Order-based Structure Learning without Score Equivalence

Hyunwoong Chang[1], James Cai[2] and Quan Zhou[1,*]

[1]Department of Statistics, Texas A&M University
[2]Department of Veterinary Integrative Bioscience, Texas A&M University

Abstract

We propose an empirical Bayes formulation of the structure learning problem, where the prior specification assumes that all node variables have the same error variance, an assumption known to ensure the identifiability of the underlying causal directed acyclic graph (DAG). To facilitate efficient posterior computation, we approximate the posterior probability of each ordering by that of a best DAG model, which naturally leads to an order-based Markov chain Monte Carlo (MCMC) algorithm. Strong selection consistency for our model in high-dimensional settings is proved under a condition that allows heterogeneous error variances, and the mixing behavior of our sampler is theoretically investigated. Further, we propose a new iterative top-down algorithm, which quickly yields an approximate solution to the structure learning problem and can be used to initialize the MCMC sampler. We demonstrate that our method outperforms other state-of-the-art algorithms under various simulation settings, and conclude the paper with a single-cell real-data study illustrating practical advantages of the proposed method.

Keywords: Directed acyclic graphs; Empirical Bayes methods; Strong selection consistency; Markov chain Monte Carlo methods; Non-decomposable scores.

1 Introduction

We consider Bayesian structure learning of a directed acyclic graph (DAG) model from observational data. Bayesian algorithms for structure learning are often classified as score-based in the literature, since they assign a posterior probability to each candidate DAG, the logarithm of which can be interpreted as a score [Drton and Maathuis, 2017]. A Markov equivalence class is a set of all DAGs that encode the same set of conditional independence relations among node variables. Without a priori knowledge, we cannot distinguish between two Markov equivalent DAGs using only observational data [Koller and Friedman, 2009]. If a Bayesian model yields the same score for DAGs in the same equivalence class, we say it is score equivalent, which is widely considered a desirable property [Andersson et al., 1997]. Most Bayesian structure learning methods used in practice are score equivalent [Geiger and Heckerman, 2002].

*Corresponding author: quan@stat.tamu.edu

Since the number of p-node DAGs grows super-exponentially with p, an exact evaluation of the posterior distribution is impossible unless p is extremely small, and Markov chain Monte Carlo (MCMC) methods are commonly employed to generate samples from the posterior distribution. As a classical example, structure MCMC, which was proposed in the seminal work of Madigan et al. [1995], is a random walk Metropolis-Hastings algorithm on the DAG space that uses single-edge addition, deletion, and reversal as proposal moves. However, it is known that this algorithm can often suffer from computational inefficiency due to the considerable time it spends sampling DAGs within the same equivalence class [Andersson et al., 1997, Chickering, 2002]. Even if the data is very informative on the conditional independence relations among all variables, we are only able to learn the equivalence class of the underlying true DAG model, which can easily be very large and takes the chain a long time to explore. In order to overcome slow mixing behavior caused by equivalence classes, many DAG MCMC samplers have been proposed, which typically introduce new DAG operations that can realize jumps between very different DAGs, enabling the chain to move more efficiently across equivalence classes [Grzegorczyk and Husmeier, 2008, Su and Borsuk, 2016]. Another strategy is to devise MCMC samplers on some other spaces that might be easier to explore than the DAG space. Indeed, one can directly search on the equivalence class space so that redundant moves between Markov equivalent DAGs are avoided [Castelletti et al., 2018, Zhou and Chang, 2021]. But this approach is not commonly used in the Bayesian literature, and one likely reason is that, unlike DAG MCMC samplers, the implementation of graph operations for equivalence classes can be highly complicated.

A more popular approach is to perform MCMC sampling on the order space [Friedman and Koller, 2003, Agrawal et al., 2018, Kuipers et al., 2022]. Due to the acyclicity constraint, every p-node DAG has at least one consistent ordering of the p nodes such that node i precedes node j whenever the edge $i \to j$ is in the DAG. Order-based MCMC methods are largely motivated by the following observation: the main computational challenge in structure learning lies in the uncertainty of order estimation, since once the ordering of variables is fixed, structure learning can be reduced to a collection of variable selection problems that are often considered to have a much smaller complexity. It is generally believed that the mixing of order MCMC is better than that of structure MCMC, because the search space is smaller and the posterior distribution on the order space tends to be smoother [Friedman and Koller, 2003]. However, the problem of traversing large equivalence classes still exists. To see this, assume again that all conditional independence relations can be learned from the data so that the posterior concentrates on one equivalence class. But any two DAGs in this equivalence class must have different orderings since at least one edge is flipped. This implies that the posterior distribution on the order space concentrates on a set at least as large as this equivalence class.

To mitigate the potential mixing problem caused by traversing large equivalence classes, we propose to impose identifiability conditions so that within each equivalence class, the posterior mass tends to concentrate on only one DAG. Consequently, the overall posterior distribution tends to have less and sharper modes. To this end, we follow the work of Peters and Bühlmann [2014] to consider Gaussian structural equation models with equal error variances. Intuitively, by assuming equal error variances, the data becomes informative on edge directions so that an MCMC sampler can quickly learn the best DAG in its equiva-

lence class. For example, consider two correlated variables $\mathsf{X}_1, \mathsf{X}_2$. The DAGs $\mathsf{X}_1 \to \mathsf{X}_2$ and $\mathsf{X}_2 \to \mathsf{X}_1$ are Markov equivalent, and in general, we cannot determine the causal direction if only observational data is available. But the equal variance assumption forces the posterior score to favor $\mathsf{X}_1 \to \mathsf{X}_2$ if X_2 has a larger marginal variance than X_1. Though a score equivalent Bayesian procedure allows us to make posterior inferences by averaging over Markov equivalent DAGs, this advantage is often merely theoretical due to its slow convergence, even when dealing with a moderately large number of node variables. Our simulation study and real data analysis will show that the use of equal variance assumption does provide practical advantages, and it improves the posterior inference accuracy unless there is a huge degree of heterogeneity among error variances.

There is a rapidly growing literature on the identifiability conditions for structure learning [Shimizu et al., 2006, Hoyer et al., 2008, Peters et al., 2011, Peters and Bühlmann, 2014, Strieder et al., 2021, Drton and Maathuis, 2017, Glymour et al., 2019]. In particular, two deterministic search algorithms have been proposed recently for structure learning with equal error variances [Ghoshal and Honorio, 2018, Chen et al., 2019], and they are shown to be advantageous in terms of computational cost and scale well to high-dimensional data. But to our knowledge, the corresponding Bayesian theory and methodology is largely underdeveloped. Aiming to fill this gap, we formulate an empirical Bayes model under the equal variance assumption and obtain a posterior score that distinguishes between Markov equivalent DAGs. We prove a strong selection consistency result for our model, which shows that the posterior probability of the true DAG tends to one in probability under mild high-dimensional conditions. In particular, while our prior distribution encodes the equal variance constraint, the consistency result holds under a weaker assumption known as the minimum-trace condition [Aragam et al., 2019]. Further, we extend the consistency result to cases where errors follow sub-Gaussian distributions, which include more interesting settings such as mixed discrete-Gaussian DAG models.

