Multi-task Learning of Order-Consistent Causal Graphs

Anonymous Author(s)
Affiliation
Address
email

Abstract

We consider the problem of discovering K related directed acyclic graph (DAG) structures, where the involved graph nodes share a consistent causal order and sparse unions of supports. Under this multi-task learning setting, we formulate the joint recovery problem as an l_1/l_2-regularized maximum likelihood estimation (MLE) for K linear structural equation models. We 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. Furthermore, 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 analyzed 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, since the search space is combinatorial, i.e., the number of possible DAG structures increases super-exponentially in the number of random variables, DAG estimation remains a very challenging problem from both the *computational* and *statistical* aspects [7].

Very often multiple related DAG structures need to be estimated from data. 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 to leverage their similarities.

Although a line of research on *undirected* graphical models has nicely characterized how multi-task learning can reduce the sample complexity for structure estimation by leveraging data across related tasks [13, 14, 15, 16, 17, 18], the joint DAG estimation problem is more challenging and less understood, largely due to the more complex combinatorial space for DAG. Thus, progress still needs to be made to characterize how exactly multi-task DAG estimation can benefit from related datasets.

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

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

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 very 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 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.

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 $p!$.

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 [19] 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 [20, 21, 22, 23]. 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 and identifiability. 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 $[p]$. We say $\widetilde{G}_0$ is **consistent** with π_0 if and only if

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

Note that if simultaneous row and column permutations are made according to π_0, then $\widetilde{G}_0$ can be permuted to be strictly lower triangular [21]. 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. However, estimating the oracle 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 can also be written as [21]

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

Furthermore, the random variable X in Eq. 1 can be equivalently described by: $X = \widetilde{G}(\pi)^\top X + W(\pi)$, where $W(\pi) \sim \mathcal{N}(0, \Omega(\pi))$. The pair $(\widetilde{G}_0, \Omega_0)$ will correspond to $(\widetilde{G}(\pi_0), \Omega(\pi_0))$ for any $\pi_0 \in \Pi_0$. Without further assumption, $\widetilde{G}_0 = \widetilde{G}(\pi_0)$ **cannot** be identified from the equivalent class

$\{\widetilde{G}(\pi) : \pi \in \mathbb{S}_p\}$ based on the distribution of X in general. Therefore, in the causal DAG learning regime, the noise variance Ω_0 is assumed to have equal diagonal entries: $\Omega_0 = \mathcal{N}(0, \sigma^2 I_p)$, which ensures that $\widetilde{G}_0$ is identifiable from Σ [24]. More precisely, $\widetilde{G}_0$ will be the minimum-trace DAG [21].

3 Joint estimation of multiple DAGs

In the multi-task setting, we consider K linear SEMs: $X^{(k)} = \widetilde{G}_0^{(k)\top} X^{(k)} + W^{(k)}$, for $k = 1, \cdots, K$. In this paper, the superscript notation $^{(k)}$ indicates the k-th model. As mentioned in Sec 2, each of such model defines a random variable $X^{(k)} \sim \mathcal{N}(0, \Sigma^{(k)})$ with $\Sigma^{(k)} = (1 - \widetilde{G}_0^{(k)})^{-\top} \Omega_0^{(k)} (1 - \widetilde{G}_0^{(k)})^{-1}$. We make a few assumptions on the identifiability of the DAGs and their similarity:

Condition 3.1 (Equal variance). For all $k = 1, \cdots, K$, the noise $W^{(k)} \sim \mathcal{N}(0, \Omega_0^{(k)})$ with $\Omega_0^{(k)} = I_p$.

Condition 3.2 (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 $\widetilde{G}_{0ij}^{(k)} \neq 0 \Rightarrow \pi_0(i) < \pi_0(j)$.

Definition 3.1 (Support union). Recall $\widetilde{G}(\pi)$ in Eq. 4. The support union of the K DAGs associated with permutation π is denoted by $S(\pi) := \big\{(i,j) : \exists k \in [K]\ s.t.\ \widetilde{G}_{ij}^{(k)}(\pi) \neq 0\big\}$. 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 $s := \sup_{\pi \in \mathbb{S}_p} |S(\pi)|$.

