measure of the sparsity level. $g_{\max}$ is bounded by the maximal entry value in the matrices $\widetilde{G}^{(k)}(\pi)$.

Condition 3.4 (Bounded spectrum). Assume for all $k = 1, \cdots, K$, the covariance matrix $\Sigma^{(k)}$ is positive definite. There exists constants $0 < \Lambda_{\min} \le \Lambda_{\max} < \infty$ such that for all $k = 1, \cdots, K$,

(a) all eigenvalues of $\Sigma^{(k)}$ are upper bounded by $\Lambda_{\max}$;
(b) all eigenvalues of $\Sigma^{(k)}$ are lower bounded by $\Lambda_{\min}$.

Condition 3.5 (Omega-min). There exists a constant $\eta_w > 0$ such that for any permutations $\pi \notin \Pi_0$,

$$\frac{1}{pK'}\sum_{k=1}^{K'}\sum_{j=1}^{p}(\sigma_j^{(k)}(\pi)^2 - \sigma_0^{(k)2})^2 > \frac{1}{\eta_w}.$$

Condition 3.5 with $K' = K = 1$ is called 'omega-min' condition in [14], so we follows this terminology. In some sense, when η_w is larger, $\sigma_j(\pi)$ with $\pi \notin \Pi_0$ is allowed to deviate less from the true variance $\sigma_j(\pi_0) = \sigma_0$ with $\pi_0 \in \Pi_0$, which will make it more difficult to separate the set Π_0 from its complement in a finite sample scenario. Ideally, we should allow η_w to be large, so that the recovery is not only restricted to easy problems.

Now we are ready to present the recovery guarantee for causal order. Theorem 3.1 is a specific statement when the regularization parameter λ follows the classic choice in (group) Lasso problems. A more general statement which allows other λ is given in Appendix B along with the proof.

Theorem 3.1 (Causal order recovery)**.** *Suppose we solve the joint optimization in Eq. 7 with specified regularization parameter $\lambda = \sqrt{\frac{p\log p}{n}}$ for a set of K problems that satisfy Condition 3.1, 3.4 (a), 3.2, 3.3 and 3.5. If the following conditions are satisfied:*

$$\theta(n, K, K', p, s) := \frac{p}{s}\sqrt{\frac{n}{p\log p}\frac{K'^2}{K}} > \kappa_1\eta_w, \tag{8}$$

$$n \ge c_1\log K + c_2(d+1)\log p, \tag{9}$$

$$K \le \kappa_2 p\log p, \tag{10}$$

then the following statements hold true:

(a) With probability at least $1 - c_3\exp(-\kappa_4(d+1)\log p) - \exp(-c_4 n)$, it holds that

$$\hat{\pi} \in \Pi_0.$$

(b) If in addition, n satisfies $n \ge \kappa_5\hat{d}(\log K + \log p)$ with $\hat{d} := \max_{j\in[p],k\in[K]}\|\widehat{g}_j^{(k)} - \widetilde{g}_j^{(k)}\|_0$, and Condition 3.4 (b) holds, then with probability at least $1 - c_3\exp(-\kappa_4(d+1)\log p) - \exp(-c_4 n) - \exp(-\log p - \log K)$,

$$\frac{1}{K}\sum_{k=1}^{K}\|\widehat{G}^{(k)} - \widetilde{G}_0^{(k)}\|_F^2 = \mathcal{O}\left(s_0\sqrt{\frac{p\log p}{nK}}\right).$$

In this statement, c_1, c_2, c_3, c_4 are universal constants (i.e., independent of $n, p, K, s, \Sigma^{(k)}$), $\kappa_1, \kappa_2, \kappa_3, \kappa_4$ are constants depending on $g_{\max}$ and $\Lambda_{\max}$, and κ_5 is a constant depending on $\Lambda_{\min}$.

Discussion on Eq. 8-Eq. 10. (i) In Eq. 8, Theorem 3.1 identifies a *sample complexity parameter* $\theta(n, K, K', p, s)$. Following the terminology in [18], our use of the term "sample complexity" for θ reflects the dominant role it plays in our analysis as the rate at which the sample size much grow in order to obtain a consistent causal order. More precisely, for scalings (n, K, K', p, s) such that $\theta(n, K, K', p, s)$ exceeds a fixed critical threshold $\kappa_1 \eta_w$, we show that the causal order can be correctly recovered with high probability.

(ii) The additional condition for the sample size n in Eq. 9 is the sample requirement for an ordinary linear regression problem. It is in general much weaker than Eq. 8, unless K grows very large.

(iii) The last condition in Eq. 10 on K could be relaxed if a tighter analysis on the distribution properties of Chi-squared distribution is available. However, it is notable that this restriction on the size of K has already been weaker than many related works on multi-task ℓ_1 sparse recovery, which either implicitly treat K as a constant [18, 19] or assume $K = o(\log p)$ [20, 21].

Recovering non-identifiable DAGs. A direct consequence of Theorem 3.1 is that as long as the number of identifiable DAGs K' is non-zero, the joint estimator can recover the causal order of non-identifiable DAGs with high probability. This is not achievable in separate estimation even with infinitely many samples. Therefore, we show that how the information of identifiable DAGs helps recover the non-identifiable ones.

Effective sample size. As indicated by $\theta(n, K, K', p, s)$ in Eq. 8, the effective sample size for recovering the correct causal order is $\frac{nK'^2}{K}$ if the support union size s is of order $\mathcal{O}(1)$ in K. To show the improvement in sample complexity, it is more fair to consider the scenario when the DAGs are identifiable, i.e., $K' = K$. In this case, it is clear that the parameter $\theta(n, K, K, p, s)$ indicates a lower sample complexity relative to separate estimation as long as s is of order $o(\sqrt{K})$.

Separate estimation. Consider the special case of a single task estimation with $K = K' = 1$, in which the joint estimator reduces to ℓ_1-penalized MLE. We discuss how our result can recover previously known results for single DAG recovery. Unfortunately, existing analyses were conducted under different frameworks with different conditions. [14] and [22] are the most comparable ones since they were based on the same omega-min condition (i.e., Condition 3.5), but they chose a smaller regularization parameter λ. In our proof, the sample complexity parameter is derived from $\theta(n, K, K', p, s) = pK'/\left(s\lambda\sqrt{K}\right)$ and Eq. 8 is $pK'/\left(s\lambda\sqrt{K}\right) > \kappa_1 \eta_w$. When $K = K' = 1$, this condition matches what is identified in [14] and [22] for recovering the order of a single DAG.

Error of $\widehat{G}^{(k)}$. To compare the estimation of $\widehat{G}^{(k)}$ to the true DAG $\widetilde{G}_0^{(k)}$, Theorem 3.1 (b) says the averaged error in F-norm goes to zero when $nK \to \infty$. It decreases in K as long as $s = o(K)$.

To summarize, Theorem 3.1 analyzes the causal order consistency for the joint estimator in Eq. 7. Order recovery is the most challenging component in DAG estimation. After $\hat{\pi} \in \Pi_0$ has been identified, the DAG estimation becomes p linear regression problems that can be solved separately. Theorem 3.1 (b) only shows the estimation error of connection matrices in F-norm. To characterize the structure error, additional conditions are required.

3.4 Support union recovery

Theorem 3.1 has shown that $\hat{\pi} = \pi_0 \in \Pi_0$ holds with high probability. Consequently, the support recovery analysis in this section is conditioned on this event. In fact, given the true order π_0, what remains is a set of p separate l_1/l_2-penalized group Lasso problems, in each of which the order π_0 plays a role of constraining the support set by the set $S_j(\pi_0) := \{i : \pi_0(i) < j\}$. However, we need to solve p such problems simultaneously where p is large. A careful analysis is required, and directly combining existing results will not give a high recovery probability.

In the following, we impose a set of conditions and definitions, which are standard in many l_1 sparse recovery analyses [18, 19], after which theorem statement follows. See Appendix C for the proof.

Definition 3.3. The union support of j-th columns of $\{\widetilde{G}_0^{(k)}\}_{k\in[K]}$ is denoted by $\mathrm{RS}_j := \{i \in [p] : \exists k \in [K]\, s.t.\, \widetilde{G}_{0ij}^{(k)} \neq 0\}$. The maximal cardinality is $r_{\max} := \sup_{j\in[p]} \left|\mathrm{RS}_j\right|$.

Definition 3.4. $\rho_u := \sup_{k\in[K], j\in[p], S=\text{RS}_j} \max_{i\in S^c} \left(\Sigma^{(k)}_{S^cS^c|S}\right)_{ii}$ is the maximal diagonal entry of the conditional covariance matrices, where $\Sigma^{(k)}_{S^cS^c|S} := \Sigma^{(k)}_{S^cS^c} - \Sigma^{(k)}_{S^cS}(\Sigma^{(k)}_{SS})^{-1}\Sigma^{(k)}_{SS^c}$.

Definition 3.5. $g_{\min} := \inf_{(i,j)\in S_0} \|\widetilde{G}^{(1:K)}_{0ij}\|_2/\sqrt{K}$ represents the signal strength.

Condition 3.6 (Irrepresentable condition). There exists a fixed parameter $\gamma \in (0, 1]$ such that $\sup_{j\in[p], S=\text{RS}(\widetilde{g}^{(1:K)}_{0j})} \|A(S)\|_\infty \le 1-\gamma$, where $A(S)_{ij} := \sup_{k\in[K]} \left|\left(\Sigma^{(k)}_{S^cS}(\Sigma^{(k)}_{SS})^{-1}\right)_{ij}\right|$.

Theorem 3.2 (Union support recovery)**.** *Assume on the subset of probability space where $\{\hat{\pi} \in \Pi_0\}$ holds, and assume Condition 3.6. Assume the following conditions are satisfied*

$$n \ge \kappa_6 r_{\max} \log p, \tag{11}$$

$$K \le c_7 \log p, \tag{12}$$

$$\sqrt{\frac{8\Lambda_{\max}\log p}{\Lambda_{\min} n}} + \frac{2}{\Lambda_{\min}}\sqrt{\frac{p\log p}{n}\frac{r_{\max}}{K}} = o(g_{\min}), \tag{13}$$

where κ_6 is a constant depending on $\gamma, \Lambda_{\min}, \rho_u, \sigma_{\max}$, and c_7 is some universal constant. Then w.p. at least $1 - r_{\max}\exp(-c_5 K\log p) - \exp(-c_6\log p) - c_8 K\exp(-c_9(n - r_{\max} - \log p))$, the support union of $\widehat{G}^{(1:K)}$ is the same as that of $\widetilde{G}^{(1:K)}_0$, and that $\left\|\widehat{G}^{(1:K)} - \widetilde{G}^{(1:K)}\right\|_{l_\infty/l_2}/\sqrt{K} = o(g_{\min})$.

Discussion on Eq. 11 and Eq. 12. *(i)* Eq. 11 poses a sample size condition. The value $r_{\max}$ is the sparsity overlap defined in Definition 3.3 (i). It takes value in the interval $[d, \min\{s_0, p, dK\}]$, depending on the similarity in sparsity pattern.

(ii) The restriction on K in Eq. 12 plays a similar role as Eq. 10 in Theorem 3.1. This a stronger restriction, but also guarantees the stronger result of support recovery. Existing analyses on l_1/l_2-penalized group Lasso were not able to relax this constraint, neither, so some of them treated K as a constant in the analysis [18, 19]. Recall that in Theorem 3.1, we were able to allow $K = \mathcal{O}(p\log p)$. Technically, this was achieved because in the proof of Theorem 3.1, we avoid analyzing the general recovery of group Lasso, but only its null-consistency (i.e., the special case of true structures having zero support), where tighter bound can be derived and it is sufficient for order recovery.

Benefit of joint estimation. Eq. 13 plays a similar role as Eq. 8 in Theorem 3.1. It specifies a rate at which the sample size must grow for successful union support recovery. As long as $r_{\max}$ is of order $o(\sqrt{K})$, K will effectively reduce the second term in Eq. 8. Apart from that, the recover probability specified in Theorem 3.2 grows in K.

4 Algorithm

Solving the optimization in Eq. 7 by searching over all permutations $\pi \in \mathbb{S}_p$ is intractable due to the large combinatorial space. Inspired by the smooth characterization of acyclic graph [16], we propose a continuous optimization problem, whose optimizer is the same as the estimator in Eq. 7. Furthermore, we will design an efficient iterative algorithm to approximate the solution.

4.1 Continuous program

We convert Eq. 7 to the following constrained continuous program

$$\min_{\substack{T\in\mathbb{R}^{p\times p}\\ G^{(1)},\cdots,G^{(K)}\in\mathbb{R}^{p\times p}}} \sum_{k=1}^{K}\frac{1}{2n}\left\|\boldsymbol{X}^{(k)} - \boldsymbol{X}^{(k)}\overline{G}^{(k)}\right\|_F^2 + \lambda\|\overline{G}^{(1:K)}\|_{l_1/l_2} + \rho\|\mathbf{1}_{p\times p} - T\|_F^2 \tag{14}$$

$$\text{subject to} \quad h(T) := \text{trace}(e^{T\circ T}) - p = 0, \tag{15}$$

where $\overline{G}^{(k)} := G^{(k)}\circ T$ is element-wise multiplication between $G^{(k)}$ and T, and $\mathbf{1}_{p\times p}$ is a $p\times p$ matrix with entries equal to one. Eq. 15 is a smooth 'DAG-ness' constraint proposed by NOTEARS [16], which ensures T is acyclic. One can also use $h(T) := \text{trace}((I + T\circ T)^p) - p$ proposed in [23].

We would like to highlight the novel and interesting design of the matrix T in Eq. 14. What makes Eq. 7 difficult to solve is the requirement that $\{G^{(k)}\}$ must be DAGs and share the same order. A straightforward idea is to apply the smooth acyclic constraint to every $G^{(k)}$, but it is not clear how to enforce their consistent topological order. Our formulation realizes this by a single matrix T.

Algorithm 1: Joint Estimation Algorithm

Hyperparameters : $\rho, \alpha, \lambda, t, \delta$
Initialize $G^{(1:K)}, T$ randomly;
for $itr = 1, \cdots, M$ **do**
  for $itr' = 1, \cdots, M'$ **do**
    $[G^{(1:K)}, T] \leftarrow \texttt{GradOptStep}\big(f; G^{(1:K)}, T, \beta\big)$; ▷ `Gradient-based update on` f
    $\forall i, j \in [p],\ G_{ij}^{(1:K)} \leftarrow \frac{G_{ij}^{(1:K)}}{\|G_{ij}^{(1:K)}\|_2} \max\left\{0, \|G_{ij}^{(1:K)}\|_2 - t\lambda|T_{ij}|\right\}$; ▷ `Proximal step`
    $\forall i, j \in [p],\ T_{ij} \leftarrow \operatorname{sign}(T_{ij}) \max\left\{0, |T_{ij}| - t\lambda\|G_{ij}^{(1:K)}\|_2\right\}$; ▷ `Proximal step`
  $\beta \leftarrow \beta + \tau h(T)$; ▷ `Dual ascent`
  $\alpha \leftarrow \alpha \cdot (1 + \delta)$; ▷ `Typical rule [23]`

To better understand the design rationale of T, recall in Eq. 3 that a matrix G is a DAG of order π *if and only if* its support set is in $\{(i, j) : \pi(i) < \pi(j)\}$. The matrix T plays a role of restricting the support set of $G^{(k)}$ by masking its entries. Two examples are shown above. However, unlike the learning of masks in other papers which allows T to have any combinations of nonzero entries, here we need T to exactly represent the support set for each π. That is T is from a space $\mathcal{T}_p$ with $p!$ elements: $\mathcal{T}_p := \{T \in \{0, 1\}^{p\times p} : T_{ij} = 1 \Leftrightarrow \pi(i) < \pi(j)\}$. Now a key question arises: *How to perform a continuous and differentiable search over $\mathcal{T}_p$?* The following finding motivates our design:

$$\pi = (1, 2, 3, 4) \quad \Updownarrow \quad T = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix} \qquad \pi = (4, 2, 1, 3) \quad \Updownarrow \quad T = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

$$T \in \mathcal{T}_p \Longleftrightarrow T \in \arg\min_{T\in\mathbb{R}^{p\times p}} \left\{\|\mathbf{1}_{p\times p} - T\|_F^2 \text{ subject to } h(T) = 0\right\}.$$

In other words, T is a continuous projection of $\mathbf{1}_{p\times p}$ to the space of DAGs. We can then optimize the mask T in the continuous space $\mathbb{R}^{p\times p}$ but the optimal solution must be an element in the discrete space $\mathcal{T}_p$. This observation also naturally leads to the design of Eq. 14.