The posterior score derived from our model is non-decomposable (see Remark 2), which is expected since, under the equal variance assumption, the marginal likelihood of a DAG model should depend on how close the residual variances of the p nodes are to each other. This poses new computational challenges and again makes our method very different from the existing Bayesian literature, where decomposable scores are almost always used because the decomposability enables one to evaluate the posterior probability of a DAG by local calculations at each node [Chickering, 2002].

To numerically evaluate the posterior distribution of our empirical Bayes model, in the same spirit of the minimal I-MAP MCMC of Agrawal et al. [2018], we approximate the posterior probability of an ordering by that of the best consistent DAG and then build a sampling algorithm on the order space. We show that, under some conditions on the edge weights, the chain will never get stuck at a sub-optimal local mode for exponentially many iterations in expectation, which partially explains why this order MCMC scheme may perform well in practice. Further, we propose a generalized iterative version of the top-down algorithm of Chen et al. [2019]. This algorithm is deterministic and quickly finds a likely ordering of the variables, which can be used as a warm start for our order MCMC sampler. When estimating edge inclusion probabilities, we tune our estimators via a conditional expectation calculation so that we can reduce the estimation variance caused by picking one single best

DAG for each ordering. Lastly, though the non-decomposable score of our model cannot be evaluated locally, we are able to devise an implementation strategy that makes the posterior evaluation for our model as efficient as that with a decomposable score. The key idea is to store the search paths of the forward-backward stepwise selection at each node, which can be reused in finding the best DAG consistent with a given ordering.

2 An empirical Bayes model for order-based structure learning

2.1 Notation and terminology

We set up the notation and terminology to be used throughout the paper. Let $G = (V, E)$ denote a DAG, where V is a node set and $E \subset V \times V$ is a set of directed edges that form no cycle. Without loss of generality, for a p-node DAG, we assume $V = [p] = \{1, \ldots, p\}$. For ease of notation, we write $\{i \to j\} \in G$ to mean that $(i, j) \in E$, and use $G \cup \{i \to j\}$ (respectively $G \setminus \{i \to j\}$) to denote the DAG obtained by adding (respectively removing) the edge $i \to j$. We use $|G|$ to denote then number of edges in G. We denote by $\mathbb{S}^p$ the set of all bijections from $[p]$ to $[p]$. An element $\sigma \in \mathbb{S}^p$ is said to be a topological ordering for a DAG G if the following holds: for any indices $k < l$, the edge between the nodes $\sigma(k)$ and $\sigma(l)$ is directed as $\sigma(k) \to \sigma(l)$, if it exists in G. Let σ^{-1} denote the inverse function of σ, and for each node $j \in [p]$, let

$$P_j^\sigma = \{i \colon \sigma^{-1}(i) < \sigma^{-1}(j)\} \tag{1}$$

denote the set of potential parents of node j under the ordering σ, i.e., all nodes preceding j in σ. Let $\mathcal{G}_p$ be the collection of all p-node DAGs and $\mathcal{G}_p^\sigma$ be the collection of all p-node DAGs consistent with topological ordering σ; that is, $\mathcal{G}_p^\sigma = \{G \in \mathcal{G}_p \colon \{i \to j\} \in G \text{ implies } \sigma^{-1}(i) < \sigma^{-1}(j)\}$. Given a node j, we use $\mathrm{Pa}_j(G)$ and $\mathrm{Ch}_j(G)$ to denote the set of its parent nodes and that of its child nodes, respectively, in the DAG G. If the underlying DAG is clear from the context, we simply write Pa_j and Ch_j. Finally, given a matrix $A \in \mathbb{R}^{a \times b}$, $j \in [b]$, $J \subseteq [b]$ and $I \subseteq [a]$, A_j denotes the j-th column of A, A_J denotes the submatrix of A containing columns indexed by J, and $A_{I,j}$ denotes the subvector of A_j with entries $\{A_{ij} \colon i \in I\}$. We use $|J|$ to denote the cardinality of the set J.

2.2 Model specification

Let $\mathsf{X} = (\mathsf{X}_1, \ldots, \mathsf{X}_p)$ denote a p-dimensional random vector, and denote by X an $n \times p$ data matrix, each row of which is an independent copy of X. For each $\sigma \in \mathbb{S}^p$ and $G \in \mathcal{G}_p^\sigma$, consider the following structural equation model for the random vector X,

$$\mathsf{X}_j = B_{\mathrm{Pa}_j(G),j}^{\mathrm{T}} \mathsf{X}_{\mathrm{Pa}_j(G)} + \mathsf{e}_j, \quad \mathsf{e}_j \mid \omega \overset{\text{i.i.d}}{\sim} N(0, \omega) \text{ for } j = 1, \ldots, p, \tag{2}$$

where $\mathrm{Pa}_j(G) \subseteq P_j^\sigma$ for each j, and B is a $p \times p$ matrix. Entries of B that are not involved in (2) are set to zero. B can be seen as the weighted adjacency matrix of the DAG G such that $\{i \to j\} \in G$ if $|B_{ij}| > 0$.

We use the following empirical prior on the parameter (σ, G, B, ω), where π_0 denotes the prior density function:

$$B_{\mathrm{Pa}_j(G),j} \mid G, \omega \overset{\text{ind}}{\sim} N_{|\mathrm{Pa}_j(G)|}\left(\hat{B}_{\mathrm{Pa}_j(G),j}, \frac{\omega}{\gamma}(X^{\mathrm{T}}_{\mathrm{Pa}_j(G)} X_{\mathrm{Pa}_j(G)})^{-1}\right), \quad \forall\, j \in [p], \tag{3}$$

$$\pi_0(\omega \mid \sigma) \propto \omega^{-\frac{\kappa}{2}-1}, \tag{4}$$

$$\pi_0(G, \sigma) \propto (p^{c_0})^{-|G|}\, \mathbb{1}_{\{\hat{G}_\sigma\}}(G), \tag{5}$$

where $\hat{B}_{\mathrm{Pa}_j(G),j}$ is the least-squares estimator of $B_{\mathrm{Pa}_j(G),j}$, c_0, γ, κ are hyperparamters of the prior, and $\hat{G}_\sigma$ in (5) is the best estimate for G among $\mathcal{G}_p^\sigma$; we will detail how to obtain $\hat{G}_\sigma$ later. This prior is doubly empirical. First, given G and ω, we use an empirical prior on $B_{\mathrm{Pa}_j(G),j}$ for each j in (3), where the conditional prior mean depends on the data. Following Martin et al. [2017] and Lee et al. [2019], when computing the posterior distribution, we raise the data likelihood to the power of α, where $\alpha \in (0, 1)$ is a constant, so that we can reduce the influence of the data that is inflated by the usage of the empirical prior. Lee et al. [2019] suggests setting α close to 1 to make the α-likelihood behave similarly to the standard likelihood in finite sample scenarios. Observe that the covariance in (3) is identical to that of Zellner's g-prior, proportional to the inverse Fisher information matrix for $B_{\mathrm{Pa}_j(G),j}$ [Tadesse and Vannucci, 2021]. An alternative approach to specifying the prior is to use the fractional Bayes factor [Carvalho and Scott, 2009, Castelletti and Consonni, 2021]. This yields a fractional posterior with the value of α determined automatically, but the resulting posterior is more difficult to calculate than the proposed posterior. Second, according to (5), the conditional prior distribution of G given σ is again empirical: it assigns unit mass to some $\hat{G}_\sigma$ that can be seen as the solution to a DAG selection problem given ordering σ. This implies that the marginal prior distribution of G has support $\hat{\mathcal{G}} = \{\hat{G}_\sigma \colon \sigma \in \mathbb{S}^p\}$. For moderately large p, searching the entire space $\mathcal{G}_p$ is impossible, but the empirical prior (5) reduces the size of the search space to that of the order space $\mathbb{S}^p$. Unfortunately, $|\mathbb{S}^p| = p!$ is still super-exponential in p, making it challenging to devise an efficient MCMC sampler.