Equal variance condition is required to ensure the **identifiability** of $\widetilde{G}_0^{(k)}$ from $\Sigma^{(k)}$. We assume all variances are 1 for simplicity, but results in this paper can be extended to the case with $\Omega_0^{(k)} = \sigma^{(k)2} I_p$.

The **similarity** among the K tasks are described in Condition 3.2 and Definition 3.1. Condition 3.2 ensures the causal orders in the K DAGs do not have conflicts. Definition 3.1 can be interpreted as 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.

3.1 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 $\{\widetilde{G}_0^{(k)}\}_{k=1}^K$:

$$\hat{\pi}, \{\widehat{G}^{(k)}\} = \underset{\pi \in \mathbb{S}_p,\, \{G^{(k)} \in \mathbb{D}(\pi)\}}{\arg\min} \sum_{k=1}^{K} \frac{1}{2n} \|\boldsymbol{X}^{(k)} - \boldsymbol{X}^{(k)} G^{(k)}(\pi)\|_F^2 + \lambda \|G^{(1:K)}(\pi)\|_{l_1/l_2}. \quad (5)$$

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. 5, which respects Condition 3.2. The group norm over the set of K matrices is defined as

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

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. 5 is used for analysis only. A continuous program with the same optimizer will be discussed in Section 4 for practical implementation.

3.2 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

$$d := \sup_{k \in [K], j \in [p], \pi \in \mathbb{S}_p} \|\widetilde{g}_j^{(k)}(\pi)\|_0, \quad g_{\max} := \sup_{\pi \in \mathbb{S}_p, (i,j) \in S(\pi)} \|\widetilde{G}_{ij}^{(1:K)}(\pi)\|_2 / \sqrt{K}.$$

Condition 3.3 (Bounded spectrum). Assume for all $k = 1, \cdots, K$, the covariance matrix $\Sigma^{(k)}$ is positive definite. There exists constants $0 < \Lambda_{\min} \leq \Lambda_{\max} < \infty$ 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 3.4 (Omega-min). There exists a constant $\eta_w > 0$ such that for any permutations $\pi \notin \Pi_0$,

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

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

Condition 3.4 with $K = 1$ is called 'omega-min' condition in [20], 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 $\sigma_j(\pi_0) = 1$ 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. 5 with specified regularization parameter $\lambda = \sqrt{\frac{p\log p}{n}}$ for a set of K problems that satisfy Condition 3.1, 3.2, 3.3 (a) and 3.4. If the following conditions are satisfied:*

$$\theta(n, K, p, s) := \frac{p}{s}\sqrt{\frac{nK}{p\log p}} > \kappa_1\eta_w, \tag{6}$$

$$n \geq c_1 \log K + c_2(d+1)\log p, \tag{7}$$

$$K \leq \kappa_2 p \log p, \tag{8}$$

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

$$\hat{\pi} \in \Pi_0 \quad \textit{and} \quad \hat{s} = \mathcal{O}(s_0),$$

where $\hat{s} := \left|\{(i,j) : \exists k \; s.t. \; \widehat{G}_{ij}^{(k)} \neq 0\}\right|$ is the sparsity of the estimated support union.

(b) If in addition, n satisfies $n \geq \kappa_5 \hat{d}(\log K + \log p)$ with $\hat{d} := \max_{j\in[p],k\in[K]} \|\widehat{g}_j^{(k)} - \widetilde{g}_j^{(k)}\|_0$, and Condition 3.3 (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 $n, p, K, s, \Sigma^{(k)}$), $\kappa_2, \kappa_1, \kappa_3, \kappa_4$ are constants depending on $g_{\max}$ and $\Lambda_{\max}$, and κ_5 is a constant depending on $\Lambda_{\min}$.

Discussion on Eq. 6-Eq. 8. (i) In Eq. 6, Theorem 3.1 identifies a *sample complexity parameter* $\theta(n, K, p, s)$. Following the terminology in [25], 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, p, s) such that $\theta(n, K, p, s)$ 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. 7 is the sample requirement for an ordinary linear regression problem. It is in general much weaker than Eq. 6, unless K grows very large.