Finally, we want to emphasize that it is important for the optimal T to have binary entries. Without this property, any nonzero value c can scale the (G, T) pair to give an equivalent masked DAG, i.e., $G \circ T = (cG) \circ (\frac{1}{c}T)$. This scaling equivalence will make the optimization hard to solve in practice.

Proofs for the above arguments and the equivalence between Eq. 7 and Eq. 14 are in Appendix F.

4.2 Iterative algorithm

We derive an efficient iterative algorithm using the Lagrangian method with quadratic penalty, which converts Eq. 14 to an unconstrained problem:

$$\min_{T, G^{(1)}, \cdots, G^{(K)} \in \mathbb{R}^{p\times p}} \max_{\beta \geq 0} \ \mathcal{L}(G^{(1:K)}, T;\ \beta) := f(G^{(1:K)}, T;\ \beta) + \lambda\|\overline{G}^{(1:K)}\|_{l_1/l_2},$$

$$\text{where } f(G^{(1:K)}, T;\ \beta) := \sum_{k=1}^{K} \frac{1}{2n} \left\|\boldsymbol{X}^{(k)} - \boldsymbol{X}^{(k)}\overline{G}^{(k)}\right\|_F^2 + \rho\|\mathbf{1}_{p\times p} - T\|_F^2 + \beta h(T) + \alpha h(T)^2,$$

β is dual variable, α is the coefficient for quadratic penalty, and f is the smooth term in the objective.

We can solve this min-max problem by alternating primal updates on $(G^{(1:K)}, T)$ and dual updates on β. Due to the non-smoothness of group norm, the primal update is based on proximal-gradient method, where the **proximal-operator** with respect to $\|\cdot\|_{l_1/l_2}$ has a closed form:

$$\left[\arg\min_{Z^{(1:K)}\in\mathbb{R}^{K\times p\times p}} \tfrac{1}{2}\textstyle\sum_{k=1}^{K} \|Z^{(k)} - \boldsymbol{X}^{(k)}\|_F^2 + c\|Z^{(1:K)}\|_{l_1/l_2}\right]_{ij}^{(1:K)}$$
$$= \frac{\boldsymbol{X}_{ij}^{(1:K)}}{\|\boldsymbol{X}_{ij}^{(1:K)}\|_2} \max\{0, \|\boldsymbol{X}_{ij}^{(1:K)}\|_2 - c\},$$

which is a group-wise soft-threshold. Since $G^{(k)}$ and T are multiplied together element-wisely inside the group norm, the proximal operator will be applied to both of them separately. Together with the

dual update for β, $\beta \leftarrow \beta + \tau h(T)$ with τ as the step size, we summarize the overall algorithm for solving Eq. 14 in Algorithm 1.

5 Related work

Single DAG estimation. Unlike the large literature of research on undirected graphical models [24, 25, 26, 27], statistical guarantees for score-based DAG estimator have been available only in recent years. [14, 15, 17] have shown the DAG estimation consistency in high-dimensions, but they do not consider joint estimation. Nevertheless, some techniques in [14, 15] are useful for our derivations.

Multi-task learning. *(i) Undirected graph estimation.* There have been extensive studies on the joint estimation of multiple undirected Gaussian graphical models [28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39]. *(ii) DAG estimation.* In contrast, not much theoretical work has been done for joint DAG learning. A few pieces of recent works addressed certain statistical aspects of multi-task DAG estimation [9, 20, 21], but [20, 21] tackle the fewer task regime with $K = o(\log p)$, and [9] assumes the causal order is given. Another related work that we notice after the paper submission is [40], which also assumes the DAGs have a consistent causal order, but it focuses on estimating the difference between two DAGs. *(iii) Linear regression.* Multi-task linear regression is also a related topic [41, 42, 43, 44, 45, 46, 47], because the techniques for analyzing group Lasso are used in our analysis [18, 19].

Practical Algorithms. Works on practical algorithm design for efficiently solving the score-based optimization are actively conducted [48, 49, 50, 51]. Our algorithm is most related to recent methods exploiting a smooth characterization of acyclicity, including NOTEARS [16] and several subsequent works [23, 52, 53, 54], but they only apply for single-task DAG estimation. Although algorithms for the multi-task counterpart were proposed a decade ago [11, 55, 56, 57], none of them leverage recent advances in characterizing DAGs and providing theoretically guarantees.

6 Experiments

6.1 Synthetic data

The set of experiments is designed to reveal the effective sample size predicted by Theorem 3.1, and demonstrate the effectiveness of the proposed algorithm. In the simulations, we randomly sample a causal order π and a union support set S_0. Then we randomly generate multiple DAGs that follow the order π and have edges contained in the set S_0. For each DAG, we construct a linear SEM with standard Gaussian noise, and sample n data points from it. On tasks with different combinations of (p, n, s, K), we exam the behavior of the joint estimator, estimated by Algorithm 1, on 64 tasks and report the statistics in the following for evaluation. In this experiment, we take $K' = K$ so that all the DAGs are identifiable. This simpler case will make it easier to verify the proposed algorithm and the rates in the theorem.

Success probability for order recovery. For each fixed tuple (p, n, s, K), we measure the sample complexity in terms of the parameter θ specified by Theorem 3.1. Fig 1 plots the success probability $\Pr[\hat{\pi} \in \Pi_0]$, versus $\theta = p/s\sqrt{nK/(p \log p)}$ at a logarithmic scale. Theorem 3.1 predicts that the success probability should transition to 1 once θ exceeds a critical threshold. Curves in Fig 1 actually have sharp transitions, showing step-function behavior. The sharpness is moderated by the logarithmic scale in x-axis. Moreover, by scaling the sample size n using θ, the curves align well as predicted by the theory and have a similar transition point, even though they correspond to very different model dimensions p.

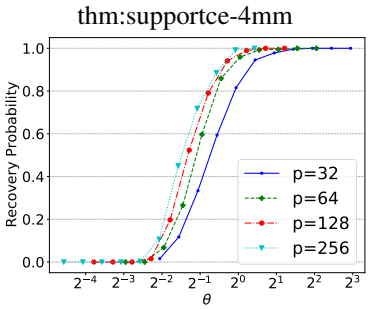

Figure 1: Success probability vs $\log \theta$.

Fig 2 shows the success probability in the form of Heat Maps, where the rows indicate an increase in per task sample size n, and the columns indicate an increase in the number of tasks K. The results show that the increases in these two quantities have similar effect to the success probability.

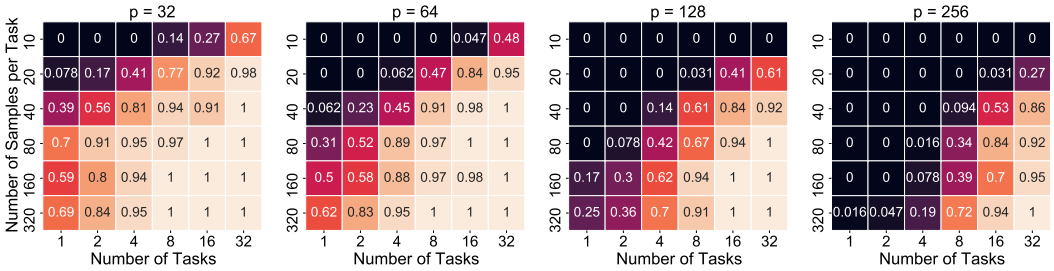

Figure 2: Heat map: Darker colors indicate lower success probability, and lighter colors are higher.

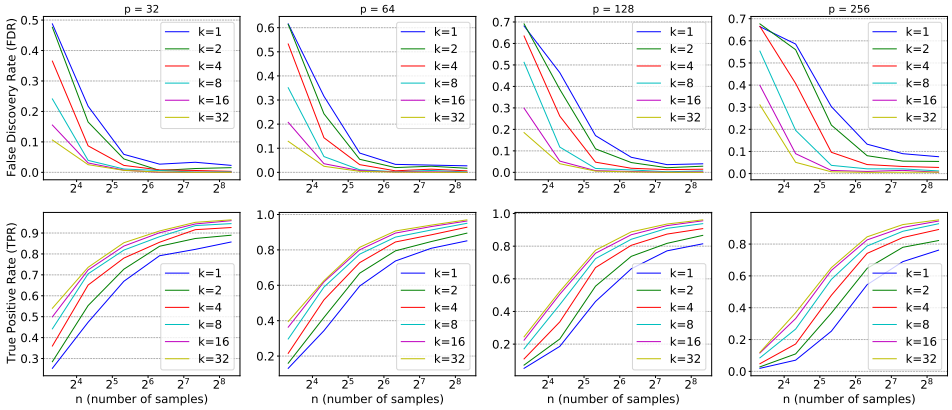

Figure 3: False discovery rate (FDR) and true positive rate (TPR) of the edges.

Effectiveness of K in structure recovery. In this experiment, we aim at verifying the effectiveness of the joint estimation algorithm for recovering the structures. For brevity, we only report the numbers for false discovery rate (FDR) and true positive rate (TPR) of edges in Fig 3, but figures for additional metrics can be found in Appendix G. In Fig 3, we can observe consistent improvements when increasing the number of tasks K. When per task sample size is small, this improvement reveals to be more obvious.

Comparison with other joint estimator. In this experiment, we compare our method MultiDAG with JointGES [21] on models with $p = 32$ and $p = 64$. Results in Table 1 show that two algorithms have similar performance when $K = 1$. However, when K increases, our method returns consistently better structures in terms of SHD.

Table 1: Comparison of MultiDAG (MD) and JointGES (JG) in SHD

	$p = 32$				$p = 64$			
	$n = 10$	$n = 20$	$n = 80$	$n = 320$	$n = 10$	$n = 20$	$n = 80$	$n = 320$
MD(k=1)	39 ± 5	25 ± 5	10 ± 4	6 ± 3	104 ± 7	77 ± 8	29 ± 6	19 ± 8
MD(k=2)	37 ± 5	22 ± 5	8 ± 3	4 ± 3	103 ± 7	68 ± 8	22 ± 6	13 ± 6
MD(k=8)	29 ± 5	13 ± 3	4 ± 2	2 ± 1	85 ± 7	42 ± 6	13 ± 3	6 ± 3
MD(k=32)	23 ± 4	11 ± 3	3 ± 2	1 ± 1	66 ± 5	35 ± 5	10 ± 3	3 ± 2
JG(k=1)	31 ± 5	19 ± 5	8 ± 4	6 ± 5	100 ± 11	53 ± 11	18 ± 11	18 ± 9
JG(k=2)	32 ± 4	19 ± 5	9 ± 5	7 ± 5	99 ± 10	51 ± 10	20 ± 9	21 ± 10
JG(k=8)	30 ± 5	19 ± 5	12 ± 4	10 ± 4	82 ± 10	42 ± 10	20 ± 5	27 ± 9
JG(k=32)	26 ± 4	18 ± 4	12 ± 3	9 ± 3	57 ± 9	36 ± 6	19 ± 5	26 ± 6

6.2 Recovery of gene regulatory network

We investigate how our joint estimator works on more realistic models, by conducting a set of experiments on realistic gene expression data generated by SERGIO [13], which models the additive

effect of cooperative transcription factor binding across multiple gene regulators in parallel with protein degradation and noisy expression.

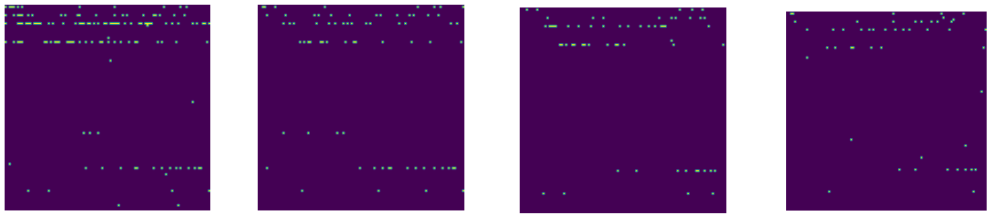

Figure 4: Visualization of the recovered DAG structures. Each light green colored pixel at position (i, j) indicates an edge from node i to j.

We conduct this experiment on the E. coli 100 network, which includes 100 known genes and 137 known regulatory interactions. To evaluate our algorithm, we generate multiple networks by rearranging and re-weighting 5 edges at random in this network without violating the topological order. We simulate the gene expression from each network using SERGIO with a universal non-cooperative hill coefficient of 0.05, which works well with our linear recovery algorithm.

Fig 4 provides a visual comparison of the recovered structures. It can be seem from the true network that there are a few key transcription factors that highlight several rows in the figure. These transcription factors are better identified by the two structures in (b) and (c), but not that clear in (d). Combining this observation with the more quantitative results in Table 2, we see that the combination $(K = 10, n = 100)$ achieves comparable performance to $K = 1$ with the same total number of samples, and outperforms the single task estimation with $n = 100$.

K	n	FDR	TPR	FPR	SHD
1	100	$0.18 \pm 3.8\text{e}{-3}$	$0.30 \pm 1.2\text{e}{-3}$	$0.001 \pm 7.3\text{e}{-7}$	$104.61 \pm 3.2\text{e}1$
10	100	$0.07 \pm 1.2\text{e}{-3}$	$0.42 \pm 5.1\text{e}{-4}$	$0.001 \pm 2.8\text{e}{-7}$	$84.0 \pm 1.5\text{e}1$
1	1000	$0.09 \pm 2.9\text{e}{-3}$	$0.50 \pm 1.4\text{e}{-3}$	$0.002 \pm 9.2\text{e}{-7}$	$76.4 \pm 5.9\text{e}1$

Table 2: Recovery across 25 independent initializations of SERGIO for each experiment. FPR and SHD stand for false positive rate and structural hamming distance, respectively.

7 Conclusion and discussion

In this paper, we have analyzed the behavior of l_1/l_2-penalized joint MLE for multiple DAG estimation tasks. Our main result is to show that its performance in recovering the causal order is governed by the sample complexity parameter $\theta(n, K, K', p, s)$ in Eq. 8. Besides, we have proposed an efficient algorithm for approximating the joint estimator via formulating a novel continuous programming, and demonstrated its effectiveness experimentally. The current work applies to DAGs that have certain similarity in sparsity pattern. It will be interesting to consider whether the joint estimation without the group-norm (and without the union support assumption) can also lead to similar improvement in causal order recovery.

Acknowledgement

Xinshi Chen is supported by the Google PhD Fellowship. This works is done partially during a visit at MBZUAI. We are grateful for the computing resources provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta. We are thankful for PACE Research Scientist Fang (Cherry) Liu's excellent HPC consulting.

References

[1] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620, 2000.

[2] James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

[3] Pedro Aguilera Aguilera, Antonio Fernández, Rosa Fernández, Rafael Rumí, and Antonio Salmerón. Bayesian networks in environmental modelling. *Environmental Modelling & Software*, 26(12):1376–1388, 2011.

[4] A Nicholson, F Cozman, M Velikova, JT Van Scheltinga, PJ Lucas, and M Spaanderman. Applications of bayesian networks exploiting causal functional relationships in bayesian network modelling for personalised healthcare. *International Journal of Approximate Reasoning*, 55(1):59–73, 2014.

[5] Bin Zhang, Chris Gaiteri, Liviu-Gabriel Bodea, Zhi Wang, Joshua McElwee, Alexei A Podtelezhnikov, Chunsheng Zhang, Tao Xie, Linh Tran, Radu Dobrin, et al. Integrated systems approach identifies genetic nodes and networks in late-onset alzheimer's disease. *Cell*, 153(3):707–720, 2013.

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

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

[8] T Tony Cai, Hongzhe Li, Weidong Liu, and Jichun Xie. Joint estimation of multiple high-dimensional precision matrices. *Statistica Sinica*, 26(2):445, 2016.

[9] Jianyu Liu, Wei Sun, and Yufeng Liu. Joint skeleton estimation of multiple directed acyclic graphs for heterogeneous population. *Biometrics*, 75(1):36–47, 2019.

[10] Teppei Shimamura, Seiya Imoto, Rui Yamaguchi, Masao Nagasaki, and Satoru Miyano. Inferring dynamic gene networks under varying conditions for transcriptomic network comparison. *Bioinformatics*, 26(8):1064–1072, 2010.

[11] Shohei Shimizu. Joint estimation of linear non-gaussian acyclic models. *Neurocomputing*, 81:104–107, 2012.

[12] Aiying Zhang, Gemeng Zhang, Vince D Calhoun, and Yu-Ping Wang. Causal brain network in schizophrenia by a two-step bayesian network analysis. In *Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications*, volume 11318, page 1131817. International Society for Optics and Photonics, 2020.

[13] Payam Dibaeinia and Saurabh Sinha. Sergio: a single-cell expression simulator guided by gene regulatory networks. *Cell Systems*, 11(3):252–271, 2020.

[14] Sara Van de Geer, Peter Bühlmann, et al. ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *Annals of Statistics*, 41(2):536–567, 2013.