Remark 1. The use of the empirical prior (5) makes our approach very different from traditional Bayesian structure learning methods, where posterior inference is performed by averaging over all DAG models that satisfy certain sparsity constraints. The seminal order-based MCMC sampler of Friedman and Koller [2003] imposes a uniform conditional prior given σ on all DAGs satisfying degree constraints in $\mathcal{G}_p^\sigma$. But calculating the un-normalized marginal posterior probability of an ordering requires summation over all possible DAGs, which is infeasible unless p is small or the degree constraint is highly demanding. Further, the technique used in Friedman and Koller [2003, Eq. (8)] to expedite this calculation is not applicable in our case since our score is not decomposable; see Remark 2. Therefore, we prefer using the empirical prior (5) for its computational efficiency. A similar approach is taken in Agrawal et al. [2018], which uses empirical conditional independence tests to construct a minimal independence map for each ordering and restricts the search space to the set of minimal independence maps. Henceforth, we will always use DAG selection to refer to the problem of identifying the best DAG with given ordering.

Let π_n denote the posterior distribution given the observed data matrix X. By a standard

normal-inverse-gamma calculation that integrates out the parameters B and ω, we get

$$\pi_n(G, \sigma) \propto e^{\phi(G)} \mathbb{1}_{\{\hat{G}_\sigma\}}(G), \tag{6}$$

where $\phi(G)$ is called the score of G and is given by

$$\begin{aligned} \phi(G) &= -|G| c_0 \log p - \frac{|G|}{2} \log[(1+\alpha/\gamma)] - \frac{\alpha p n + \kappa}{2} \log \left(\sum_{j=1}^{p} \mathrm{RSS}_j(G) \right), \\ \text{where } \mathrm{RSS}_j(G) &= X_j^{\mathrm{T}} \Phi^{\perp}_{\mathrm{Pa}_j(G)} X_j, \quad \Phi^{\perp}_S = I - X_S (X_S^{\mathrm{T}} X_S)^{-1} X_S. \end{aligned} \tag{7}$$

We will also sometimes refer to $\phi(G)$ as the posterior score. For a detailed derivation of (6), see Section B.7 in the supplementary material. The marginal posterior probability of an ordering σ and that of a DAG G are

$$\pi_n(\sigma) \propto e^{\phi(\hat{G}_\sigma)}, \quad \pi_n(G) \propto e^{\phi(G)} \sum_{\sigma \in \mathbb{S}^p} \mathbb{1}_{\{\hat{G}_\sigma\}}(G). \tag{8}$$

For our model, $\pi_n(G)$ is not exactly proportional to the exponentiation of the score of G due to the factor $\sum_{\sigma \in \mathbb{S}^p} \mathbb{1}_{\{\hat{G}_\sigma\}}(G)$, and in our high-dimensional analysis we will show this term is negligible under mild assumptions.

In the rest of this work, we consider the following choice for $\hat{G}_\sigma$,

$$\hat{G}^{\mathrm{MAP}}_\sigma(d_{\mathrm{in}}) = \arg\max_{G \in \mathcal{G}^\sigma_p(d_{\mathrm{in}})} \phi(G), \quad \forall\, \sigma \in \mathbb{S}^p, \tag{9}$$

where $\mathcal{G}^\sigma_p(d_{\mathrm{in}}) = \{G \in \mathcal{G}^\sigma_p : |\mathrm{Pa}_j(G)| \le d_{\mathrm{in}} \text{ for all } j \in [p]\}$ is the collection of all p-node DAGs with maximum in-degree bounded by d_{in}. For our high-dimensional analysis, we will impose the condition $d_{\mathrm{in}} \log p = o(n)$, which is commonly used in the literature on high-dimensional DAG selection [Cao et al., 2019, Lee et al., 2019]. The superscript MAP indicates that $\hat{G}^{\mathrm{MAP}}_\sigma$ is the DAG with the largest posterior score among $\mathcal{G}^\sigma_p(d_{\mathrm{in}})$, i.e., the maximum a posteriori estimate.

Remark 2. In most existing methods for Bayesian structure learning, the posterior score of a DAG G takes a decomposable form in the sense that it can be written as the sum of p terms, where the i-th term only involves node i and its parent set and thus can be evaluated locally. But our posterior score given in (7) is not decomposable due to the equal variance assumption used in the prior: integrating out ω results in the logarithm of the sum of p residual sum of squares (RSS) terms in (7). This non-decomposable score is able to discriminate between Markov equivalent DAGs, and as we will prove shortly, given sufficiently large sample size, the posterior distribution of our model concentrates on only the unique true DAG.

2.3 Strong model selection consistency

We consider a high-dimensional setting where n tends to infinity and both $p = p(n)$ and $d_{\mathrm{in}} = d_{\mathrm{in}}(n)$ may grow with n. Strong model selection consistency means that the posterior probability of the true model converges to 1 in probability with respect to the true probability measure from which the data is generated. This is often regarded as one of the most important theoretical guarantees for a high-dimensional Bayesian model selection procedure. In the DAG literature, it was proven for DAG selection with known ordering [Cao et al., 2019,

Lee et al., 2019] and structure learning up to equivalence class [Zhou and Chang, 2021]. To the best of our knowledge, there is no strong selection consistency result on Bayesian structure learning under an identifiability condition.

Though the equal variance assumption was used in the prior specification, for our consistency analysis, we consider a more general setting. Assume the data is generated according to the structural equation model

$$\mathsf{X}_j = (B^*_{\mathrm{Pa}_j(G^*),j})^{\mathrm{T}} \mathsf{X}_{\mathrm{Pa}_j(G^*)} + \mathsf{e}_j, \quad \mathsf{e}_j \sim N(0, \omega_j^*) \text{ for } j = 1, \ldots, p, \tag{10}$$

where $G^*, B^*, \{\omega_j^*\}_{j=1}^p$ denote the true parameter values, and we assume $B^*_{ij} \neq 0$ if and only if $\{i \to j\} \in G$. Define $\Omega^* = \mathrm{diag}(\omega_1^*, \ldots, \omega_p^*)$. Let $[\sigma^*]$ denote the set of all orderings consistent with G^*, where σ^* is some element in $[\sigma^*]$ interpreted as the true ordering. Thus, $G^* \in \mathcal{G}_p^\sigma$ if and only if $\sigma \in [\sigma^*]$. Let $\mathbb{P}^*$ denote the probability measure corresponding to the structural equation model (10). Observe that the covariance matrix of the random vector X can be written as $\Sigma^* = \Sigma(B^*, \Omega^*)$, where

$$\Sigma(B, \Omega) = (I_p - B^{\mathrm{T}})^{-1} \Omega (I_p - B)^{-1}. \tag{11}$$

This is known as the modified Cholesky decomposition. This decomposition of Σ^* is not unique, as we explain in the following remark.

