High-Dimensional Joint Estimation of Multiple Directed Gaussian Graphical Models

Yuhao Wang*

Statistical Laboratory, University of Cambridge, Cambridge, UK
e-mail: yw505@cam.ac.uk

Santiago Segarra*

Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA
e-mail: segarra@rice.edu

Caroline Uhler

Laboratory for Information & Decision Systems and Institute for Data, Systems, and Society
Massachusetts Institute of Technology, Cambridge, MA, USA
e-mail: cuhler@mit.edu

Abstract: We consider the problem of jointly estimating multiple related directed acyclic graph (DAG) models based on high-dimensional data from each graph. This problem is motivated by the task of learning gene regulatory networks based on gene expression data from different tissues, developmental stages or disease states. We prove that under certain regularity conditions, the proposed ℓ_0-penalized maximum likelihood estimator converges in Frobenius norm to the adjacency matrices consistent with the data-generating distributions and has the correct sparsity. In particular, we show that this joint estimation procedure leads to a faster convergence rate than estimating each DAG model separately. As a corollary, we also obtain high-dimensional consistency results for causal inference from a mix of observational and interventional data. For practical purposes, we propose *jointGES* consisting of Greedy Equivalence Search (GES) to estimate the union of all DAG models followed by variable selection using lasso to obtain the different DAGs, and we analyze its consistency guarantees. The proposed method is illustrated through an analysis of simulated data as well as epithelial ovarian cancer gene expression data.

MSC 2010 subject classifications: Primary 62F12; secondary 62F30.
Keywords and phrases: Causal inference, linear structural equation model, high-dimensional statistics, graphical model.

Contents

*At the time this research was completed, Yuhao Wang and Santiago Segarra were at the Massachusetts Institute of Technology.

1. Introduction

Directed acyclic graph (DAG) models, also known as Bayesian networks, are widely used to model causal relationships in complex systems across various fields such as computational biology, epidemiology, sociology, and environmental management [1, 12, 31, 36, 40]. In these applications we often encounter *high-dimensional* datasets where the number of variables or nodes greatly exceeds the number of observations. While the problem of structure identification for undirected graphical models in the high-dimensional setting is quite well understood [35, 26, 11, 5, 46], such results are just starting to become available for directed graphical models. The difficulty in identifying DAG models can be attributed to the fact that searching over the space of DAGs is NP-complete in general [6].

Methods for structure identification in directed graphical models can be divided into two categories and hybrids of these categories. *Constraint-based methods*, such as the prominent PC algorithm, first learn an undirected graph from conditional independence relations and in a second step orient some of the edges [15, 40]. *Score-based methods*, on the other hand, posit a scoring criterion for each DAG model, usually a penalized

likelihood score, and then search for the network with the highest score given the observations. An example is the celebrated Greedy Equivalence Search (GES) algorithm, which can be used to greedily optimize the ℓ_0-penalized likelihood such as the Bayesian Information Criterion (BIC) [7]. High-dimensional consistency guarantees were recently obtained for the PC algorithm [20] and for score-based methods [24, 29, 45].

Existing methods have focused on estimating a single directed graphical model. However, in many applications we have access to data from related classes, such as gene expression data from different tissues, cell types or states [25, 38], different developmental stages [3], different disease states [42], or from different perturbations such as knock-out experiments [9]. In all these applications, one would expect that the underlying regulatory networks are similar to each other, since they stem from the same species, individual or cell type, but also have important differences that drive differentiation, development or a certain disease. This raises an important statistical question, namely how to jointly estimate related directed graphical models in order to effectively make use of the available data.

Various methods have been proposed for jointly estimating *undirected* Gaussian graphical models. To preserve the common structure, Guo et al. [16] suggested to use a hierarchical penalty and Danaher et al. [8] suggested the use of a generalized fused lasso or group lasso penalty. While both approaches achieve the same convergence rate as the individual estimators, Cai et al. [4] were able to improve the asymptotic convergence rate of joint estimation using a weighted constrained ℓ_∞/ℓ_1 minimization approach. Bayesian methods have been proposed for this problem as well [33]. Related works also include [28], where it is assumed that the networks differ only locally in a few nodes and [22, 39], where the assumption is that the networks are ordered and related by continuously changing edge weights.

In this paper, we propose a framework based on ℓ_0-penalized maximum likelihood estimation for jointly estimating related *directed* Gaussian graphical models. We show that the joint ℓ_0-penalized maximum likelihood estimator (MLE) achieves a faster asymptotic convergence rate as compared to the individual estimators. In addition, by viewing interventional data as data coming from a related network, we show that the interventional BIC scoring function proposed in [17] can be obtained as a special case of the joint ℓ_0-penalized maximum likelihood approach presented here. Our theoretical consistency guarantees also explain the empirical findings of [17], namely that estimating a DAG model from interventional data usually leads to better recovery rates as compared to estimating a DAG model from the same amount of purely observational data. These theoretical results are based on the global optimum of ℓ_0-penalized maximum likelihood estimation. To overcome the computational bottleneck of this optimization problem we propose a greedy approach (*jointGES*) for solving this problem by extending GES [7] to the joint estimation setting. We analyze its properties from a theoretical point of view and test its performance on synthetic data and gene expression data from epithelial ovarian cancer.

The remainder of this paper is structured as follows. In Section 2, we review some relevant background related to DAG models and introduce notation for the joint DAG estimation problem studied in this paper. In Section 3, we present the joint ℓ_0-penalized maximum likelihood estimator and jointGES, an adaptation of GES for solving this optimization problem. Section 4 establishes results regarding the statistical consistency

of the ℓ_0-penalized MLE and jointGES. Section 5 presents the implications for learning DAG models from a mix of observational and interventional data. In Section 6, we illustrate the performance of our proposal in a simulation study and an application to the analysis of gene expression data. We conclude with a short discussion in Section 7. The proofs of supporting results are contained in the Appendix.

2. Preliminaries

In Section 2.1 we introduce DAG models, in particular linear structural equation models, and discuss statistical features enjoyed by random vectors following these models. In Section 2.2 we briefly review existing approaches for estimating a *single* directed graphical model from observational data. Finally, Section 2.3 describes a setting where *multiple* related directed graphical models exist.

2.1. Directed acyclic graphs and linear structural equation models

Let $\mathcal{G} = (V, E)$ denote a DAG with vertices $V = [p] = \{1, \cdots, p\}$ and directed edges $E \subseteq V \times V$, where $|\mathcal{G}|$ denotes the cardinality of E. Let $A \in \mathbb{R}^{p\times p}$ be the adjacency matrix specifying the edge weights of the underlying DAG $\mathcal{G}$, i.e., $A_{ij} \neq 0$ if and only if $(i, j) \in E$. Also, let $\epsilon \sim \mathcal{N}(0, \Omega)$ denote a p-dimensional multivariate Gaussian random variable with zero mean and diagonal covariance matrix Ω. In this work, we assume that the observed random vector $X = (X_1, \cdots, X_p) \in \mathbb{R}^p$ is generated according to the following linear structural equation model (SEM).

$$X = A^T X + \epsilon. \tag{1}$$

Hence X follows a multivariate Gaussian distribution with zero mean and covariance matrix Σ, where the inverse covariance (or *precision*) matrix $\Theta = \Sigma^{-1}$ is given by

$$\Theta = (I - A)\Omega^{-1}(I - A)^T. \tag{2}$$

Let $\mathrm{Pa}_j(\mathcal{G})$ denote the parents of node j in $\mathcal{G}$;then it follows from (1) that the distribution of X factorizes as

$$\mathbb{P}(X) = \prod_{j=1}^{p} \mathbb{P}(X_j | X_{\mathrm{Pa}_j(\mathcal{G})}).$$

Such a factorization of $\mathbb{P}$ according to $\mathcal{G}$ is equivalent to the *Markov assumption* with respect to $\mathcal{G}$ [23, Theorem 3.27]. Formally, given $j, k \in V$ and an arbitrary subset of nodes $S \subset V \setminus \{j, k\}$, then

$$j \text{ is d-separated from } k \mid S \text{ in } \mathcal{G} \quad \Rightarrow \quad X_j \perp\!\!\!\perp X_k | X_S \text{ in } \mathbb{P}. \tag{3}$$

If the implication (3) holds bidirectionally, then $\mathbb{P}$ is said to be *faithful* [40] with respect to $\mathcal{G}$. Note that there exist DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ that encode the same d-separations and hence the same conditional independence relations. Such DAGs are said to belong to the same *Markov equivalence class*.

A consequence of the acyclicity of $\mathcal{G}$ is that there exists at least one permutation π of $[p]$ such that $A_{ij} = 0$ for all $\pi(i) \geq \pi(j)$. Putting it differently, if the rows and columns of A are reordered according to π, then the resulting matrix is strictly upper triangular. Hence, if such a permutation π is known a priori, one can obtain the SEM parameters (A, Ω) from Θ according to the following steps [cf. (2)]. First, we reorder Θ according to π. Then, we perform on the reordered Θ an upper-triangular-plus-diagonal Cholesky decomposition to obtain (A', Ω'). Finally, we revert the ordering by permuting the rows and columns of A' and Ω' according to π^{-1} and obtain the sought (A, Ω). For an arbitrary permutation π and a given Θ, we denote by (A_π, Ω_π) the Cholesky decomposition parameters obtained from the procedure just described. Alternatively, one can obtain (A_π, Ω_π) by solving p linear regressions [cf. (1)]. More precisely, we can obtain each column of A_π by regressing X_j only on those X_i such that $\pi(i) < \pi(j)$ for all j. Once A_π is obtained, one can estimate the variance of ϵ in (1) to get Ω_π. In the remainder of the paper, we denote by (A_0, Ω_0) and Θ_0 the true parameters of the data-generating SEM and the associated precision matrix, respectively. Moreover, we denote by $(A_{0\pi}, \Omega_{0\pi})$ the SEM parameters obtained from the described procedure when the true precision matrix Θ_0 is used. Notice that $(A_0, \Omega_0) = (A_{0\pi}, \Omega_{0\pi})$ if π is *any* permutation consistent with the true underlying DAG $\mathcal{G}_0$. The DAG $\mathcal{G}_\pi$ corresponding to the non-zero entries of A_π is known as the *minimal* I-MAP (independence map) with respect to π. The minimal I-MAP with the fewest number of edges is called minimal-edge I-MAP [45]. If $\mathbb{P}$ is faithful with respect to a DAG $\mathcal{G}$, then $\mathcal{G}$ is a minimal-edge I-MAP of $\mathbb{P}$ [34, 45].

Furthermore, it has been shown in [32] that all DAGs in a Markov equivalence class share the same *skeleton* – i.e., the set of edges when directions are ignored – and *v-structures*. A v-structure is a triplet $(j, k, \ell) \subseteq V$ such that $(j, k), (\ell, k) \in E$ but j and ℓ are not connected in either direction. This motivates the representation of a Markov equivalence class as a *completely partially directed acyclic graph* (*CPDAG*), which is a graph containing both directed and undirected edges [2]. A directed edge means that all DAGs in the Markov equivalence class share the same direction for this edge whereas an undirected edge means that both directions for that specific edge are present within the class. In the same way, one can represent a subset of a Markov equivalence class via a *partially directed acyclic graph* (*PDAG*), where the directions of the edges are only determined by the graphs within the subset. In particular, some undirected edges in a CPDAG would become directed edges in a PDAG representing a subset of the class. Notice that both DAGs and CPDAGs are special cases of PDAGs, where the former represents a single graph and the latter represents the whole equivalence class.

To consistently estimate causal DAG models in high dimensions, the ℓ_0-penalized maximum likelihood estimation approach [45], the high-dimensional PC method [20] and the ARGES method [29] have been proposed. These methods have high-dimensional guarantees under different conditions, and are thus not directly comparable. In particular, the theoretical guarantees of ℓ_0-penalized maximum likelihood estimation requires the so-called "beta-min" condition [45]; the high-dimensional PC algorithm requires the "strong faithfulness" condition [20]; and ARGES requires the "strong faithfulness" condition as well as additional conditions. For further discussions on the strength of different conditions, especially the "beta-min" and "strong faithfulness" conditions, we refer the readers to Remark 4.10 and [45, Section 4.3.2].

2.2. ℓ_0-penalized maximum likelihood estimation for a single DAG model

We denote by $\hat{X} \in \mathbb{R}^{n\times p}$ the observed data, where each row of $\hat{X}$ represents a realization of the random vector X. We say that we are in the *low-dimensional* setting if asymptotically p remains a constant as $n \to \infty$. By contrast, whenever $p \to \infty$ as $n \to \infty$, we say that we are in the *high-dimensional* setting. Assuming faithfulness, Chickering [7] shows that GES outputs a consistent estimator in the low-dimensional setting by optimizing the following objective – also known as the Bayesian information criterion (BIC) –

$$(\hat{A}, \hat{\Omega}) := \underset{A\in\mathcal{A},\, \Omega\in\mathcal{D}_+}{\operatorname{argmax}} \ell_n(\hat{X}; A, \Omega) - \lambda^2\|A\|_0, \tag{4}$$

where $\lambda^2 = \frac{1}{2}\frac{\log n}{n}$, $\mathcal{A}$ denotes the set of all valid adjacency matrices associated with DAGs, $\mathcal{D}_+$ is the set of non-negative diagonal matrices, and ℓ_n is the likelihood function

$$\begin{aligned}\ell_n(\hat{X}; A, \Omega) \;:=\; &-\text{trace}\left(\frac{\hat{X}^T\hat{X}}{n}\cdot(I-A)\Omega^{-1}(I-A)^T\right)\\ &+\log\det\left((I-A)\Omega^{-1}(I-A)^T\right).\end{aligned} \tag{5}$$

In the high-dimensional setting, van de Geer and Bühlmann [45] give consistency guarantees for the global optimum of (4) when the collection $\mathcal{A}$ is further constrained to contain only adjacency matrices with at most d incoming edges for each node, where $d = \mathcal{O}(n/\log p)$. More precisely, they show that there exists some parameter $\lambda^2 \asymp \frac{\log p}{n}$ such that the optimum $(\hat{A}, \hat{\Omega})$ in (4) converges in Frobenius norm to $(A_{0\hat{\pi}}, \Omega_{0\hat{\pi}})$ for increasing n and p, where $\hat{\pi}$ is a permutation consistent with $\hat{A}$, i.e.,

$$\|\hat{A} - A_{0\hat{\pi}}\|_F^2 + \|\hat{\Omega} - \Omega_{0\hat{\pi}}\|_F^2 = \mathcal{O}\left(\lambda^2|\mathcal{G}_0|\right). \tag{6}$$

Notice, however, that (6) does not guarantee statistical consistency since $\hat{\pi}$ need not be a permutation consistent with the true underlying DAG. Moreover, (6) does not hold for *every* permutation $\hat{\pi}$ consistent with $\hat{A}$, but [45] shows the existence of at least one such permutation. In addition, it is shown in [45] that the number of non-zero elements in $\hat{A}$, $A_{0\hat{\pi}}$, and A_0 are all of the same order of magnitude, i.e., $|\hat{\mathcal{G}}| \asymp |\mathcal{G}_{0\hat{\pi}}| \asymp |\mathcal{G}_0|$.

2.3. Collection of DAGs

Consider the setting where not all the observed data comes from the same DAG, but rather from a collection of DAGs $\{\mathcal{G}^{(k)} = (V, E^{(k)})\}_{k=1}^K$ that share the same node set $V = [p]$. In addition, we assume that all DAGs in a collection are consistent with some permutation π. This precludes a scenario where $(i, j) \in E^{(k)}$ and $(j, i) \in E^{(k')}$ for some $k \neq k'$. This is a reasonable assumption in, e.g., the analysis of gene expression data, where regulatory links may appear or disappear, but they in general do not change direction.

Denote by $\{(A^{(k)}, \Omega^{(k)})\}_{k=1}^K$ a set of SEMs on the K DAGs $\{\mathcal{G}^{(k)}\}_{k=1}^K$ and by $\{\hat{X}^{(k)}\}_{k=1}^K$ the data generated from each SEM, where we observe n_k realizations for

each DAG $\mathcal{G}^{(k)}$. In this way, each row of the data matrix $\hat{X}^{(k)} \in \mathbb{R}^{n_k \times p}$ corresponds to a realization of the random vector $X^{(k)}$ defined as

$$X^{(k)} = {A^{(k)}}^T X^{(k)} + \epsilon^{(k)} \quad \text{with} \quad \epsilon^{(k)} \sim \mathcal{N}(0, \Omega^{(k)}).$$

Collections of DAGs arise for example naturally when considering data from *perfect* (also known as *hard*) *interventions* [10]. Consider a non-intervened DAG $\mathcal{G}$ with SEM parameters (A, Ω) [cf. (1)]. Then a perfect intervention on a subset of nodes $I_k \subset V$ gives rise to the interventional distribution

$$X^{I_k} = {A^{I_k}}^T X^{I_k} + \epsilon^{I_k} \quad \text{with} \quad \epsilon^{I_k} \sim \mathcal{N}(0, \Omega^{I_k}),$$

where $A^{I_k}_{ij} = 0$ if $j \in I_k$ and $A^{I_k}_{ij} = A_{ij}$ otherwise, and the diagonal matrix Ω^{I_k} satisfies $\Omega^{I_k}_{ii} = \Omega_{ii}$ if $i \notin I_k$ [17, 18]. We denote the DAG given by the non-zero entries of A^{I_k} by $\mathcal{G}^{I_k}$.

In accordance with the notation introduced in Section 2.1, we denote by $\mathcal{G}_0^{(k)}$ and $(A_0^{(k)}, \Omega_0^{(k)})$ the true data-generating DAG and SEM parameters for class k, respectively, and by π_0 a permutation that is consistent with $A_0^{(k)}$ for all classes $k \in [K]$. Moreover, we denote by $\mathcal{G}_{0\pi}^{(k)}$ and $(A_{0\pi}^{(k)}, \Omega_{0\pi}^{(k)})$ the DAG and SEM parameters obtained from the Cholesky decomposition of the true precision matrix $\Theta_0^{(k)}$ when permuted by π. We denote by $\Sigma_0^{(k)}$ the true covariance matrix of the SEM for class k, i.e., the inverse of $\Theta_0^{(k)}$. Finally, we define $\mathcal{G}_0^{\text{union}}$ as the union of all $\mathcal{G}_0^{(k)}$ – i.e., an edge appears in $\mathcal{G}_0^{\text{union}}$ if it appears in *any* $\mathcal{G}_0^{(k)}$ – and $\mathcal{G}_{0\pi}^{\text{union}}$ as the union of $\mathcal{G}_{0\pi}^{(k)}$. For interventional data, we use $(A_0^{I_k}, \Omega_0^{I_k})$ to denote the true SEM parameters after intervening on targets I_k.

3. Joint estimation of multiple DAGs

We first present a penalized maximum likelihood estimator that is the natural extension of (4) for the case where a collection of DAGs is being estimated. Since this involves minimizing $\|\cdot\|_0$, we then discuss a greedy approach that alleviates the computational complexity of this estimator.

3.1. Joint ℓ_0-penalized maximum likelihood estimator

With d denoting a pre-specified sparsity level and $w_k = n_k/n$ indicating the proportion of observed data from DAG k, we propose the following estimator:

$$\begin{aligned}\left\{\hat{\pi}, \{(\hat{A}^{(k)}, \hat{\Omega}^{(k)})\}_{k=1}^K\right\} \\ := \underset{\pi, \{(A^{(k)}, \Omega^{(k)})\}_{k=1}^K}{\operatorname{argmax}} & \sum_{k=1}^K w_k \ell_{n_k}(\hat{X}^{(k)}; A^{(k)}, \Omega^{(k)}) - \lambda^2 \left\| \sum_{k=1}^K |A^{(k)}| \right\|_0 \\ \text{subject to} \quad & A^{(k)} \in \mathcal{A}_\pi, \ \|A^{(k)}\|_{\infty,0} \le d, \ \Omega^{(k)} \in \mathcal{D}_+ \ \forall k,\end{aligned} \tag{7}$$

Algorithm 1 JointGES for joint ℓ_0-penalized maximum likelihood estimation of multiple DAGs.