[15] Bryon Aragam, Arash A. Amini, and Qing Zhou. Learning directed acyclic graphs with penalized neighbourhood regression, 2017.

[16] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31:9472–9483, 2018.

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

[18] Guillaume Obozinski, Martin J Wainwright, Michael I Jordan, et al. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39(1):1–47, 2011.

[19] Weiguang Wang, Yingbin Liang, and Eric P Xing. Collective support recovery for multi-design multi-response linear regression. *IEEE Transactions on Information Theory*, 61(1):513–534, 2014.

[20] Yuhao Wang, Santiago Segarra, Caroline Uhler, et al. High-dimensional joint estimation of multiple directed gaussian graphical models. *Electronic Journal of Statistics*, 14(1):2439–2483, 2020.

[21] Kyoungjae Lee and Xuan Cao. Bayesian joint inference for multiple directed acyclic graphs. *arXiv preprint arXiv:2008.06190*, 2020.

[22] Magali Champion, Victor Picheny, and Matthieu Vignes. Inferring large graphs using l (1)-penalized likelihood (vol 28, pg 905, 2018). *STATISTICS AND COMPUTING*, 28(6):1231–1231, 2018.

[23] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pages 7154–7163. PMLR, 2019.

[24] DR Cox and Nanny Wermuth. *Multivariate Dependencies: Models, Analysis and Interpretation*, volume 67. CRC Press, 1996.

[25] Nicolai Meinshausen, Peter Bühlmann, et al. High-dimensional graphs and variable selection with the lasso. *Annals of statistics*, 34(3):1436–1462, 2006.

[26] Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[27] Pradeep Ravikumar, Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Model selection in gaussian graphical models: High-dimensional consistency of l1-regularized mle. In *NIPS*, pages 1329–1336, 2008.

[28] Le Song, Mladen Kolar, and Eric P Xing. Keller: estimating time-varying interactions between genes. *Bioinformatics*, 25(12):i128–i136, 2009.

[29] Jean Honorio and Dimitris Samaras. Multi-task learning of gaussian graphical models. In *ICML*, 2010.

[30] Jian Guo, Elizaveta Levina, George Michailidis, and Ji Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.

[31] Tony Cai, Weidong Liu, and Xi Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

[32] Diane Oyen and Terran Lane. Leveraging domain knowledge in multitask bayesian network structure learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, 2012.

[33] Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 76(2):373, 2014.

[34] Mladen Kolar, Le Song, Amr Ahmed, Eric P Xing, et al. Estimating time-varying networks. *Annals of Applied Statistics*, 4(1):94–123, 2010.

[35] Karthik Mohan, Palma London, Maryam Fazel, Daniela Witten, and Su-In Lee. Node-based learning of multiple gaussian graphical models. *The Journal of Machine Learning Research*, 15(1):445–488, 2014.

[36] Christine Peterson, Francesco C Stingo, and Marina Vannucci. Bayesian inference of multiple gaussian graphical models. *Journal of the American Statistical Association*, 110(509):159–174, 2015.

[37] Sen Yang, Zhaosong Lu, Xiaotong Shen, Peter Wonka, and Jieping Ye. Fused multiple graphical lasso. *SIAM Journal on Optimization*, 25(2):916–943, 2015.

[38] André R Gonçalves, Fernando J Von Zuben, and Arindam Banerjee. Multi-task sparse structure learning with gaussian copula models. *The Journal of Machine Learning Research*, 17(1):1205–1234, 2016.

[39] Burak Varici, Saurabh Sihag, and Ali Tajer. Learning shared subgraphs in ising model pairs. In *International Conference on Artificial Intelligence and Statistics*, pages 3952–3960. PMLR, 2021.

[40] Asish Ghoshal, Kevin Bello, and Jean Honorio. Direct learning with guarantees of the difference dag between structural equation models. *arXiv preprint arXiv:1906.12024*, 2019.

[41] Francis R Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9(6), 2008.

[42] Sahand Negahban and Martin J Wainwright. Joint support recovery under high-dimensional scaling: Benefits and perils of ℓ_1/ℓ_∞-regularization. *Advances in Neural Information Processing Systems*, 21:1161–1168, 2008.

[43] Peng Zhao, Guilherme Rocha, Bin Yu, et al. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497, 2009.

[44] Karim Lounici, Massimiliano Pontil, Alexandre B Tsybakov, and Sara Van De Geer. Taking advantage of sparsity in multi-task learning. *arXiv preprint arXiv:0903.1468*, 2009.

[45] Junzhou Huang, Tong Zhang, et al. The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978–2004, 2010.

[46] Jianhui Chen, Jiayu Zhou, and Jieping Ye. Integrating low-rank and group-sparse structures for robust multi-task learning. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 42–50, 2011.

[47] Lu Tian, Pan Xu, and Quanquan Gu. Forward backward greedy algorithms for multi-task learning with faster rates. In *UAI*, 2016.

[48] Mauro Scanagatta, Antonio Salmerón, and Fabio Stella. A survey on bayesian network structure learning from data. *Progress in Artificial Intelligence*, 8(4):425–439, 2019.

[49] Ignavier Ng, Zhuangyan Fang, Shengyu Zhu, Zhitang Chen, and Jun Wang. Masked gradient-based causal structure learning. *arXiv preprint arXiv:1910.08527*, 2019.

[50] Shengyu Zhu, Ignavier Ng, and Zhitang Chen. Causal discovery with reinforcement learning. In *International Conference on Learning Representations*, 2020.

[51] Hasan Manzour, Simge Küçükyavuz, Hao-Hsiang Wu, and Ali Shojaie. Integer programming for learning directed acyclic graphs from continuous data. *Informs Journal on Optimization*, 3(1):46–73, 2021.

[52] Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. In *International Conference on Learning Representations*, 2020.

[53] Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425. PMLR, 2020.

[54] Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pages 1595–1605. PMLR, 2020.

[55] Robert Tillman and Peter Spirtes. Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 3–15. JMLR Workshop and Conference Proceedings, 2011.

[56] Robert E Tillman. Structure learning with independent non-identically distributed data. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1041–1048, 2009.

[57] Lele Xu, Tingting Fan, Xia Wu, KeWei Chen, Xiaojuan Guo, Jiacai Zhang, and Li Yao. A pooling-lingam algorithm for effective connectivity analysis of fmri data. *Frontiers in computational neuroscience*, 8:125, 2014.

A List of definitions and notations

For the convenience of the reader, we summarize a list of notations blow.

1. $\widehat{G}_j(\hat{\pi}) := [\widehat{g}_j^{(1)}(\hat{\pi}), \cdots, \widehat{g}_j^{(K)}(\hat{\pi})]$ and $\widetilde{G}_j(\hat{\pi}) := [\widetilde{g}_j^{(1)}(\hat{\pi}), \cdots, \widetilde{g}_j^{(K)}(\hat{\pi})]$.
2. For all $k \in [K]$, the eigenvalues of $\Sigma^{(k)}$ are in $[\Lambda_{\min}, \Lambda_{\max}]$, for some constants $0 < \Lambda_{\min} \leq \Lambda_{\max} < \infty$.
3. $\sigma_{\max} = \sup_{k\in[K], j\in[p], \pi\in\mathbb{S}_p} |\sigma_j^{(k)}(\pi)|$. Note that $\sigma_{\max}^2 \leq \Lambda_{\max}$.
4. $c_{\max} := \sup_{k,j,\pi} |1 - \sigma_j^{(k)}(\pi)^2|$. Note that $c_{\max} \leq 1 + \sigma_{\max}^2 \leq 1 + \Lambda_{\max}$.
5. $\rho = \sup_{k\in[K], j\in[p]} \Sigma_{jj}^{(k)}$. Note that $\rho \leq \Lambda_{\max}$.
6. $s(\pi) := |S(\pi)|$ where $S(\pi) := \cup_{k\in[K]} \text{supp}(\widetilde{G}^{(k)}(\pi))$. $s_0 := s(\pi_0)$. $s := \sup_{\pi\in\mathbb{S}_p} s(\pi)$.
7. $g_{\max} := \sup_{\pi\in\mathbb{S}_p, (i,j)\in S(\pi)} \left\|\widetilde{G}_{ij}^{(1:K)}(\pi)\right\|_2 / \sqrt{K}$.
8. $g_{\min} := \inf_{(i,j)\in S(\pi_0)} \left\|\widetilde{G}_{0ij}^{(1:K)}\right\|_2 / \sqrt{K}$.
9. $\text{RS}_j := \{i \in [p] : \exists k \in [K]\, s.t.\, \widetilde{G}_{0ij}^{(k)} \neq 0\}$.
10. $r_{\max} := \sup_{j\in[p]} \left|\text{RS}_j\right|$.
11. $D_{\max} := \sup_{j\in[p], S=\text{RS}_j, k\in[K]} \left\|(\Sigma_{SS}^{(k)})^{-1}\right\|_\infty$.
12. $\rho_u := \sup_{j\in[p], S=\text{RS}_j, k\in[K]} \max_{i\in S^c} \left(\Sigma_{S^cS^c|S}^{(k)}\right)_{ii}$.
13. $S_j(\pi) := \{i : \pi(i) < \pi(j)\}$.
14. $U_j(\pi) := \cup_{k\in[K]} \text{supp}\big(\widetilde{g}_j^{(k)}(\pi)\big) = \{i : \exists k \in [K]\, s.t.\, \widetilde{G}_{ij}^{(k)}(\pi) \neq 0\}$.
15. $d_j := \sup_{\pi\in\mathbb{S}_p} |U_j(\pi)|$. $d = \sup_{j\in[p]} d_j$.

B Details of Theorem 3.1: causal order recovery

In Appendix B.1, we present a general statement of Theorem 3.1 (a) along with its proof. Proof of part (b) in Theorem 3.1 is given in Appendix B.3.

B.1 Order recovery: proof of Theorem 3.1 (a) (Theorem B.1)

Theorem 3.1 (a) states the order recovery guarantee for a specified parameter $\lambda = \sqrt{\frac{p\log p}{n}}$. In the following, we will present a more general statement of Theorem 3.1 (a) that does not specify the choice of λ, after which we will present the proof.

Theorem B.1 (General statement of Theorem 3.1 (a))**.** *For any $\delta_1, \delta_2, \delta_3, \delta_4, \delta_5 \in (0, 1)$, if the following conditions are satisfied*

$$
\begin{aligned}
K &\leq \frac{\delta_2^2\delta_3^2\delta_4^2\lambda^2 n}{64\rho\sigma_{\max}^2}, \\
n &\geq \left(\left(4(1-\delta_3)\delta_3^{-2} - \delta_5\right)\right)^{-1} \left(\log K + (d+1)\log p\right), \\
\frac{1}{\eta_w} &> \left(\frac{16\sigma_{\max}^8}{\delta_1(1-\delta_1)} + \frac{4\sigma_{\max}^4 c_{\max}}{1-\delta_1}\right)\frac{2\log p}{nK'} + \frac{4\sigma_{\max}^4 g_{\max}}{\delta_1}\frac{\lambda\left(s(\pi_0) + \delta_2 s(\hat{\pi})\right)}{p}\sqrt{\frac{K}{K'^2}} \\
&\quad + \frac{8\sigma_{\max}^6}{\delta_1}\sqrt{\frac{2\log p}{n}\frac{K-K'}{K'^2}},
\end{aligned}
$$

then $\hat{\pi} = \pi_0$ with probability at least

$$
1 - \exp\left(-t^*(1-\delta_4) + (d+2)\log p\right) - \exp\left(-\delta_5 n\right) - 2\exp\left(-p\log p\right),
$$

where $t^* := \frac{\delta_2^2\delta_3^2\lambda^2 n}{16\rho\sigma_{\max}^2}$.

Proof outline. By optimality of the joint estimator, (for simplicity, we write $\widehat{G}^{(k)} := \widehat{G}^{(k)}(\hat{\pi})$)

$$
\begin{aligned}
&\sum_{k=1}^{K} \frac{1}{2n} \|\boldsymbol{X}^{(k)}\widehat{G}^{(k)} - \boldsymbol{X}^{(k)}\widetilde{G}^{(k)}(\hat{\pi})\|_F^2 + \lambda \|\widehat{G}^{(1:K)}\|_{l_1/l_2} \\
\leq &\underbrace{\sum_{k=1}^{K} \frac{1}{2n} \left(\|\boldsymbol{X}^{(k)} - \boldsymbol{X}^{(k)}\widetilde{G}_0^{(k)}\|_F^2 - \|\boldsymbol{X}^{(k)} - \boldsymbol{X}^{(k)}\widetilde{G}^{(k)}(\hat{\pi})\|_F^2 \right)}_{(I)} \\
&+ \underbrace{\sum_{k=1}^{K} \frac{1}{n} \langle \boldsymbol{X}^{(k)} - \boldsymbol{X}^{(k)}\widetilde{G}^{(k)}(\hat{\pi}), \boldsymbol{X}^{(k)}\widehat{G}^{(k)} - \boldsymbol{X}^{(k)}\widetilde{G}^{(k)}(\hat{\pi}) \rangle_F}_{(II)} + \lambda \|\widetilde{G}_0^{(1:K)}\|_{l_1/l_2}.
\end{aligned}
$$

The proof is based on a bound for the term (I) and a bound for the term (II).

To bound (I), we show that the empirical variances of the error terms are close to their expectations, which is achieved mainly by a concentration bound on a linear combination of Chi-squared random variables.

To bound (II), we show the following inequality holds true for all $j \in [p]$ and $\hat{\pi} \in \mathbb{S}_p$ with high probability (where $\widetilde{\varepsilon}_j^{(k)}(\hat{\pi})$ is the empirical error):

$$
\sup_{\{\beta^{(k)} \in \mathbb{R}^m\}} \frac{1}{n} \sum_{k=1}^{K} \langle \widetilde{\varepsilon}_j^{(k)}(\hat{\pi}), \boldsymbol{X}_{S_j}^{(k)} \beta^{(k)} \rangle - \frac{\delta}{2n} \sum_{k=1}^{K} \|\boldsymbol{X}_{S_j}^{(k)} \beta^{(k)}\|_2^2 - \delta\lambda \|[\beta^{(1)}, \cdots, \beta^{(K)}]\|_{l_1/l_2} \leq 0.
$$

We highlight two technical aspects in bounding (II):

- For each fixed j and $\hat{\pi}$, the above inequality is proved by showing the **null-consistency** of l_1/l_2-penalized group Lasso problem (see Appendix B.2.3). Null-consistency means successfully recovering the true linear regression model when the true parameters have null support (all parameters are zeros). Technically, the improvement in sample complexity for recovering multiple DAGs partially comes from the benefit of a larger K for guaranteeing the null-consistency.

- We need to insure the bounds hold uniformly over all permutations $\hat{\pi} \in \mathbb{S}_p$ and $j \in [p]$. To avoid using a naive union bound over $p!$ many permutations, we leverage the sparsity of the graph structures and prove that the number of elements in the set $\{\widetilde{G}^{(1:K)}(\pi) : \pi \in \mathbb{S}_p\}$ can be fewer than $p!$ (see Appendix E), so that we can take a uniform control over this smaller set instead.

We summarize the bounds for (I) and (II) in Lemma B.1 and Lemma B.2, which can be found in Appendix B.2.1 and Appendix B.2.2.

Detailed proof of Theorem B.1. Collecting the results in Lemma B.1 and Lemma B.2 and reorganizing the terms in the inequalities, we have the following conclusion.