Remark 3. For each ordering $\sigma \in \mathbb{S}^p$, there exists a unique tuple $(B^*_\sigma, \Omega^*_\sigma)$ such that B^*_σ is the weighted adjacency matrix of a DAG in $\mathcal{G}_p^\sigma$, Ω^*_σ is a diagonal matrix with all diagonal entries being strictly positive, and $\Sigma^* = \Sigma(B^*_\sigma, \Omega^*_\sigma)$. Write $\Omega^*_\sigma = \mathrm{diag}(\omega_1^\sigma, \ldots, \omega_p^\sigma)$ and use G^*_σ to denote the DAG with edge set $E^*_\sigma = \{(i,j) : |(B^*_\sigma)_{ij}| > 0\}$ and define

$$d^* = \max_{\sigma \in \mathbb{S}^p} \max_{j \in [p]} |\mathrm{Pa}_j(G^*_\sigma)|. \tag{12}$$

To prove that the empirical Bayes model specified in Section 2 has strong model selection consistency in high-dimensional settings, we make the following two assumptions.

Assumption A (Minimum-trace condition). There exists a universal constant $\eta \in (0, \infty)$ such that $\min_{\sigma \notin [\sigma^*]} \mathrm{tr}(\Omega^*_\sigma)/\mathrm{tr}(\Omega^*) > 1 + \eta^{-1}$, where tr denotes the trace.

Assumption B (Consistency of DAG selection given true ordering). The estimator $\hat{G}_\sigma$ satisfies $\mathbb{P}^*(\cap_{\sigma \in [\sigma^*]} \{\hat{G}_\sigma = G^*\}) \geq 1 - \zeta(p)$ for some $\zeta(p) \to 0$.

The first assumption includes the equal variance assumption as a special case. To see this, suppose that $\Omega^* = \mathrm{diag}(\omega^*, \ldots, \omega^*)$ for some $\omega^* > 0$. Since the determinant of Σ^* satisfies $\det(\Sigma^*) = (\omega^*)^p = \prod_{j=1}^p \omega_j^\sigma$ for all $\sigma \in \mathbb{S}^p$, we have $p\omega^* \leq \sum_{j=1}^p \omega_j^\sigma$ by the inequality of arithmetic and geometric means. That is, the true ordering σ^* satisfies $\mathrm{tr}(\Omega^*_{\sigma^*}) = \min_\sigma \mathrm{tr}(\Omega^*_\sigma)$. Hence, there always exists some $\eta(n)$ such that $\min_{\sigma \notin [\sigma^*]} \mathrm{tr}(\Omega^*_\sigma)/\mathrm{tr}(\Omega^*) > 1 + \eta(n)^{-1}$. Assumption A just requires that $\eta(n)^{-1}$ can be bounded away from zero so that we can replace it with some universal constant η. Under the equal variance assumption, we can rewrite Assumption A as follows, which has been used in Van de Geer and Bühlmann [2013] and is known as the omega-min condition.

Assumption A' (Assumption A with equal variances). Suppose $\Omega^* = \mathrm{diag}(\omega^*, \ldots, \omega^*)$, where $\omega^* > 0$ is the error variance shared by all node variables. There exists a universal constant $\eta \in (0, \infty)$ such that $\min_{\sigma \notin [\sigma^*]} p^{-1} \sum_{j=1}^p (\omega_j^\sigma / \omega^*) > 1 + \eta^{-1}$.

Remark 4. Recall our score function given in (7) and that RSS_j/n is an estimate of the error variance ω_j^σ. So our method essentially aims to select the DAG that provides the tightest fit to the data. More precisely, the score (7) aims to learn the best DAG in $\mathcal{G}_p^\sigma$ where σ minimizes $\mathrm{tr}(\Omega_\sigma^*)$, the sum of error variances; such a DAG is called the minimum-trace DAG. Our strong consistency result, which only requires Assumption A instead of Assumption A', confirms that though the equal variance assumption was used to derive (7), our method has the theoretical guarantee under a more general setting. We refer readers to Aragam et al. [2019] for a general theory on structure learning using minimum-trace DAGs.

Remark 5. An interesting open question is, without the equal variance assumption, what choices of (B^*, Ω^*) can satisfy the minimum-trace condition so that the true model is identifiable. We conjecture that if for some $\sigma^* \in \mathbb{S}^p$, we have $\omega_{\sigma^*(1)}^{\sigma^*} \le \omega_{\sigma^*(2)}^{\sigma^*} \le \cdots \le \omega_{\sigma^*(p)}^{\sigma^*}$, then $\mathrm{tr}(\Omega_{\sigma^*}^*) = \min_\sigma \mathrm{tr}(\Omega_\sigma^*)$. This weakly increasing variance condition falls under the broader identifiability conditions presented in Park [2020], which extend beyond the equal variance assumption. We have conducted extensive numerical experiments, which suggest that the conjecture is likely to be true, but a proof for every $p \ge 2$ seems highly challenging. Simulation studies are presented in Section C.2 of the supplement.

The second assumption says that when we are given an ordering $\sigma \in [\sigma^*]$, the pre-specified DAG selection procedure is able to identify the true DAG with high probability. This is a very mild assumption since if the ordering is known, one can often apply an existing consistent algorithm for high-dimensional variable selection to select the parent set of node j for each $j \in [p]$ separately [Ben-David et al., 2011, Yu and Bien, 2017, Shojaie and Michailidis, 2010, Cao et al., 2019, Lee et al., 2019]. We do not need any assumption on the behavior of $\hat{G}_\sigma$ when $\sigma \notin [\sigma^*]$. Among many possible DAG selection methods, we use the estimator defined in (9) for the following reason. If some other DAG selection method is used, for any $\sigma \notin [\sigma^*]$, there is no guarantee that $\hat{G}_\sigma$ has a sufficiently large posterior score compared with other DAGs in $\mathcal{G}_p^\sigma$, and the resulting posterior distribution on the order space $\mathbb{S}^p$ could be very irregular and contain more sub-optimal local modes. However, no existing consistency result can be readily applied to the estimator (9) due to the non-decomposable posterior score it uses. We prove in the following proposition that it does have strong consistency for DAG selection, and it satisfies Assumption B with $\zeta(p) = 4p^{-1}$. All the three conditions assumed in Proposition 1 are commonly used in the literature: (C1) is known as the restricted eigenvalue condition, (C2) assumes prior parameters are properly chosen, and (C3) is often called the β-min condition [Lee et al., 2019]. Except universal constants, all parameters are allowed to depend on n.

Proposition 1. Suppose $\max_j |\mathrm{Pa}_j(G^*)| \le d_{\mathrm{in}}$, and the following conditions hold.