Input: Collection of observations $\hat{X}^{(1)} \in \mathbb{R}^{n_1 \times p}, \cdots, \hat{X}^{(K)} \in \mathbb{R}^{n_K \times p}$, sparsity bound d, penalization parameters λ_1 and λ_2

Output: Collection of weighted adjacency matrices $\hat{A}^{(1)}, \cdots, \hat{A}^{(K)}$

1: Apply GES to find $\hat{\mathcal{G}}^{\text{union}}$, an approximate solution to the following optimization problem

$$\begin{aligned} &\underset{\mathcal{G}}{\operatorname{argmin}} \sum_{j=1}^{p} \left(\sum_{k=1}^{K} w_k \left[\min_{a \in \mathbb{R}^{|\mathrm{Pa}_j(\mathcal{G})|}} \log\left(\|\hat{X}_j^{(k)} - \hat{X}^{(k)}_{\mathrm{Pa}_j(\mathcal{G})}\, a\|_2^2 \right) \right] + \lambda_1^2 |\mathrm{Pa}_j(\mathcal{G})| \right) \\ &\text{subject to} \quad \max_j |\mathrm{Pa}_j(\mathcal{G})| \leq d \end{aligned} \tag{8}$$

2: Estimate the weighted adjacency matrices $\{\hat{A}^{(k)}\}_{k=1}^K$ consistent with $\hat{\mathcal{G}}^{\text{union}}$ by solving Kp sparse regressions of the form

$$\hat{a}_j^{(k)} = \underset{a \,|\, \mathrm{supp}(a) \subseteq \mathrm{Pa}_j(\hat{\mathcal{G}}^{\text{union}})}{\operatorname{argmin}} \frac{1}{n_k} \|\hat{X}_j^{(k)} - \hat{X}^{(k)} a\|_2^2 + \lambda_2^2 \|a\|_1.$$

where $\mathcal{A}_\pi$ is the set of all adjacency matrices consistent with permutation π and the matrix norm $\|\cdot\|_{\infty,0}$ computes the maximum ℓ_0-norm across the rows of the argument matrix. The optimization problem in (7) seeks to maximize a weighted log-likelihood of the observations (where more weight is given to SEMs with more realizations) penalized by the support of the union of all estimated DAGs. To see why this is true, notice that $\|\sum_{k=1}^K |A^{(k)}|\|_0$ counts the number of (i,j) entries for which $A_{ij}^{(k)} \neq 0$ for at least one graph k. This penalization on the union of estimated DAGs promotes overlap in the supports of the different $A^{(k)}$. Regarding the constraints in (7), the first constraint imposes that all estimated DAGs are consistent with the same permutation π, which is itself an optimization variable. This constraint is in accordance with our assumption in Section 2.3 and drastically reduces the search space of DAGs. The second constraint ensures that the maximum in-degree in all graphs is at most d, and the last constraint imposes the natural requirement that all noise covariances are diagonal and non-negative.

Notice that (7) is a natural extension of (4). Indeed, for the case $K = 1$ the objective in (7) immediately boils down to that in (4). Moreover, when there is only one graph and π can be selected freely, the constraint $A^{(1)} \in \mathcal{A}_\pi$ is effectively identical to $A^{(1)} \in \mathcal{A}$, i.e., the constraint in (4). Finally, observe that in (7) we have included the additional maximum in-degree constraint required in the high-dimensional setting [cf. discussion after (5)].

3.2. JointGES: Joint greedy equivalence search

The ℓ_0 norm as well as the optimization over all permutations π render the problem of (7) non-convex, thus, hard to solve efficiently. In this section, we present a greedy approach to find a computationally tractable approximation to a solution to (7). The algorithm, which we term *JointGES*, is succinctly presented in Algorithm 1 and consists of two steps.

In the first step of Algorithm 1 we recover $\hat{\mathcal{G}}^{\text{union}}$, our estimate of the union of all the DAGs to be inferred. We do this by finding an approximate solution to (8) via the implementation of GES [7]. The objective (scoring function) in (8) consists of two terms. The first term is given by the sum of the log-likelihoods of the achievable residues when regressing the jth column of $X^{(k)}$, denominated as $X_j^{(k)}$, on $X^{(k)}_{\text{Pa}_j(\mathcal{G})}$ for each node j and DAG k. In [45], van de Geer and Bühlmann show that if we keep the underlying DAG $\mathcal{G}$ fixed, the maximum likelihood estimator proposed in (4) is equivalent to optimizing $\sum_{j=1}^{p}\left(\min_{a\in\mathbb{R}^{|\text{Pa}_j(\mathcal{G})|}}\log\left(\|\hat{X}_j-\hat{X}_{\text{Pa}_j(\mathcal{G})}\,a\|_2^2\right)\right)$. Thus, the first term in (8) corresponds to the first term in the objective of (7). The second term penalizes the size of the parent set of each node in the graph to be recovered, effectively penalizing the number of edges in the graph. In this way, the scoring function in (8) promotes a sparse $\mathcal{G}$ in the same way that the objective of (7) promotes the union of all K recovered graphs to have a sparse support. Additionally, it is immediate to see that the scoring function in (8) is *decomposable* [7], a key feature that enables the implementation of GES to find an approximate solution. Once we have obtained the union of all sought DAGs $\hat{\mathcal{G}}^{\text{union}}$ from step 1, in the second step of our algorithm we estimate the DAGs $\hat{\mathcal{G}}^{(1)},\cdots,\hat{\mathcal{G}}^{(K)}$ by searching over the subDAGs of $\hat{\mathcal{G}}^{\text{union}}$. More precisely, for each node j we estimate its parents in $\hat{\mathcal{G}}^{(k)}$ by regressing $X_j^{(k)}$ on $X^{(k)}_{\text{Pa}_j(\hat{\mathcal{G}}^{\text{union}})}$ using lasso, where the support of $\hat{a}_j^{(k)}$ corresponds to the set of parents of j in $\hat{\mathcal{G}}^{(k)}$.

To summarize, Algorithm 1 recovers K DAGs by first estimating the union of all these DAGs $\hat{\mathcal{G}}^{\text{union}}$ using GES and then inferring the specific weight adjacency matrices $\hat{A}^{(k)}$ via a lasso regression, while ensuring consistency with the previously estimated $\hat{\mathcal{G}}^{\text{union}}$.

4. Consistency guarantees

The main goal of this section is to provide theoretical guarantees on the consistency of the solution to Problem (7) in the high-dimensional setting. Our main result is presented in Theorem 4.9; in Section 4.3 we present a laxer statement of consistency based on milder conditions.

4.1. Statistical consistency of the joint ℓ_0-penalized MLE

A series of conditions must hold for our main result to be valid. We begin by stating these conditions followed by the formal consistency result in Theorem 4.9. The rationale behind these conditions and their implications are discussed after the theorem in Section 4.2.

Condition 4.1. *All DAGs $\mathcal{G}_0^{(1)},\cdots,\mathcal{G}_0^{(K)}$ are minimal-edge I-MAPs.*

Condition 4.2. *There exists a constant σ_0^2 that bounds the variance of all the observed processes, i.e., $\max_{k,i}[\Sigma_0^{(k)}]_{ii}\leq\sigma_0^2$.*

Condition 4.3. *The smallest eigenvalues of all $\Sigma_0^{(k)}$ are non-zero, i.e. $\min_k\Lambda_{\min}(\Sigma_0^{(k)})=\Lambda_{\min}>0$.*

Condition 4.4. *There exists some constant α such that, for all k, the maximum allowable in-degree d in the objective function (7) is bounded as $d \leq \alpha n_k / \log p$.*

Condition 4.5. *For all π and j there exist some constants $\tilde{\alpha}$ and $c_s > 2$ such that*

$$|\mathrm{Pa}_j(\mathcal{G}_{0\pi}^{union})| + c_s \leq \tilde{\alpha} \left(\min \left\{ \left(\frac{n}{K^7 (\log p)^3} \right)^{\frac{1}{3}}, \frac{n}{K^7 (\log n)^2 \log p} \right\} \right).$$

Condition 4.6. *The number of DAGs K satisfies $K = o(\log p)$ and the amount of data associated with each DAG is comparable in the sense that $n_1 \asymp n_2 \asymp \cdots \asymp n_K$.*

Condition 4.7. *There exists some constant $c_t > 0$ such that $|\mathcal{G}_{0\pi}^{union}| \leq c_t \sum_{k=1}^{K} w_k |\mathcal{G}_{0\pi}^{(k)}|$ for any permutation π.*

Condition 4.8. *There exist constants η_0 and η_1 such that $0 \leq \eta_1 < 1$, $0 < \eta_0^2 < (1 - \eta_1)/c_t$, and*

$$\sum_{i,j} \mathbf{1} \left\{ \left| [A_{0\pi}^{(k)}]_{i,j} \right| > \frac{\sqrt{\log p / n}}{\eta_0} \left(\sqrt{p / |\mathcal{G}_0^{union}|} \vee 1 \right) \right\} \geq (1 - \eta_1) |\mathcal{G}_{0\pi}^{(k)}|, \tag{9}$$

for all permutations π and graphs $k \in [K]$, where $\mathbf{1}\{\cdot\}$ denotes the indicator function and c_t is as in Condition 4.7.

With the above conditions in place, the following result can be shown.

Theorem 4.9. *If Conditions 4.1-4.8 hold and λ is chosen such that*

$$\lambda^2 \asymp \frac{\log p}{n} \left(\frac{p}{|\mathcal{G}_0^{union}|} \vee 1 \right),$$

then there exists a constant $c > 0$, that depends on c_s, such that with probability $1 - \exp(-cp)$ the solution to (7) satisfies

$$\sum_{k=1}^{K} w_k \|\hat{A}^{(k)} - A_{0\hat{\pi}}^{(k)}\|_F^2 + \sum_{k=1}^{K} w_k \|\hat{\Omega}^{(k)} - \Omega_{0\hat{\pi}}^{(k)}\|_F^2 = \mathcal{O}\left(\lambda^2 |\mathcal{G}_0^{union}| \right). \tag{10}$$

Furthermore, denoting by $\hat{\mathcal{G}}$ the union of the graphs $\hat{\mathcal{G}}^{(k)}$ associated with the K recovered adjacency matrices $\hat{A}^{(k)}$, we have that

$$|\hat{\mathcal{G}}| \asymp |\mathcal{G}_{0\hat{\pi}}^{union}| \asymp |\mathcal{G}_0^{union}|. \tag{11}$$

The proof of Theorem 4.9 is given in Appendix A.2. To intuitively grasp the result in the above theorem, assume that the number of edges in $\mathcal{G}_0^{\text{union}}$ is proportional to the number of nodes p so that $\lambda^2 |\mathcal{G}_0^{\text{union}}| \to 0$ for increasing n as long as $n > p \log p$. Hence, under these conditions, (10) guarantees that for the recovered permutation $\hat{\pi}$, the estimated adjacency matrix $\hat{A}^{(k)}$ converges to $A_{0\hat{\pi}}^{(k)}$ in Frobenius norm for all k. This not only implies that both adjacency matrices have similar structure, but also that the edge weights are similar. Moreover, from (11) it follows that the number of edges in the

estimated graph $\hat{\mathcal{G}}$, i.e., $|\hat{\mathcal{G}}|$ is similar to the number of edges in the union of all minimal I-MAPs with permutation $\hat{\pi}$, i.e., $|\mathcal{G}_{0\hat{\pi}}^{\text{union}}|$. More importantly, $|\hat{\mathcal{G}}|$ is also similar to the number of edges in the true union graph $|\mathcal{G}_0^{\text{union}}|$. Despite these guarantees, it should be noted that similar to the results in [45], the permutation $\hat{\pi}$ need not coincide with the permutation π of the true graphs to be recovered.

We now assess the benefits of performing joint estimation of the K DAGs as opposed to estimating them separately. To do so, we compare the guarantees in Theorem 4.9 to those developed in [45] for separate estimation. The application of the consistency bound reviewed in (6) yields that for the separate estimation of K DAGs, when we are in the setting where all K DAGs are highly overlapping (cf. Condition 4.7), by choosing λ such that $\lambda^2 \asymp \frac{\log p}{n}\left(\frac{p}{|\mathcal{G}_0^{\text{union}}|} \vee 1\right)$, one can guarantee that

$$\sum_{k=1}^{K} w_k \|\hat{A}^{(k)} - A^{(k)}_{0\hat{\pi}^{(k)}}\|_F + \sum_{k=1}^{K} w_k \|\hat{\Omega}^{(k)} - \Omega^{(k)}_{0\hat{\pi}^{(k)}}\|_F^2 = \mathcal{O}\left(K\lambda^2 \max_{k\in[K]} |\mathcal{G}_0^{(k)}|\right), \tag{12}$$

where it should be noted that in the separate estimation the recovered permutation $\hat{\pi}^{(k)}$ can vary with k. A direct comparison of (10) and (12) reveals that performing joint estimation improves the accuracy by a factor of K from $\Omega(K\frac{\log p}{n})$ to $\Omega(\frac{\log p}{n})$. Hence, for joint estimation the accuracy scales with the total number of samples n, showing that our procedure yields maximal gain from each observation, even if the data is generated from K different DAGs. Moreover, the result in (10) holds under slightly milder conditions than those needed for (12) to hold since Condition 4.8 is a relaxed version of the *beta-min condition* in [45]. A more detailed discussion about the conditions of Theorem 4.9 is given next.

4.2. Conditions for Theorem 4.9

It has been shown in [34] that if a data-generating distribution is faithful with respect to $\mathcal{G}$, then $\mathcal{G}$ must be a minimal-edge I-MAP. By enforcing the latter for every true graph, Condition 4.1 imposes a milder requirement compared to the well-established faithfulness assumption [40]. Conditions 4.2-4.4 ensure that we avoid overfitting and provide bounds for the noise variances. These are direct adaptations from Conditions 3.1-3.3 in [45]. Condition 4.5 is required to bound the difference between the sample variances of our observations and the true variances, and is related to Condition 3.4 in [45] but adapted to our joint inference setting. Notice that Condition 4.5 is trivially satisfied when $p = \mathcal{O}\left(\frac{n^{1/3}}{K^{7/3}\log n}\right)$. Condition 4.6 follows from the bounds for sample variances shown in [4]. Intuitively, we are imposing the natural restriction that the number of DAGs is small compared to the number of nodes p in each DAG and the total number of observations n. Moreover, given that our objective is to draw estimation power from the joint inference of multiple graphs, we require that each DAG is associated with a non-vanishing fraction of the total observations.

Condition 4.7 enforces that, for every permutation π, the number of edges in the union of all recovered graphs is proportional to the weighted sum of the edges in every

graph as $K \to \infty$. In particular, this requires the individual graphs $\mathcal{G}_{0\pi}^{(k)}$ to be highly overlapping. To see why this is the case, notice that $\sum_{k=1}^{K} w_k |\mathcal{G}_{0\pi}^{(k)}|$ is upper bounded by the maximum number of edges across graphs $\mathcal{G}_{0\pi}^{(k)}$. Consequently, Condition 4.7 enforces the number of edges in the union of graphs to be proportional to the number of edges in the single graph with most edges, thus requiring a high level of overlap. Imposing high overlap for all permutations π might seem too restrictive in some settings. Nonetheless, Condition 4.7 can sometimes be derived from apparently less restrictive conditions as the following example illustrates.

Consider the more relaxed bound $|\mathcal{G}_0^{\text{union}}| \leq c_t \sum_{k=1}^{K} w_k |\mathcal{G}_0^{(k)}|$, which is equivalent to requiring Condition 4.7 to hold but *only* for permutations consistent with the true graph $\mathcal{G}_0^{\text{union}}$. In the following example, we show that this might be sufficient for Condition 4.7 to hold. Suppose that $\mathcal{G}_0^{\text{union}}$ consists of two connected components $\mathcal{G}'^{\text{union}}_0$ and $\mathcal{G}''^{\text{union}}_0$ respectively defined on the subsets of nodes V_1 and V_2. Moreover, assume that the subDAGs of $\mathcal{G}_0^{(k)}$ over V_1 (denoted by $\mathcal{G}'^{(k)}_0$) are identical for all k. Putting it differently, the differences between the DAGs $\mathcal{G}_0^{(k)}$ are limited to the second connected component. In addition, assume that for all possible permutations π_2 of nodes V_2 we have that $|\mathcal{G}''^{\text{union}}_{0\pi_2}| \leq |\mathcal{G}'^{\text{union}}_0|$. Then, for any permutation π, where we denote by π_1 (respectively π_2) the restriction of π to the node set V_1 (respectively V_2), we have

$$|\mathcal{G}_{0\pi}^{\text{union}}| = |\mathcal{G}'^{\text{union}}_{0\pi_1}| + |\mathcal{G}''^{\text{union}}_{0\pi_2}| \leq \sum_{k=1}^{K} w_k |\mathcal{G}'^{(k)}_{0\pi_1}| + \sum_{k=1}^{K} w_k |\mathcal{G}'^{(k)}_{0}| \leq 2 \sum_{k=1}^{K} w_k |\mathcal{G}_{0\pi}^{(k)}|,$$

which shows that Condition 4.7 is satisfied for $c_t = 2$. This example shows that learning the structure of large components that are common across the different DAGs is not affected by the changes in the smaller components of these DAGs. Beyond this example, in Section 6.2, we also provide simulation results to study the strength of Condition 4.7 for sparse DAG models. Our simulation analysis shows that, when the $\mathcal{G}_0^{(k)}$'s are highly overlapping (recall that this corresponds to a more relaxed scenario than Condition 4.7 that requires high overlap across $\mathcal{G}_{0\pi}^{(k)}$'s for all π's), Condition 4.7 is naturally satisfied with a reasonably small c_t. Despite the above example as well as the empirical analysis, Condition 4.7 might still be too restrictive for some applications; we discuss a relaxed requirement and its implications on the consistency guarantees in Section 4.3.

Condition 4.8 requires that, for every permutation π and every graph k, the value of at least a fixed proportion $(1-\eta_1)$ of the edges in $\mathcal{G}_{0\pi}^{(k)}$ is above the *'noise level'*, i.e., the lower bound within the indicator function in (9). Intuitively, if the true weight of many edges is close to zero then correct inference of the graphs would be impossible since the true edges would be mistaken with spurious ones. Thus, it is expected that the weights of a sufficiently large fraction of the edges have to be sufficiently large. Condition 4.8 is the right formalization of this intuition. Moreover, notice that a straightforward replication of the beta-min condition introduced in [45] would have required the *'noise level'* to scale with $\sqrt{\log p / n_k}$, instead of the smaller scaling of $\sqrt{\log p / n}$ required in (9). In this sense, Condition 4.8 (together with Condition 4.1) is a relaxed version of the extension of the beta-min condition to the setting of joint graph estimation.

Remark 4.10 (Strength of assumptions). Requiring strong assumptions for consistent estimation is a common theme in existing methods for causal inference. For example, the PC algorithm requires the strong faithfulness assumption [20], which has been shown to be a very restrictive assumption for high-dimensional causal graphical models [44]. For a discussion on the comparison between the strong faithfulness assumption and the beta-min condition for estimating a single DAG model, see [45, Section 4.3.2]. In this context, the assumptions presented here are in line with or slight relaxations (Conditions 4.1 and 4.8) of those in state-of-the-art approaches. While it would be interesting in future work to formally compare Conditions 4.1 and 4.8 to strong faithfulness, our goal here is not to relax existing assumptions for the estimation of DAG models, but to show that *joint* estimation can result in faster rates than *separate* estimation of multiple DAGs under comparable assumptions.