For any $\delta_1, \delta_2, \delta_3, \delta_4, \delta_5 \in (0, 1)$ and $t^* := \frac{\delta_2^2 \delta_3^2 \lambda^2 n}{16 \rho \sigma_{\max}^2}$, if the following conditions are satisfied:

$$
\begin{aligned}
K &\leq \frac{\delta_4^2}{4} t^* = \frac{\delta_2^2 \delta_3^2 \delta_4^2 \lambda^2 n}{64 \rho \sigma_{\max}^2} \\
\left(4(1-\delta_3)\delta_3^{-2} - \delta_5\right) n &\geq \log K + (d+1) \log p,
\end{aligned}
$$

then with probability at least $1-\exp\left(-t^*(1-\delta_4)+(d+2)\log p\right)-\exp\left(-\delta_5 n\right)-2\exp\left(-p\log p\right)$, it holds for all $\hat{\pi}\in\mathbb{S}_p$ that

$$
\begin{aligned}
&\frac{1-\delta_2}{2n}\sum_{k=1}^{K}\|\boldsymbol{X}^{(k)}\widehat{G}^{(k)}-\boldsymbol{X}^{(k)}\widetilde{G}^{(k)}(\hat{\pi})\|_F^2+\lambda\|\widehat{G}^{(1:K)}\|_{l_1/l_2}\\
&\quad+\frac{\delta_1}{4\sigma_{\max}^4}\sum_{k=1}^{K'}\sum_{j=1}^{p}\left(\sigma_j^{(k)}(\hat{\pi})^2-\sigma_j^{(k)}(\pi_0)^2\right)^2\\
&\le\left(\frac{4\sigma_{\max}^4}{1-\delta_1}+2\sigma_{\max}^2\right)\frac{2p\log p}{n}+2\sigma_{\max}^2\sqrt{\frac{(K-K')2p^2\log p}{n}} \qquad (16)\\
&\quad+\delta_2\lambda\|\widehat{G}^{(1:K)}-\widetilde{G}^{(1:K)}(\hat{\pi})\|_{l_1/l_2}+\lambda\|\widetilde{G}_0^{(1:K)}\|_{l_1/l_2}\\
&\le\left(\frac{4\sigma_{\max}^4}{1-\delta_1}+2\sigma_{\max}^2\right)\frac{2p\log p}{n}+2\sigma_{\max}^2\sqrt{\frac{(K-K')2p^2\log p}{n}}\\
&\quad+\delta_2\lambda\|\widehat{G}^{(1:K)}\|_{l_1/l_2}+\delta_2\lambda\|\widetilde{G}^{(1:K)}(\hat{\pi})\|_{l_1/l_2}+\lambda\|\widetilde{G}_0^{(1:K)}\|_{l_1/l_2}. \qquad (17)
\end{aligned}
$$

Suppose $\hat{\pi}\neq\pi_0$. Condition 3.5 implies

$$
\begin{aligned}
\frac{\delta_1}{4\sigma_{\max}^4}\frac{pK'}{\eta_w}\le&\left(\frac{4\sigma_{\max}^4}{1-\delta_1}+2\sigma_{\max}^2\right)\frac{2p\log p}{n}+\lambda\|\widetilde{G}_0^{(1:K)}\|_{l_1/l_2}+\delta_2\lambda\|\widetilde{G}^{(1:K)}(\hat{\pi})\|_{l_1/l_2}\\
&+2\sigma_{\max}^2\sqrt{\frac{(K-K')2p^2\log p}{n}}.
\end{aligned}
$$

Divide both sides by pK', it implies

$$
\begin{aligned}
\frac{\delta_1}{4\sigma_{\max}^4}\frac{1}{\eta_w}\le&\left(\frac{4\sigma_{\max}^4}{1-\delta_1}+2\sigma_{\max}^2\right)\frac{2\log p}{nK'}+\frac{\lambda\|\widetilde{G}_0^{(1:K)}\|_{l_1/l_2}}{pK'}+\delta_2\frac{\lambda\|\widetilde{G}^{(1:K)}(\hat{\pi})\|_{l_1/l_2}}{pK'}\\
&+2\sigma_{\max}^2\sqrt{\frac{(K-K')2\log p}{nK'^2}}\\
\le&\left(\frac{4\sigma_{\max}^4}{1-\delta_1}+c_{\max}\right)\frac{2\log p}{nK'}+\frac{\lambda\left(s(\pi_0)+\delta_2 s(\hat{\pi})\right)g_{\max}\sqrt{K}}{pK'}\\
&+2\sigma_{\max}^2\sqrt{\frac{(K-K')2\log p}{nK'^2}}.
\end{aligned}
$$

The last inequality uses the fact that $\|\widetilde{G}^{(1:K)}(\pi)\|_{l_1/l_2}\le s(\pi)\sqrt{K}g_{\max}$. It contradicts with the condition

$$
\begin{aligned}
\frac{1}{\eta_w}>&\left(\frac{16\sigma_{\max}^8}{\delta_1(1-\delta_1)}+\frac{4\sigma_{\max}^4 c_{\max}}{1-\delta_1}\right)\frac{2\log p}{nK'}+\frac{4\sigma_{\max}^4 g_{\max}}{\delta_1}\frac{\lambda\left(s(\pi_0)+\delta_2 s(\hat{\pi})\right)}{p}\sqrt{\frac{K}{K'^2}}\\
&+\frac{8\sigma_{\max}^6}{\delta_1}\sqrt{\frac{2\log p}{n}\frac{K-K'}{K'^2}}.
\end{aligned}
$$

Therefore, $\hat{\pi}\in\Pi_0$.

Theorem 3.1 (a) is straightforward by taking $\lambda=\sqrt{p\log p/n}$.

B.2 Key lemmas for proving Theorem 3.1 (a)

B.2.1 Lemma B.1: Analysis of (I)

Lemma B.1. *Denote* $\sigma_{\max} = \sup_{k\in[K], j\in[p], \pi\in\mathbb{S}_p} |\sigma_j^{(k)}(\pi)|$. *With probability at least* $1 - 2e^{-p\log p}$, *it holds for any* $\delta_1 \in (0,1)$ *and any permutations* $\hat{\pi} \in \mathbb{S}_p$ *that,*

$$\begin{aligned}
&\sum_{k=1}^{K} \frac{1}{2n} \left(\|\boldsymbol{X}^{(k)} - \boldsymbol{X}^{(k)}\widetilde{G}_0^{(k)}\|_F^2 - \|\boldsymbol{X}^{(k)} - \boldsymbol{X}^{(k)}\widetilde{G}^{(k)}(\hat{\pi})\|_F^2 \right) \\
\leq &- \frac{\delta_1}{4\sigma_{\max}^4} \sum_{k=1}^{K'} \sum_{j=1}^{p} \left(\sigma_j^{(k)}(\pi_0)^2 - \sigma_j^{(k)}(\hat{\pi})^2 \right)^2 \\
&+ \left(\frac{\sigma_{\max}^4}{1-\delta_1} + 2\sigma_{\max}^2 \right) \frac{2p\log p}{n} + 2\sigma_{\max}^2 \sqrt{\frac{(K-K')2p^2\log p}{n}}.
\end{aligned}$$

We now state the proof of this Lemma. Denote the j-th column of $\widetilde{G}^{(k)}(\pi)$ as $\widetilde{g}_j^{(k)}(\pi)$, and the noise as

$$\widetilde{\boldsymbol{\varepsilon}}_j^{(k)}(\pi) := \boldsymbol{X}_j^{(k)} - \boldsymbol{X}^{(k)}\widetilde{g}_j^{(k)}(\pi) \in \mathbb{R}^n. \tag{18}$$

Then we can rewrite the term (I) as follows.

$$\begin{aligned}
(I) &= \sum_{k=1}^{K} \frac{1}{2n} \left(\|\boldsymbol{X}^{(k)} - \boldsymbol{X}^{(k)}\widetilde{G}_0^{(k)}\|_F^2 - \|\boldsymbol{X}^{(k)} - \boldsymbol{X}^{(k)}\widetilde{G}^{(k)}(\hat{\pi})\|_F^2 \right) \\
&= \frac{1}{2} \sum_{k=1}^{K} \left[\sum_{j=1}^{p} \frac{\frac{1}{n}\|\widetilde{\boldsymbol{\varepsilon}}_j^{(k)}(\hat{\pi})\|_2^2}{\sigma_j^{(k)}(\hat{\pi})^2} \sigma_j^{(k)}(\pi_0)^2 - \sum_{j=1}^{p} \frac{1}{n}\|\widetilde{\boldsymbol{\varepsilon}}_j^{(k)}(\hat{\pi})\|_2^2 \right] \\
&= \frac{1}{2} \sum_{k=1}^{K} \sum_{j=1}^{p} \left(\sigma_j^{(k)}(\pi_0)^2 - \sigma_j^{(k)}(\hat{\pi})^2 \right) \left(\frac{\frac{1}{n}\|\widetilde{\boldsymbol{\varepsilon}}_j^{(k)}(\hat{\pi})\|_2^2}{\sigma_j^{(k)}(\hat{\pi})^2} - 1 \right) + \frac{1}{2} \sum_{k=1}^{K} \sum_{j=1}^{p} \left(\sigma_j^{(k)}(\pi_0)^2 - \sigma_j^{(k)}(\hat{\pi})^2 \right) \\
&\leq \frac{1}{2} \sum_{k=1}^{K} \sum_{j=1}^{p} \left(\sigma_j^{(k)}(\pi_0)^2 - \sigma_j^{(k)}(\hat{\pi})^2 \right) \left(\frac{\frac{1}{n}\|\widetilde{\boldsymbol{\varepsilon}}_j^{(k)}(\hat{\pi})\|_2^2}{\sigma_j^{(k)}(\hat{\pi})^2} - 1 \right) \\
&\quad - \frac{1}{4\sigma_{\max}^4} \sum_{k=1}^{K'} \sum_{j=1}^{p} \left(\sigma_j^{(k)}(\hat{\pi})^2 - \sigma_j^{(k)}(\pi_0)^2 \right)^2
\end{aligned}$$

The last inequality holds because for $k = 1, \cdots, K'$, $\sum_{j=1}^{p} \left(\sigma_j^{(k)}(\pi_0)^2 - \sigma_j^{(k)}(\hat{\pi})^2 \right) \leq -\frac{1}{2\sigma_{\max}^4} \sum_{j=1}^{p} \left(\sigma_j^{(k)}(\hat{\pi})^2 - \sigma_j^{(k)}(\pi_0)^2 \right)^2$, and that for $k > K'$, $\sum_{j=1}^{p} \left(\sigma_j^{(k)}(\pi_0)^2 - \sigma_j^{(k)}(\hat{\pi})^2 \right) \leq 0$. Then we bound the first term using the concentration bound on Chi-squared random variables.

$$\begin{aligned}
&\sum_{k=1}^{K} \sum_{j=1}^{p} \left(\sigma_j^{(k)}(\pi_0)^2 - \sigma_j^{(k)}(\hat{\pi})^2 \right) \left(\frac{\frac{1}{n}\|\widetilde{\boldsymbol{\varepsilon}}_j^{(k)}(\hat{\pi})\|_2^2}{\sigma_j^{(k)}(\hat{\pi})^2} - 1 \right) \\
&\overset{d.}{=} \sum_{k=1}^{K} \sum_{j=1}^{p} \left(\sigma_j^{(k)}(\pi_0)^2 - \sigma_j^{(k)}(\hat{\pi})^2 \right) \left(\frac{1}{n}\xi_j^2 - 1 \right) = \frac{1}{n} \sum_{k=1}^{K} \sum_{j=1}^{p} \left(\sigma_j^{(k)}(\pi_0)^2 - \sigma_j^{(k)}(\hat{\pi})^2 \right) \left(\xi_j^2 - n \right),
\end{aligned}$$

where $\xi_j^2 \sim \chi^2(n)$ are i.i.d. Chi-squared random variables of degree n.

By Lemma H.1, for any fixed $\hat{\pi} \in \mathbb{S}_p$ and for any $t > 0$, it holds with probability at least $1 - e^{-t}$ that

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{K'}\sum_{j=1}^{p}\left(\sigma_j^{(k)}(\pi_0)^2 - \sigma_j^{(k)}(\hat{\pi})^2\right)\left(\xi_j^2 - n\right) \\
&\leq 2\sqrt{\frac{\sum_{k=1}^{K'}\sum_{j=1}^{p}\left(\sigma_j^{(k)}(\pi_0)^2 - \sigma_j^{(k)}(\hat{\pi})^2\right)^2}{n}}\sqrt{t} + \frac{2\sigma_{\max}^2}{n}t \\
&\leq \delta\sum_{k=1}^{K'}\sum_{j=1}^{p}\left(\sigma_j^{(k)}(\pi_0)^2 - \sigma_j^{(k)}(\hat{\pi})^2\right)^2 + \left(\frac{1}{\delta} + 2\sigma_{\max}^2\right)\frac{t}{n},
\end{aligned}$$

The last inequality holds because $2ab \leq \delta a^2 + \frac{1}{\delta}b^2$ for any $\delta > 0$. Now it remains to take a union bound over the permutation $\hat{\pi} \in \mathbb{S}_p$. There are $p!$ many permutations. Take an uniform control over all possible $\hat{\pi} \in \mathbb{S}_p$. It implies that with probability at least $1 - (p!)e^{-t}$, the above inequality holds for all $\hat{\pi} \in \mathbb{S}_p$. Equivalently, we can say it holds with probability at least $1 - e^{-t}$ that it holds for all $\hat{\pi}$ that

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{K'}\sum_{j=1}^{p}\left(\sigma_j^{(k)}(\pi_0)^2 - \sigma_j^{(k)}(\hat{\pi})^2\right)\left(\xi_j^2 - n\right) \\
&\leq \delta\sum_{k=1}^{K'}\sum_{j=1}^{p}\left(\sigma_j^{(k)}(\pi_0)^2 - \sigma_j^{(k)}(\hat{\pi})^2\right)^2 + \left(\frac{1}{\delta} + 2\sigma_{\max}^2\right)\frac{t + p\log p}{n}.
\end{aligned}$$

For the non-identifiable models, we can use Lemma H.1 in a similar way to obtain that with probability at least $1 - e^{-t}$, the following holds for all $\hat{\pi}$,

$$\begin{aligned}
&\frac{1}{n}\sum_{k=K'+1}^{K}\sum_{j=1}^{p}\left(\sigma_j^{(k)}(\pi_0)^2 - \sigma_j^{(k)}(\hat{\pi})^2\right)\left(\xi_j^2 - n\right) \\
&\leq 2\sqrt{\frac{\sum_{k=K'+1}^{K}\sum_{j=1}^{p}\left(\sigma_j^{(k)}(\pi_0)^2 - \sigma_j^{(k)}(\hat{\pi})^2\right)^2}{n}}\sqrt{t} + \frac{2\sigma_{\max}^2}{n}t \\
&\leq 2\sigma_{\max}^2\sqrt{\frac{(K-K')p(t + p\log p)}{n}} + \frac{2\sigma_{\max}^2}{n}(t + p\log p).
\end{aligned}$$

Putting the above results back into the term (I), taking $\delta' = 2\delta$, and taking $t = p\log p$, we have with probability at least $1 - 2e^{-p\log p}$ that

$$\begin{aligned}
(I) \leq\ & \frac{\delta'}{4}\sum_{k=1}^{K'}\sum_{j=1}^{p}\left(\sigma_j^{(k)}(\pi_0)^2 - \sigma_j^{(k)}(\hat{\pi})^2\right)^2 - \frac{1}{4\sigma_{\max}^4}\sum_{k=1}^{K'}\sum_{j=1}^{p}\left(\sigma_j^{(k)}(\hat{\pi})^2 - \sigma_j^{(k)}(\pi_0)^2\right)^2 \\
&+ \left(\frac{1}{\delta'} + 2\sigma_{\max}^2\right)\frac{2p\log p}{n} + 2\sigma_{\max}^2\sqrt{\frac{(K-K')2p^2\log p}{n}}
\end{aligned}$$

Finally, take $\delta_1 = 1 - \sigma_{\max}^4\delta'$ so that $\delta' = \frac{1}{\sigma_{\max}^4}(1 - \delta_1)$. Then for any $\delta_1 \in (0, 1)$, the following inequality holds with probability at least $1 - e^{-p\log p}$ for all $\hat{\pi} \in \mathbb{S}_p$:

$$\begin{aligned}
(I) \leq\ & -\frac{\delta_1}{4\sigma_{\max}^4}\sum_{k=1}^{K'}\sum_{j=1}^{p}\left(\sigma_j^{(k)}(\pi_0)^2 - \sigma_j^{(k)}(\hat{\pi})^2\right)^2 \\
&+ \left(\frac{\sigma_{\max}^4}{1-\delta_1} + 2\sigma_{\max}^2\right)\frac{2p\log p}{n} + 2\sigma_{\max}^2\sqrt{\frac{(K-K')2p^2\log p}{n}}.
\end{aligned}$$

B.2.2 Lemma B.2: Analysis of (II)

Lemma B.2. *Denote* $\rho := \sup_{k\in[K],j\in[p]} \Sigma_{jj}^{(k)}$ *and* $\sigma_{\max} := \sup_{k\in[K],j\in[p],\pi\in\mathbb{S}_p} |\sigma_j^{(k)}(\pi)|$. *For any* $\delta_2, \delta_3 \in (0,1)$, *assume* λ *satisfies* $t^* := \frac{\delta_2^2\delta_3^2\lambda^2 n}{16\rho\sigma_{\max}^2} > K$. *With probability at least*

$$1 - \exp\left(-t^*\left[1 - 2\sqrt{\frac{K}{t^*}}\right] + (d+2)\log p\right) - \exp\left(-\frac{4(1-\delta_3)}{\delta_3^2}n + \log K + (d+1)\log p\right),$$

the following inequality holds true:

$$\begin{aligned}
&\frac{1}{2n}\sum_{k=1}^{K}\langle \boldsymbol{X}^{(k)} - \boldsymbol{X}^{(k)}\widetilde{G}^{(k)}(\hat{\pi}), \boldsymbol{X}^{(k)}\widehat{G}^{(k)} - \boldsymbol{X}^{(k)}\widetilde{G}^{(k)}(\hat{\pi})\rangle \\
&\leq \frac{\delta_2}{2n}\sum_{k=1}^{K}\|\boldsymbol{X}^{(k)}\widehat{G}^{(k)} - \boldsymbol{X}^{(k)}\widetilde{G}^{(k)}(\hat{\pi})\|_F^2 + \delta_2\lambda\|\widehat{G}^{(1:K)} - \widetilde{G}^{(1:K)}(\hat{\pi})\|_{l_1/l_2}.
\end{aligned}$$