(C1) There exist $\underline{\nu}, \overline{\nu} > 0$ and a universal constant $\delta > 0$ such that

$$\frac{\underline{\nu}}{(1-\delta)^2} \le \lambda_{\min}(\Sigma^*) \le \lambda_{\max}(\Sigma^*) \le \frac{\overline{\nu}}{(1+\delta)^2},$$

where $\lambda_{\min}, \lambda_{\max}$ are the smallest and largest eigenvalues, respectively.

(C2) The sparsity parameter d_{in} satisfies $d_{\mathrm{in}} \log p = o(n)$, and prior parameters satisfy that $\kappa \le np, 0 \le \alpha/\gamma \le p^2 - 1$, $c_0 > \rho(\alpha+1) \max_{i \ne j}(\omega_j^*/\omega_i^*)$, and $\rho > 4d_{\mathrm{in}} + 6$.

(C3) For the true weighted adjacency matrix B^*,

$$C_{\min} = \min\{|(B^*)_{ij}|^2 : (B^*)_{ij} \neq 0\} \geq 16c_0 \frac{\overline{\nu}^2 \log p}{\alpha \underline{\nu}^2 n}.$$

Consider the posterior score given in (7) and the estimator defined in (9). For sufficiently large n, with probability at least $1 - 4p^{-1}$, all the following three events happen.

(i) For any $\sigma \in [\sigma^*]$, $G \in \mathcal{G}_p^\sigma(2d_{\text{in}})$, $j \in [p]$ such that $\text{Pa}_j(G^*) \subset \text{Pa}_j(G)$, there exists some $G' \in \mathcal{G}_p^\sigma$ such that $\phi(G') > \phi(G)$ and $G' = G \setminus \{i \to j\}$ for some $i \in [p]$.

(ii) For any $\sigma \in [\sigma^*]$, $G \in \mathcal{G}_p^\sigma(2d_{\text{in}})$, $j \in [p]$ such that $\text{Pa}_j(G^*) \not\subseteq \text{Pa}_j(G)$, there exists some $G' \in \mathcal{G}_p^\sigma$ such that $\phi(G') > \phi(G)$ and $G' = G \cup \{i \to j\}$ for some $i \in [p]$.

(iii) For any $\sigma \in [\sigma^*]$, $\hat{G}_\sigma^{\text{MAP}} = G^*$.

Proof. See Section B.2 in the supplementary material. □

Remark 6. For computational efficiency, to estimate $\hat{G}_\sigma^{\text{MAP}}$, one may use a forward-backward stepwise selection to find Pa_j for each j separately. This is outlined in Algorithm 4 in Section A.2 of the supplementary material. Since the posterior score is not decomposable, the stepwise selection at node j depends on the values of $\{\text{RSS}_i : i \neq j\}$. A simple solution is to estimate RSS_i by $X_i^{\mathrm{T}} X_i$ for each $i \neq j$. Then, parts (i) and (ii) of Proposition 1 imply that this procedure is consistent as long as for each j, $|\text{Pa}_j|$ is bounded by d_{in} at the end of the forward phase in Algorithm 4. As shown in An et al. [2008] and Zhou [2010], this condition on the output of forward selection can often be satisfied, with high probability, by choosing some $d_{\text{in}} = O(\max_j |\text{Pa}_j(G^*)|)$; i.e., d_{in} has the same order as the maximum in-degree of G^*. Actually, Proposition 1 implies that the following procedure is also consistent: starting from an arbitrary DAG G with maximum in-degree bounded by d_{in}, one performs stepwise selection at each node j by setting $\text{RSS}_i = \text{RSS}_i(G)$ for each $i \neq j$.

Remark 7. An alternative approach to performing forward-backward DAG selection with given ordering is to consider all the p nodes jointly; see Algorithm 5 in Section A.3 of the supplementary material. In the forward phase, we add one best edge consistent with the given ordering in each iteration, while in the backward phase, we remove one edge in each iteration. Proposition 1 implies that this algorithm is also consistent for $\sigma \in [\sigma^*]$, provided that the maximum in-degree of any DAG on the search path is bounded by d_{in}.

The main result of this section is given in the following theorem.

Theorem 1 (Strong selection consistency). Suppose Assumption A, B hold, and assume that $d^* \leq d_{\text{in}}$ and $d_{\text{in}} \log p = o(n)$. Then $\pi_n(G^*)$ converges in probability to 1 with respect to $\mathbb{P}^*$, where π_n is as given in (8).

Proof. See Section B.3 in the supplementary material. □

Remark 8. The proof can be further extended to cases where the errors e_j, $j = 1, \ldots, p$ in (10) follow a sub-Gaussian distribution. As any bounded random variable is sub-Gaussian, this relaxation covers scenarios where some variables are normally distributed and others are discrete and bounded [Lauritzen, 1992]. The proof is given in Section B.4 in the supplementary material. Some inequalities cannot be obtained as sharply as in the Gaussian case, because zero correlation does not imply independence in the sub-Gaussian case.

Consider the marginal posterior distribution on the order space $\mathbb{S}^p$. The following corollary shows that the posterior mass concentrates on the set of orderings consistent with G^*, and the posterior probabilities of all other orderings vanish.

Corollary 1. Under the setting of Theorem 1, $\pi_n([\sigma^*])$ converges in probability to 1 with respect to $\mathbb{P}^*$.

Proof. This follows from Theorem 1 and $\pi_n(G^*) = \sum_{\sigma\in[\sigma^*]} \pi_n(G^*, \sigma) = \sum_{\sigma\in[\sigma^*]} \pi_n(\sigma)$. □

3 Posterior sampling via order MCMC

3.1 Metropolis-Hastings algorithms on the order space

To generate posterior samples for our model, we use random walk Metropolis-Hastings algorithms on the order space $\mathbb{S}^p$. For each $\sigma \in \mathbb{S}^p$, let $\mathbf{K}(\sigma, \cdot)$ denote the proposal distribution at state σ. We consider three types of random walk proposals: adjacent transposition, which is a standard choice for order-based MCMC methods [Friedman and Koller, 2003, Agrawal et al., 2018], random transpositions and random-to-random shuffles, which are more commonly seen in the literature on random walks on symmetric groups [Levin and Peres, 2017, Bernstein and Nestoridi, 2019]. All three types of proposals correspond to defining $\mathbf{K}(\sigma, \cdot)$ by

$$\mathbf{K}(\sigma, A) = \frac{|\mathcal{N}(\sigma) \cap A|}{|\mathcal{N}(\sigma)|}, \quad \forall\, A \subseteq \mathbb{S}^p, \tag{13}$$