(iii) The last condition in Eq. 8 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 [25, 26] or assume $K = o(\log p)$ [27, 28].

Effective sample size. As indicated by $\theta(n, K, p, s)$ in Eq. 6, the effective sample size is nK if the support union size s is of order $\mathcal{O}(1)$ in K. An extreme example of $s = \mathcal{O}(1)$ in K is when all DAGs

share the same support set. In fact, the parameter $\theta(n, K, p, s)$ indicates a lower sample complexity relative to separate estimation as long as s is of order $o(\sqrt{K})$. However, if K grows so large that $(\kappa_1 \eta_w s)^2 \log p/(pK) < c_1 \log K + c_2(d+1)\log p$, Eq. 7 will dominate the sample complexity because now the linear regression error becomes the bottleneck of this estimation problem.

Separate estimation. Consider the special case of a single task estimation with $K = 1$, 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. [20] and [29] are the most comparable ones since they were based on the same omega-min condition as ours (i.e., Condition 3.4), but they chose a smaller regularization parameter λ. In our proof, the sample complexity parameter is derived from $\theta(n, K, p, s) = p\sqrt{K}/(s\lambda)$ and the condition in Eq. 6 is $p\sqrt{K}/(s\lambda) > \kappa_1 \eta_w$. When $K = 1$, this condition matches what is identified in [20] and [29] for recovering the order of a single DAG.

Sparsity and 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 is noticeable that the additional condition on n in Theorem 3.1 (b) is slightly stronger or even weaker than Eq. 7, because the result $\hat{s} = \mathcal{O}(s_0)$ from (a) suggests $\widehat{G}^{(1:K)}$ is sparse, which in turns implies $\hat{d}$ should not be large.

To summarize, Theorem 3.1 analyzes the causal order consistency for the joint estimator in Eq. 5. 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.

3.3 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 [25, 26], 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 $\text{RS}_j := \{i \in [p] : \exists k \in [K]\ s.t.\ \widetilde{G}_{0ij}^{(k)} \neq 0\}$. The maximal cardinality is $r_{\max} := \sup_{j\in[p]} \left|\text{RS}_j\right|$.

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

Condition 3.5 (Self-incoherence). $\sup_{k\in[K], j\in[p], S=\text{RS}(\widetilde{g}_{0j}^{(1:K)})} \|(\Sigma^{(k)}_{SS})^{-1}\|_\infty \leq D_{\max} < \infty$.

Condition 3.6 (Irrepresentable condition). There exists a fixed parameter $\gamma \in (0, 1]$ such that $\sup_{j\in[p], S=\text{RS}(\widetilde{g}_{0j}^{(1:K)})} \|A(S)\|_\infty \leq 1 - \gamma$, 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 3.5 and 3.6. If the following conditions are satisfied*

$$n \geq \kappa_6 r_{\max} \log p, \tag{9}$$

$$K \leq \kappa_7 \log p, \tag{10}$$

$$\sqrt{\frac{8\Lambda_{\max}\log p}{C_{\min} n}} + \lambda\left(\frac{D_{\max}}{\sqrt{K}} + \frac{12 r_{\max}}{C_{\min}\sqrt{nK}}\right) = o(g_{\min}), \tag{11}$$

then with probability at least $1 - c_5 K \exp\left(-(K-1)\log p\right) - \exp\left(-c_6 \log p\right)$, the support union of $\widehat{G}^{(1:K)}$ is the same as the support union of $\widetilde{G}_0^{(1:K)}$, and that $\left\|\widehat{G} - \widetilde{G}\right\|_{l_\infty/l_2}/\sqrt{K} = o(g_{\min})$.

Discussion on Eq. 9 and Eq. 10. *(i)* Eq. 9 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. 10 plays a similar role as Eq. 8 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 [25, 26]. 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. 11 plays a similar role as Eq. 6 in Theorem 3.1. It specifies a rate at which the sample size must grow for successful union support recovery. K appears in two terms in the left hand side of Eq. 11. Apparently, a larger K can decrease the first term $D_{\max}/\sqrt{K}$. Besides, as long as $r_{\max}$ is of order $o(\sqrt{K})$, K will also be effective in the second term $12r_{\max}/(C_{\min}\sqrt{nK})$. Apart from that, the recover probability specified in Theorem 3.2 grows in K.