We now state the proof of this Lemma. To show the inequality in Lemma B.2 holds true, it is sufficient to show the following inequality holds true for all j and $\hat{\pi}$:

$$\begin{aligned}
\frac{1}{n}\sum_{k=1}^{K}\langle \widetilde{\boldsymbol{\varepsilon}}_j^{(k)}(\hat{\pi}), \boldsymbol{X}^{(k)}\left(\widehat{g}_j^{(k)}(\hat{\pi}) - \widetilde{g}_j^{(k)}(\hat{\pi})\right)\rangle_F \leq & \frac{\delta}{2n}\sum_{k=1}^{K}\|\boldsymbol{X}^{(k)}\left(\widehat{g}_j^{(k)}(\hat{\pi}) - \widetilde{g}_j^{(k)}(\hat{\pi})\right)\|_2^2 \\
& - \delta\lambda\|\widehat{g}_j^{(1:K)}(\hat{\pi}) - \widetilde{g}_j^{(1:K)}(\hat{\pi})\|_{l_1/l_2}, \qquad (19)
\end{aligned}$$

where we denote

$$\widehat{G}_j(\hat{\pi}) := [\widehat{g}_j^{(1)}(\hat{\pi}), \cdots, \widehat{g}_j^{(K)}(\hat{\pi})] \quad \text{and} \quad \widetilde{G}_j(\hat{\pi}) := [\widetilde{g}_j^{(1)}(\hat{\pi}), \cdots, \widetilde{g}_j^{(K)}(\hat{\pi})].$$

Now consider a fixed j and a fixed $\hat{\pi}$. Recall $S_j(\hat{\pi})$ which denotes the set of ancestors of the node j specified by the permutation $\hat{\pi}$, and let $m = |S_j(\hat{\pi})| \in [0, p-1]$ be its cardinality. Let $\boldsymbol{X}_{S_j}^{(k)} = \boldsymbol{X}^{(k)}|_{S_j(\hat{\pi})}$ denote the submatrix of $\boldsymbol{X}^{(k)}$ whose column indices are in $S_j(\hat{\pi})$. We define the event

$$\mathcal{E}(\delta, \lambda; \widetilde{\boldsymbol{\varepsilon}}_j^{(k)}(\hat{\pi})) :=$$
$$\left\{\sup_{\{\beta^{(k)}\in\mathbb{R}^m\}} \frac{1}{n}\sum_{k=1}^{K}\langle \widetilde{\boldsymbol{\varepsilon}}_j^{(k)}(\hat{\pi}), \boldsymbol{X}_{S_j}^{(k)}\beta^{(k)}\rangle - \frac{\delta}{2n}\sum_{k=1}^{K}\|\boldsymbol{X}_{S_j}^{(k)}\beta^{(k)}\|_2^2 - \delta\lambda\|[\beta^{(1)}, \cdots, \beta^{(K)}]\|_{l_1/l_2} \leq 0\right\}.$$

It's easy to see that with probability at least $\Pr\left[\mathcal{E}(\delta, \lambda; \widetilde{\boldsymbol{\varepsilon}}_j^{(1:K)}(\hat{\pi}))\right]$, the inequality in Eq. 19 holds true. Therefore, we need to derive the probability of the joint event $\cap_{j\in[p],\hat{\pi}\in\mathbb{S}_p}\mathcal{E}(\delta, \lambda; \widetilde{\boldsymbol{\varepsilon}}_j^{(1:K)}(\hat{\pi}))$ in this proof. Observe that:

$$\mathcal{E}(\delta, \lambda; \widetilde{\boldsymbol{\varepsilon}}_j^{(1:K)}(\hat{\pi})) \subseteq$$
$$\left\{\sup_{\{\beta^{(k)}\in\mathbb{R}^m\}} \frac{1}{2n}\sum_{k=1}^{K}\|\frac{\widetilde{\boldsymbol{\varepsilon}}_j^{(k)}(\hat{\pi})}{\delta}\|_2^2 - \frac{1}{2n}\sum_{k=1}^{K}\|\frac{\widetilde{\boldsymbol{\varepsilon}}_j^{(k)}(\hat{\pi})}{\delta} - \boldsymbol{X}_{S_j}^{(k)}\beta^{(k)}\|_2^2 - \lambda\|[\beta^{(1)}, \cdots, \beta^{(K)}]\|_{l_1/l_2} \leq 0\right\}$$
$$= \left\{\boldsymbol{0} \in \operatorname*{arg\,min}_{\{\beta^{(k)}\in\mathbb{R}^m\}} \frac{1}{2n}\sum_{k=1}^{K}\|\frac{\widetilde{\boldsymbol{\varepsilon}}_j^{(k)}(\hat{\pi})}{\delta} - \boldsymbol{X}_{S_j}^{(k)}\beta^{(k)}\|_2^2 + \lambda\|[\beta^{(1)}, \cdots, \beta^{(K)}]\|_{l_1/l_2}\right\}$$

Therefore, we resort to bound the probability of the above event, which is the **null-consistency** of l_1/l_2-penalized group Lasso problem. We present the null-consistency analysis by Lemma B.3, and its proof is given in Sec B.2.3.

Note that in the event $\mathcal{E}(\delta, \lambda; \widetilde{\boldsymbol{\varepsilon}}_j^{(1:K)}(\hat{\pi}))$, the variance is $\frac{\widetilde{\boldsymbol{\varepsilon}}_j^{(k)}(\hat{\pi})}{\delta} \overset{d.}{=} \boldsymbol{w}_k$ with

$$\boldsymbol{w}_k \sim \mathcal{N}\left(0, \left(\frac{\sigma_j^{(k)}(\hat{\pi})}{\delta}\right)^2 I_n\right).$$

Therefore, we can take σ_0 in Lemma B.3 to be $\sigma_0 = \frac{\sigma_{\max}}{\delta}$, which implies for any $\delta' \in (0,1)$, if

$$t^* := \frac{\delta'^2 \delta^2 \lambda^2 n}{16\rho\sigma_{\max}^2} > K,$$

then

$$\Pr\left[\mathcal{E}(\delta, \lambda; \widetilde{\boldsymbol{\varepsilon}}_j^{(1:K)}(\hat{\pi}))\right] \geq 1 - p\exp\left(-t^*\left[1 - 2\sqrt{\frac{K}{t^*}}\right]\right) - K\exp\left(-\frac{4(1-\delta')}{\delta'^2}n\right).$$

Now what remains is to take a uniform control over all events $\cap_{j\in[p],\hat{\pi}\in\mathbb{S}_p}\mathcal{E}(\delta, \lambda; \widetilde{\boldsymbol{\varepsilon}}_j^{(1:K)}(\hat{\pi}))$. A naive way is to enumerate over all permutations $\hat{\pi}$ and all j which will constitute $p \cdot p!$ events. However, recall $\widetilde{\boldsymbol{\varepsilon}}_j^{(k)}(\pi) := \boldsymbol{X}_j^{(k)} - \boldsymbol{X}^{(k)}\widetilde{g}_j^{(k)}(\pi)$. Then it is enough to take a uniform control over the set $\{\widetilde{g}_j^{(1:K)}(\pi) : \pi \in \mathbb{S}_p, j \in [p]\}$. By Eq. 35, this set contains at most $p \cdot p^d$ many elements. Therefore,

$$\begin{aligned}
&\Pr\left[\cap_{j\in[p],\hat{\pi}\in\mathbb{S}_p}\mathcal{E}(\delta, \lambda; \widetilde{\boldsymbol{\varepsilon}}_j^{(1:K)}(\hat{\pi}))\right] \\
&\geq 1 - p^2 p^d \exp\left(-t^*\left[1 - 2\sqrt{\frac{K}{t^*}}\right]\right) - Kp \cdot p^d \exp\left(-\frac{4(1-\delta')}{\delta'^2}n\right) \\
&\geq 1 - \exp\left(-t^*\left[1 - 2\sqrt{\frac{K}{t^*}}\right] + (d+2)\log p\right) - \exp\left(-\frac{4(1-\delta')}{\delta'^2}n + \log K + (d+1)\log p\right)
\end{aligned}$$

which implies Eq. 19 holds with the above probability.

B.2.3 Lemma B.3: Null Consistency

Lemma B.3 (Null-consistency)**.** *Let $S \subseteq [p]$ be a set of m indices. Consider the following linear regression model with zero vector as the true parameters:*

$$\boldsymbol{y}^{(k)} = \boldsymbol{X}_S^{(k)}\mathbf{0} + \boldsymbol{w}^{(k)}, \quad \text{for } k \in [K]$$

*where $\boldsymbol{y}^{(k)} = \boldsymbol{w}^{(k)} \in \mathbb{R}^n$, $\boldsymbol{X}_S^{(k)} \in \mathbb{R}^{n\times m}$ and $\mathbf{0} \in \mathbb{R}^m$. Assume that for each k, the row vectors of $\boldsymbol{X}^{(k)}$ are i.i.d. sampled from $\mathcal{N}(0, \Sigma^{(k)})$ and the noise is sampled from $\boldsymbol{w}^{(k)} \sim \mathcal{N}(0, \sigma_W^{(k)2} I_n)$. Denote $\rho := \max_{k\in[K]} \Sigma_{jj}^{(k)}$ and $\sigma_0 := \max_{k\in[K]} \sigma_W^{(k)}$. Consider the following l_1/l_2-**regularized Lasso problem**:*

$$\widehat{B} = \underset{B\in\mathbb{R}^{m\times K}}{\arg\min} \frac{1}{2n}\sum_{k=1}^{K}\|\boldsymbol{y}^{(k)} - \boldsymbol{X}_S^{(k)}\beta^{(k)}\|_2^2 + \lambda\|B\|_{l_1/l_2}, \tag{20}$$

where $B = [\beta^{(1)}, \cdots, \beta^{(k)}]$. For any $\delta \in (0,1)$, if

$$t^* = \frac{\delta^2\lambda^2 n}{16\rho\sigma_0^2} > K,$$

then with probability at least

$$1 - m\exp\left(-t^*\left[1 - 2\sqrt{\frac{K}{t^*}}\right]\right) - K\exp\left(-\frac{4(1-\delta)}{\delta^2}n\right),$$

$\widehat{B} = \mathbf{0}$ is an optimal solution to the problem in Eq. 20.

The proof of this lemma is stated below, in which we simply use the notation $\boldsymbol{X}^{(k)}$ to replace $\boldsymbol{X}_S^{(k)}$.

Lemma B.4. *Suppose there exists a primal-dual pair $(\widehat{B}, \widehat{Z}) \in \mathbb{R}^{m\times K} \times \mathbb{R}^{m\times K}$ which satisfies the following conditions:*

$$\widehat{Z} \in \partial\|\widehat{B}\|_{l_1/l_2}, \tag{21a}$$

$$-\frac{1}{n}\boldsymbol{X}^{(k)\top}\left(\boldsymbol{y}^{(k)} - \boldsymbol{X}^{(k)}\widehat{\beta}^{(k)}\right) + \lambda\widehat{\boldsymbol{z}}^{(k)} = 0, \quad \forall k \in [K], \tag{21b}$$

$$\|\widehat{Z}\|_{l_\infty/l_2} < 1, \tag{21c}$$

where $[\widehat{\beta}^{(1)}, \cdots, \widehat{\beta}^{(K)}]$ are the columns of $\widehat{B}$ and $[\widehat{\boldsymbol{z}}^{(1)}, \cdots, \widehat{\boldsymbol{z}}^{(K)}]$ are the columns of $\widehat{Z}$. Then $\mathbf{0}$ is the solution to the problem in Eq. 7 and it is the only solution.

Proof. Straightforward by Lemma 1 in [19]. □

Therefore, to show $\mathbf{0} \in \arg\min_{B \in \mathbb{R}^{m \times K}} \frac{1}{2n} \sum_{k=1}^{K} \|\boldsymbol{y}^{(k)} - \boldsymbol{X}^{(k)} \beta^{(k)}\|_2^2 + \lambda \|B\|_{l_1/l_2}$, it is sufficient to show the existence of $(\widehat{B}, \widehat{Z})$ which satisfies the conditions in Eq. 21. We construct such a pair by the following definitions:

$$\widehat{B} := \mathbf{0}, \tag{22}$$

$$\widehat{\boldsymbol{z}}^{(k)} := \frac{1}{\lambda n} \boldsymbol{X}^{(k)\top} \boldsymbol{y}^{(k)}. \tag{23}$$

Clearly, they satisfy Eq. 21b. If we can show $\|\widehat{Z}\|_{l_\infty/l_2} < 1$, then Eq. 21 holds. Therefore, in this proof, the main goal is the analyze $\Pr\left[|\widehat{Z}\|_{l_\infty/l_2} < 1\right]$.

Denote the row vectors of $\widehat{Z}$ as $\widehat{Z}_j := [\widehat{\boldsymbol{z}}_j^{(1)}, \cdots, \widehat{\boldsymbol{z}}_j^{(K)}]$. Then $\|\widehat{Z}\|_{l_\infty/l_2} = \max_{j \in [m]} \|\widehat{Z}_j\|_2$. By definition in Eq. 23,

$$\widehat{\boldsymbol{z}}_j^{(k)} = \frac{1}{\lambda n} \boldsymbol{X}_j^{(k)\top} \boldsymbol{y}^{(k)} = \frac{1}{\lambda n} \boldsymbol{X}_j^{(k)\top} \boldsymbol{w}^{(k)}.$$

Since the linear combination of Gaussian distribution is still Gaussian, then given $\boldsymbol{w}^{(1:K)}$, the variable $\|\widehat{Z}_j\|_2^2$ is equivalent to a Chi-squared random variable in distribution:

$$\begin{aligned}
\|\widehat{Z}_j\|_2^2 \mid \boldsymbol{w}^{(1:K)} &= \frac{1}{\lambda^2 n^2} \sum_{k=1}^{K} \left(\boldsymbol{X}_j^{(k)\top} \boldsymbol{w}^{(k)} \right)^2 \mid \boldsymbol{w}^{(1:K)} \\
&\overset{d.}{=} \frac{1}{\lambda^2 n^2} \sum_{k=1}^{K} \Sigma_{jj}^{(k)} \|\boldsymbol{w}^{(k)}\|_2^2 \xi_{jk}^2 \quad \text{where } \xi_{jk} \sim \mathcal{N}(0,1) \\
&\leq \frac{1}{\lambda^2 n^2} \max_{k \in [K]} \Sigma_{jj}^{(k)} \max_{k \in [K]} \|\boldsymbol{w}^{(k)}\|_2^2 \sum_{k=1}^{K} \xi_{jk}^2
\end{aligned}$$

By Lemma H.2, for any $\delta > 0$,

$$\Pr\left[\max_{k \in [K]} \|\boldsymbol{w}^{(k)}\|_2^2 \geq \sigma_W^{(k)2} 2n(1+\delta)\right] \leq K \exp\left(-n(1+\delta)\left[1 - 2\sqrt{\frac{1}{1+\delta}}\right]\right).$$

Therefore, for all $\delta > 0$,

$$\Pr\left[\max_j \|\widehat{Z}_j\|_2 < 1\right] \geq \Pr\left[\max_{j \in [p]} \sum_{k=1}^{K} \xi_{jk}^2 < \frac{\lambda^2 n}{2(1+\delta)\rho\sigma_0^2}\right] \Pr\left[\max_k \|\boldsymbol{w}^{(k)}\|_2^2 < 2n(1+\delta)\right]$$

where

$$\rho := \max_{k \in [K]} \Sigma_{jj}^{(k)} \quad \text{and} \quad \sigma_0 := \max_{k \in [K]} \sigma_W^{(k)}.$$

Take $t^* = \frac{\lambda^2 n}{4(1+\delta)\rho\sigma_0^2}$, then if $t^* > K$, we have

$$\Pr\left[\max_{j \in [p]} \sum_{k=1}^{K} \xi_{jk}^2 < 2t^*\right] \geq 1 - p \exp\left(-t^*\left[1 - 2\sqrt{\frac{K}{t^*}}\right]\right).$$