for some set $\mathcal{N}(\sigma) \subset \mathbb{S}^p$. We refer to $\mathcal{N}(\sigma)$ as the neighborhood of σ, and now we formally define this set for each type of proposal. Let $(\cdot)_{\mathrm{c}}$ denote an ordering in the cycle notation; for example, $\mu = (a, b, c)_{\mathrm{c}}$ is the ordering given by $\mu(a) = b, \mu(b) = c, \mu(c) = a$ and $\mu(k) = k$ for every $k \notin \{a, b, c\}$. Let $\circ$ denote the composition of two orderings; that is, $\tau = \sigma \circ \mu$ is defined by $\tau(i) = \sigma(\mu(i))$. Then, we can use $\sigma \circ (i, j)_{\mathrm{c}}$ to denote the ordering obtained by interchanging the i-th and the j-th elements of σ while keeping the others unchanged. Let $\sigma \circ \xi(i, j)$ denote the ordering obtained by inserting the i-th element of σ to the j-th position, where $\xi(i, j)$ is defined by $\xi(i, j) = (i, i + 1, \ldots, j)_{\mathrm{c}}$ if $i < j$, and $\xi(i, j) = (i, i - 1, \ldots, j)_{\mathrm{c}}$ if $i > j$. Define the adjacent transposition neighborhood by

$$\mathcal{N}_{\mathrm{adj}}(\sigma) \ = \{\sigma' \in \mathbb{S}^p \mid \sigma' = \sigma \circ (i, i + 1)_{\mathrm{c}},\ i \in [p - 1]\};$$

that is, $\mathcal{N}_{\mathrm{adj}}(\sigma)$ is the set of all orderings that can be obtained from σ by one adjacent transposition. Similarly, we denote the neighborhood corresponding to random transpositions by $\mathcal{N}_{\mathrm{rtp}}$ and that corresponding to random-to-random shuffles by $\mathcal{N}_{\mathrm{rrs}}$, which are defined by

$$\begin{aligned}
\mathcal{N}_{\mathrm{rtp}}(\sigma) \ &= \{\sigma' \in \mathbb{S}^p \mid \sigma' = \sigma \circ (i, j)_{\mathrm{c}},\ i < j, \text{ and } i, j \in [p]\}, \\
\mathcal{N}_{\mathrm{rrs}}(\sigma) \ &= \{\sigma' \in \mathbb{S}^p \mid \sigma' = \sigma \circ \xi(i, j),\ i \neq j, \text{ and } i, j \in [p]\}.
\end{aligned}$$

We provide an illustration of the three proposals in the supplementary material A.4. Observe that all the three neighborhood relations defined above are symmetric: if $\sigma' \in \mathcal{N}(\sigma)$, then $\sigma \in \mathcal{N}(\sigma')$. Therefore, by the Metropolis rule, the transition matrix of the algorithm can be

calculated by

$$\mathbf{P}(\sigma, \sigma') = \begin{cases} \mathbf{K}(\sigma, \sigma') \min\left\{1, \frac{\pi_n(\sigma')\mathbf{K}(\sigma', \sigma)}{\pi_n(\sigma)\mathbf{K}(\sigma, \sigma')}\right\}, & \text{if } \sigma' \neq \sigma, \\ 1 - \sum_{\tau \neq \sigma} \mathbf{P}(\sigma, \tau), & \text{if } \sigma' = \sigma, \end{cases} \tag{14}$$

where $\pi_n(\sigma)$ is the marginal posterior probability and also the stationary probability of σ. The Hastings ratio $\mathbf{K}(\sigma', \sigma)/\mathbf{K}(\sigma, \sigma') = 1$ for all the three neighborhood relations we consider. As explained in Section 2, once we select an ordering $\sigma \in \mathbb{S}^p$, we can find the associated $\hat{G}_\sigma$ by a pre-specified DAG selection method. Further, given a stationary Markov chain $(\sigma_t)_{t \geq 1}$ with transition matrix $\mathbf{P}$, $\{\hat{G}_{\sigma_t}\}_{t \geq 1}$ can be seen as correlated samples drawn from the marginal posterior distribution on the DAG space given in (8), which is just the pushforward of the marginal posterior distribution on $\mathbb{S}^p$ under the mapping $\sigma \mapsto \hat{G}_\sigma$.

The choice of the neighborhood $\mathcal{N}(\cdot)$ may affect the mixing of the chain significantly. In order to achieve efficient local exploration, the neighborhood size needs to be small. All the three types of proposals considered are desirable in this regard, since the corresponding neighborhood sizes grow at most quadratically in p: $|\mathcal{N}_{\text{adj}}(\sigma)| = p - 1$, and $|\mathcal{N}_{\text{rtp}}(\sigma)| = |\mathcal{N}_{\text{rrs}}(\sigma)| = p(p-1)/2$. However, if the neighborhood size is too small, the chain might get stuck at sub-optimal local modes, where a local mode refers to a state with posterior probability larger than that of any neighboring state. We will present a simulation study in Section 4.1 which confirms that all three proposals yield good mixing of the sampler for moderately large p.

In general, theoretical analysis of the mixing behavior of order-based MCMC methods is very difficult. Existing results on the mixing of MCMC for high-dimensional model selection problems suggest that if the posterior distribution is unimodal and tails decay sufficiently fast, an MCMC sampler is expected to mix rapidly [Yang et al., 2016, Zhou and Chang, 2021, Chang et al., 2022]; this intuition is highly similar to the rapid mixing of the algorithms with log-concave targets on continuous spaces [Mangoubi and Smith, 2017, Dwivedi et al., 2018]. However, to rigorously prove a rapid mixing result for our problem seems very difficult. One possible strategy is to assume a permutation β-min condition [Aragam et al., 2019], but such a permutation β-min condition is very restrictive since it requires all nonzero edge weights to be sufficiently large no matter what topological ordering we assume; in our context, this condition means that $\hat{G}_\sigma$ is equal to G^*_σ for any $\sigma \in \mathbb{S}^p$. Here we choose to consider a contrasting setting where all the edge weights of the true DAG G^* are not too large. This is probably more realistic and complements the existing theory, though still being moderately restrictive; see Remark 10 below. We are able to prove that the acceptance probability cannot be extremely small for any state proposed from $\mathcal{N}_{\text{adj}}(\cdot)$; see Remark 9. That is, by using adjacent transpositions, the chain is able to escape from any sub-optimal local mode, if there is any, in a relatively short amount of time. Observe that for any $\sigma \in \mathbb{S}^p$, $\mathcal{N}_{\text{adj}}(\sigma)$ is a proper subset of both $\mathcal{N}_{\text{rtp}}(\sigma)$ and $\mathcal{N}_{\text{rrs}}(\sigma)$. Hence, our result partly explains why all the three proposals appear to work well.

Proposition 2. Assume (C1) in Proposition 1 and the following conditions hold.

(C1') The true covariance matrix $\Omega^* = \text{diag}(\omega^*, \ldots, \omega^*)$ for some universal constant $\omega^* > 0$,

and the edge weights of G^* satisfy

$$\max_{i,j\in[p]} |B^*_{ij}|^2 = O\left(\frac{\overline{\nu}^2 \log p}{\underline{\nu}^2 n}\right).$$

(C2') The parameter d_{in} satisfies $d^* \leq d_{\text{in}}$ and

$$d_{\text{in}}^2 \frac{\overline{\nu}^2 \log p}{\underline{\nu}^2 n} \to 0 \text{ as } n \to \infty.$$

Let $\mathcal{N}_{\text{rev}}(G)$ denote the set of all DAGs that can be obtained by applying one edge reversal to G, and $c > 0$ be an arbitrary universal constant. Then, for sufficiently large n,

$$\max_{\sigma\in\mathbb{S}^p} \max_{G_1\in\mathcal{G}_p^\sigma(d_{\text{in}})} \max_{G_2\in\mathcal{N}_{\text{rev}}(G_1)} \frac{\exp(\phi(G_1))}{\exp(\phi(G_2))} \leq p^{c\overline{\nu}^2/\underline{\nu}^3},$$

with probability at least $1 - 6p^{-1}$.