4.3. Consistency under milder conditions

As previously discussed, in some settings Condition 4.7 might be too restrictive. Hence, in this section we present a consistency statement akin to Theorem 4.9 that holds for a milder version of Condition 4.7:

Condition 4.7'. *Let $c_t(\pi)$ be some function of π that scales as a constant for permutations consistent with $\mathcal{G}_0^{union}$ and scales as $o(K)$ for all other permutations such that $|\mathcal{G}_{0\pi}^{union}| \leq c_t(\pi) \sum_{k=1}^{K} w_k |\mathcal{G}_{0\pi}^{(k)}|$ for all π.*

Observe that for permutations π consistent with the true union graph $\mathcal{G}_0^{\text{union}}$, Condition 4.7' boils down to the previously discussed Condition 4.7. However, for all other permutations, $c_t(\pi)$ need not be a constant and is allowed to grow with K. Intuitively, for all permutations *not* consistent with $\mathcal{G}_0^{\text{union}}$ we are *not* requiring a high level of overlap among all the graphs $\mathcal{G}_{0\pi}^{(k)}$. Nonetheless, since $c_t(\pi) = o(K)$ we *do* require $\mathcal{G}_{0\pi}^{\text{union}}$ to be *'sparser'* than the extreme case in which all graphs $\mathcal{G}_{0\pi}^{(k)}$ are disjoint.

In order to account for the fact that c_t depends on the permutation π in Condition 4.7', we have to modify Condition 4.8 accordingly, resulting in the following alternative statement.

Condition 4.8'. *Let $C_{\max} := \max_{\pi} c_t(\pi)$, then there exist constants η_0 and η_1 such that $0 \leq \eta_1 < 1$, $0 < \eta_0^2 < (1 - \eta_1)$, and*

$$\sum_{i,j} \mathbf{1}\left\{ \left| [A_{0\pi}^{(k)}]_{i,j} \right| > \frac{\sqrt{C_{\max} \log p/n}}{\eta_0} \left(\sqrt{p/|\mathcal{G}_0^{union}|} \vee 1 \right) \right\} \geq (1 - \eta_1)|\mathcal{G}_{0\pi}^{(k)}|,$$

for all permutations π and graphs k, where $\mathbf{1}\{\cdot\}$ denotes the indicator function.

The following consistency result holds for the alternative set of conditions.

Theorem 4.11. *Under Conditions 4.1-4.6, 4.7' and 4.8' and with λ such that $\lambda^2 \asymp C_{\max} \frac{\log p}{n} \left(\frac{p}{|\mathcal{G}_0^{union}|} \vee 1 \right)$, then there exists a constant $c > 0$ that depends on c_s such that with probability $1 - \exp(-cp)$, the solution to (7) satisfies that, at least for one*

$k \in [K]$,

$$\|\hat{A}^{(k)} - A^{(k)}_{0\hat{\pi}}\|_F^2 + \|\hat{\Omega}^{(k)} - \Omega^{(k)}_{0\hat{\pi}}\|_F^2 = \mathcal{O}\left(\lambda^2 |\mathcal{G}_0^{(k)}|\right). \tag{13}$$

Furthermore, denoting by $\hat{\mathcal{G}}^{(k)}$ the graph associated with $\hat{A}^{(k)}$ for the $k \in [K]$ satisfying (13), we have that

$$|\hat{\mathcal{G}}^{(k)}| \asymp |\mathcal{G}^{(k)}_{0\hat{\pi}}| \asymp |\mathcal{G}^{(k)}_0|. \tag{14}$$

The proof is given in Appendix A.3. Condition 4.7' is milder than Condition 4.7 and this relaxation entails a corresponding loss in the guarantees of recovery: Comparing (13) and (14) with (10) and (11) immediately reveals that what could be guaranteed for the ensemble of graphs in Theorem 4.9 can only be guaranteed for a single graph in Theorem 4.11, thereby explaining the trade-off in relaxing the conditions.

However, the result in Theorem 4.11 still draws inference power from the joint estimation of multiple graphs since neither (13) nor (14) can be shown using existing results for separate estimation. To be more precise, as discussed in Section 4.2, when performing separate estimation, theoretical guarantees are based on the assumption that at least a fixed proportion of the edge weights are above the *'noise level'*, which scales as $\sqrt{\log p/n_k}$. However, Condition 4.8' requires the noise level to scale with $\sqrt{C_{\max} \log p/n}$ which, given the fact that $C_{\max} = o(K)$, is not large enough to achieve the guarantee needed for separate estimation. In addition, the convergence rate of $\Omega(C_{\max} \frac{\log p}{n})$ in (13) is still faster than the corresponding convergence rate of $\Omega(K \frac{\log p}{n})$ associated with separate estimation [cf. discussion after (12)]. A potential limitation of Theorem 4.11 is that, since one cannot know which of the K DAGs achieves such $\sqrt{C_{\max} \log p/n}$ rate and which remains at $\sqrt{\log p/n_k}$, the result may be of limited utility for practitioners. However, note that the much weaker Condition 4.7' helps to illustrate that, even in such scenario, joint estimation can be helpful compared with separate estimation. In addition, it also clarifies which guarantees are lost with respect to the more stringent scenario when Condition 4.7 holds. In this sense, although not fully interpretable, this intermediate case provides an idea of how the guarantees degrade as we start to soften the assumptions.

We end this section with the following remark discussing the consistency guarantees of jointGES.

Remark 4.12 (Consistency of jointGES). In the low-dimensional setting, by choosing $\lambda_1^2 = \sum_{k=1}^K w_k \frac{\log n_k}{2n_k}$, assuming faithfulness and assuming that GES finds the global optimum of (8), it can be inferred from [7] that, in the limit of large data, the first step in Algorithm 1 is guaranteed to produce a Markov equivalence class (MEC) $\hat{\mathcal{M}}$ that is within the following set of MECs:

$$\mathcal{M}^* := \left\{\mathcal{M} : \text{ there exists a } \pi \in \Pi \text{ such that } \mathcal{M} = \mathcal{M}(\mathcal{G}^{\text{union}}_{0\pi})\right\}$$

where

$$\Pi := \left\{\pi : \forall k \in [K], \mathcal{G}^{(k)}_{0\pi} \in \mathcal{M}(\mathcal{G}^{(k)}_0)\right\}.$$

This allows us to recover $\{\mathcal{M}(\mathcal{G}_0^{(k)})\}_{k=1}^K$ by successively considering all DAGs in $\hat{\mathcal{M}}$ as inputs to the second step of Algorithm 1, and selecting the DAG $\hat{\mathcal{G}}^{\text{union}} \in \hat{\mathcal{M}}$ whose output $\{\hat{\mathcal{G}}^{(k)}\}_{k=1}^K$ from step 2 is the sparsest. Then the MECs $\{\mathcal{M}(\hat{\mathcal{G}}^{(k)})\}_{k=1}^K$ produced from step 2 asymptotically coincide with $\{\mathcal{M}(\mathcal{G}_0^{(k)})\}_{k=1}^K$. Note that no matter which MEC is chosen from the set $\mathcal{M}^*$, by doing edge reductions in step 2, the final result is always guaranteed to asymptotically converge to $\{\mathcal{M}(\mathcal{G}_0^{(k)})\}_{k=1}^K$. In Example 4.13, we show how Algorithm 1 works for a particular 3-node instance. In the high-dimensional setting, where even the global optimum of (7) is not guaranteed to recover the true $\mathcal{G}_0^{\text{union}}$ (cf. Theorems 4.9 and 4.11), jointGES is in general not consistent. Recently, Maathuis et al. [29] showed consistency of GES for single DAG estimation in the high-dimensional setting under more restrictive assumptions than the ones considered here. Although of potential interest, further strengthening the presented conditions to guarantee consistency of jointGES also in the high-dimensional setting is not pursued in the current paper.

Example 4.13. We present an example to illustrate how the output from Algorithm 1 works in the low-dimensional regime. Consider the setting where we have data collected from two causal DAG models on 3 nodes, namely $1 \to 2$ and $2 \leftarrow 3$. In the first step, Algorithm 1 will produce either the PDAG $1 \to 2 \leftarrow 3$ or $1 - 2 - 3$. Then, no matter which of the two PDAGs is learned in the first step, by taking it to Step 2, it is guaranteed to asymptotically converge to the desired output $1 \to 2$ (or $1 \leftarrow 2$) as well as $2 \to 3$ (or $2 \leftarrow 3$), of which the MECs are $1 - 2$ and $2 - 3$, respectively.

5. Extension to interventions

In this section, we show how our proposed method for joint estimation can be extended to learn DAGs from interventional data. It is natural to consider learning from interventional data as a special case of joint estimation since the DAGs associated with interventions are different but closely related. In this section, we mimic some of the developments of Sections 3 and 4 but specialized for the case of interventional data. More precisely, we first propose an optimization problem akin to (7) and then state the consistency guarantees in the high-dimensional setting of the associated optimal solution.

Recall from Section 2.3 that the true adjacency matrix $A_0^{I_k}$ of the SEM associated with an intervention on the nodes I_k is identical to the true adjacency matrix A_0 of the non-intervened model except that $[A_0^{I_k}]_{ij} = 0$ for all $j \in I_k$. In this way, our assumption that there exists a common permutation π consistent with all DAGs under consideration (cf. Section 2.3) is automatically satisfied for interventional data. Additionally, assuming that we observe samples $\hat{X}^{I_k}$ from K different models corresponding to the respective intervention on the nodes in $\{I_k\}_{k=1}^K$, the knowledge of the intervened nodes

can be incorporated into our optimization problem as follows [cf. (7)].

$$\left\{\hat{\pi},\hat{A},\hat{\Omega},\{(\hat{A}^{I_k},\hat{\Omega}^{I_k})\}_{k=1}^{K}\right\}$$

$$= \underset{\pi,A,\Omega,\{(A^{I_k},\Omega^{I_k})\}_{k=1}^{K}}{\operatorname{argmax}} \quad \sum_{k=1}^{K} w_k \ell_{n_k}(\hat{X}^{I_k}; A^{I_k}, \Omega^{I_k}) - \lambda^2 \|A\|_0 \tag{15a}$$

$$\text{subject to} \quad A \in \mathcal{A}_\pi, \quad \|A\|_{\infty,0} \leq d, \quad \Omega \in \mathcal{D}_+, \tag{15b}$$

$$A_{ij}^{I_k} = A_{ij} \;\; \forall j \notin I_k, \quad A_{ij}^{I_k} = 0 \;\; \forall j \in I_k, \tag{15c}$$

$$\Omega_{jj}^{I_k} = \Omega_{jj} \;\; \forall j \notin I_k, \quad \Omega^{I_k} \in \mathcal{D}_+. \tag{15d}$$

From the solution of (15) we obtain an estimate for the non-intervened SEM $(\hat{A},\hat{\Omega})$ as well as K estimates for the corresponding intervened models $(\hat{A}^{I_k},\hat{\Omega}^{I_k})$. The objective in (15a) is equivalent to that in (7) where we leverage the fact that the union of all intervened graphs results in the non-intervened one under the implicit assumption that no single node has been intervened in every experiment. Alternatively, if some nodes were intervened in all experiments, objective (15a) would still be valid since enforcing zeros in the unobservable portions of A does not affect the recovery of the intervened adjacency matrices A^{I_k}. The constraints in (15b) impose that A has to be consistent with permutation π and with bounded in-degree, and Ω has to be a valid covariance matrix for uncorrelated noise. Putting it differently, (15b) enforces for the non-intervened SEM what we impose separately for all SEMs in (7). The constraints in (15c) impose the known relations between the intervened and the non-intervened adjacency matrices. Finally, (15d) constrains the matrices Ω^{I_k} to be consistent with the base model on the non-intervened nodes while still being a valid covariance on the intervened ones.

Even though it might seem that in (15) we are estimating $K+1$ SEMs (the base case plus the K intervened ones), from the previous reasoning it follows that the effective number of optimization variables is significantly smaller. To be more specific, for a given π, once A is fixed then all the adjacency matrices A^{I_k} are completely determined. Moreover, for a fixed Ω, the only freedom in Ω^{I_k} corresponds to the diagonal entries associated with intervened nodes in I_k. In this way, it is expected that for a given number of samples, the joint estimation of K SEMs obtained from interventional data [cf. (15)] outperforms the corresponding estimation from purely observational data [cf. (7)].

Recalling that we denote by $(A_{0\hat{\pi}}^{I_k},\Omega_{0\hat{\pi}}^{I_k})$ the parameters recovered from the Cholesky decomposition of the true precision matrix $\Theta_0^{I_k}$ under the assumption of consistency with permutation $\hat{\pi}$, the following result holds.

Corollary 5.1. *If Conditions 4.1-4.8 hold and λ is chosen as $\lambda^2 \asymp \frac{\log p}{n}\left(\frac{p}{|\mathcal{G}_0|} \vee 1\right)$, then there exist constants $c_1, c_2 > 0$ such that with probability $1 - c_1\exp(-c_2 p)$, the solution to (15) satisfies*

$$\sum_{k=1}^{K} w_k \|\hat{A}^{I_k} - A_{0\hat{\pi}}^{I_k}\|_F^2 = \mathcal{O}(\lambda^2|\mathcal{G}_0|). \tag{16}$$

Furthermore, denoting by $\hat{\mathcal{G}}$ the graph associated with the recovered adjacency matrix

$\hat{A}$ for the non-intervened model, we have that

$$|\hat{\mathcal{G}}| \asymp |\mathcal{G}_{0\hat{\pi}}| \asymp |\mathcal{G}_0|. \tag{17}$$

The proof is given in Appendix A.4. A quick comparison of Theorem 4.9 and Corollary 5.1 seems to indicate that the consistency guarantees of observational and interventional data are very similar. However, recovery from interventional data is strictly better as we argue next. As discussed after Theorem 4.9, the results presented do not guarantee that the permutation $\hat{\pi}$ recovered coincides with the true permutation of the nodes. In principle, one could recover a spurious permutation $\hat{\pi}$ (different from the true permutation π) that correctly explains the observed data [cf. (10) and (16)] and leads to sparse graphs [cf. (11) and (17)]. However, the more interventions we have, the smaller the set of spurious permutations $\hat{\pi}$ that can be recovered, as we illustrate in the following example. Figure 1 portrays the existence of a spurious permutation that could be recovered from observational data but cannot be recovered from interventional data. More precisely, Figure 1(a) presents the two true DAGs that we aim to recover, where the second one is obtained by intervening on node 2. By contrast, Figure 1(b) shows the DAGs that are obtained when performing Cholesky decompositions on the true precision matrices under the spurious permutation π_1. Notice that the sparsity levels of the DAGs in both figures are the same. In general, one could recover π_1 instead of π_0 from observational data, but one would never recover π_1 from interventional data. To see this, simply notice from the figure that $[A_{0\pi_1}^{I_2}]_{32} \neq 0$ whereas for the interventional estimate $[\hat{A}^{I_2}]_{32} = 0$ [cf. (15c)], thus, the error terms in (16) cannot vanish for $\hat{\pi} = \pi_1$. This example also indicates that it is preferable to intervene on multiple targets in the same experiment instead of doing interventions one at a time. This observation is in accordance with new genetic perturbation techniques, such as Perturb-seq [9].

From a practical perspective, the objective in (15) corresponds to the same scoring function as GIES [17]. Therefore, GIES can be used to obtain an approximate solution to (15). A simulation study using GIES was performed in [17, Section 5.2] showing that in line with the theoretical results obtained in this section, not only identifiability increases, but also the estimates obtained using interventional data are better than with the same amount of purely observational data.

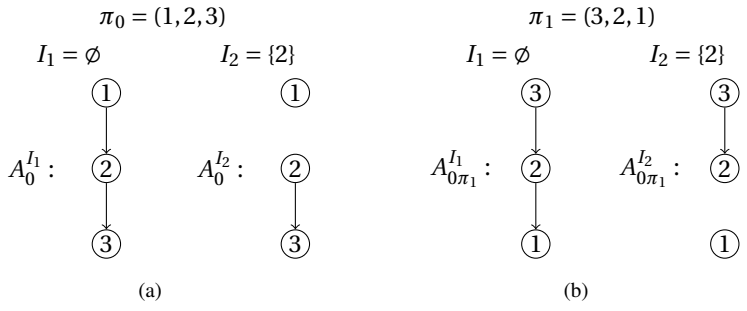


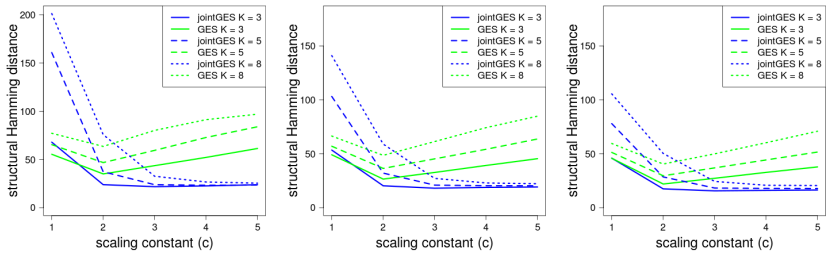
FIG 1. *Interventional data can avoid the recovery of spurious permutations. (a) True DAGs to be recovered. (b) DAGs obtained from the Cholesky decomposition consistent with π_1. The spurious permutation π_1 does not satisfy (16) for cases where node 2 is intervened.*

6. Experiments

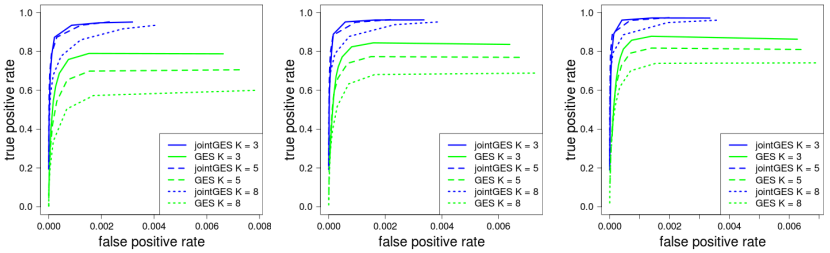
In this section, we present numerical experiments on both synthetic (Section 6.1) and real (Section 6.3) data that support our theoretical findings. We also provide an empirical analysis to study the strength of Condition 4.7 for sparse DAG models in Section 6.2.