Rewrite $\delta = \frac{4}{\delta'^2} - 1$ for some $\delta' \in (0,1)$ so that $1 + \delta = \frac{4}{\delta'^2} > 4$. If λ is taken to be some value that satisfies the condition

$$t^* = \frac{\delta'^2 \lambda^2 n}{16\rho\sigma_0^2} > K,$$

then

$$\Pr\left[\max_j \|\widehat{Z}_j\|_2 < 1\right] \geq 1 - p \exp\left(-t^*\left[1 - 2\sqrt{\frac{K}{t^*}}\right]\right) - K \exp\left(-\frac{4(1-\delta')}{\delta'^2} n\right).$$

B.3 Proof of error in F-norm

Denote the error vector as $\Delta_j^{(k)} := \widehat{g}_j^{(k)}(\hat{\pi}) - \widetilde{g}_j^{(k)}(\hat{\pi})$. Then

$$\sum_{k=1}^{K} \frac{1}{2n} \|\boldsymbol{X}^{(k)}\widehat{G}^{(k)} - \boldsymbol{X}^{(k)}\widetilde{G}^{(k)}(\hat{\pi})\|_F^2 = \sum_{k=1}^{K}\sum_{j=1}^{p} \frac{1}{2n}\|\boldsymbol{X}^{(k)}\Delta_j^k\|_2^2$$

By Theorem 7.3 in [14], with probability at least $1 - \exp(-\log p - \log K)$, it holds for all $k \in [K]$ and $j \in [p]$ that

$$\begin{aligned}\frac{1}{\sqrt{n}}\|\boldsymbol{X}^{(k)}\Delta_j^{(k)}\|_2 &\geq \left(3/4\Lambda_{\text{min}} - 3\sigma_{\text{max}}\sqrt{\frac{\hat{d}\,(\log p + \log K)}{n}} - \sqrt{\frac{4(\log p + \log K)}{n}}\right)\|\Delta_j^{(k)}\|_2 \\ &\geq \left(3/4\Lambda_{\text{min}} - 3\sigma_{\text{max}}\sqrt{\frac{\hat{d}\,(\log p + \log K)}{n}} - c\right)\|\Delta_j^{(k)}\|_2\end{aligned}$$

where $\hat{d} := \sup_{j,k}\|\Delta_j^{(k)}\|_0$. If the sample size n satisfies the condition with a suitable constant $\kappa(\Lambda_{\text{min}})$:

$$n \geq \kappa(\Lambda_{\text{min}})\hat{d}\,(\log p + \log K)\,,$$

then $\frac{1}{\sqrt{n}}\|\boldsymbol{X}^{(k)}\Delta_j^{(k)}\|_2 \geq \kappa'(\Lambda_{\text{min}})\|\Delta_j^{(k)}\|_2$ for some constant $\kappa'(\Lambda_{\text{min}})$. Therefore,

$$\begin{aligned}&\frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{p}\|\Delta_j^{(k)}\|_2^2 \leq \frac{2}{\kappa'(\Lambda_{\text{min}})^2}\sum_{k=1}^{K}\frac{1}{2nK}\|\boldsymbol{X}^{(k)}\widehat{G}^{(k)} - \boldsymbol{X}^{(k)}\widetilde{G}^{(k)}(\hat{\pi})\|_F^2 \\ &\leq \frac{2}{\kappa'(\Lambda_{\text{min}})^2}\left(\kappa(\sigma_{\text{max}})\frac{p\log p}{nK} + cg_{\text{max}}\frac{s_0\lambda}{\sqrt{K}}\right),\end{aligned}$$

which implies

$$\frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{p}\|\Delta_j^{(k)}\|_2^2 = \mathcal{O}\left(\frac{p\log p}{nK} + \frac{s_0\lambda}{\sqrt{K}}\right) = \mathcal{O}\left(\frac{s_0\lambda}{\sqrt{K}}\right).$$

The last equation holds for the case when $\lambda = \sqrt{\frac{p\log p}{nK}}$.

C Proof of Theorem 3.2: Support Recovery

We are interested in showing that the support union of $\widehat{G}^{(1:K)}$ is the same as that of $\widetilde{G}_0^{(1:K)}$. To prove this, we can equivalently show the support union of $\widehat{g}_j^{(1:K)}$ is the same as that of $\widetilde{g}_j^{(1:K)}$ for any $j \in [p]$.

Now we state the proof of Theorem 3.2.

Proof. Given a permutation $\pi_0 \in \Pi_0$, the DAG structure learning problem is equivalent to solving p separate group Lasso problems, where for each j, the following l_1/l_2-penalized group Lasso is solved:

$$\widehat{g}_j^{(1:K)}|_{S_j(\pi_0)} = \underset{B\in\mathbb{R}^{|S_j(\pi_0)|\times K}}{\arg\min}\sum_{k=1}^{K}\frac{1}{2n}\|\boldsymbol{X}_j^{(k)} - \left(\boldsymbol{X}^{(k)}|_{S_j(\pi_0)}\right)\beta^{(k)}\|_2^2 + \lambda\|B\|_{l_1/l_2}, \tag{24}$$

where $S_j(\pi_0) := \{i : \pi_0(i) < \pi_0(j)\}$, and we denote the columns of B as $B = \left[\beta^{(1)}, \cdots, \beta^{(K)}\right]$. The proof in this section is based on a uniform control over all j. For each j, the estimation in form of Eq. 24 is called a *multi-design multi-response* (or multivariate) regression problem, which has been studied in the last decade [18, 19]. Our support recovery analysis is based on techniques for

analyzing multivariate regression problem in existing literature, but careful adaptation is needed to simultaneously handle a set of p problems where p is the dimension of the problem.

More precisely, we present the analysis of multivariate regression problems in Appendix D, where the main results are summarized in Theorem D.1. Then the results in Theorem 3.2 can be obtained with direct computations by applying Theorem D.1 to p separate problems defined by Eq. 24 and taking a union bound over all $j \in [p]$.

□

D Support Union Recovery for Multi-Design Multi-response Regression

The analysis in this section can be independent of the other content in this paper. We first introduce the multivariate regression setting and notations below.

Problem setting and assumptions. Consider the following K linear regression models

$$\boldsymbol{y}^{(k)} = \boldsymbol{X}^{(k)}\beta^{*(k)} + \boldsymbol{w}^{(k)}, \quad \text{for } k = 1, \cdots, K,$$

where $\boldsymbol{y}^{(k)} \in \mathbb{R}^n$, $\boldsymbol{X}^{(k)} \in \mathbb{R}^{n\times p}$, $\beta^{*(k)} \in \mathbb{R}^p$, and $\boldsymbol{w}^{(k)} \in \mathbb{R}^n$. Assume that for each k, the row vectors of $\boldsymbol{X}^{(k)}$ are i.i.d. sampled from $\mathcal{N}(0, \Sigma^{(k)})$ and the noise is sampled from $\boldsymbol{w}^{(k)} \sim \mathcal{N}(0, \sigma^{(k)2} I_n)$. Denote S as the support union of true parameters $\{\beta^{*(k)}\}_{k\in[K]}$, i.e., $S := \{j : \exists k \in [K]\ s.t.,\ \beta_j^{*(k)} \neq 0\}$, and $s = |S|$ as its size. Note that the s in this section has a different meaning from s in other sections. Furthermore, for the true parameters, we denote $B^* = [\beta^{*(1)}, \cdots, \beta^{*(K)}]$ as the matrix whose columns are $\beta^{*(k)}$. Besides, we use B_j^* to denote the j-th row of B^*.

Assumptions and definitions. Consider the following list of assumptions and definitions:

1. There exists $\gamma \in (0, 1]$ such that $\|A\|_\infty \leq 1 - \gamma$, where $A_{js} = \max_{k\in[K]} \left| \left(\Sigma_{S^cS}^{(k)} (\Sigma_{SS}^{(k)})^{-1} \right)_{js} \right|$ for $j \in S^c$ and $s \in S$.
2. There exist constants $0 < \Lambda_{\min} \leq \Lambda_{\max} < \infty$ such that all eigenvalues of $\Sigma_{SS}^{(k)}$ are in $[\Lambda_{\min}, \Lambda_{\max}]$ for all $k = 1, 2, \cdots, K$.
3. $\rho_u := \max_{j\in S^c, k\in[K]} \left(\Sigma_{S^cS^c|S}^{(k)} \right)_{jj}$
4. $\sigma_{\max} := \max_{k\in[K]} \sigma^{(k)}$
5. $b_{\min} := \min_{j\in S} \left\| B_j^* \right\|_2 / \sqrt{K}$.

With the above assumptions, we are ready to present the theorem.

Theorem D.1. *Assume the problem setting and assumptions in this section stated above. Consider the following l_1/l_2-**regularized Lasso problem**:*

$$\min_{B\in\mathbb{R}^{p\times K}} \frac{1}{2n} \sum_{k=1}^{K} \|\boldsymbol{y}^{(k)} - \boldsymbol{X}^{(k)}\beta^{(k)}\|_2^2 + \lambda \|B\|_{l_1/l_2}, \tag{25}$$

where $B = [\beta^{(1)}, \cdots, \beta^{(k)}]$. If the following condition holds

$$n \geq \kappa_6 s \log p,$$
$$K \leq c_0 \log p$$
$$\sqrt{\frac{8\sigma_{\max}^2 \log p}{\Lambda_{\min} n}} + \frac{2}{\Lambda_{\min}} \sqrt{\frac{sp \log p}{nK}} = o\left(b_{\min}^*\right),$$

then Eq. 25 has a unique solution $\widehat{B}$, and that with probability at least

$$1 - c_1 K \exp\left(-c_2(n - s)\right) - \exp\left(-c_3 \log p\right) - s \exp\left(-c_4 K \log p\right),$$

the support union of $\widehat{B}$ is the same as S, and that $\left\| \widehat{B} - B^* \right\|_{l_\infty/l_2} / \sqrt{K} = o(b_{\min})$.

In this statement, κ_6 is a constant depending on $\gamma, \Lambda_{\min}, \rho_u, \sigma_{\max}$ and c_i are universal constants.

D.1 Proof of Theorem D.1

The proof is based on a constructive procedure as specified by Lemma D.1, which characterizes an optimal primal-dual pair for which the primal solution $\widehat{B}$ correctly recovers the support set S.

Lemma D.1. *Define a pair $\widehat{B} = [\widehat{\beta}^{(1)}, \cdots, \widehat{\beta}^{(K)}]$ and $\widehat{Z} = [\widehat{\boldsymbol{z}}^{(1)}, \cdots, \widehat{\boldsymbol{z}}^{(K)}]$ as follows.*

$$\widehat{B}_{S^c} := 0 \tag{26a}$$

$$\widehat{B}_S := \underset{B_S \in \mathbb{R}^{s\times K}}{\arg\min} \frac{1}{2n} \sum_{k=1}^{K} \|\boldsymbol{y}^{(k)} - \boldsymbol{X}_S^{(k)} \beta_S^{(k)}\|_2^2 + \lambda \|B_S\|_{l_1/l_2} \tag{26b}$$

$$\widehat{\boldsymbol{z}}_S^{(k)} := -\lambda^{-1} \left(\widehat{\Sigma}_{SS}^{(k)} (\widehat{\beta}_S^{(k)} - \beta_S^{*(k)}) - \frac{1}{n} \boldsymbol{X}_S^{(k)\top} \boldsymbol{w}^{(k)} \right) \tag{26c}$$

$$\widehat{\boldsymbol{z}}_{S^c}^{(k)} := -\lambda^{-1} \left(\widehat{\Sigma}_{S^cS}^{(k)} (\widehat{\beta}_S^{(k)} - \beta_S^{*(k)}) - \frac{1}{n} \boldsymbol{X}_{S^c}^{(k)\top} \boldsymbol{w}^{(k)} \right) \tag{26d}$$

The following statements hold true.

(a) If the matrix $\widehat{Z}_{S^c} := [\widehat{\boldsymbol{z}}_{S^c}^{(1)}, \cdots, \widehat{\boldsymbol{z}}_{S^c}^{(K)}]$ defined by Eq. 26d satisfies

$$\|\widehat{Z}_{S^c}\|_{l_\infty/l_2} < 1, \tag{27}$$

then $(\widehat{B}, \widehat{Z})$ is a primal-dual optimal solution to the l_1/l_2-regularized Lasso problem in Eq. 25. Furthermore, any optimal solution $\widehat{B}$ to Eq. 25 satisfies $\widehat{B}_{S^c} = \mathbf{0}$.

(b) Define a matrix $U_S = [\boldsymbol{u}_S^{(1)}, \cdots, \boldsymbol{u}_S^{(K)}]$ whose column vectors are

$$\boldsymbol{u}_S^{(k)} := \widehat{\beta}_S^{(k)} - \beta_S^{*(k)} = (\widehat{\Sigma}_{SS}^{(k)})^{-1} \left(\frac{1}{n} X_S^{(k)\top} \boldsymbol{w}^{(k)} - \lambda \widehat{\boldsymbol{z}}_S^{(k)} \right).$$

If the conditions in (a) are satisfied, and furthermore, U_S satisfies

$$\frac{\|U_S\|_{l_\infty/l_2}}{\sqrt{K}} \le \frac{1}{2} b_{\min}^*, \tag{28}$$

then $\widehat{B}$ correctly recovers the union support S. That is,

$$\left\{ i \in [p] : \exists k \in [K] \; s.t. \; \widehat{\beta}_i^{(k)} \neq 0 \right\} = S.$$

Remark D.1. Note that the matrix $\widehat{Z}_S := [\widehat{\boldsymbol{z}}_S^{(1)}, \cdots, \widehat{\boldsymbol{z}}_S^{(K)}]$ defined by Eq. 26c is a dual solution to the restricted optimization in Eq. 26b, and thus satisfies $\widehat{Z}_S \in \partial \|\widehat{B}_S\|_{l_1/l_2}$.

Proof. The proof of Lemma D.1 (a) is similar to Lemma 1 in [19] and Lemma 2 in [18]. The proof of Lemma D.1 (b) is straightforward from the condition in Eq. 28. By definition of $b_{\min}$, Eq. 28 implies $\|\widehat{\beta}_j^{(1:K)}\|_2 \ge \|\beta_j^{*(1:K)}\|_2 - \|\widehat{\beta}_j^{(k)} - \beta_j^{*(k)}\|_2 \ge \frac{1}{2}\sqrt{K} b_{\min}^* > 0$ for any $j \in S$. □

Based on Lemma D.1, if we can show the primal-dual pair defined in its statement can satisfy both conditions in Eq. 27 and Eq. 28, then the support recovery guarantee is proved. We provide the analysis of these two conditions in Appendix D.1.1 and Appendix D.1.4 respectively.

Collecting the results in Appendix D.1.1 and Appendix D.1.4, we conclude that, if the following conditions are satisfied:

$$n \ge \kappa_6 s \log p, \quad K \le \frac{5}{64} \log p,$$

$$\sqrt{\frac{8\sigma_{\max}^2 \log p}{\Lambda_{\min} n}} + \frac{2}{\Lambda_{\min}} \sqrt{\frac{sp \log p}{nK}} = o\left(b_{\min}^*\right),$$

then with probability at least

$$
\begin{aligned}
&1 - 3K\exp\left(-\frac{n}{2}\left(\frac{1}{4}-\sqrt{\frac{s}{n}}\right)_+^2\right) - K\exp\left(-(1+\delta)(n-s)\left[1-2\sqrt{\frac{1}{1+\delta}}\right]\right)\\
&-\exp\left(-2\frac{3}{4}\log p\right) - s\exp\left(-2K\log p\left(1-2\sqrt{\frac{1}{2\log p}}\right)\right)\\
&\geq 1 - c_1 K\exp\left(-c_2(n-s)\right) - \exp\left(-c_3\log p\right) - s\exp\left(-c_4 K\log p\right),
\end{aligned}
$$

conditions in Eq. 27 and Eq. 28 are satisfied and therefore $\widehat{B}$ correctly recovers the union support S.

D.1.1 No false recovery: $\|\widehat{Z}_{S^c}\|_{l_\infty/l_2} < 1$

Denote the row vectors of $\widehat{Z}_{S^c}$ as $\widehat{Z}_j := [\widehat{\boldsymbol{z}}_j^{(1)}, \cdots, \widehat{\boldsymbol{z}}_j^{(k)}]$. Then

$$\|\widehat{Z}_{S^c}\|_{l_\infty/l_2} = \max_{j\in S^c}\|\widehat{Z}_j\|_2.$$

Since

$$
\begin{aligned}
\widehat{Z}_j =& \underbrace{\mathbb{E}[\widehat{Z}_j \mid \boldsymbol{X}_S^{(1:K)}]}_{T_{j1}} + \underbrace{\mathbb{E}[\widehat{Z}_j \mid \boldsymbol{X}_S^{(1:K)}, \boldsymbol{w}^{(1:K)}] - \mathbb{E}[\widehat{Z}_j \mid \boldsymbol{X}_S^{(1:K)}]}_{T_{j2}}\\
&+ \underbrace{\widehat{Z}_j - \mathbb{E}[\widehat{Z}_j \mid \boldsymbol{X}_S^{(1:K)}, \boldsymbol{w}^{(1:K)}]}_{T_{j3}},
\end{aligned}
$$

to prove $\|\widehat{Z}_{S^c}\|_{l_\infty/l_2} < 1$, we resort to bound $\max_{j\in S^c}\|T_{ja}\|_2$ for $a = 1, 2, 3$ separately. The analyses of T_{j1} and T_{j2} largely follow the arguments in [19] and [18], so details are omitted for brevity. We summarize the results of these two terms below, after which we present the detailed analysis for T_{j3}.