Proof. See Section B.5 in the supplementary material. □

Remark 9. To see the implication of this result on the mixing of our order MCMC, consider $\sigma = (1, 2, \ldots, p)$, and let $\tau = \sigma \circ (i, i+1)_{\text{c}}$ for some i. Recall that we use $\hat{G}_\sigma = \hat{G}^{\text{MAP}}_\sigma$ where $\hat{G}^{\text{MAP}}_\sigma$ is defined in (9). Hence, $\pi_n(\sigma)/\pi_n(\tau) \leq \exp(\phi(\hat{G}_\sigma))/\exp(\phi(G'))$ where G' is the DAG that results from reversing the edge $i \to (i+1)$ of $\hat{G}_\sigma$; if the edge does not exist, then $G' = \hat{G}_\sigma$. Assuming $\overline{\nu}, \underline{\nu}$ are bounded, Proposition 2 implies that with high probability $\pi_n(\sigma)/\pi_n(\tau)$ is bounded from above by p^c where $c > 0$ is arbitrary, as long as $G' \in \mathcal{G}_p^\tau(d_{\text{in}})$. For the schemes we propose on $\mathbb{S}^p$, this further implies that an adjacent transposition proposal has acceptance probability greater than p^{-c}, and thus the chain cannot get trapped at a local mode for exponentially many iterations in expectation.

Remark 10. The purpose of Proposition 2 is to theoretically analyze the posterior landscape when we probably do not have posterior concentration at the true model and Proposition 1 no longer holds. In particular, Proposition 2 does not require any assumption on the hyperparameters of our model, so the nonzero entries in B^* may or may not be detected, depending on the choice of c_0. Condition (C1') essentially requires that no signal size has a strictly larger order than the detection threshold given in condition (C3) of Proposition 1. This is restrictive but arguably represents a scenario of more practical interest than Proposition 1, since in reality signals of small or moderate sizes are common. It is possible to construct a scenario where the assumptions of Propositions 1 and 2 both hold. For example, assume $d^* = O(1)$, which is referred to as the ultra-high sparsity regime in the literature [Van de Geer and Bühlmann, 2013]. Then we can set $d_{\text{in}} = O(1)$, which implies that we can choose $c_0 = O(1)$ to satisfy condition (C2) of Proposition 1. Assuming $\overline{\nu}, \underline{\nu}$ are bounded for convenience, in order to satisfy condition (C3) of Proposition 1 and condition (C1') of Proposition 2, we just need to require that the order of any nonzero entry B^*_{ij} is exactly given by $n^{-1} \log p$.

3.2 Iterative top-down initialization

Standard theory yields that the Markov chain defined in (14) converges to the marginal posterior distribution on $\mathbb{S}^p$ in total variation distance regardless of the initial state. However, the actual mixing rate of the chain we observe depends on the initial state [Sinclair, 1992, Proposition 1], and in general, it is desirable to start the chain at a state with reasonably high posterior probability. Since the size of $\mathbb{S}^p$ grows super-exponentially in p, choosing a warm start for our sampler can significantly improve the performance of posterior estimation with MCMC samples. We propose an initialization method for our order MCMC sampler, called iterative top-down, which aims to quickly find the topological ordering of the true data-generating DAG G^*.

Our method is based on the top-down method proposed by Chen et al. [2019], which we now briefly explain. We say a node in a DAG is a source if the node has no parents. If the data is generated according to (2), due to the equal variance assumption, a source node always has the smallest marginal variance, and any node with at least one parent has a strictly larger marginal variance. The top-down method first identifies a source node of G^*, which always exists, sets it to $\hat{\sigma}(1)$ and then removes it from G^*. The resulting subgraph is also a DAG, and thus we can set $\hat{\sigma}(2)$ to a source node of this subDAG; how to identify the source node is explained in the next paragraph. Repeating this procedure p times, we obtain $\hat{\sigma}$, the top-down estimator for the ordering.

Suppose that in the first k iterations of the top-down method we have identified $\sigma(j) = j$ for $j = 1, \ldots, k$. Then in the $(k+1)$-th iteration, we need to estimate the variance of each remaining node that cannot be explained by the first k nodes, and pick the node with the smallest unexplained variance, which we infer as a source node of the subDAG of the remaining $p - k$ nodes. Chen et al. [2019] estimated the unexplained variance of the node j (assuming $j > k$) by $\min_{S \subseteq [k], |S| = d_{\text{in}}} X_j^{\mathrm{T}} \Phi_S^{\perp} X_j$, but they noted that a variable selection procedure may be applied as well. Since our purpose is to find a warm start for our order MCMC sampler, we estimate the unexplained variance of a node by performing a variable selection procedure that aims to maximize the score (7). One caveat is that since our score is non-decomposable, when inferring the parent set of node j, we need to know the residual

Algorithm 1: Score-based top-down algorithm

Input: A positive vector $\mathrm{RSS} = (\mathrm{RSS}_1, \ldots, \mathrm{RSS}_p)$ (for all displayed algorithms, we assume the data X and parameters $(c_0, \gamma, \alpha, \kappa, d_{\text{in}})$ are given).

$\hat{\sigma} \leftarrow \arg\min_{j \in [p]} \mathrm{RSS}_j$

while $|\hat{\sigma}| < p$ **do**

  for $j \in [p] \backslash \hat{\sigma}$ **do**

    $S \leftarrow \arg\max_{S_j \subset \hat{\sigma} :\, |S_j| \le d_{\text{in}}} \phi_j(S_j, \sum_{i \neq j} \mathrm{RSS}_i)$

    // $\phi_j(S, R) = -|S| \log \left\{ p^{c_0} \sqrt{(1 + \alpha/\gamma)} \right\} - \frac{\alpha p n + \kappa}{2} \log \left(R + X_j^{\mathrm{T}} \Phi_S^{\perp} X_j \right)$

    $\mathrm{RSS}_j \leftarrow X_j^{\mathrm{T}} \Phi_S^{\perp} X_j$

  $j_0 \leftarrow \arg\min_{j \in [p] \backslash \hat{\sigma}} \mathrm{RSS}_j$

  $\hat{\sigma} \leftarrow (\hat{\sigma}, j_0)$

Output: An ordering $\hat{\sigma}$, a vector of estimated residual sums of squares RSS.