4 Algorithm

Solving the optimization in Eq. 5 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 [22], we propose a continuous optimization problem, whose optimizer is the same as the estimator in Eq. 5. Furthermore, we will design an efficient iterative algorithm to approximate the solution.

4.1 Continuous program

We convert Eq. 5 to the following constrained continuous program

$$\min_{\substack{T \in \mathbb{R}^{p\times p} \\ G^{(1)},\cdots,G^{(K)} \in \mathbb{R}^{p\times p}}} \sum_{k=1}^{K} \frac{1}{2n} \left\| X^{(k)} - X^{(k)} \overline{G}^{(k)} \right\|_F^2 + \lambda \|\overline{G}^{(1:K)}\|_{l_1/l_2} + \rho \|\mathbf{1}_{p\times p} - T\|_F^2 \quad (12)$$

$$\text{subject to} \quad h(T) := \text{trace}(e^{T\circ T}) - p = 0, \quad (13)$$

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. 13 is a smooth 'DAG-ness' constraint proposed by NOTEARS [22], which ensures T is acyclic. One can also use $h(T) := \text{trace}((I + T \circ T)^p) - p$ proposed in [30].

We would like to highlight the novel and interesting design of the matrix T in Eq. 12. What makes Eq. 5 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 realize this by a single matrix T.

$$\pi = (1,2,3,4) \qquad \pi = (4,2,1,3)$$

$$\Updownarrow \qquad\qquad \Updownarrow$$

$$T = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix} \quad T = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

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 allow 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:

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

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. 12.

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 [30]`

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. 5 and Eq. 12 are in Appendix G.

4.2 Iterative algorithm

We derive an efficient iterative algorithm using the Lagrangian method with quadratic penalty, which converts Eq. 12 to an unconstrained problem:

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

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

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:

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

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. 12 in Algorithm 1.

5 Related work

Single DAG estimation. Unlike the large literature of research on undirected graphical models [31, 32, 33, 34], statistical guarantees for score-based DAG estimator have been available only in recent years. [20, 21, 23] have shown the DAG estimation consistency in high-dimensions, but they do not consider joint estimation. Nevertheless, some techniques in [20, 21] are useful for our derivations.

Practical Algorithms. Works on practical algorithm design for efficiently solving the score-based optimization are more actively conducted [35, 36, 37, 38]. Our algorithm is most related to recent methods exploiting a smooth characterization of acyclicity, including NOTEARS [22] and several subsequent works [30, 39, 40, 41], but they only apply for single-task DAG estimation. Although algorithms for the multi-task counterpart were proposed a decade ago [11, 42, 43, 44], none of them leverage recent advances in characterizing DAGs and providing theoretically guarantees.

Multi-task learning. There have been extensive studies on the joint estimation of multiple undirected Gaussian graphical models [13, 14, 15, 16, 17, 18, 45, 46, 47, 48, 49]. 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, 27, 28], but [27, 28] tackle the fewer task regime with $K = o(\log p)$, and [9] assumes the causal order is given. Multi-task linear regression are also related [50, 51, 52, 53, 54, 55, 56], especially the techniques for analyzing group Lasso [25, 26].

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.

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.

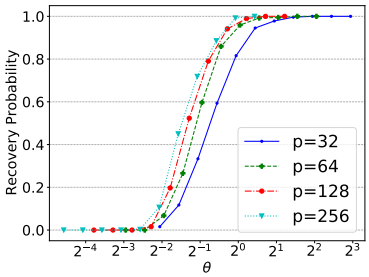

Figure 1: Success probability vs $\log \theta$.

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.