D.1.2 Analysis of T_{j1} and T_{j2}

T_{j1}: Following the same arguments as the derivations of Equation (28) in [19] and Equation (39) in [18], we have $\max_{j\in S^c}\|T_{j1}\|_2 \leq 1-\gamma$.

T_{j2}: Following the same arguments as the derivations of Equation (32) in [19], we have that

$$\max_{j\in S^c}\|T_{j2}\|_2 \leq (1-\gamma)\|\widehat{Z}_S - Z_S^*\|_{l_\infty/l_2} + (1-\gamma)\mathbb{E}[\|\widehat{Z}_S - Z_S^*\|_{l_\infty/l_2} \mid \boldsymbol{X}_S^{(1:K)}],$$

where the rows of Z_S^* are defined as $Z_i^* := B_i^*/\|B_i^*\|$ for $i \in S$. Define the matrix $\Delta \in \mathbb{R}^{s\times K}$ with rows $\Delta_i := (\widehat{B}_i - B_i^*)/\|B_i^*\|_2$. By Lemma H.3, if $\|\Delta\|_{l_\infty/l_2} < 1/2$, then it holds true that

$$\max_{j\in S^c}\|T_{j2}\|_2 \leq 4(1-\gamma)\left(\|\Delta\|_{l_\infty/l_2} + \mathbb{E}[\|\Delta\|_{l_\infty/l_2} \mid \boldsymbol{X}_S^{(1:K)}]\right).$$

We will show later in the analysis pf U_S that $\|\Delta\|_{l_\infty/l_2}$ is of order $o(1)$ with high probability.

D.1.3 Analysis of T_{j3}

Following the same arguments as the derivations of Equation (36) in [19] and Equation (42) [18], we have that for each $j \in S^c$,

$$
\begin{aligned}
&\text{given } \left(\boldsymbol{X}_S^{(1:K)}, \boldsymbol{w}^{(1:K)}\right),\\
&\widehat{\boldsymbol{z}}_j^{(k)} - \mathbb{E}[\widehat{\boldsymbol{z}}_j^{(k)} \mid \boldsymbol{X}_S^{(1:K)}, \boldsymbol{w}^{(1:K)}] \overset{d.}{=} \sigma_{jk}\xi_{jk},
\end{aligned}
$$

where

$$
\begin{cases}
\xi_{jk} \sim \mathcal{N}(0,1),\\
\sigma_{jk}^2 := (\Sigma_{S^cS^c|S}^{(k)})_{jj} M_k \leq \rho_u M_k,\\
M_k := \frac{1}{n}\widehat{\boldsymbol{z}}_S^{(k)\top}(\widehat{\Sigma}_{SS}^{(k)})^{-1}, \widehat{\boldsymbol{z}}_S^{(k)} - \frac{1}{n^2\lambda^2}\boldsymbol{w}^{(k)\top}(\Pi_S^{(k)} - I_n)\boldsymbol{w}^{(k)}.
\end{cases}
$$

Therefore,

$$\text{given } \left(\boldsymbol{X}_S^{(1:K)}, \boldsymbol{w}^{(1:K)}\right),$$

$$\max_{j\in S^c} \|\widehat{\boldsymbol{z}}_j^{(k)} - \mathbb{E}[\widehat{\boldsymbol{z}}_j^{(k)} \mid \boldsymbol{X}_S^{(1:K)}, \boldsymbol{w}^{(1:K)}]\|_2^2 \overset{d.}{=} \max_{j\in S^c} \sum_{k=1}^K \sigma_{jk}^2 \xi_{jk}^2$$

$$\leq \rho_u \max_{k\in[K]} |M_k| \max_{j\in S^c} \sum_{k=1}^K \xi_{jk}^2 \tag{29}$$

(1) Bounding $\max_{k\in[K]} |M_k|$.

Bound the term $\frac{1}{n}\widehat{\boldsymbol{z}}_S^{(k)\top}(\widehat{\Sigma}_{SS}^{(k)})^{-1}, \widehat{\boldsymbol{z}}_S^{(k)}$ is based on the following relations:

$$\max_{k\in[K]} \|\boldsymbol{z}_S^{*(k)}\|_2 \leq \sqrt{s} \quad \text{(by definition)},$$

$$\max_{k\in[K]} \|(\widehat{\Sigma}_{SS}^{(k)})^{-1}\|_2 \leq \frac{2}{\Lambda_{\min}} \quad \text{w.p.} \geq 1 - K\exp\left(-\frac{n}{2}\left(\frac{1}{4} - \sqrt{\frac{s}{n}}\right)_+^2\right) \quad \text{(Lemma 10 in [18])}.$$

Therefore, with probability $\geq 1 - K\exp\left(-\frac{n}{2}\left(\frac{1}{4} - \sqrt{\frac{s}{n}}\right)_+^2\right)$,

$$\frac{1}{n}\left|\widehat{\boldsymbol{z}}_S^{(k)\top}(\widehat{\Sigma}_{SS}^{(k)})^{-1}\widehat{\boldsymbol{z}}_S^{(k)}\right| \leq \frac{1}{n}\|(\widehat{\Sigma}_{SS}^{(k)})^{-1}\|_2 \|\widehat{\boldsymbol{z}}_S^{(k)}\|_2^2 \leq \frac{1}{n}\frac{2s}{\Lambda_{\min}},$$

For the second term in $\max_{k\in[K]} |M_k|$, note that

$$\boldsymbol{w}^{(k)\top}(I_n - \Pi_S^{(k)})\boldsymbol{w}^{(k)} \overset{d.}{=} \sigma^{(k)2} \sum_{j=1}^{n-s} \zeta_{jk}^2 \quad \text{with } \zeta_{jk} \sim \mathcal{N}(0,1).$$

By Lemma H.2, for any $\delta > 0$,

$$\Pr\left[\max_{k\in[K]} \sum_{j=1}^{n-s} \zeta_{jk}^2 \geq 2(1+\delta)(n-s)\right] \leq K\exp\left(-(1+\delta)(n-s)\left[1 - 2\sqrt{\frac{1}{1+\delta}}\right]\right).$$

To summarize, with probability at least

$$1 - K\exp\left(-\frac{n}{2}\left(\frac{1}{4} - \sqrt{\frac{s}{n}}\right)_+^2\right) - K\exp\left(-(1+\delta)(n-s)\left[1 - 2\sqrt{\frac{1}{1+\delta}}\right]\right),$$

it holds that

$$\max_{k\in[K]} |M_k| < \frac{1}{n}\frac{2s}{\Lambda_{\min}} + \frac{2\sigma_{\max}^2(1+\delta)(n-s)}{n^2\lambda^2}.$$

(2) Bounding $\max_{j\in S^c} \sum_{k=1}^K \xi_{jk}^2$.

Combining the bound on $\max_{k\in[K]} |M_k|$ with Eq. 29, it implies

$$\max_{j\in S^c} \|T_{j3}\|_2^2 \leq \rho_u \left(\frac{1}{n}\frac{2s}{\Lambda_{\min}} + \frac{2\sigma_{\max}^2(1+\delta)(n-s)}{n^2\lambda^2}\right) \max_{j\in S^c} \sum_{k=1}^K \xi_{jk}^2$$

$$\Longrightarrow \left\{\max_{j\in S^c} \|T_{j3}\|_2 < \gamma\right\} \subseteq \left\{\max_{j\in S^c} \sum_{k=1}^K \xi_{jk}^2 < \frac{\gamma^2}{2\rho_u} \frac{\lambda^2 n}{\lambda^2 s/\Lambda_{\min} + \sigma_{\max}^2(1+\delta)\frac{n-s}{n}}\right\}.$$

What's left is to bound the term $\sum_{k=1}^K \xi_{jk}^2$. Take $t^* = \frac{\gamma^2}{4\rho_u}\frac{\lambda^2 n}{\lambda^2 s/\Lambda_{\min} + \sigma_{\max}^2(1+\delta)\frac{n-s}{n}}$. If $t^* > K$, by Lemma H.2 and a union bound over $j \in S^c$, we have that

$$\Pr[\max_{j\in S^c} \sum_{k=1}^K \xi_{jk}^2 \geq 2t^*] \leq (p-s)\exp\left(-t^*\left[1 - 2\sqrt{\frac{K}{t^*}}\right]\right).$$

(3) Collecting all results.

To conclude, for any $\delta > 0$, if the following conditions are satisfied:

$$\|\Delta\|_{l_\infty/l_2} < \frac{1}{2},$$
$$t^* = \frac{\gamma^2}{4\rho_u} \frac{\lambda^2 n}{\lambda^2 s/\Lambda_{\min} + \sigma_{\max}^2(1+\delta)\frac{n-s}{n}} > K, \tag{30}$$

then for any $\delta > 0$, with probability at least

$$1 - K\exp\left(-\frac{n}{2}\left(\frac{1}{4} - \sqrt{\frac{s}{n}}\right)_+^2\right) - K\exp\left(-(1+\delta)(n-s)\left[1 - 2\sqrt{\frac{1}{1+\delta}}\right]\right)$$
$$- (p-s)\exp\left(-t^*\left[1 - 2\sqrt{\frac{K}{t^*}}\right]\right),$$

it holds that

$$\max_{j\in S^c} \|T_{j3}\|_2 < \gamma.$$

(4) Condition in Eq. 30.

If we assume that

$$n \geq Cs\log p$$

for some constant C. Then

$$t^* \geq \frac{\gamma^2}{4\rho_u} \frac{\lambda^2 n}{\lambda^2 n/(C\Lambda_{\min}\log p) + \sigma_{\max}^2(1+\delta)\frac{n-s}{n}} = \frac{\gamma^2}{4\rho_u} \frac{1}{(C\Lambda_{\min}\log p)^{-1} + \sigma_{\max}^2(1+\delta)\frac{n-s}{\lambda^2 n^2}}.$$

With the specified choice of parameter $\lambda = \sqrt{p\log p/n}$, it implies

$$t^* \geq \frac{\gamma^2}{4\rho_u} \frac{\log p}{(C\Lambda_{\min})^{-1} + \sigma_{\max}^2(1+\delta)\frac{n-s}{np}}.$$

Assume C is chosen such that $C \geq \Lambda_{\min}^{-1}\left(\frac{\gamma^2}{20\rho_u} - \sigma_{\max}^2(1+\delta)\frac{n-s}{np}\right)^{-1}$ which can be easily satisfied since $\frac{n-s}{np} < 1$. Then it implies $t^* \geq 5\log p$. To satisfy the condition in Eq. 30, it is sufficient to assume $K \leq \frac{5}{64}\log p$, which implies $K \leq \frac{1}{64}t^*$. Furthermore, it implies $\exp\left(-t^*\left[1 - 2\sqrt{\frac{K}{t^*}}\right]\right) < \exp\left(-3\frac{3}{4}\log p\right)$.

To conclude, if $\|\Delta\|_{l_\infty/l_2} < \frac{1}{2}$ and that

$$n \geq \kappa_6 s\log p, \quad K \leq \frac{5}{64}\log p,$$

then $\max_{j\in S^c}\|T_{j3}\|_2 < \gamma$ holdes with probability at least

$$1 - K\exp\left(-\frac{n}{2}\left(\frac{1}{4} - \sqrt{\frac{s}{n}}\right)_+^2\right) - K\exp\left(-(1+\delta)(n-s)\left[1 - 2\sqrt{\frac{1}{1+\delta}}\right]\right)$$
$$- (p-s)\exp\left(-3\frac{3}{4}\log p\right).$$

D.1.4 No exclusion: $\frac{\|U_S\|_{l_\infty/l_2}}{\sqrt{K}} \leq \frac{1}{2}b_{\min}^*$

Eq. 26c implies that

$$\widehat{\beta}_S^{(k)} - \beta_S^{*(k)} = \left(\widehat{\Sigma}_{SS}^{(k)}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}_S^{(k)\top}\boldsymbol{w}^{(k)} - \lambda\widehat{\boldsymbol{z}}_S^{(k)}\right).$$

Define $\overline{\boldsymbol{w}}^{(k)} := \frac{1}{\sqrt{n}}(\widehat{\Sigma}_{SS}^{(k)})^{-1/2}\boldsymbol{X}_S^{(k)\top}\boldsymbol{w}^{(k)} \overset{d.}{=} \sigma^{(k)}\xi_k$ with $\xi_k \sim \mathcal{N}(0, I_p)$. Then

$$\widehat{\beta}_S^{(k)} - \beta_S^{*(k)} \overset{d.}{=} \underbrace{(\widehat{\Sigma}_{SS}^{(k)})^{-1/2}\frac{\overline{\boldsymbol{w}}^{(k)}}{\sqrt{n}}}_{A^{(k)}} - \underbrace{(\widehat{\Sigma}_{SS}^{(k)})^{-1}\lambda\widehat{\boldsymbol{z}}_S^{(k)}}_{B^{(k)}}.$$

Denote the i-the entry in the vector $A^{(k)}$ as $A_i^{(k)}$. Then for a fixed $i \in S$, the entry $\{A_i^{(k)}\}_{k\in[K]}$ are independent. Its easy to see that

$$A_i^{(k)} \mid X^{(1:K)} \overset{d.}{=} \frac{\sigma^{(k)}}{\sqrt{n}}\sqrt{\left((\widehat{\Sigma}_{SS}^{k})^{-1}\right)_{ii}}\xi_{ik} \quad \text{with } \xi_{ik} \sim \mathcal{N}(0,1) \text{ and } \mathrm{Cov}(\xi_{ik}, \xi_{ik'}) = 0.$$

Therefore,

$$\max_{i\in S}\sum_{k=1}^{K} A_i^{(k)2} \mid X^{(1:K)} \leq \frac{\sigma_{\max}^2}{n}\max_{k\in[K]}\left\|(\widehat{\Sigma}_{SS}^{(k)})^{-1}\right\|_2 \max_{i\in S}\sum_{k=1}^{K}\xi_{ik}^2.$$

Since we have

$$\Pr\left[\max_{k\in[K]}\left\|(\widehat{\Sigma}_{SS}^{(k)})^{-1}\right\|_2 \leq \frac{2}{\Lambda_{\min}}\right] \geq 1 - K\exp\left(-\frac{n}{2}\left(\frac{1}{4} - \sqrt{\frac{s}{n}}\right)_+^2\right) \quad \text{by Lemma 10 in [18]},$$

$$\Pr\left[\max_{i\in S}\sum_{k=1}^{K}\xi_{ik}^2 \leq 4K\log p\right] \geq 1 - s\exp\left(-2K\log p\left(1 - 2\sqrt{\frac{1}{2\log p}}\right)\right) \quad \text{by Lemma H.2},$$

then with probability at least $1 - K\exp\left(-\frac{n}{2}\left(\frac{1}{4} - \sqrt{\frac{s}{n}}\right)_+^2\right) - s\exp\left(-2K\log p\left(1 - 2\sqrt{\frac{1}{2\log p}}\right)\right)$, it holds that

$$\max_{i\in S}\sqrt{\sum_{k=1}^{K} A_i^{(k)2}} \leq \sqrt{\frac{8\sigma_{\max}^2 K\log p}{\Lambda_{\min} n}}.$$

Tuning now to the term $B^{(k)}$:

$$\begin{aligned}
\max_{i\in S}\sqrt{\sum_{k=1}^{K} B_i^{(k)2}} &= \lambda\max_{i\in S}\sqrt{\sum_{k=1}^{K}\left(\boldsymbol{e}_i^\top(\widehat{\Sigma}_{SS}^{(k)})^{-1}\widehat{\boldsymbol{z}}_S^{(k)}\right)^2} \\
&\leq \lambda\max_{i\in S}\sqrt{\sum_{k=1}^{K}\|(\widehat{\Sigma}_{SS}^{(k)})^{-T}\boldsymbol{e}_i\|_2^2\|\widehat{\boldsymbol{z}}_S^{(k)}\|_2^2} \quad \text{by Cauchy-Schwarz inequality} \\
&\leq \lambda\max_{i\in S}\max_{k\in[K]}\|(\widehat{\Sigma}_{SS}^{(k)})^{-T}\boldsymbol{e}_i\|_2\sqrt{\sum_{k=1}^{K}\|\widehat{\boldsymbol{z}}_S^{(k)}\|_2^2} \leq \lambda\max_{k\in[K]}\|(\widehat{\Sigma}_{SS}^{(k)})^{-1}\|_2\sqrt{s}.
\end{aligned}$$

The last inequality holds because $\|\widehat{Z}_S\|_{l_\infty/l_2} \leq 1$. Applying Lemma 10 in [18] to $\max_{k\in[K]}\|(\widehat{\Sigma}_{SS}^{(k)})^{-1}\|_2$ again, with probability at least $1 - K\exp\left(-\frac{n}{2}\left(\frac{1}{4} - \sqrt{\frac{s}{n}}\right)_+^2\right)$, it holds that

$$\max_{i\in S}\sqrt{\sum_{k=1}^{K} B_i^{(k)2}} \leq \frac{2\lambda\sqrt{s}}{\Lambda_{\min}}.$$

To conclude, with probability at least

$$1 - 2K\exp\left(-\frac{n}{2}\left(\frac{1}{4} - \sqrt{\frac{s}{n}}\right)_+^2\right) - s\exp\left(-2K\log p\left(1 - 2\sqrt{\frac{1}{2\log p}}\right)\right), \tag{31}$$

it holds that (with specified $\lambda = \sqrt{\frac{p \log p}{n}}$)

$$\|\widehat{B}_S - B^*_S\|_{l_\infty/l_2} \leq \sqrt{\frac{8\sigma^2_{\max} K \log p}{\Lambda_{\min} n}} + \frac{2}{\Lambda_{\min}} \sqrt{\frac{sp \log p}{n}}$$

$$\implies \frac{\|U_S\|_{l_\infty/l_2}}{\sqrt{K}} \leq \sqrt{\frac{8\sigma^2_{\max} \log p}{\Lambda_{\min} n}} + \frac{2}{\Lambda_{\min}} \sqrt{\frac{sp \log p}{nK}}.$$

Therefore, if

$$\sqrt{\frac{8\sigma^2_{\max} \log p}{\Lambda_{\min} n}} + \frac{2}{\Lambda_{\min}} \sqrt{\frac{sp \log p}{nK}} \leq \frac{1}{2} b^*_{\min},$$

then Eq. 28 is satisfied with probability specified in Eq. 31.