6.1. Performance evaluation of joint causal inference

We analyze the performance of the joint recovery of K different DAGs where we vary $K \in \{3, 5, 8\}$ and $n \in \{600, 900, 1200\}$. For all experiments, we set the number of nodes $p = 100$. In addition, we selected the number of samples from each DAG to be the same, i.e., $n_1 = \ldots = n_K = n/K$. For each experiment, the true DAGs were constructed in two steps. First, we generated a *core* graph that is shared among the K DAGs under consideration. We did this by generating a random graph from an Erdős-Rényi model with 100 edges in expectation, and then oriented the edges according to a random permutation of the nodes. Then we sampled $e_{\text{private}} \in \{30, 60\}$ additional *private* edges uniformly at random. Each such edge was assigned uniformly at random to one of the K DAGs, thereby keeping the total number of private edges across all K DAGs

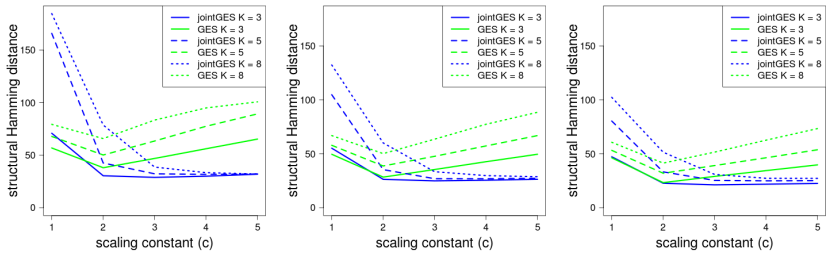


(a) Average SHD, $n = 600$ (b) Average SHD, $n = 900$ (c) Average SHD, $n = 1200$

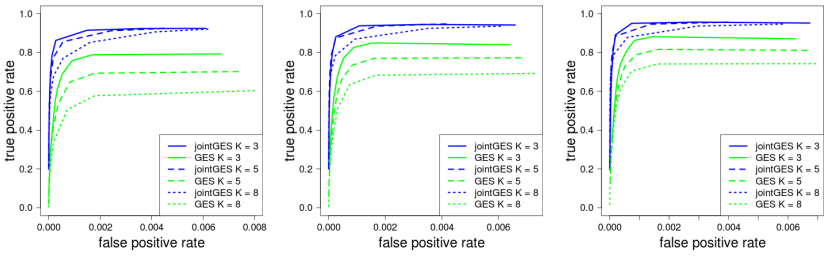


(d) Average ROC, $n = 600$ (e) Average ROC, $n = 900$ (f) Average ROC, $n = 1200$

FIG 2. *Simulation results when we set the private-to-core edge ratio to* 0.3. *(a) - (c) Average SHD as a function of the scaling constant* c *for joint and separate GES with* $n = 600, 900, 1200$ *respectively; (d) - (f) Average ROC curve obtained by varying the tuning parameters with* $n = 600, 900, 1200$. *JointGES consistently achieves a better performance across all settings.*



(a) Average SHD, $n = 600$ (b) Average SHD, $n = 900$ (c) Average SHD, $n = 1200$



(d) Average ROC, $n = 600$ (e) Average ROC, $n = 900$ (f) Average ROC, $n = 1200$

FIG 3. *Simulation results when we set the private-to-core edge ratio to* 0.6. *(a) - (c) Average SHD as a function of the scaling constant c for joint and separate GES with* $n = 600, 900, 1200$ *respectively; (d) - (f) Average ROC curve obtained by varying the tuning parameters with* $n = 600, 900, 1200$. *JointGES consistently achieves a better performance across all settings.*

to be e_{private}. This procedure results in the generation of a collection of true underlying DAGs $\mathcal{G}_0^{(1)}, \ldots, \mathcal{G}_0^{(K)}$ with a private-to-core edge ratio of 0.3 and 0.6 respectively. Associated with each DAG, we generated a true adjacency matrix $A_0^{(k)}$ and a true diagonal error covariance matrix $\Omega_0^{(k)}$. For the latter, we drew each error variance independently and uniformly from the interval $[1, 2.25]$. Regarding the adjacency matrices, we drew the edge weights independently and uniformly from $[-1, -0.1] \cup [0.1, 1]$ to ensure that they are bounded away from zero. Notice that we did not put any constraints on the edge weights that are in the shared core structure for different DAGs: the same edge can change its weight in different DAGs, or even flip sign.

We randomly generated 100 collections of DAGs and data associated with them. We then estimated the DAGs from the data via two different methods: a joint estimation procedure using jointGES presented in Algorithm 1 and a separate estimation procedure using the well-established GES method [7].

To assess performance of the two algorithms, we considered two standard measures, namely the structural Hamming distance (SHD) [43] and the receiver operating characteristic (ROC) curve. SHD is a commonly used metric based on the number of operations needed to transform the estimated DAG into the true one [20, 43]. Hence, a smaller SHD value indicates better performance. The ROC curve plots the true positive rate against the false positive rate for different choices of tuning parameters. The

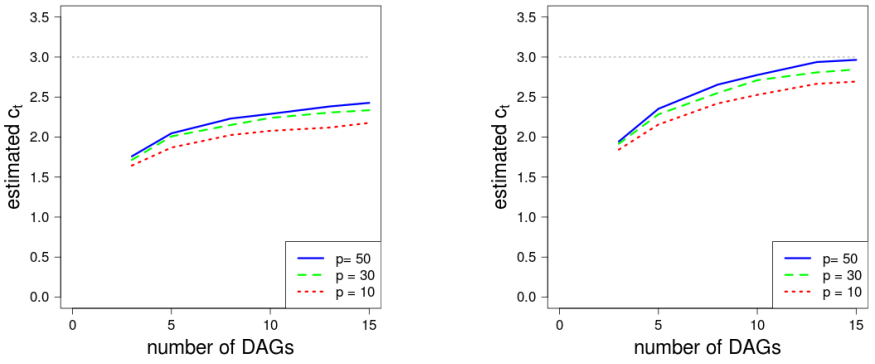
results are shown in Figures 2 and 3. Notice that for plotting SHD, we selected the ℓ_0-penalization parameter $\lambda_1^2 = c\,\frac{\log p}{n}$ with *scaling constant* $c \in \{1, 2, 3, 4, 5\}$ in both joint and separate estimation and then plotted average SHD as a function of the scaling constant c averaged over the K DAGs to be recovered and the 100 realizations. The penalization parameter λ_2 in the second step of the joint estimation procedure was chosen based on 10-fold cross validation. We plotted the average ROC curve where for each choice of tuning parameter, we computed the true positive and false positive rates by averaging over the K DAGs to be recovered and the 100 realizations. It is clear from the two figures that in general joint inference achieves better performance, which matches our theoretical results in Section 4.

However, Figures 2 (a)-(c) and 3 (a)-(c) show also that jointGES performs worse than separate estimation for small scaling constants ($c = 1$). Note that this is in line with our theoretical findings in Theorem 4.11, which imply that whenever Condition 4.7 – which sometimes is a restrictive assumption – is violated, we need to choose a larger penalization parameter.

6.2. *Simulation analysis of Condition 4.7*

In this section we provide simulation results to empirically study the strength of Condition 4.7. We follow the same procedure as in Section 6.1 to generate a collection of DAGs, except that we set $p \in \{10, 30, 50\}$ and $K \in \{3, 5, 8, 10, 13, 15\}$. Then for each randomly generated $\{\mathcal{G}_0^{(k)}\}_{k=1}^K$, we randomly select 10000 permutations and estimate the corresponding c_t by $c_t := \max_{\pi\in\Pi} |\mathcal{G}_{0\pi}^{\text{union}}| / \big(\frac{1}{K}\sum_{k=1}^K |\mathcal{G}_{0\pi}^{(k)}|\big)$.

In Figure 4 we present the estimated value of c_t as a function of K for $p = 10, 30, 50$. Note that for each setting, we plotted the estimated c_t by averaging over 100 realizations. The figure shows that if the private-to-core edge ratio is below 0.6, Condition 4.7



(a) private-to-core edge ratio 0.3 (b) private-to-core edge ratio 0.6

FIG 4. *Averaged c_t across 100 realizations. (a) is the curve when the private-to-core edge ratio is chosen as 0.3; (b) corresponds to the curve with private-to-core edge ratio of 0.6.*

holds with $c_t \leq 3$, even for very large $K = 15$. This implies that when the true DAGs $\{\mathcal{G}_0^{(k)}\}_{k=1}^K$ are highly overlapping, Condition 4.7 can usually be satisfied with a reasonably small constant c_t. This is in line with the results from Section 6.1, where we showed that `jointGES` significantly outperforms the separate estimation approaches.

6.3. Gene regulatory networks of ovarian cancer subtypes

To assess the performance of the proposed joint ℓ_0-penalized maximum likelihood method on real data, we analyzed gene expression microarray data of patients with ovarian cancer [42]. Patients were divided into six subtypes of ovarian cancer, labeled as C1-C6, where C1 is characterized by significant differential expression of genes associated with stromal and immune cell types and with a lower survival rate as compared to the other 5 subtypes. We hence grouped the subtypes C2-C6 together and our goal was to infer the differences in terms of gene regulatory networks in ovarian cancer that could explain the different survival rates. The gene expression data in [42] includes the expression profile of $n = 83$ patients with C1 subtype and $n = 168$ patients with other subtypes. We implemented our jointGES algorithm (Algorithm 1) to jointly learn *two* gene regulatory networks: one corresponding to the C1 subtype $\mathcal{G}_{\mathrm{C1}}$ and another corresponding to the other five subtypes together $\mathcal{G}_{\mathrm{C2-6}}$. In addition, as in [4], we focused on a particular pathway, namely the *apoptosis* pathway. Using the KEGG database [21, 30] we selected the genes in this pathway that were associated with at most two microarray probes, resulting in a total of $p = 76$ genes.

Table 1 lists the number of edges discovered by jointGES as well as for two separate estimation methods, namely using the GES [7] and PC [40] algorithms. All three methods were combined with stability selection [27] in order to increase robustness of the output and provide a fair comparison. As expected, the two graphs inferred using jointGES share a significant proportion of edges, whereas the overlap is markedly smaller for the two separate estimation methods. Interestingly, under all estimation methods the network for the C1 subtype contains fewer edges than the network of the other subtypes, thereby suggesting that $\mathcal{G}_{\mathrm{C1}}$ could lack some important links that are associated with patient survival.

To obtain more insights into the relevance of the obtained networks, we analyzed the inferred hub nodes in the three networks. For our analysis we defined as hub nodes those nodes for which the sum of the in- and out-degree was larger than some threshold T in the union of the two DAGs, where T was chosen such that there are at most 5 hub nodes discovered by each method. For jointGES, this union is given by $\hat{\mathcal{G}}^{\mathrm{union}}$, the graph identified in the first step of Algorithm 1. The hub nodes identified by jointGES are CAPN1, CTSD, LMNB1, CSF2RB, BIRC3. Among these, CAPN1 [13], CTSD [41], LMNB1 [37], and BIRC3 [19] have been reported as being central to ovarian cancer in the existing literature. In addition, CSF2RB was also discovered by joint estimation of undirected graphical models on this data set [4]. The hub nodes discovered by GES are ATF4, BIRC2, CSF2RB, TUBA1C, MAPK3, while PC only discovered the hub node CSF2RB. While we were not able to validate the relevance of any of these genes for ovarian cancer in the literature, interestingly, CSF2RB has been identified as a hub node by all methods, thereby suggesting this gene as an interesting candidate for future

Method	$\lvert\mathcal{G}_{\mathrm{C1}}\rvert$	$\lvert\mathcal{G}_{\mathrm{C2-C6}}\rvert$	$\lvert\mathcal{G}_{\mathrm{C1}} \cap \mathcal{G}_{\mathrm{C2-C6}}\rvert$
JointGES	50	73	48
GES	68	101	32
PC	14	30	9

TABLE 1

Number of edges in the DAGs estimated by different methods for the gene regulatory network of subtype C1 ($\lvert\mathcal{G}_{\mathrm{C1}}\rvert$) and all other subtypes ($\lvert\mathcal{G}_{\mathrm{C2-6}}\rvert$). The last column shows the number of edges shared between both inferred graphs.

genetic intervention experiments.

7. Discussion

In this paper we presented jointGES, an algorithm for the joint estimation of multiple *related* DAG models from independent realizations. Joint estimation is of particular interest in applications where data is collected not from a single DAG, but rather multiple related DAGs, such as gene expression data from different tissues, cell types or from different interventional experiments. JointGES first estimates the union of DAGs $\hat{\mathcal{G}}^{\text{union}}$ by applying a greedy search to approximate the joint ℓ_0-penalized maximum likelihood estimator and then it uses variable selection to discover each DAG as a sub-DAG of $\hat{\mathcal{G}}^{\text{union}}$. From an algorithmic perspective, jointGES is to the best of our knowledge the first method to jointly estimate related DAG models in the high-dimensional setting. From a theoretical perspective, we presented consistency guarantees on the joint ℓ_0-penalized maximum likelihood estimator, and showed that the accuracy bound scales with the total number of samples, rather than the number of samples from each DAG that would be achieved by separately estimating each DAG. As a corollary to this result, we obtained consistency guarantees for ℓ_0-penalized maximum likelihood estimation of a causal graph from a mix of observational and interventional data. Finally, we validated our results via numerical experiments on simulated and real-world data, showing that the proposed jointGES algorithm for joint inference outperforms separate-inference approaches using well-established algorithms such as PC and GES.

The present work serves as a platform for the potential development of multiple future directions. These directions include: i) relaxing the assumption that all DAGs must be consistent with the same underlying permutation; ii) extending jointGES to the setting where the samples come from K related DAGs but it is unknown a priori which particular DAG each sample comes from; this is for example of interest in the analysis of gene expression data from tumors or tissues that consist of a mix of cell types; iii) extending jointGES to allow for latent confounders.

Acknowledgments

The authors thank two anonymous reviewers for their thoughtful comments, which helped improve our paper. Santiago Segarra was supported by an MIT IDSS seed grant. Caroline Uhler was partially supported by NSF (DMS-1651995), ONR (N00014-17-1-2147 and N00014-18-1-2765), IBM, a Sloan Fellowship and a Simons Investigator Award.

Appendix A: Theoretical analysis of statistical consistency results

In the following, we develop the proofs of Theorems 4.9 and 4.11. To facilitate understanding, we first provide a high-level explanation of the rationale behind the proof. If we have data generated from a single DAG and we are given a permutation π consistent with the true DAG a priori, then we can estimate $(\hat{A}, \hat{\Omega})$ by performing p regressions as explained in Section 2.1. By contrast, when the permutation is unknown and we need to consider all the possible permutations, the total number of regressions to run increases to $p \cdot p!$. However, these regressions are not independent and, intuitively, by bounding the noise level of a subset of these regressions, we can derive bounds for the noise on the other ones. We characterize the 'noise level' of these regressions by analyzing the asymptotic properties of three random events. More precisely, whenever these events hold – and we show that they hold with high probability –, the noise is small enough so that the error of the ℓ_0-penalized maximum likelihood estimator can also be bounded. Finally, we use this upper bound in the error to show that the recovered graph converges to a minimal I-MAP that is as sparse as the true DAG.

The remainder of the appendix is organized as follows. In Section A.1 we define the three random events previously mentioned and show that each of them holds with high probability. Section A.2 then leverages the definition of these events to prove Theorem 4.9, our main result. In Section A.3 we prove Theorem 4.11, which relaxes some of the conditions of Theorem 4.9, but uses similar proof techniques. Finally, Section A.4 fleshes out the proof of Corollary 5.1, our result applicable to the setting for interventional data.

Throughout the appendix, we use the following notation. Let $\hat{a}_j$ denote the j-th column of $\hat{A}$ and $\hat{\omega}_j$ denote the j-th diagonal entry of $\hat{\Omega}$. Also, denote by $a_{0j\pi}$ and $\omega_{0j\pi}$ the j-th column of $A_{0\pi}$ and the j-th diagonal entry of $\Omega_{0\pi}$, respectively.

A.1. Random events

As in [45], our proofs of Theorems 4.9 and 4.11 are based on a set of random events. However, the events considered here differ from those in [45] since, as explained in Section 4.1, a naive application of the guarantees in [45] to the joint estimation scenario does not achieve the desired learning rates [cf. discussion after (12)]. Intuitively, the rate gain achieved here comes from the assumption that all DAGs are consistent with a permutation, allowing us to effectively reduce the size of the search space.

In our proofs we consider three random events $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$ that are respectively stated – along with proofs showing that they hold with high probability – in Sections A.1.1, A.1.2, and A.1.3.

A.1.1. Random event $\mathcal{E}_1$

Let $\epsilon_{j\pi}^{(k)} \in \mathbb{R}^n$ denote the residual when regressing $X_j^{(k)}$ on $X_S^{(k)}$ with $S = \{i \,|\, \pi(i) < \pi(j)\}$, i.e., $\epsilon_{j\pi}^{(k)} := X_j^{(k)} - {X^{(k)}}^T a_{0j\pi}^{(k)}$. Similarly, let $\hat{\epsilon}_{j\pi}^{(k)} \in \mathbb{R}^n$ denote the regression residual from the sampled data, i.e., $\hat{\epsilon}_{j\pi}^{(k)} := \hat{X}_j^{(k)} - \hat{X}^{(k)} a_{0j\pi}^{(k)}$. Consider a generic

set $\{A^{(k)}\}_{k=1}^K$ of adjacency matrices consistent with a given permutation π, where the columns of $A^{(k)}$ are denoted by $A^{(k)} := (a_1^{(k)}, \ldots, a_p^{(k)})$, and let $\mathcal{G}^{\text{union}}$ denote the union of the support of $A^{(1)}, \ldots, A^{(K)}$. Then, event $\mathcal{E}_1$ is defined as

$$\begin{aligned}\mathcal{E}_1 := \Bigg\{ & 2\sum_{j=1}^{p}\sum_{k=1}^{K}\frac{w_k}{n_k}\left|\hat{\epsilon}_{j\pi}^{(k)T}\hat{X}^{(k)}(a_j^{(k)} - a_{0j\pi}^{(k)})\right| \\ & \leq \delta_1 \sum_{j=1}^{p}\sum_{k=1}^{K}\frac{w_k}{n_k}\left\|\hat{X}^{(k)}(a_j^{(k)} - a_{0j\pi}^{(k)})\right\|_2^2 + \lambda_1^2\left(|\mathcal{G}^{\text{union}}| + |\mathcal{G}_{0\pi}^{\text{union}}|\right)/\delta_1, \\ & \forall \text{ permutations } \pi \text{ , and } \forall \{A^{(k)}\}_{k=1}^K \text{ consistent with } \pi \Bigg\},\end{aligned} \tag{18}$$

for some constant $\delta_1 > 0$ and some $\lambda_1 \asymp \sqrt{(\log p)/n}$. On random event $\mathcal{E}_1$ a uniform inequality holds across all K DAGs for the *sample correlation* between the regression residual $\epsilon_{j\pi}^{(k)}$ and any random variable spanned by the random vector $X_S^{(k)}$, i.e., ${X^{(k)}}^T v$ for any $v \in \mathbb{R}^p$ with $v_i = 0$ for all $i \notin S$. Notice that for convenience for further steps of the analysis, this generic vector v is written as $a_j^{(k)} - a_{0j\pi}^{(k)}$ in (18). Furthermore, for simplicity in the rest of this appendix, we denominate the space spanned by $X_S^{(k)}$ as the *projection space* of $\epsilon_{j\pi}^{(k)}$. Intuitively, one could foresee $\mathcal{E}_1$ holding since the expected correlation between the regression residual $\epsilon_{j\pi}^{(k)}$ and X_S is equal to zero, and therefore the sample correlation can be upper bounded by a sum of terms that converge to zero as $n \to \infty$ as in (18).

We now show that random event $\mathcal{E}_1$ holds with high probability, a result stated in Theorem A.2. Essential towards proving this result is the observation that, since the random variable ${X^{(k)}}^T(a_j^{(k)} - a_{0j\pi}^{(k)})$ lies within the projection space of $\epsilon_{j\pi}^{(k)}$, these two random variables are independent. We can therefore deal with the randomness in $\hat{\epsilon}_{j\pi}^{(k)}$ and $\hat{X}_S^{(k)}$ separately, one at a time. To formally leverage this observation, we rely on Lemma 7.4 of [45], stated next for completeness.

Lemma A.1. *[45, Lemma 7.4] Let Z be a fixed $n \times m$ matrix and $e_1, \cdots, e_n$ be independent $\mathcal{N}(0, \sigma_e^2)$-distributed random variables. Then for all $t > 0$*

$$\mathbb{P}\left(\sup_{\|Za\|_2^2/n \leq 1} |e^T Za|/n \geq \sigma_e(\sqrt{2m/n} + \sqrt{2t/n})\right) \leq \exp(-t).$$

Based on the above lemma and recalling that $\mathcal{A}_\pi$ denotes the set of adjacency matrices consistent with a given permutation π, we can show the following result.