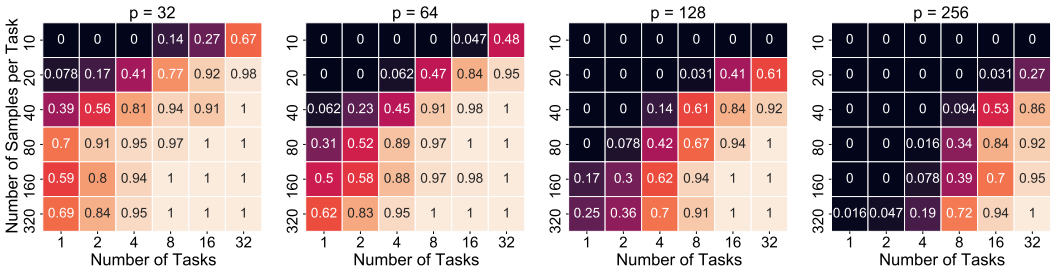

Figure 2: Heat map: Darker colors indicate lower success probability, and lighter colors are higher.

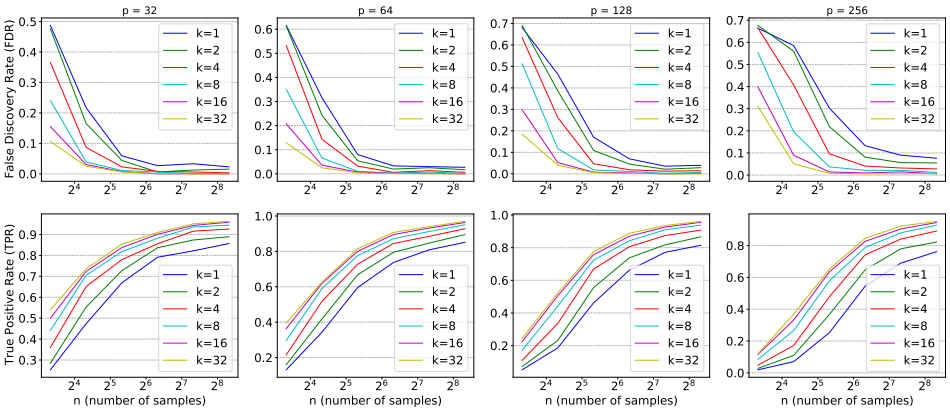

Figure 3: False discovery rate (FDR) and true positive rate (TPR) of the edges.

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 H. 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.

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 [19], which models the additive effect of cooperative transcription factor binding across multiple gene regulators in parallel with protein degradation and noisy expression.

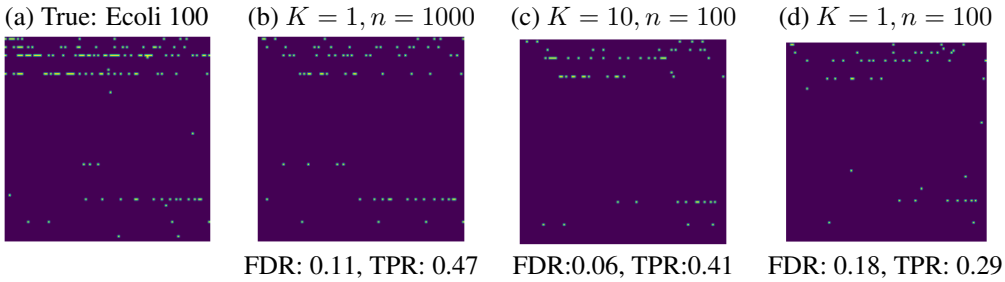

Figure 4: Visualization of the recovered DAG structures. Each light green colored pixel at position (i, j) indicates an edge from node i to j.

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 1, 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 1: 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 $\theta(n, K, p, s)$ in Eq. 6. 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.

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

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

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

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

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

[22] 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.

[23] 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.

[24] Jonas Peters and Peter Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

[30] 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.

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

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

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

[34] 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.

[35] 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.

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

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

[38] 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.

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

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

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

[46] 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.

[47] 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.

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

[49] 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.

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

[51] 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.

[52] 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.

[53] 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.

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

[55] 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.

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

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] . We did not anticipate any potential negative impacts of our work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]

(b) Did you include complete proofs of all theoretical results? [Yes]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]