E Invariant Sets

We often need to take a union control over all permutations $\pi \in \mathbb{S}_p$ in the proofs. However, in these steps, we often care about the connection matrices $\widetilde{G}^{(1:K)}(\pi)$ instead of the permutation π itself. Therefore, we want to see whether we can control a fewer number of events instead of enumerating over all $p!$ many permutations. Alternatively, we want to should that, given $\Sigma^{(1:K)}$, the number of elements in the set $\{\widetilde{G}^{(1:K)}(\pi) : \pi \in \mathbb{S}_p\}$ can be fewer than $p!$.

We start with the following definition which specifies the population-level quantity that we are interested in.

Definition E.1 (Population SEM). For any $S \subseteq [p] \setminus \{j\}$, let

$$g_j^{(k)}(S) := \underset{g \in \mathbb{R}^p, \text{supp}(g) \subseteq S}{\arg\min} \mathbb{E}[X_j^{(k)} - g^\top X^{(k)}]^2, \tag{32}$$

where $X^{(k)}$ is the random variable that follows $\mathcal{N}(0, \Sigma^{(k)})$.

$g_j^{(k)}(S)$ is called the SEM coefficients for variable X_j regressed on the nodes in S [15, 17]. It is a population-level quantity that depends on $\Sigma^{(k)}$, but not on the sample $\boldsymbol{X}^{(k)}$. In [15, 17], this quantity is used for a similar purpose on the single DAG estimation task. It is easy to verify that $g_j^{(k)}(S_j(\pi)) = \widetilde{g}_j^{(k)}(\pi)$. Lemma E.1 summarizes the key observations.

Lemma E.1. *Let $S_j(\pi) = \{i : \pi(i) < \pi(j)\}$ and $\widetilde{g}_j^{(k)}(\pi)$ is the j-th column of $\widetilde{G}^{(k)}(\pi)$. Then*

$$g_j^{(k)}(S) = \widetilde{g}_j^{(k)}(\pi)$$

for any set S such that

$$\textit{supp}\big(\widetilde{g}_j^{(k)}(\pi)\big) \subseteq S \subseteq S_j(\pi). \tag{33}$$

Since the set of union parents $U_j(\pi) := \cup_{k \in [K]} \textit{supp}\big(\widetilde{g}_j^{(k)}(\pi)\big)$ satisfies Eq. 33, it implies that

$$g_j^{(k)}(U_j(\pi)) = \widetilde{g}_j^{(k)}(\pi), \quad \forall k \in [K]. \tag{34}$$

A direct consequence of this Lemma E.1 is that:

$$\Big|\{\widetilde{g}_j^{(1:K)}(\pi) : \pi \in \mathbb{S}_p\}\Big| = \Big|\{\widetilde{g}_j^{(1:K)}(U_j(\pi)) : \pi \in \mathbb{S}_p\}\Big| \leq \Big|\{U_j(\pi) : \pi \in \mathbb{S}_p\}\Big|.$$

Recall that $d_j := \max_{\pi \in \mathbb{S}_p} |U_j(\pi)|$. Then there are at most $\sum_{0 \leq m \leq d_j} \binom{p}{m}$ many elements in this set. Note that

$$\sum_{0 \leq m \leq d_j} \binom{p}{m} \leq \sum_{0 \leq m \leq d} \binom{p}{m} \leq p^d.$$

The last inequality holds for all $p \geq d \geq 2$. Therefore,

$$\Big|\{\widetilde{g}_j^{(1:K)}(\pi) : \pi \in \mathbb{S}_p\}\Big| \leq \Big|\{U_j(\pi) : \pi \in \mathbb{S}_p\}\Big| \leq \sum_{0 \leq m \leq d_j} \binom{p}{m} \leq p^d. \tag{35}$$

F Details of Continuous Formulation

We first prove that $\widehat{T} \in \mathcal{T}_p$:

Lemma F.1. *If $(\widehat{G}^{(1:K)}, \widehat{T})$ is a pair of optimal solution to Eq. 14, then $\widehat{T}$ is in the discrete space $\mathcal{T}_p$. Equivalently, if $\widehat{T}$ is an optimal solution, then there exists a permutation $\hat{\pi} \in \mathbb{S}_p$ such that*

$$\widehat{T}_{ij} = \begin{cases} 1 & \text{if } \hat{\pi}(i) < \hat{\pi}(j), \\ 0 & \text{otherwise}. \end{cases} \tag{36}$$

Proof. By the constraint $h(T) = 0$ in Eq. 15, the graph structure induced by the matrix $\widehat{T}$ must be acyclic. Therefore, $\widehat{T}$ represents a DAG and has an associated topological order (causal order). Denote this order by $\hat{\pi}$. What remains is to show Eq. 36 holds true.

Assume there exists an entry (i', j') such that $\hat{\pi}(i') < \hat{\pi}(j')$ and $\widehat{T}_{i'j'} \neq 1$. We construct the following pair of $(T, G^{(1:K)})$:

$$T := \begin{cases} T_{i'j'} = 1 \\ T_{ij} = \widehat{T}_{ij} \text{ for } (i,j) \neq (i',j') \end{cases} \qquad G^{(k)} := \begin{cases} G^{(k)}_{i'j'} = \widehat{T}_{i'j'} \cdot \widehat{G}^{(1:K)}_{i'j'} \\ G^{(k)}_{ij} = \widehat{G}^{(k)}_{ij} \text{ for } (i,j) \neq (i',j') \end{cases} \quad \forall k \in [K].$$

It is constructed by modifying the (i', j')-th entries in the solution pair $(\widehat{T}, \widehat{G}^{(1:K)})$. It is easy to see that: (i) this constructed pair $(T, G^{(1:K)})$ is a feasible solution to Eq. 14; and (ii) $(T, G^{(1:K)})$ achieves a smaller objective value than $(\widehat{T}, \widehat{G}^{(1:K)})$.

The reason for (ii) is that after the modification, the matrix $\overline{G}^{(1:K)}$ remains unchanged. That is, $\widehat{T} \circ \widehat{G}^{(k)} = T \circ G^{(k)}$. Therefore, the squared loss and the group-norm in the objective remain unchanged. However, the term $\|\mathbf{1}_{p\times p} - T\|_F^2$ has been reduced by setting $T_{i'j'} = 1$.

This makes a contradiction to the optimality of $(\widehat{T}, \widehat{G}^{(1:K)})$. Therefore, the assumption is not true and we conclude that:

$$\hat{\pi}(i) < \hat{\pi}(j) \Longrightarrow \widehat{T}_{ij} = 1.$$

Finally, since $\widehat{T}$ is consistent with $\hat{\pi}$, by definition, $\widehat{T}_{ij} = 0$ if $\hat{\pi}(i) \geq \hat{\pi}(j)$. □

We now start to show the equivalence between the optimization in Eq. 7 and in Eq. 14.

Firstly, the solution search spaces are the same. We have shown that $\widehat{T} \in \mathcal{T}_p$. For each element in $\mathcal{T}_p$, we denote it by $\widehat{T}(\pi)$ based on its associated order π. Since $\widehat{T}(\pi)$ is a *dense* DAG with topological order π, it is easy to see the space $\{G \circ \widehat{T}(\pi) : G \in \mathbb{R}^{p\times p}\}$ includes all DAGs that are consistent with $\hat{\pi}$ and excludes any DAGs that are not. In other words, $\{G \circ \widehat{T}(\pi) : G \in \mathbb{R}^{p\times p}\} = \mathbb{D}(\pi)$. Therefore, the solution search spaces of these two optimization problems are equivalent.

Secondly, the optimization objectives are the same. Again, since $\widehat{T} \in \mathcal{T}_p$, the term $\rho\|\mathbf{1}_{p\times p} - \widehat{T}\|_F^2$ is a constant with a fixed value $\frac{\rho(p-1)p}{2}$. The remaining two terms in the objective are the same as the objective in Eq. 7.

Since both the solution search space and the optimization objectives are equivalent, these two optimizations are equivalent.

G Details of Synthetic Experiments in Sec 6.1

G.1 Evaluation of structure prediction

We classify the positive predictions in three types:

- True Positive: predicted association exists in correct direction.
- Reverse: predicted association exists in opposite direction.

- False Positive: predicted association does not exist

Based on them, we use five metrics:

- False Discovery Rate (FDR): (reverse + false positive) / (true positive + reverse + false positive)
- True Positive Rate (TPR): (true positive) / (ground truth positive)
- False Positive Rate (FPR): (false positive) / (ground truth positive)
- Structure Hamming Distance (SHD): (false negative + reverse + false positive)
- Number of Non-Zero (NNZ): (true positive + reverse + false positive)

G.2 A more complete result for Fig 3

We demonstrate our methods on synthetic data with $(p, s) \in \{(32, 40), (64, 96), (128, 224), (256, 512)\}$, $K \in \{1, 2, 4, 8, 16, 32\}$, $n \in \{10, 20, 40, 80, 160, 320\}$. For each $\{p, s, K, n\}$, we run experiments on 64 graphs. We report the full results in Fig.5.

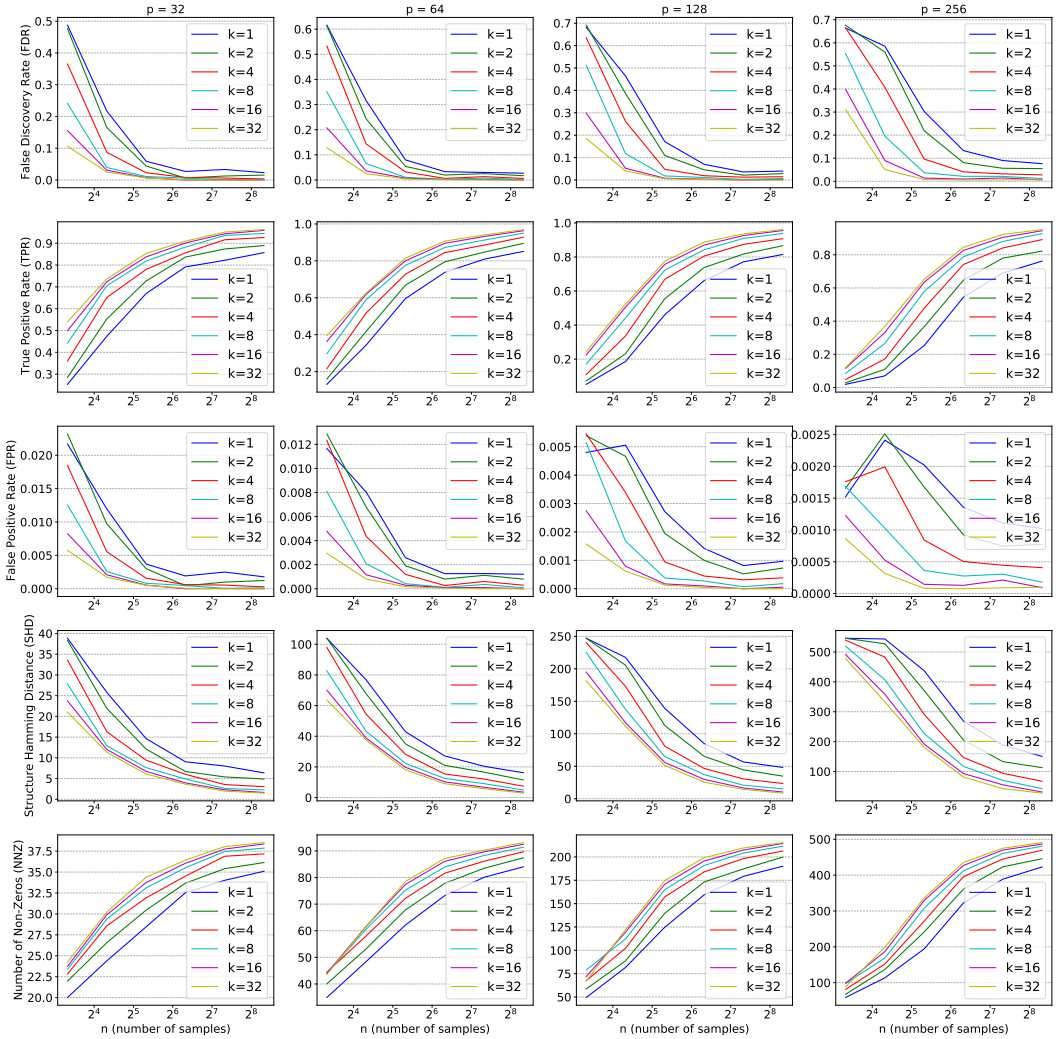

Figure 5: Full results of FDR, TPR, FPR, SHD, and NNZ.

G.3 Computing resources

Since we need to run a large set of experiments spanning different values of $\{p, s, K, n\}$, the synthetic experiments are run on a CPU cluster containing 416 nodes. On each node, there are 24 CPUs (Xeon 6226 CPU @ 2.70GHz) with 192 GB memory. Each individual experiment is run on 4 CPUs. It takes about 10 hours to finish a complete set of experiments on about 400 CPUs.

H Useful Results In Existing Works

Lemma H.1 (Laurent-Massart)**.** *Let $a_1, \cdots, a_m$ be nonnegative, and set*

$$\|a\|_\infty = \sup_{i\in[m]} |a_i|, \quad \|a\|_2^2 = \sum_{i=1}^m a_i^2.$$

For i.i.d $Z_i \sim \mathcal{N}(0,1)$, the following inequalities hold for any positive t:

$$\Pr\left[\sum_{i=1}^m a_i(Z_i^2 - 1) \geq 2\|a\|_2\sqrt{t} + 2\|a\|_\infty t\right] \leq e^{-t},$$

$$\Pr\left[\sum_{i=1}^m a_i(Z_i^2 - 1) \leq -2\|a\|_2 t\right] \leq e^{-t}.$$

Lemma H.2. *[18] Let Z be a central Chi-squared distributed random variable with the degree m. Then for all $t > m$, we have*

$$\Pr[Z \geq 2t] \leq \exp\left(-t\left[1 - 2\sqrt{\frac{m}{t}}\right]\right).$$

Lemma H.3. *[18] Consider the matrix $\Delta \in \mathbb{R}^{s\times K}$ with rows $\Delta_i := (\widehat{B}_i - B_i^*)/\|B_i^*\|_2$. If $\|\Delta\|_{l_\infty/l_2} < \frac{1}{2}$, then $\|\widehat{Z}_S - Z_S^*\|_{l_\infty/l_2} \leq 4\|\Delta\|_{l_\infty/l_2}$.*