Theorem A.2. *Assume that Conditions 4.2 and 4.6 hold, then for all $t > 0$ and all*

$\delta_1 > 0$,

$$\mathbb{P}\Bigg(\max_{\pi} \sup_{\{A^{(k)}\}_{k=1}^K \in \mathcal{A}_\pi} 2\sum_{j=1}^{p}\sum_{k=1}^{K} \frac{w_k}{n_k} \left| \hat{\epsilon}_{j\pi}^{(k)T} \hat{X}^{(k)} (a_j^{(k)} - a_{0j\pi}^{(k)}) \right| - \delta_1 \sum_{j=1}^{p}\sum_{k=1}^{K} \frac{w_k}{n_k} \left\| \hat{X}^{(k)} (a_j^{(k)} - a_{0j\pi}^{(k)}) \right\|_2^2 \geq \frac{16\sigma_0^2 (t + 2\log p)(|\mathcal{G}^{union}| + |\mathcal{G}_{0\pi}^{union}|)}{n\delta_1} \Bigg) \leq \exp(-t).$$

Proof. Let $\hat{\epsilon}_{j\pi}$ and $a_{0j\pi}$ be the concatenated vectors $\hat{\epsilon}_{j\pi} := (\hat{\epsilon}_{j\pi}^{(1)T}, \ldots, \hat{\epsilon}_{j\pi}^{(K)T})^T$ and $a_{0j\pi} := (a_{0j\pi}^{(1)T}, \ldots, a_{0j\pi}^{(K)T})^T$. Also, define the block diagonal matrix

$$\hat{X} := \text{diag}(\hat{X}^{(1)}, \cdots, \hat{X}^{(K)}).$$

We denote by $\mathcal{A}_{j\pi} \subset \mathbb{R}^{pK}$ the set containing all vectors that can be formed by vertically concatenating the jth columns $a_j^{(k)}$ for all k and satisfy

$$\mathcal{A}_{j\pi} := \left\{ a_j \in \mathbb{R}^{pK} \mid \forall i, \text{if } \exists k \text{ such that } a_{i,j}^{(k)} \neq 0, \text{ then } X_i^{(k)} \perp\!\!\!\perp \epsilon_{j\pi}^{(k)} \text{ for all } k \right\}.$$

Based on this, and recalling that $\text{Pa}_j(\cdot)$ denotes the set of parent nodes of j in the argument graph, we define the random event $\mathcal{B}_{j\pi}$ as

$$\mathcal{B}_{j\pi} := \Bigg\{ \exists a_j \in \mathcal{A}_{j\pi} : \sup_{\|\hat{X}(a_j - a_{0j\pi})\|_2^2/n \leq 1} \left| \hat{\epsilon}_{j\pi}^T \hat{X} (a_j - a_{0j\pi}) \right| / n \geq \sigma_0 \left(\sqrt{\frac{2K(|\text{Pa}_j(\mathcal{G}^{\text{union}})| + |\text{Pa}_j(\mathcal{G}_{0\pi}^{\text{union}})|)}{n}} + \sqrt{\frac{2(t + |\text{Pa}_j(\mathcal{G}_{0\pi}^{\text{union}})| \log p + 2\log p)}{n}} \right) \Bigg\}. \tag{19}$$

Combining the facts that: i) $a_j - a_{0j\pi}$ may have at most $K(|\text{Pa}_j(\mathcal{G}^{\text{union}})| + |\text{Pa}_j(\mathcal{G}_{0\pi}^{\text{union}})|)$ non-zero entries, and ii) the variance of each element of $\hat{\epsilon}_{j\pi}$ is upper bounded by σ_0^2 (cf. Condition 4.2), we may apply Lemma A.1 to show that

$$\mathbb{P}(\mathcal{B}_{j\pi}) \leq \exp\left(-t - |\text{Pa}_j(\mathcal{G}_{0\pi}^{\text{union}})| \log p - 2\log p\right).$$

As can be seen from (19), event $\mathcal{B}_{j\pi}$ depends exclusively on the set of parents of node j in $\mathcal{G}_{0\pi}^{\text{union}}$. Putting it differently, if for two permutations π_1 and π_2 node j has the same set of parents in $\mathcal{G}_{0\pi_1}^{\text{union}}$ and $\mathcal{G}_{0\pi_2}^{\text{union}}$, then the random events $\mathcal{B}_{j\pi_1}$ and $B_{j\pi_2}$ coincide, since $\mathcal{A}_{j\pi}$, $a_{0j\pi}$ and $\hat{\epsilon}_{j\pi}$ would all be the same for $\pi \in \{\pi_1, \pi_2\}$. If we denote by $\Pi_j(m)$ the set of permutations where node j has exactly m parents in $\mathcal{G}_{0\pi}^{\text{union}}$, then there are at most $\binom{p}{m}$ unique events $\mathcal{B}_{j\pi}$ for all $\pi \in \Pi_j(m)$. We therefore obtain that

$$\mathbb{P}\left(\bigcup_{\pi \in \Pi_j(m)} \mathcal{B}_{j\pi} \right) \leq \binom{p}{m} \exp\left(-t - m\log p - 2\log p\right) \leq \exp\left(-t - 2\log p\right).$$

Applying a union bound on the events $\mathcal{B}_{j\pi}$ across all nodes j and permutations π yields that

$$\mathbb{P}\left(\bigcup_j \bigcup_\pi \mathcal{B}_{j\pi}\right) \leq \sum_{j=1}^{p} \sum_{m=1}^{p-1} \mathbb{P}\left(\bigcup_{\pi \in \Pi_j(m)} \mathcal{B}_{j\pi}\right) \leq p^2 \exp\left(-t - 2\log p\right) = \exp(-t). \tag{20}$$

Combining (20) and (19) it follows that with probability at least $1 - \exp(-t)$, for all j, π and all $a_j \in \mathcal{A}_{j\pi}$,

$$\frac{|\hat{\epsilon}_{j\pi}^T \hat{X}(a_j - a_{0j\pi})|}{\|\hat{X}(a_j - a_{0j\pi})\|_2} \leq \sigma_0 \left(\sqrt{2K(|\mathrm{Pa}_j(\mathcal{G}^{\text{union}})| + |\mathrm{Pa}_j(\mathcal{G}_{0\pi}^{\text{union}})|)} + \sqrt{2(t + |\mathrm{Pa}_j(\mathcal{G}_{0\pi}^{\text{union}})| \log p + 2\log p)} \right).$$

Based on the collection of adjacency matrices $\{A^{(k)}\}_{k=1}^K \in \mathcal{A}_\pi$ we define another collection $\{A'^{(k)}\}_{k=1}^K$ where each column $a'^{(k)}_j$ is given by

$$a'^{(k)}_j = \begin{cases} a_j^{(k)} & \text{if } \hat{\epsilon}_{j\pi}^{(k)T} \hat{X}^{(k)}(a_j^{(k)} - a_{0j\pi}^{(k)}) \geq 0, \\ 2a_{0j\pi}^{(k)} - a_j^{(k)} & \text{otherwise.} \end{cases}$$

Notice that the positions of the non-zero entries in $a'^{(k)}_j - a_{0j\pi}^{(k)}$ coincide with those in $a_j^{(k)} - a_{0j\pi}^{(k)}$. By also using the fact that

$$\begin{aligned} \|\hat{X}(a_j - a_{0j\pi})\|_2^2 &= \sum_{k=1}^K \|\hat{X}^{(k)}(a_j^{(k)} - a_{0j\pi}^{(k)})\|_2^2 \\ &= \sum_{k=1}^K \|\hat{X}^{(k)}(a'^{(k)}_j - a_{0j\pi}^{(k)})\|_2^2 = \|\hat{X}(a'_j - a_{0j\pi})\|_2^2, \end{aligned}$$

we have that for all j and π with probability at least $1 - \exp(-t)$, it holds that

$$\begin{aligned} \frac{\sum_{k=1}^K |\hat{\epsilon}_{j\pi}^{(k)T} \hat{X}^{(k)}(a_j^{(k)} - a_{0j\pi}^{(k)})|}{\|\hat{X}(a_j - a_{0j\pi})\|_2} &= \frac{\sum_{k=1}^K |\hat{\epsilon}_{j\pi}^{(k)T} X^{(k)}(a'^{(k)}_j - a_{0j\pi}^{(k)})|}{\|\hat{X}(a'_j - a_{0j\pi})\|_2} \\ = \frac{|\hat{\epsilon}_{j\pi}^T \hat{X}(a'_j - a_{0j\pi})|}{\|\hat{X}(a'_j - a_{0j\pi})\|_2} &\leq \sigma_0 \left(\sqrt{2K(|\mathrm{Pa}_j(\mathcal{G}^{\text{union}})| + |\mathrm{Pa}_j(\mathcal{G}_{0\pi}^{\text{union}})|)} \right. \\ &\quad \left. + \sqrt{2(t + |\mathrm{Pa}_j(\mathcal{G}_{0\pi}^{\text{union}})| \log p + 2\log p)} \right). \end{aligned}$$

In the above expression we may use that $ab \leq \delta_1 a^2 + b^2/\delta_1$ for all $\delta_1, a, b > 0$ in order

to obtain

$$2|\hat{\epsilon}_{j\pi}^T \hat{X}(a'_j - a_{0j\pi})| \leq \delta_1 \|\hat{X}(a'_j - a_{0j\pi})\|_2^2 + 4\sigma_0^2 \left(\sqrt{2K(|\mathrm{Pa}_j(\mathcal{G}^{\text{union}})| + |\mathrm{Pa}_j(\mathcal{G}_{0\pi}^{\text{union}})|)} + \sqrt{2(t + |\mathrm{Pa}_j(\mathcal{G}_{0\pi}^{\text{union}})| \log p + 2\log p)} \right)^2 / \delta_1. \tag{21}$$

By combining this with the fact that $\forall a, b > 0$, $(a+b)^2 \leq 2(a^2+b^2)$ and the fact that $K = o(\log p)$ from Condition 4.6, we get that

$$\left(\sqrt{2K(|\mathrm{Pa}_j(\mathcal{G}^{\text{union}})| + |\mathrm{Pa}_j(\mathcal{G}_{0\pi}^{\text{union}})|)} + \sqrt{2(t + |\mathrm{Pa}_j(\mathcal{G}_{0\pi}^{\text{union}})| \log p + 2\log p)} \right)^2 \leq 4(t + 2\log p)(|\mathrm{Pa}_j(\mathcal{G}^{\text{union}})| + |\mathrm{Pa}_j(\mathcal{G}_{0\pi}^{\text{union}})|). \tag{22}$$

By replacing (22) back into (21), we obtain that

$$2|\hat{\epsilon}_{j\pi}^T \hat{X}(a'_j - a_{0j\pi})|/n \leq \delta_1 \|\hat{X}(a'_j - a_{0j\pi})\|_2^2/n + \frac{16\sigma_0^2(t + 2\log p)(|\mathrm{Pa}_j(\mathcal{G}^{\text{union}})| + |\mathrm{Pa}_j(\mathcal{G}_{0\pi}^{\text{union}})|)}{n\delta_1}.$$

Rewriting the absolute value in the left-hand side as the sum of the corresponding K absolute values and adding the above inequality for $j = 1, \ldots, p$ we get that, with probability at least $1 - \exp(-t)$,

$$2\sum_{j=1}^{p}\sum_{k=1}^{K} w_k |\hat{\epsilon}_{j\pi}^{(k)T} \hat{X}^{(k)}(a_j^{(k)} - a_{0j\pi}^{(k)})|/n_k \leq \sum_{j=1}^{p} \left(\delta_1 \|\hat{X}(a_j - a_{0j\pi})\|_2^2/n + \frac{16\sigma_0^2(t+2\log p)(|\mathrm{Pa}_j(\mathcal{G}^{\text{union}})| + |\mathrm{Pa}_j(\mathcal{G}_{0\pi}^{\text{union}})|)}{n\delta_1} \right).$$

Finally, by noticing that $\|\hat{X}(a_j - a_{0j\pi})\|_2^2/n = \sum_{k=1}^{K} w_k \|\hat{X}^{(k)}(a_j^{(k)} - a_{0j\pi}^{(k)})\|_2^2/n_k$ we recover the statement of the theorem, thereby concluding the proof. □

A.1.2. Random event $\mathcal{E}_2$

Let $\omega_{0j\pi}^{(k)}$ denote the j-th diagonal entry of $\Omega_{0\pi}^{(k)}$, then $\mathcal{E}_2$ holds whenever the empirical variances of all $\epsilon_{j\pi}^{(k)}$, i.e., $\|\hat{\epsilon}_{j\pi}^{(k)}\|_2^2/n_k$ are close to the true variances $\omega_{0j\pi}^{(k)}$, where

$$\mathcal{E}_2 := \left\{ \sum_{j=1}^{p}\sum_{k=1}^{K} w_k \left(\frac{\|\hat{\epsilon}_{j\pi}^{(k)}\|_2^2/n_k - \omega_{0j\pi}^{(k)}}{\omega_{0j\pi}^{(k)}} \right)^2 \leq 4\,\lambda_2^2 \left(p + |\mathcal{G}_{0\pi}^{\text{union}}|\right) \right\}, \tag{23}$$

for some $\lambda_2 \asymp \sqrt{(\log p)/n}$. Mimicking the development for event $\mathcal{E}_1$, we now show that $\mathcal{E}_2$ also holds with high probability. This result is stated in Theorem A.4. Similar

to the proof of Theorem A.2, we first consider the asymptotic property for a particular node j and permutation π, and then leverage this to get a uniform bound across all permutations and nodes. For this proof, we use the following lemma stated in [4], which also follows from [47]. After the lemma, we formally state our result.

Lemma A.3. *[4, Lemma 2] Suppose $X_1, \cdots, X_n$ are K-dimensional random vectors satisfying $\mathbb{E}X_i = 0$ and $\|X_i\|_2 \leq M$ for $1 \leq i \leq n$. We have for any $\beta > 0$ and $x > \beta$*

$$\begin{aligned}\mathbb{P}(\|\textstyle\sum_{i=1}^n X_i\|_2 \geq x) \leq & \mathbb{P}\left(\|N\|_2 \geq (x-\beta)/\lambda_{\max}^{1/2}\right) \\ & + c_1 K^{5/2} \exp(-c_2 K^{-5/2}\beta/M),\end{aligned}$$

where $\lambda_{\max}$ is the largest eigenvalue of Cov$(\sum_{i=1}^n X_i)$, N is a K-dimensional standard normal random vector and c_1, c_2 are positive constants.

Theorem A.4. *Assume Conditions 4.5 and 4.6 hold, then there exist constants $c_1, c_2, c_3 > 0$ such that*

$$\begin{aligned}\mathbb{P}\left(\exists \pi : \sum_{j=1}^p \sum_{k=1}^K w_k \left(\frac{\|\hat{\epsilon}_{j\pi}^{(k)}\|_2^2/n_k - \omega_{0j\pi}^{(k)}}{\omega_{0j\pi}^{(k)}}\right)^2 \geq c_1 \frac{c_s p \log p + |\mathcal{G}_{0\pi}^{union}| \log p}{n}\right) \\ \leq c_2 \exp(-c_3 \log p),\end{aligned}$$

where c_s is the constant defined in Condition 4.5.

Proof. We begin by analyzing the asymptotic properties of $\epsilon_{j\pi}^{(k)}$ for all k given a fixed permutation π and node j. More specifically, consider the following random event

$$\mathcal{C}_{j\pi} := \left\{ \sum_{k=1}^K w_k \left(\frac{\|\hat{\epsilon}_{j\pi}^{(k)}\|_2^2/n_k - \omega_{0j\pi}^{(k)}}{\omega_{0j\pi}^{(k)}}\right)^2 \geq c_1^2 \frac{\log p \left(|\mathrm{Pa}_j(\mathcal{G}_{0\pi}^{\text{union}})| + c_s\right)}{n} \right\}, \quad (24)$$

for some positive constant c_1. Following the proof of Theorem 1 in [4], we define $u_t^{(k)}$ as follows:

$$u_t^{(k)} := \begin{cases} \frac{\sqrt{w_k}}{n_k}\left(\frac{[\hat{\epsilon}_{j\pi}^{(k)}]_t^2 - \omega_{0j\pi}^{(k)}}{\omega_{0j\pi}^{(k)}}\right) & \text{if } t \leq n_k, \\ 0 & \text{otherwise.}\end{cases}$$

Denoting by $u_t = (u_t^{(1)}, \cdots, u_t^{(K)})^T$ the random vector collecting all $u_t^{(k)}$, by definition it follows that

$$\sum_{k=1}^K w_k \left(\frac{\|\hat{\epsilon}_{j\pi}^{(k)}\|_2^2/n_k - \omega_{0j\pi}^{(k)}}{\omega_{0j\pi}^{(k)}}\right)^2 = \left\|\sum_{t=1}^n u_t\right\|_2^2.$$

A straightforward substitution in (24) allows us to rewrite the event $\mathcal{C}_{j\pi}$ as

$$\mathcal{C}_{j\pi} := \left\{ \left\|\sum_{t=1}^n u_t\right\|_2 \geq c_1 \lambda_n \right\} \quad \text{with} \quad \lambda_n^2 := \frac{\log p \left(|\mathrm{Pa}_j(\mathcal{G}_{0\pi}^{\text{union}})| + c_s\right)}{n}. \quad (25)$$

Notice that in order to apply Lemma A.3 to bound the probability of occurrence of $\mathcal{C}_{j\pi}$, we would need $\|u_t\|_2$ to be smaller than a constant M, which is not true in general. Hence, we split $\mathcal{C}_{j\pi}$ into two subevents, enabling the utilization of Lemma A.3. More precisely, whenever the following random event holds

$$\mathcal{F}_a := \left\{ |u_t^{(k)}| \leq \lambda_n^{-1} K^{1/2-a}/n, \quad \forall t, k \right\},$$

then the ℓ_2 norm of u_t is bounded as follows:

$$\|u_t\|_2 \leq M_a := \lambda_n^{-1} K^{1-a}/n,$$

where a is a free parameter that will be fixed later in the proof. In detail, we bound the probability of $\mathcal{C}_{j\pi}$ according to

$$\mathbb{P}(\mathcal{C}_{j\pi}) \leq \mathbb{P}(\mathcal{C}_{j\pi}|\mathcal{F}_a)\mathbb{P}(\mathcal{F}_a) + \mathbb{P}(\neg\mathcal{F}_a), \tag{26}$$

where we use Lemma A.3 to bound $\mathbb{P}(\mathcal{C}_{j\pi}|\mathcal{F}_a)$ and where $\mathbb{P}(\neg\mathcal{F}_a)$ can be estimated from the chi-squared tail bound [14].

We first focus on bounding $\mathbb{P}(\mathcal{C}_{j\pi}|\mathcal{F}_a)$. For this, we introduce a new variable $\tilde{u}_t$ obtained by truncating the tail of u_t. Formally, recalling that $\mathbf{1}\{\cdot\}$ denotes the indicator function, we have that $\tilde{u}_t := (\tilde{u}_t^{(1)}, \cdots, \tilde{u}_t^{(K)})$ where

$$\tilde{u}_t^{(k)} := u_t^{(k)} \mathbf{1}\left\{ |u_t^{(k)}| \leq \lambda_n^{-1} K^{1/2-a}/n \right\} - \mathbb{E}\left[u_t^{(k)} \mathbf{1}\left\{ |u_t^{(k)}| \leq \lambda_n^{-1} K^{1/2-a}/n \right\} \right].$$

Notice that whenever the random event $\mathcal{F}_a$ holds, then u_t and $\tilde{u}_t$ follow the same distribution except for a shift $v_t^{(k)} := \mathbb{E}\left[u_t^{(k)} \mathbf{1}\left\{ |u_t^{(k)}| \leq \lambda_n^{-1} K^{1/2-a}/n \right\} \right]$. Putting it differently, the distribution of $u_t^{(k)} - \tilde{u}_t^{(k)}$ is a constant $v_t^{(k)}$ whenever we are on the random event $\mathcal{F}_a$. This implies that $\|\sum_{t=1}^n u_t\|_2 \leq n \max_{t,k} |v_t^{(k)}| + \|\sum_{t=1}^n \tilde{u}_t\|_2$.

Therefore, if we can guarantee that $n|v_t^{(k)}| = o(1)\lambda_n$ for all t and k, then there must exist a constant $0 < \delta < 1$ such that if we are on the random event $\mathcal{F}_a$, $\|\sum_{t=1}^n u_t\|_2 \geq c_1\lambda_n$ implies that $\|\sum_{t=1}^n \tilde{u}_t\|_2 \geq (1-\delta)c_1\lambda_n$. Equivalently, we may write

$$\begin{aligned} \mathbb{P}(\mathcal{C}_{j\pi}|\mathcal{F}_a)\mathbb{P}(\mathcal{F}_a) &\leq \mathbb{P}\left(\left\| \sum_{t=1}^n \tilde{u}_t \right\|_2 \geq (1-\delta)c_1\lambda_n \middle| \mathcal{F}_a \right) \mathbb{P}(\mathcal{F}_a) \\ &\leq \mathbb{P}\left(\left\| \sum_{t=1}^n \tilde{u}_t \right\|_2 \geq (1-\delta)c_1\lambda_n \right), \end{aligned} \tag{27}$$

where the second inequality follows from Bayes' theorem, and we can bound the last term by applying Lemma A.3 since $\tilde{u}_t$ is bounded by definition. We now show that, indeed, $n|v_t^{(k)}| = o(1)\lambda_n$ for all t and k. From the cumulative tail bound of a chi-squared random variable with one degree of freedom we have that $\mathbb{P}(u_t^{(k)} \geq l) \leq \exp(-\eta n l/\sqrt{K})$ for some constant $\eta > 0$. Based on this, we can estimate the scale of $v_t^{(k)}$ with respect to p and n as

$$\begin{aligned} |v_t^{(k)}| &= \left| \mathbb{E}\left[u_t^{(k)} \mathbf{1}\left\{ |u_t^{(k)}| \leq \lambda_n^{-1} K^{1/2-a}/n \right\} \right] \right| \\ &= \left| \mathbb{E}\left[u_t^{(k)} \mathbf{1}\left\{ |u_t^{(k)}| > \lambda_n^{-1} K^{1/2-a}/n \right\} \right] \right| \leq \exp\left(-\eta' \lambda_n^{-1} K^{-a}\right) \end{aligned}$$

for some $0 < \eta' < \eta$, where the second equality follows from the fact that $u_t^{(k)}$ has zero mean. Notice that in the last inequality, $u_t^{(k)}$ has been absorbed into the exponential term. As $|v_t^{(k)}|$ decays exponentially with respect to λ_n^{-1}, we have that $n|v_t^{(k)}| = o(1)\lambda_n$. Having justified this, we may now apply Lemma A.3 to the rightmost term in (27) in order to bound $\mathbb{P}(\mathcal{C}_{j\pi}|\mathcal{F}_a)\mathbb{P}(\mathcal{F}_a)$. From the definition of $\tilde{u}_t$ it follows that $\lambda_{\max}\{\mathrm{Cov}(\sum_{t=1}^n \tilde{u}_t)\} \leq \lambda_{\max}\{\mathrm{Cov}(\sum_{t=1}^n u_t)\}$. Furthermore, since the variables $u_t^{(k)}$ are independently distributed for all t, we have that

$$\mathrm{var}\left(\sum_{t=1}^n u_t^{(k)}\right) = \frac{w_k}{n_k}\mathrm{var}\left(\frac{[\hat{\epsilon}_{j\pi}^{(k)}]_t^2 - \omega_{0j\pi}^{(k)}}{\omega_{0j\pi}^{(k)}}\right) \leq c_2/n$$

for some constant $c_2 > 0$. This also implies that $\lambda_{\max}\{\mathrm{Cov}(\sum_{t=1}^n \tilde{u}_t)\} \leq c_2/n$. Applying Lemma A.3, where we select $x = (1-\delta)c_1\lambda_n$, $\beta = \delta' x$ for some arbitrary positive constant $0 < \delta' < 1$ and, $M = \lambda_n^{-1}K^{1-a}/n$, it follows that

$$\mathbb{P}\left(\left\|\sum_{t=1}^n \tilde{u}_t\right\|_2 \geq (1-\delta)c_1\lambda_n\right) \leq \exp(-c_3(n\lambda_n^2 - K)) + \exp\left(\frac{2}{5}\log K - c_4 K^{a-\frac{7}{2}} n\lambda_n^2\right),$$

for some constants $c_3, c_4 > 0$ that increase if constant c_1 is increased. In addition, by choosing $a = 7/2$, there must exist a large enough c_1 such that $c_3, c_4 > 1$ and therefore

$$\mathbb{P}\left(\left\|\sum_{t=1}^n \tilde{u}_t\right\|_2 \geq (1-\delta)c_1\lambda_n\right) \leq \exp(-n\lambda_n^2) = \exp(-|\mathrm{Pa}_j(\mathcal{G}^{\text{union}})|\log p - c_s \log p). \tag{28}$$

Replacing (28) into (27) gives us the sought exponential bound for the first summand in (26).

We are now left with the task of finding a bound for $\mathbb{P}(\neg\mathcal{F}_a)$. By relying on the fact that $n^{(1)} \asymp \cdots \asymp n^{(K)}$ (cf. Condition 4.6), we get that $(\max_k\ n_k)\, K \leq c_5 n$ for some constant c_5 and therefore

$$\mathbb{P}(\neg\mathcal{F}_a) \leq c_5 n \max_{1\leq k\leq K, 1\leq l\leq n} \mathbb{P}\left\{|u_t^{(k)}| \geq \lambda_n^{-1}K^{1/2-a}/n\right\}.$$

Following this, in order to bound the probability that $\neg\mathcal{F}_a$ holds we further rely on the tail bound of the chi-squared random variable $u_t^{(k)}$ to obtain

$$\mathbb{P}(\neg\mathcal{F}_a) \leq c_5 n \exp\left(-\eta\lambda_n^{-1}K^{-a}\right) = c_5 \exp\left(\log n - \eta\lambda_n^{-1}K^{-a}\right).$$

Recalling the definition of λ_n from (25), Condition 4.5 implies that

$$\frac{\lambda_n^{-1}}{K^{7/2}} \geq \log p(|\mathrm{Pa}_j(\mathcal{G}^{\text{union}})| + c_s)/\tilde{\alpha}^{\frac{3}{2}} \quad \text{and} \quad \sqrt{\tilde{\alpha}}\frac{\lambda_n^{-1}}{K^{7/2}} \geq \log n. \tag{29}$$

By recalling that we have fixed $a = \frac{7}{2}$, it follows that there exists a constant η' such that

$$\begin{aligned}\mathbb{P}(\neg\mathcal{F}_a) &\leq c_5 \exp\left(\log n - \eta\lambda_n^{-1}K^{-a}\right) \\ &\leq c_5 \exp\left(\sqrt{\tilde{\alpha}}\lambda_n^{-1}K^{-a} - \eta\lambda_n^{-1}K^{-a}\right) \leq c_5 \exp\left(-\eta'\lambda_n^{-1}K^{-a}\right),\end{aligned}$$

where we have used the second inequality in (29). Furthermore, by leveraging the first inequality in (29) we obtain that

$$\mathbb{P}(\neg\mathcal{F}_a) \leq c_5 \exp\left(-\eta'\lambda_n^{-1}K^{-a}\right) \leq c_5 \exp\left(-\eta' \log p \left(|\mathrm{Pa}_j(\mathcal{G}^{\text{union}})| + c_s\right)/\tilde{\alpha}^{\frac{3}{2}}\right),$$

thus obtaining an exponential bound for the second summand in (26).

Having found exponential bounds for both summands in (26), it follows that for $c_1 > 0$ sufficiently large and $\tilde{\alpha}$ sufficiently small we have that

$$\mathbb{P}(\mathcal{C}_{j\pi}) \leq (1 + c_5)\exp(-|\mathrm{Pa}_j(\mathcal{G}^{\text{union}})|\log p - c_s \log p).$$

Following an argument based on union bounds similar to the one presented in the proof of Theorem A.2, we have that

$$\begin{aligned}\mathbb{P}\left(\sum_{k=1}^{K} w_k \left(\frac{\|\hat{\epsilon}_{j\pi}^{(k)}\|_2^2/n_k - \omega_{0j\pi}^{(k)}}{\omega_{0j\pi}^{(k)}}\right)^2 \leq c_1 \frac{\log p \left(|\mathrm{Pa}_j(\mathcal{G}^{\text{union}})| + c_s\right)}{n}, \quad \forall\, j, \pi\right) \\ \geq 1 - \mathbb{P}\left(\bigcup_{j=1}^{p}\bigcup_{m=1}^{p}\bigcup_{\pi\in\Pi_j(m)} \mathcal{C}_{j\pi}\right) \geq 1 - (1 + c_5)\exp(-(c_s - 2)\log p)\end{aligned}$$

for some constant $c_1 > 0$. It is immediately implied from the previous expression that

$$\begin{aligned}\mathbb{P}\left(\exists\pi : \sum_{j=1}^{p}\sum_{k=1}^{K} w_k \left(\frac{\|\hat{\epsilon}_{j\pi}^{(k)}\|_2^2/n_k - \omega_{0j\pi}^{(k)}}{\omega_{0j\pi}^{(k)}}\right)^2 \geq c_1 \frac{c_s p \log p + |\mathcal{G}_{0\pi}^{\text{union}}|\log p}{n}\right) \\ \leq (1 + c_5)\exp(-(c_s - 2)\log p),\end{aligned}$$

thus recovering the statement of the theorem (since $c_s > 2$ by Assumption 4.5) after accordingly renaming the constants on the right-hand side. □

A.1.3. Random event $\mathcal{E}_3$

Event $\mathcal{E}_3$ is defined as the intersection of $2K$ events that we denote by $\{\mathcal{E}_{3a}^{(k)}\}_{k=1}^K$ and $\{\mathcal{E}_{3b}^{(k)}\}_{k=1}^K$, where events $\mathcal{E}_{3a}^{(k)}$ and $\mathcal{E}_{3b}^{(k)}$ are specific to the kth SEM. More specifically, event $\mathcal{E}_{3a}^{(k)}$ ensures that all the estimated noise variances $\hat{\omega}_j^{(k)}$ associated with the kth SEM are finite and bounded away from zero. Formally, we define the following events for $k = 1, \ldots, K$:

$$\mathcal{E}_{3a}^{(k)} := \left\{\min\left(\hat{\omega}_j^{(k)}, 1/\hat{\omega}_j^{(k)}\right) \geq 1/\beta^2, \quad \text{for } j = 1, \ldots, p\right\}, \tag{30}$$

for some $\beta > 0$. Event $\mathcal{E}_{3b}^{(k)}$ imposes a universal lower bound on the norm achievable by any linear combinations of the data associated with the k-th DAG. Mathematically, we consider the ensuing events for $k = 1, \ldots, K$:

$$\mathcal{E}_{3b}^{(k)} := \left\{ \|\hat{X}^{(k)} v\|_2 / \sqrt{n_k} \geq \left(\delta_3 - \lambda_3^{(k)} \sqrt{\|v\|_0} \right) \|v\|_2, \quad \forall v \in \mathbb{R}^p \right\}, \tag{31}$$

for some $\delta_3 > 0$ and $\lambda_3^{(k)} \asymp \sqrt{(\log p)/n_k}$. Based on (30) and (31) we define events $\mathcal{E}_3^{(k)} := \mathcal{E}_{3a}^{(k)} \cap \mathcal{E}_{3b}^{(k)}$, and

$$\mathcal{E}_3 := \bigcap_{k=1}^{K} \mathcal{E}_3^{(k)}. \tag{32}$$

Leveraging the fact that Condition 4.4 enforces the maximum in-degree of each $\mathcal{G}_{0\pi}^{(k)}$ to be at most $\alpha n_k / \log p$ for some positive constant α, we can generalize Lemmas 7.5 and 7.7 from [45] into the following lemma.

Lemma A.5. *[45, Lemmas 7.5 and 7.7] Assume Conditions 4.2, 4.3, 4.4 and 4.5 hold and that*

$$3\sqrt{\Lambda_{\min}}/4 - \sqrt{\frac{2(t + \log p)}{n}} - 3\sigma_0 \sqrt{\alpha + \tilde{\alpha}} \geq 1/\beta > 0,$$

for some $t > 0$. Based on this, define

$$\lambda_3^{(k)} := 3\sigma_0 \sqrt{\frac{\log p}{n_k}}, \quad \text{and} \quad \delta_3 := 3\sqrt{\Lambda_{\min}}/4 - \sqrt{\frac{2(t + \log p)}{n_k}}.$$

Then $\mathbb{P}(\mathcal{E}_3^{(k)}) \geq 1 - 5\exp(-t)$ and on $\mathcal{E}_3^{(k)}$ it holds that

$$\|\hat{X}^{(k)}(a_j^{(k)} - a_{0j\hat{\pi}}^{(k)})\|_2 / \sqrt{n_k} \geq \|a_j^{(k)} - a_{0j\hat{\pi}}^{(k)}\|_2 / \beta^2. \tag{33}$$

Lemma A.5 shows that under certain conditions the events $\mathcal{E}_3^{(k)}$ hold with high probability, thus playing a role analogous to that of Theorem A.2 for event $\mathcal{E}_1$ and Theorem A.4 for event $\mathcal{E}_2$.

With the events $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$ defined and having shown under which conditions these hold with high probability, in the next section we leverage these events to prove Theorem 4.9.

A.2. Proof of Theorem 4.9

A.2.1. Bounds on new probability space

Through direct manipulation of the likelihood function, in Lemma A.6 we show that the global optimum $(\hat{A}^{(k)}, \hat{\Omega}^{(k)})_{k=1}^K$ converges to the SEMs $(A_{0\hat{\pi}}^{(k)}, \Omega_{0\hat{\pi}}^{(k)})_{k=1}^K$, where $\hat{\pi}$ is some permutation consistent with the estimated adjacency matrices $\hat{A}^{(1)}, \cdots, \hat{A}^{(K)}$.

Lemma A.6. *Assume we are on $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ and Condition 4.2 holds. Consider a regularizer in (7) satisfying $\lambda^2 > \lambda_1^2/\delta_1 + \lambda_2^2/\delta_2$ with $0 < \delta_1 < 1/\beta^2$ and $0 < \delta_2 < 1/(2\beta^2\sigma_0^2)$. Then, $\left\{\hat{\pi}, \{(\hat{A}^{(k)}, \hat{\Omega}^{(k)})\}_{k=1}^K\right\}$ the global optimum of (7), satisfies*

$$\begin{aligned}
&\left(\frac{1}{\beta^2} - \delta_1\right) \sum_{j=1}^{p} \sum_{k=1}^{K} w_k \|X^{(k)}(\hat{a}_j^{(k)} - a_{0j\hat{\pi}}^{(k)})\|_2^2/n_k \\
&+ \left(\frac{1}{2\beta^4\sigma_0^4} - \delta_2\right) \sum_{j=1}^{p} \sum_{k=1}^{K} w_k \left(\frac{\hat{\omega}_j^{(k)} - \omega_{0j}^{(k)}}{\hat{\omega}_j^{(k)}}\right)^2 \\
&+ \left(\lambda^2 - \frac{\lambda_1^2}{\delta_1} - \frac{\lambda_2^2}{\delta_2}\right) |\hat{\mathcal{G}}| \leq \lambda^2 |\mathcal{G}_0^{union}| + \frac{\lambda_2^2(p + |\mathcal{G}_{0\hat{\pi}}^{union}|)}{\delta_2} + \frac{\lambda_1^2 |\mathcal{G}_{0\hat{\pi}}^{union}|}{\delta_1}.
\end{aligned} \tag{34}$$

Proof. By definition, the global optimum must satisfy

$$\sum_{k=1}^{K} w_k \ell_{n_k}(\hat{X}^{(k)}; \hat{A}^{(k)}, \hat{\Omega}^{(k)}) - \lambda^2 |\hat{\mathcal{G}}| \geq \sum_{k=1}^{K} w_k \ell_{n_k}(\hat{X}^{(k)}; A_0^{(k)}, \Omega_0^{(k)}) - \lambda^2 |\mathcal{G}_0^{\text{union}}|. \tag{35}$$

Let $\hat{\pi}$ denote any permutation consistent with all $\hat{A}^{(k)}$. Since the value of the likelihood $\ell_{n_k}(\hat{X}^{(k)}; A_0^{(k)}, \Omega_0^{(k)})$ is completely determined by the precision matrices $\{\Theta_0^{(k)}\}_{k=1}^K$, it then follows that the likelihood function $\ell_{n_k}(\hat{X}^{(k)}; A_0^{(k)}, \Omega_0^{(k)})$ and the function $\ell_{n_k}(\hat{X}^{(k)}; A_{0\hat{\pi}}^{(k)}, \Omega_{0\hat{\pi}}^{(k)})$ achieve the same value. We therefore replace the former by the latter in (35) and expand the definition of the likelihood function in (5) to obtain

$$\begin{aligned}
&p + \sum_{j=1}^{p} \sum_{k=1}^{K} w_k \log \hat{\omega}_j^{(k)} + \lambda^2 |\hat{\mathcal{G}}| \\
&\leq \sum_{j=1}^{p} \sum_{k=1}^{K} w_k \frac{\|\hat{\epsilon}_{j\hat{\pi}}^{(k)}\|_2^2/n_k}{\omega_{0j\hat{\pi}}^{(k)}} + \sum_{j=1}^{p} \sum_{k=1}^{K} w_k \log \omega_{0j\hat{\pi}}^{(k)} + \lambda^2 |\mathcal{G}_0^{\text{union}}|.
\end{aligned}$$

Basic manipulations transform the above expression into the following inequality

$$\sum_{j=1}^{p} \sum_{k=1}^{K} w_k \log\left(\frac{\hat{\omega}_j^{(k)}}{\omega_{0j\hat{\pi}}^{(k)}}\right) + \lambda^2 |\hat{\mathcal{G}}| \leq \sum_{j=1}^{p} \sum_{k=1}^{K} w_k \left(\frac{\|\hat{\epsilon}_{j\hat{\pi}}^{(k)}\|_2^2/n_k}{\omega_{0j\hat{\pi}}^{(k)}} - 1\right) + \lambda^2 |\mathcal{G}_0^{\text{union}}|. \tag{36}$$

Since we are on $\mathcal{E}_3$, we have that $1/\hat{\omega}_j^{(k)} \leq \beta^2$ [cf. (30)]. By combining this with Condition 4.2 we can further bound $\omega_{0j\hat{\pi}}^{(k)}/\hat{\omega}_j^{(k)} \leq \beta^2\sigma_0^2$. Then, using the Taylor expansion $\log(1+x) \leq x - x^2/(2(1+t)^2)$, for $-1 < x \leq t$, we can further replace $\log\left(\frac{\hat{\omega}_j^{(k)}}{\omega_{0j\hat{\pi}}^{(k)}}\right)$

in (36) to obtain

$$\sum_{j=1}^{p}\sum_{k=1}^{K} w_k \left(\frac{\hat{\omega}_j^{(k)} - \omega_{0j\hat{\pi}}^{(k)}}{\hat{\omega}_j^{(k)}}\right) + \frac{1}{2\beta^4\sigma_0^4}\left(\frac{\omega_{0j\hat{\pi}}^{(k)}}{\hat{\omega}_j^{(k)}} - 1\right)^2 + \lambda^2|\hat{\mathcal{G}}| \leq \sum_{j=1}^{p}\sum_{k=1}^{K} w_k \left(\frac{\|\hat{\epsilon}_{j\hat{\pi}}^{(k)}\|_2^2/n_k}{\omega_{0j\hat{\pi}}^{(k)}} - 1\right) + \lambda^2|\mathcal{G}_0^{\text{union}}|. \tag{37}$$

Finally, using the fact that $\hat{X}_j^{(k)} = \hat{\epsilon}_{j\hat{\pi}}^{(k)} + \hat{X}^{(k)} a_{0j\hat{\pi}}^{(k)}$, we also rewrite $\hat{\omega}_j^{(k)}$ as

$$\begin{aligned}\hat{\omega}_j^{(k)} &= \|\hat{X}_j^{(k)} - \hat{X}^{(k)}\hat{a}_j^{(k)}\|_2^2/n_k \\ &= \|\hat{X}^{(k)}(\hat{a}_j^{(k)} - a_{0j\hat{\pi}}^{(k)})\|_2^2/n_k - 2\hat{\epsilon}_{j\hat{\pi}}^{(k)T}\hat{X}^{(k)}(\hat{a}_j^{(k)} - a_{0j\hat{\pi}}^{(k)})/n_k + \|\hat{\epsilon}_j^{(k)}\|_2^2/n_k.\end{aligned}$$

By replacing the above into (37), we get that

$$\begin{aligned}&\sum_{j=1}^{p}\sum_{k=1}^{K} w_k \frac{\|\hat{X}^{(k)}(\hat{a}_j^{(k)} - a_{0j\hat{\pi}}^{(k)})\|_2^2/n_k}{\hat{\omega}_j^{(k)}} + \frac{1}{2\beta^4\sigma_0^4}\sum_{j=1}^{p}\sum_{k=1}^{K} w_k \left(\frac{\hat{\omega}_j^{(k)} - \omega_{0j\hat{\pi}}^{(k)}}{\hat{\omega}_j^{(k)}}\right)^2 + \lambda^2|\hat{\mathcal{G}}| \\ &\leq 2\sum_{j=1}^{p}\sum_{k=1}^{K} w_k \frac{\hat{\epsilon}_{j\hat{\pi}}^{(k)T}\hat{X}^{(k)}(\hat{a}_j^{(k)} - a_{0j\hat{\pi}}^{(k)})/n_k}{\hat{\omega}_j^{(k)}} + \sum_{j=1}^{p}\sum_{k=1}^{K} w_k \left(\frac{\|\hat{\epsilon}_{j\hat{\pi}}^{(k)}\|_2^2/n_k}{\omega_{0j\hat{\pi}}^{(k)}} - 1\right) \\ &\quad - \sum_{j=1}^{p}\sum_{k=1}^{K} w_k \left(\frac{\|\hat{\epsilon}_{j\hat{\pi}}^{(k)}\|_2^2/n_k - \omega_{0j\hat{\pi}}^{(k)}}{\hat{\omega}_j^{(k)}}\right) + \lambda^2|\mathcal{G}_0^{\text{union}}|.\end{aligned} \tag{38}$$

In order to further bound the expression in (38), notice that the first summand in the right hand side of the inequality corresponds to the sum of all empirical correlation coefficients. Leveraging that we are under the assumption that $\mathcal{E}_1$ holds [cf. (18)], we have that

$$2\sum_{j=1}^{p}\sum_{k=1}^{K} w_k \frac{\hat{\epsilon}_{j\hat{\pi}}^{(k)T}\hat{X}^{(k)}(\hat{a}_j^{(k)} - a_{0j\hat{\pi}}^{(k)})/n_k}{\hat{\omega}_j^{(k)}} \leq \delta_1 \sum_{j=1}^{p}\sum_{k=1}^{K} w_k \|\hat{X}^{(k)}(\hat{a}_j^{(k)} - a_{0j\hat{\pi}}^{(k)})\|_2^2/n_k + \frac{\lambda_1^2}{\delta_1}|\hat{\mathcal{G}}|. \tag{39}$$

In order to bound the second and third terms, we first restate their difference as follows

$$\begin{aligned}&\sum_{j=1}^{p}\sum_{k=1}^{K} w_k \left(\frac{\|\hat{\epsilon}_{j\hat{\pi}}^{(k)}\|_2^2/n_k - \omega_{0j\hat{\pi}}^{(k)}}{\hat{\omega}_j^{(k)}}\right) - w_k \left(\frac{\|\hat{\epsilon}_{j\hat{\pi}}^{(k)}\|_2^2/n_k}{\omega_{0j\hat{\pi}}^{(k)}} - 1\right) \\ &= \sum_{j=1}^{p}\sum_{k=1}^{K} w_k \left(\frac{\|\hat{\epsilon}_{j\hat{\pi}}^{(k)}\|_2^2/n_k - \omega_{0j\hat{\pi}}^{(k)}}{\omega_{0j\hat{\pi}}^{(k)}}\right)\left(\frac{\omega_{0j\hat{\pi}}^{(k)} - \hat{\omega}_j^{(k)}}{\hat{\omega}_j^{(k)}}\right).\end{aligned} \tag{40}$$

Next, by using Cauchy-Schwarz inequality, i.e.

$$|\sum_{i=1}^{n} u_i v_i|^2 \leq \sum_{j=1}^{n} |u_j|^2 \sum_{k=1}^{n} |v_k|^2,$$

we further bound (40) as

$$\begin{aligned}
&\left| \sum_{j=1}^{p} \sum_{k=1}^{K} w_k \left(\frac{\|\hat{\epsilon}_{j\hat{\pi}}^{(k)}\|_2^2/n_k - \omega_{0j\hat{\pi}}^{(k)}}{\hat{\omega}_j^{(k)}} \right) - \sum_{j=1}^{p} \sum_{k=1}^{K} w_k \left(\frac{\|\hat{\epsilon}_{j\hat{\pi}}^{(k)}\|_2^2/n_k}{\omega_{0j\hat{\pi}}^{(k)}} - 1 \right) \right| \\
&\leq \left(\sum_{j=1}^{p} \sum_{k=1}^{K} w_k \left(\frac{\|\hat{\epsilon}_{j\hat{\pi}}^{(k)}\|_2^2/n_k - \omega_{0j\hat{\pi}}^{(k)}}{\omega_{0j\hat{\pi}}^{(k)}} \right)^2 \right)^{1/2} \left(\sum_{j=1}^{p} \sum_{k=1}^{K} w_k \left(\frac{\omega_{0j\hat{\pi}}^{(k)} - \hat{\omega}_j^{(k)}}{\hat{\omega}_j^{(k)}} \right)^2 \right)^{1/2}.
\end{aligned} \tag{41}$$

From the fact that event $\mathcal{E}_2$ holds [cf. (23)], we can upper bound the first of the two factors in the right-hand side of (41) by $2\sqrt{\lambda_2^2(p + |\mathcal{G}_0^{\text{union}}(\hat{\pi})|)}$. Further, relying on the inequality $2ab \leq a^2/\delta_2 + \delta_2 b^2$ for any $\delta_2 > 0$, it follows that

$$\begin{aligned}
&\left| \sum_{j=1}^{p} \sum_{k=1}^{K} w_k \left(\frac{\|\hat{\epsilon}_{j\hat{\pi}}^{(k)}\|_2^2/n_k - \omega_{0j\hat{\pi}}^{(k)}}{\hat{\omega}_j^{(k)}} \right) - \sum_{j=1}^{p} \sum_{k=1}^{K} w_k \left(\frac{\|\hat{\epsilon}_{j\hat{\pi}}^{(k)}\|_2^2/n_k}{\omega_{0j\hat{\pi}}^{(k)}} - 1 \right) \right| \\
&\qquad \leq \frac{\lambda_2^2(p + |\mathcal{G}_0^{\text{union}}(\hat{\pi})|)}{\delta_2} + \delta_2 \sum_{j=1}^{p} \sum_{k=1}^{K} w_k \left(\frac{\omega_{0j\hat{\pi}}^{(k)} - \hat{\omega}_j^{(k)}}{\hat{\omega}_j^{(k)}} \right)^2.
\end{aligned} \tag{42}$$

By replacing (39) and (42) into (38), we recover (34), as we wanted to show. □

From Lemma A.6 it follows that the global optimum of (7) corresponds to a minimal I-MAP, but no claim is made about the sparsity level of this I-MAP. In order to show that the solution is indeed sparse, we must rely on Conditions 4.7 and 4.8. In Lemma A.7 we show that $|\mathcal{G}_{0\hat{\pi}}^{\text{union}}|$ cannot be much larger than $|\hat{\mathcal{G}}|$. Then, in Thm. A.8 we further show how to cancel out $|\mathcal{G}_{0\hat{\pi}}^{\text{union}}|$ with $|\hat{\mathcal{G}}|$ in (34) to obtain our main result.

Lemma A.7. *Assume Condition 4.7 holds and let $\tilde{\lambda} > 0$ be such that*

$$\sum_{k=1}^{K} w_k \|\hat{A}^{(k)} - A_{0\hat{\pi}}^{(k)}\|_F^2 \leq \tilde{\lambda}^2 |\mathcal{G}_{0\hat{\pi}}^{\text{union}}|. \tag{43}$$

Consider constants η_1, η_2 with $0 \leq \eta_1 < 1$ and $0 < \eta_2^2 c_t < 1 - \eta_1$ such that $\sum_{i,j} \mathbf{1}\left\{ \left| [A_{0\hat{\pi}}^{(k)}]_{i,j} \right| \geq \tilde{\lambda}/\eta_2 \right\} \geq (1 - \eta_1)|\mathcal{G}_{0\hat{\pi}}^{(k)}|$. Then, it follows that

$$|\hat{\mathcal{G}}| \geq \frac{1 - \eta_1 - \eta_2^2 c_t}{c_t} |\mathcal{G}_{0\hat{\pi}}^{\text{union}}|.$$

Proof. Let $\mathcal{N}^{(k)}$ and $\mathcal{M}^{(k)}$ be the sets of entries satisfying

$$\mathcal{N}^{(k)} := \{(i,j) : |[A_{0\hat{\pi}}^{(k)}]_{i,j}| \geq \tilde{\lambda}/\eta_2\}$$

and

$$\mathcal{M}^{(k)} := \{(i,j) : |[\hat{A}^{(k)}]_{i,j} - [A^{(k)}_{0\hat{\pi}}]_{i,j}| \geq \tilde{\lambda}/\eta_2\}.$$

From these definitions it follows that

$$\begin{aligned}\sum_{k=1}^{K} w_k |\mathcal{N}^{(k)} \cap \mathcal{M}^{(k)}| \frac{\tilde{\lambda}^2}{\eta_2^2} &\leq \sum_{k=1}^{K} w_k \sum_{(i,j) \in \mathcal{N}^{(k)} \cap \mathcal{M}^{(k)}} |[\hat{A}^{(k)}]_{i,j} - [A^{(k)}_{0\hat{\pi}}]_{i,j}|^2 \\ &\leq \sum_{k=1}^{K} w_k \|\hat{A}^{(k)} - A^{(k)}_{0\hat{\pi}}\|_F^2.\end{aligned}$$

Leveraging inequality (43) and Condition 4.7 we further have that

$$\sum_{k=1}^{K} w_k |\mathcal{N}^{(k)} \cap \mathcal{M}^{(k)}| \leq \eta_2^2 |\mathcal{G}^{\text{union}}_{0\hat{\pi}}| \leq \eta_2^2 c_t \sum_{k=1}^{K} w_k |\mathcal{G}^{(k)}_{0\hat{\pi}}|. \tag{44}$$

Notice that for all (i,j)-th entries in the set $\mathcal{N}^{(k)} \cap \mathcal{M}^{(k)^{\mathcal{C}}}$ it must be that $|[\hat{A}^{(k)}]_{i,j}| > 0$. Hence, $\mathcal{N}^{(k)} \cap \mathcal{M}^{(k)^{\mathcal{C}}}$ corresponds to a subset of non-zero entries of $\hat{A}^{(k)}$, which in turn corresponds to a subset of edges in $\hat{\mathcal{G}}$. From this we can infer that

$$\begin{aligned}|\hat{\mathcal{G}}| = \sum_{k=1}^{K} w_k |\hat{\mathcal{G}}| &\geq \sum_{k=1}^{K} w_k |\mathcal{N}^{(k)} \cap \mathcal{M}^{(k)^{\mathcal{C}}}| = \sum_{k=1}^{K} w_k (|\mathcal{N}^{(k)}| - |\mathcal{N}^{(k)} \cap \mathcal{M}^{(k)}|) \\ &\geq (1 - \eta_1 - \eta_2^2 c_t) \sum_{k=1}^{K} w_k |\mathcal{G}^{(k)}_{0\hat{\pi}}|,\end{aligned}$$

where the last inequality follows by combining (44) with the definition of η_1 in the statement of the lemma. The proof concludes by replacing Condition 4.7 in the above inequality. □

Theorem A.8. *Assume Conditions 4.1, 4.7 and 4.8 hold, and suppose that there exist constants δ_B and $0 < \delta_s < 1$ as well as λ and λ_0 that scale as $\lambda^2 \asymp \lambda_0^2 \asymp \frac{\log p}{n}(p/|\mathcal{G}_0^{union}| \vee 1)$ such that*

$$\delta_B \sum_{k=1}^{K} w_k \|\hat{A}^{(k)} - A^{(k)}_{0\hat{\pi}}\|_F^2 + \lambda^2 \delta_s |\hat{\mathcal{G}}| \leq \lambda^2 |\mathcal{G}_0^{union}| + \lambda_0^2 |\mathcal{G}^{union}_{0\hat{\pi}}|. \tag{45}$$

If the constant η_0 in Condition 4.8 is sufficiently small, then there exist constants $\delta_s', c_g, c_g' > 0$ such that

$$\delta_B \sum_{k=1}^{K} w_k \|\hat{A}^{(k)} - A^{(k)}_{0\hat{\pi}}\|_F^2 + (\lambda^2 \delta_s - \lambda_0^2 \delta_s')|\hat{\mathcal{G}}| \leq \lambda^2 |\mathcal{G}_0^{union}| \tag{46}$$

and

$$|\hat{\mathcal{G}}| \geq c_g |\mathcal{G}^{union}_{0\hat{\pi}}| \geq c_g' |\mathcal{G}_0^{union}|. \tag{47}$$

Proof. Using Conditions 4.1 and 4.7, we have that $|\mathcal{G}_0^{\text{union}}| \leq c_t \max_k |\mathcal{G}_0^{(k)}| \leq c_t\, |\mathcal{G}_{0\hat{\pi}}^{\text{union}}|$. Suppose that for some $\tilde{\lambda} > 0$ one has that $\tilde{\lambda}^2 \asymp \lambda^2 \asymp \lambda_0^2$ and $\tilde{\lambda}^2 \delta_B \geq \lambda^2 c_t + \lambda_0^2$, then it follows from (45) that

$$\delta_B \sum_{k=1}^{K} w_k \|\hat{A}^{(k)} - A_{0\hat{\pi}}^{(k)}\|_F^2 \leq \lambda^2 |\mathcal{G}_0^{\text{union}}| + \lambda_0^2 |\mathcal{G}_{0\hat{\pi}}^{\text{union}}| \leq \tilde{\lambda}^2 \delta_B |\mathcal{G}_{0\hat{\pi}}^{\text{union}}|.$$

Let η_2 be a constant defined as $\eta_2 := \eta_0 \tilde{\lambda} / \sqrt{\frac{\log p}{n}(p/|\mathcal{G}_0^{\text{union}}| \vee 1)}$, then we can rewrite Condition 4.8 as

$$\sum_{i,j} \mathbf{1}\left\{ \left| [A_{0\hat{\pi}}^{(k)}]_{i,j} \right| \geq \tilde{\lambda}/\eta_2 \right\} \geq (1 - \eta_1) |\mathcal{G}_{0\hat{\pi}}^{(k)}|.$$

Moreover, for η_0 sufficiently small, η_2 is also guaranteed to satisfy $0 < \eta_2^2 c_t < 1 - \eta_1$. We could therefore apply Lemma A.7 and get that $|\hat{\mathcal{G}}| \geq \frac{1-\eta_1-\eta_2^2 c_t}{c_t} |\mathcal{G}_{0\hat{\pi}}^{\text{union}}|$, which completes the proof of (47) by choosing $c_g = \frac{1-\eta_1-\eta_2^2 c_t}{c_t}$ and $c_g' = c_g \cdot c_t$. Notice that in order to apply Lemma A.7, it is required that $\eta_2^2 c_t < 1 - \eta_1$, which is guaranteed by the assumption that η_0 is sufficiently small. Leveraging the first inequality in (47), we can replace $\lambda_0^2 |\mathcal{G}_{0\hat{\pi}}^{\text{union}}|$ in (45) by $\frac{\lambda_0^2 c_t}{1-\eta_1-\eta_2^2 c_t} |\hat{\mathcal{G}}|$ in order to obtain

$$\delta_B \sum_{k=1}^{K} w_k \|\hat{A}^{(k)} - A_{0\hat{\pi}}^{(k)}\|_F^2 + \left(\lambda^2 \delta_s - \frac{\lambda_0^2 c_t}{1 - \eta_1 - \eta_2^2 c_t} \right) |\hat{\mathcal{G}}| \leq \lambda^2 |\mathcal{G}_0^{\text{union}}|. \tag{48}$$

Notice that (48) coincides with the sought expression (46) upon substituting $\delta_s' = c_t/(1 - \eta_1 - c_t \eta_2^2)$. □

A.2.2. *Proof of Theorem 4.9*

It follows from Theorem A.2, Theorem A.4, and Lemma A.5 that there exist constants λ_1, λ_2 with $\lambda_1^2 \asymp \lambda_2^2 \asymp \frac{\log p}{n}$ as well as some ${\lambda_3^{(k)}}^2 \asymp \frac{\log p}{n_k}$ for all k such that with probability $1 - \exp(-c \log p)$ for some constant $c > 0$, the random event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ occurs.

We may then apply Lemma A.6 to show that there exist constants δ_B, δ_W such that with high probability, for any $\lambda > 0$ satisfying $\lambda^2 > \lambda_1^2/\delta_1 + \lambda_2^2/\delta_2$, it holds that

$$\begin{aligned} \delta_B \beta^2 \sum_{j=1}^{p} \sum_{k=1}^{K} w_k \|X^{(k)}(\hat{a}_j^{(k)} - a_{0j\hat{\pi}}^{(k)})\|_2^2 / n_k + \delta_W \sum_{k=1}^{K} w_k \|\hat{\Omega}^{(k)} - \Omega_{0\hat{\pi}}^{(k)}\|_F^2 + \lambda^2 \delta_s |\hat{\mathcal{G}}| \\ \leq \lambda^2 |\mathcal{G}_0^{\text{union}}| + \frac{\lambda_2^2 (p + |\mathcal{G}_{0\hat{\pi}}^{\text{union}}|)}{\delta_2} + \frac{\lambda_1^2 |\mathcal{G}_{0\hat{\pi}}^{\text{union}}|}{\delta_1}. \end{aligned} \tag{49}$$

Note that compared with Lemma A.6, we have replaced $\sum_{j=1}^{p} \left(\frac{\hat{\omega}_j^{(k)} - {\omega_0}_j^{(k)}}{\hat{\omega}_j^{(k)}} \right)^2$ by $\|\hat{\Omega}^{(k)} - \Omega_{0\hat{\pi}}^{(k)}\|_F^2$. This follows from combining the facts that $\sum_{j=1}^{p} (\hat{\omega}_j^{(k)} - {\omega_0}_j^{(k)})^2$

is equal to $\|\hat{\Omega}^{(k)} - \Omega^{(k)}_{0\hat{\pi}}\|_F^2$ and that $\frac{1}{\hat{\omega}_j^{(k)}}$ is bigger than $1/\beta^2$ on the random event $\mathcal{E}_3$. In addition, the constants β, σ_0, δ_1 and δ_2 in Lemma A.6 have been absorbed into the new constants δ_B and δ_W. We also replaced $\lambda^2 - \frac{\lambda_1^2}{\delta_1} - \frac{\lambda_2^2}{\delta_2}$ by $\lambda^2\delta_s$ for some $0 < \delta_s < 1$.

By applying Lemma A.5 [cf. (33)] we may bound the first summand on the left-hand side of (49) by $\delta_B \sum_{k=1}^K w_k \|\hat{A}^{(k)} - A^{(k)}_{0\hat{\pi}}\|_F^2$. Furthermore, replacing λ_1 and λ_2 by some λ_0 that scales as $\lambda_0^2 \asymp \frac{\log p}{n}(p/|\mathcal{G}_0^{\text{union}}| \vee 1)$, we obtain that

$$\delta_B \sum_{k=1}^K w_k \|\hat{A}^{(k)} - A^{(k)}_{0\hat{\pi}}\|_F^2 + \delta_W \sum_{k=1}^K w_k \|\hat{\Omega}^{(k)} - \Omega^{(k)}_{0\hat{\pi}}\|_F^2 + \lambda^2\delta_s|\hat{\mathcal{G}}| \leq \lambda^2|\mathcal{G}_0^{\text{union}}| + \lambda_0^2|\mathcal{G}_{0\hat{\pi}}^{\text{union}}|. \tag{50}$$

By applying (50), we have that for a constant η_0 small enough, (46) and (47) in Theorem A.8 hold by choosing λ such that $\lambda^2 \asymp \frac{\log p}{n}(p/|\mathcal{G}_0^{\text{union}}| \vee 1)$ and $\lambda^2\delta_s > \lambda_0^2\delta_s'$. Moreover, from (46) we further infer that

$$\delta_B \sum_{k=1}^K w_k \|\hat{A}^{(k)} - A^{(k)}_{0\hat{\pi}}\|_F^2 + \lambda^2\delta_s''|\hat{\mathcal{G}}| \leq \lambda^2|\mathcal{G}_0^{\text{union}}|, \tag{51}$$

where the constant δ_s'' is chosen such that $\lambda^2\delta_s'' = \lambda^2\delta_s - \lambda_0^2\delta_s'$ in (46). From (51) it can thus be inferred that $|\hat{\mathcal{G}}| \leq |\mathcal{G}_0^{\text{union}}|/\delta_s''$. Combining this with (47) and the fact that $|\mathcal{G}_{0\hat{\pi}}^{\text{union}}| \geq c_g'/c_g|\mathcal{G}_0^{\text{union}}|$, we recover the first part of (11) in the statement of the theorem, i.e., $|\hat{\mathcal{G}}| \asymp |\mathcal{G}_{0\hat{\pi}}^{\text{union}}|$. For the relation between $|\mathcal{G}_{0\hat{\pi}}^{\text{union}}|$ and $|\mathcal{G}_0^{\text{union}}|$, we use that $|\mathcal{G}_{0\hat{\pi}}^{\text{union}}| \leq |\hat{\mathcal{G}}|/c_g \leq |\mathcal{G}_0^{\text{union}}|/(\delta_s'' \cdot c_g)$ and $|\mathcal{G}_{0\hat{\pi}}^{\text{union}}| \geq c_g'/c_g|\mathcal{G}_0^{\text{union}}|$. Finally, to recover (10) we combine (50) with (11), which concludes the proof. □

A.3. *Proof of Theorem 4.11*

We first introduce a lemma that will be instrumental in proving Theorem 4.11 and that can be obtained directly from Lemmas 7.2 and 7.3 in [45].

Lemma A.9. *[45, Lemmas 7.2 and 7.3] Suppose for some $\delta_B, \delta_s, \lambda_0, \lambda > 0$ one has that $\delta_B\|\hat{A}^{(k)} - A^{(k)}_{0\hat{\pi}}\|_F^2 + \lambda^2\delta_s|\hat{\mathcal{G}}^{(k)}| \leq \lambda^2|\mathcal{G}_0^{(k)}| + \lambda_0^2|\mathcal{G}_{0\hat{\pi}}^{(k)}|$. Let $\tilde{\lambda}^2\delta_B \geq \lambda^2 + \lambda_0^2$ and assume that $\sum_{i,j} \mathbf{1}\left\{\left|[A^{(k)}_{0\hat{\pi}}]_{i,j}\right| \geq \tilde{\lambda}/\eta_2\right\} \geq (1-\eta_1)|\mathcal{G}_{0\hat{\pi}}^{(k)}|$. Then*

$$\delta_B\|\hat{A}^{(k)} - A^{(k)}_{0\hat{\pi}}\|_F^2 + \left(\lambda^2\delta_s - \frac{\lambda_0^2}{1-\eta_1-\eta_2^2}\right)|\hat{\mathcal{G}}^{(k)}| \leq \lambda^2|\mathcal{G}_0^{(k)}|$$

and

$$|\hat{\mathcal{G}}^{(k)}| \geq (1-\eta_1-\eta_2^2)|\mathcal{G}_{0\hat{\pi}}^{(k)}| \geq (1-\eta_1-\eta_2^2)|\mathcal{G}_0^{(k)}|.$$

In order to show Theorem 4.11, we begin the proof just like for Theorem 4.9 in Section A.2.2 until we get to expression (50). It then follows from Condition 4.7' and

$|\hat{\mathcal{G}}| \geq \sum_{k=1}^{K} w_k |\hat{\mathcal{G}}^{(k)}|$ that there exists some $\lambda_0'^2 \asymp C_{\max} \frac{\log p}{n}(p/|\mathcal{G}_0^{\text{union}}| \vee 1)$, where $C_{\max}$ is defined in Condition 4.8, such that for any $\lambda > 0$,

$$\delta_B \sum_{k=1}^{K} w_k \|\hat{A}^{(k)} - A_{0\hat{\pi}}^{(k)}\|_F^2 + \delta_W \sum_{k=1}^{K} w_k \|\hat{\Omega}^{(k)} - \Omega_{0\hat{\pi}}^{(k)}\|_F^2 + \lambda^2 \delta_s \sum_{k=1}^{K} w_k |\hat{\mathcal{G}}^{(k)}|$$
$$\leq \lambda^2 c_t(\pi_0) \sum_{k=1}^{K} w_k |\mathcal{G}_0^{(k)}| + \lambda_0'^2 \sum_{k=1}^{K} w_k |\mathcal{G}_{0\hat{\pi}}^{(k)}|.$$

Let $\lambda'^2 := \lambda^2 \cdot c_t(\pi_0)$ and $\delta_s' := \delta_s / c_t(\pi_0)$, it then follows that there must exist at least one k such that for any $\lambda' > 0$,

$$\delta_B \|\hat{A}^{(k)} - A_{0\hat{\pi}}^{(k)}\|_F^2 + \delta_W \|\hat{\Omega}^{(k)} - \Omega_{0\hat{\pi}}^{(k)}\|_F^2 + \lambda'^2 \delta_s' |\hat{\mathcal{G}}^{(k)}| \leq \lambda'^2 |\mathcal{G}_0^{(k)}| + \lambda_0'^2 |\mathcal{G}_{0\hat{\pi}}^{(k)}|$$

Since according to Condition 4.7', $c_t(\pi)$ scales as a constant for permutations consistent with $\mathcal{G}_0^{\text{union}}$, we have that δ_s' is still a constant and $\lambda' \asymp \lambda$. In this case, it follows from Lemma A.9 and Condition 4.8' that there exists some constant $0 < \delta_s < 1$ and $\delta_s' > 0$ such that by choosing λ' such that $\lambda' \asymp C_{\max} \frac{\log p}{n}(p/|\mathcal{G}_0^{\text{union}}| \vee 1)$ and $\lambda'^2 \delta_s > \lambda_0'^2 \delta_s'$, it holds that

$$\delta_B \|\hat{A}^{(k)} - A_{0\hat{\pi}}^{(k)}\|_F^2 + \left(\lambda'^2 \delta_s - \lambda_0'^2 \delta_s'\right) |\hat{\mathcal{G}}^{(k)}| \leq \lambda'^2 |\mathcal{G}_0^{(k)}|.$$

It also follows from Lemma A.9 that $|\hat{\mathcal{G}}^{(k)}| \geq c_g |\mathcal{G}_{0\hat{\pi}}^{(k)}| \geq c_g |\mathcal{G}_0^{(k)}|$ for some positive constant c_g. Mimicking the arguments employed in the proof of Theorem 4.9 from (51) until the end of the proof, one can show that expressions (13) and (14) in the statement of Theorem 4.11 hold true, which completes the proof. □

A.4. *Proof of Corollary 5.1*

The following lemma is instrumental in proving the corollary.

Lemma A.10. *[18, Lemma 4] Given fixed $\mathcal{G}$, the maximum likelihood estimator in (15) can be written as*

$$p + \sum_{j=1}^{p} \left(\min_{a \in \mathbb{R}^{|\text{Pa}_j(\mathcal{G})|}} \frac{n_{-j}}{n} \log \left(\sum_{k: j \notin I_k} \frac{n_k}{n_{-j}} \|\hat{X}_j^{(k)} - \hat{X}_{\text{Pa}_j(\mathcal{G})}^{(k)} \cdot a\|_2^2 / n_k \right) \right.$$
$$\left. + \sum_{k: j \in I_k} w_k \log \left(\|\hat{X}_j^{(k)}\|_2^2 / n_k \right) \right)$$

where n_{-j} is the total number of samples where node j is not intervened on, i.e., $n_{-j} = \sum_{k: j \notin I_k} n_k$.

Recall that in the interventional setting $\mathcal{G}_0^{\text{union}}$ is given by the true graph $\mathcal{G}_0$ of the non-intervened model, and that the K models $(A_0^{(k)}, \Omega_0^{(k)})$ to be inferred correspond to the interventional models $(A_0^{I_k}, \Omega_0^{I_k})$. Denoting by $(\hat{\pi}, \hat{A}, \hat{\Omega})$ the (non-intervened)

global optimum of (15), let $\hat{\omega}_j^{(k)}$ denote the empirical variance of the random variable $X_j^{(k)} - X^{(k)}\hat{a}_j$ if $j \in I_k$ and the empirical variance of $X_j^{(k)}$ otherwise. It follows from Lemma A.10 that the global optimum satisfies

$$p + \sum_{j=1}^{p} \left(\frac{n_{-j}}{n} \log \left(\sum_{k: j \notin I_k} \frac{n_k}{n_{-j}} \hat{\omega}_j^{(k)} \right) + \sum_{k: j \in I_k} w_k \log \hat{\omega}_j^{(k)} \right) + \lambda^2 |\hat{\mathcal{G}}|$$
$$\leq \sum_{j=1}^{p} \sum_{k=1}^{K} w_k \frac{\|\hat{\epsilon}_{j\hat{\pi}}^{(k)}\|_2^2 / n_k}{\omega_{0j\hat{\pi}}^{(k)}} + \sum_{j=1}^{p} \sum_{k=1}^{K} w_k \log \omega_{0j\hat{\pi}}^{(k)} + \lambda^2 |\mathcal{G}_0^{\text{union}}|.$$

Then, applying the inequality $\log(\sum_{k=1}^{K} w_k a_k) \geq \sum_{k=1}^{K} w_k \log a_k$ for any choices of $a_1, \ldots, a_K > 0$ and $w_1, \ldots, w_K > 0$ with $\sum_{k=1}^{K} w_k = 1$, we obtain

$$p + \sum_{j=1}^{p} \sum_{k=1}^{K} w_k \log \hat{\omega}_j^{(k)} + \lambda^2 |\hat{\mathcal{G}}| \leq \sum_{j=1}^{p} \sum_{k=1}^{K} w_k \frac{\|\hat{\epsilon}_{j\hat{\pi}}^{(k)}\|_2^2 / n_k}{\omega_{0j\hat{\pi}}^{(k)}}$$
$$+ \sum_{j=1}^{p} \sum_{k=1}^{K} w_k \log \omega_{0j\hat{\pi}}^{(k)} + \lambda^2 |\mathcal{G}_0^{\text{union}}|.$$

Hence Corollary 5.1 directly follows from the proof of Theorem 4.9. □

References

[1] P. A. Aguilera, A. Fernández, R. Fernández, R. Rumí, and A. Salmerón. Bayesian networks in environmental modelling. *Environmental Modelling & Software*, 26(12):1376–1388, 2011.

[2] S. A. Andersson, D. Madigan, and M. D. Perlman. A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541, 1997.

[3] M. N. Arbeitman, E. EM. Furlong, F. Imam, E. Johnson, B. H. Null, B. S. Baker, M. A. Krasnow, M. P. Scott, R. W. Davis, and K. P. White. Gene expression during the life cycle of Drosophila melanogaster. *Science*, 297(5590):2270–2275, 2002.

[4] T. T. Cai, H. Li, W. Liu, and J. Xie. Joint estimation of multiple high-dimensional precision matrices. *Statistica Sinica*, 26(2):445, 2016.

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

[6] D. M. Chickering. Learning Bayesian networks is NP-complete. In *Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*, 1995.

[7] D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

[8] P. Danaher, P. Wang, and D. 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):373–397, 2014.

[9] A. Dixit, O. Parnas, B. Li, J. Chen, C. P. Fulco, L. Jerby-Arnon, N. D. Marjanovic, D. Dionne, T. Burks, R. Raychowdhury, B. Adamson, T. M. Norman, E. S. Lander, J. S. Weissman, N. Friedman, and A. Regev. Perturb-seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866, 2016.

[10] F. Eberhardt, C. Glymour, and R. Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*, pages 178–184. AUAI Press, 2005.

[11] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[12] N. Friedman, M. Linial, I. Nachman, and D. Peter. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3-4):601–620, 2000.

[13] D. M. Gau, J. L. Lesnock, B. L. Hood, R. Bhargava, M. Sun, K. Darcy, S. Luthra, U. Chandran, T. P. Conrads, R. P. Edwards, J. L. Kelley, T. C. Krivak, and P. Roy. BRCA1 deficiency in ovarian cancer is associated with alteration in expression of several key regulators of cell motility–a proteomics study. *Cell Cycle*, 14(12):1884–1892, 2015.

[14] B. George. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

[15] C. Glymour, R. Scheines, P. Spirtes, and K. Kelly. *Discovering Causal Strucure*. Academic Press, 1987.

[16] J. Guo, E. Levina, G. Michailidis, and J. Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15, 2011.

[17] A. Hauser and P. Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13(Aug):2409–2464, 2012.

[18] A. Hauser and P. Bühlmann. Jointly interventional and observational data: Estimation of interventional Markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1):291–318, 2015.

[19] J. Jönsson, K. Bartuma, M. Dominguez-Valentin, K. Harbst, Z. Ketabi, S. Malander, M. Jönsson, A. Carneiro, A. Måsbäck, G. Jönsson, and M. Nilbert. Distinct gene expression profiles in ovarian cancer linked to Lynch syndrome. *Familial Cancer*, 13:537–545, 2014.

[20] M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8(Mar):613–636, 2007.

[21] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and T. Mao. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1):D109–D114, 2011.

[22] M. Kolar, L. Song, A. Ahmed, and E. P. Xing. Estimating time-varying networks. *The Annals of Applied Statistics*, 4(1):94–123, 2010.

[23] S. L. Lauritzen. *Graphical Models*, volume 17. Clarendon Press, 1996.

[24] P. Loh and P. Bühlmann. High-dimensional learning of linear causal networks

via inverse covariance estimation. *Journal of Machine Learning Research*, 15(1):3065–3105, 2014.

[25] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[26] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[27] N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[28] K. Mohan, P. London, M. Fazel, D. Witten, and S. Lee. Node-based learning of multiple Gaussian graphical models. *The Journal of Machine Learning Research*, 15(1):445–488, 2014.

[29] P. Nandy, A. Hauser, and M. H. Maathuis. High-dimensional consistency in score-based and hybrid structure learning. *The Annals of Statistics*, 46(6A):3151–3183, 2018.

[30] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 27(1):29–34, 1999.

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

[32] J. Pearl and T. S. Verma. Equivalence and synthesis of causal models. In *Proceedings of Sixth Conference on Uncertainty in Artijicial Intelligence*, pages 220–227, 1991.

[33] C. Peterson, F. C. Stingo, and M. Vannucci. Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association*, 110(509):159–174, 2015.

[34] G. Raskutti and C. Uhler. Learning directed acyclic graphs based on sparsest permutations. *Stat*, 7:e183, 2018.

[35] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

[36] J. M. Robins, M. A. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

[37] A. D. Santin, F. Zhan, S. Bellone, M. Palmieri, S. Cane, E. Bignotti, S. Anfossi, M. Gokden, D. Dunn, J. J. Roman, T. J. O'Brien, E. Tian, M. J. Cannon, J. Shaughnessy, and S. Pecorelli. Gene expression profiles in primary ovarian serous papillary tumors and normal ovarian epithelium: identification of candidate molecular markers for ovarian cancer diagnosis and therapy. *International Journal of Cancer*, 112(1):14–25, 2004.

[38] A. K. Shalek, R. Satija, J. Shuga, J. J. Trombetta, D. Gennert, D. Lu, P. Chen, R. S. Gertner, J. T. Gaublomme, N. Yosef, S. Schwartz, B. Fowler, S. Weaver, J. Wang, X. Wang, R. Ding, R. Raychowdhury, N. Friedman, N. Hacohen, H. Park, A. P. May, and A. Regev. Single cell RNA Seq reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505):363, 2014.

[39] L. Song, M. Kolar, and E. P. Xing. Keller: estimating time-varying interactions between genes. *Bioinformatics*, 25(12):i128–i136, 2009.

[40] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. MIT Press, 2000.

[41] E. A. Stronach, G. C. Sellar, C. Blenkiron, G. J. Rabiasz, K. J. Taylor, E. P. Miller, C. E. Massie, A. Al-Nafussi, J. F. Smyth, D. J. Porteous, and H. Gabra. Identification of clinically relevant genes on chromosome 11 in a functional model of ovarian cancer tumor suppression. *Cancer Research*, 63(24):8648–8655, 2003.

[42] R. W. Tothill, A. V. Tinker, J. George, R. Brown, S. B. Fox, S. Lade, D. S. Johnson, M. K. Trivett, D. Etemadmoghadam, B. Locandro, N. Traficante, S. Fereday, J. A. Hung, Y. Chiew, I. Haviv, Australian Ovarian Cancer Study Group, D. Gertig, A. deFazio, and D. D.L. Bowtell. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clinical Cancer Research*, 14(16):5198–5208, 2008.

[43] I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

[44] C. Uhler, G. Raskutti, P. Bühlmann, and B. Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, 41(2):436–463, 2013.

[45] S. van de Geer and P. Bühlmann. ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, 41(2):536–567, 04 2013.

[46] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[47] A. Y. Zaitsev. On the Gaussian approximation of convolutions under multidimensional analogues of S.N. Bernstein's inequality conditions. *Probability Theory and Related Fields*, 74(4):535–566, 1987.