Algorithm 2: Iterative top-down algorithm

$(\hat{\sigma}^{\mathrm{ITD}}, \mathrm{RSS}) \leftarrow \mathrm{STD}(X_1^{\mathrm{T}} X_1, \ldots, X_p^{\mathrm{T}} X_p)$ `// STD refers to Algorithm 1`

while 1 **do**

  $(\tilde{\sigma}, \mathrm{RSS}') \leftarrow \mathrm{STD}(\mathrm{RSS})$

  if $\hat{\sigma}^{\mathrm{ITD}} \neq \tilde{\sigma}$ **then**

    $\mathrm{RSS} \leftarrow \mathrm{RSS}'$

    $\hat{\sigma}^{\mathrm{ITD}} \leftarrow \tilde{\sigma}$

  else

    return $\hat{\sigma}^{\mathrm{ITD}}$

Output: An ordering $\hat{\sigma}^{\mathrm{ITD}}$

sums of squares of all the other $p-1$ nodes. This motivates us to propose the iterative top-down method, detailed in Algorithm 2, which iteratively applies the top-down procedure and updates all the p residual sums of squares. We prove below that under a condition similar to that of Chen et al. [2019, Theorem 2], the iterative top-down algorithm identifies an ordering consistent with G^* with high probability. In our simulation studies, we observe that the algorithm usually converges within 5 iterations.

Theorem 2. Suppose the conditions in Proposition 1 hold, and let $\epsilon \in (0,1)$. If

$$n > \{\overline{\nu}(d_{\mathrm{in}}+1)(\underline{\nu} + 3\omega^*(1+1/C_{\mathrm{min}}))/\underline{\nu}^2\}^2 3200(\log(p^2-p) - \log(\epsilon/4)),$$

then for sufficiently large n, Algorithm 2 returns an ordering in $[\sigma^*]$ with probability at least $1-\epsilon$.

Proof. See Section B.6 in the supplementary material. □

3.3 Reducing variance of edge estimation

One potential limitation of our order MCMC sampler is that it does not take into account the uncertainty in DAG selection with given ordering. So we propose to estimate edge posterior inclusion probabilities using a conditioning scheme. Let $\sigma^{(t)}$ denote the t-th sample from our order MCMC sampler, and $\Gamma^{(t)}$ denote the adjacency matrix of the DAG $G^{(t)} = \hat{G}_{\sigma^{(t)}}$ such that $\Gamma_{ij}^{(t)} = 1$ if $\{i \to j\} \in G^{(t)}$ and $\Gamma_{ij}^{(t)} = 0$ otherwise. The posterior inclusion probability of edge $i \to j$ can be estimated by $T^{-1}\sum_{t=1}^{T} \Gamma_{ij}^{(t)}$ where T denotes the number of MCMC samples. To improve this estimator, for each pair $(\sigma^{(t)}, G^{(t)})$, we calculate $\hat{\Gamma}^{(t)} = \hat{\Gamma}(\sigma^{(t)}, G^{(t)})$, where the function $\hat{\Gamma}$ is given by

$$\hat{\Gamma}_{ij}(\sigma, G) = \frac{e^{\phi(G \cup \{i \to j\})}}{e^{\phi(G \cup \{i \to j\})} + e^{\phi(G \setminus \{i \to j\})}} \mathbb{1}_{P_j^\sigma}(i), \quad \forall\, i, j \in [p]. \tag{15}$$

We can now estimate the posterior inclusion probability of edge $i \to j$ by $\hat{\Gamma}_{ij}^{\mathrm{RB}} = T^{-1}\sum_{t=1}^{T} \hat{\Gamma}_{ij}^{(t)}$. The superscript RB indicates that, in a general sense, this can be seen as a Rao-Blackwellized-type estimator [Robert and Roberts, 2021]. In our numerical experiments, we find this scheme helps reduce the variance of edge posterior inclusion probability estimates.

4 Simulation studies

4.1 Mixing behavior

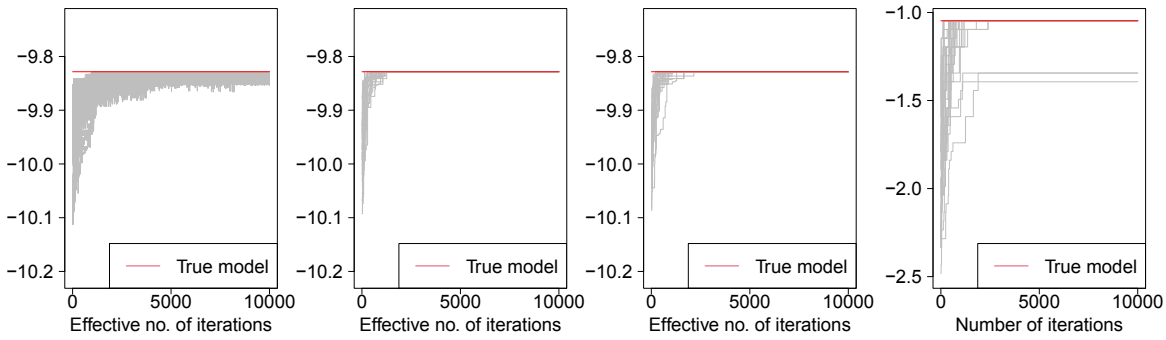
We first present a numerical example which illustrates how the choice of neighborhood and score equivalence property affect the mixing behavior of order MCMC samplers. We generate a 20-node random DAG G^* where any two distinct nodes are connected by an edge with probability 0.1, and sample the edge weight B^*_{ij} for each $i \to j$ in G^* uniformly from $[-1, -0.5] \cup [0.5, 1]$. Then, we simulate the data matrix X using the structural equation model in (2) with $n = 1,000$ and error variance $\omega^* = 1$.

We implement the order MCMC sampler described in Section 3 with $\mathcal{N} = \mathcal{N}_{\text{adj}}, \mathcal{N}_{\text{rtp}}$ or $\mathcal{N}_{\text{rrs}}$. To impartially compare the three types of proposal, we need to take into account the computational complexity of sampling from each type of neighborhood. Consider a proposal move from σ to $\sigma' = \sigma \circ (i, j)_{\text{c}}$ for some $i < j$. In Section A.3 of the supplementary material, we present a stepwise procedure for selecting the parent set of a given node in Algorithm 4, and describe how to efficiently obtain $\hat{G}_{\sigma'}$ from $\hat{G}_\sigma$ by applying Algorithm 4 at nodes $\sigma(i), \sigma(i+1), \ldots, \sigma(j)$. Hence, an adjacent transposition always requires performing Algorithm 4 at two nodes, while for a random transposition, which randomly samples σ' from $\mathcal{N}_{\text{rtp}}(\sigma)$ with equal probability, on average we need to perform Algorithm 4 at $(p+4)/3 \approx p/3$ nodes, and the same holds true for a random-to-random shuffle. So, when we run the sampler defined in (14) for T iterations, we say the effective number of iterations is $2T$ if $\mathcal{N} = \mathcal{N}_{\text{adj}}$, and $pT/3$ if $\mathcal{N} = \mathcal{N}_{\text{rtp}}$ or $\mathcal{N} = \mathcal{N}_{\text{rrs}}$. We let the effective number of iterations be $10,000$ for all three samplers in our simulation; that is, we run our sampler with $\mathcal{N} = \mathcal{N}_{\text{adj}}$ for $5,000$ iterations, and the samplers with $\mathcal{N} = \mathcal{N}_{\text{rtp}}$ and $\mathcal{N} = \mathcal{N}_{\text{rrs}}$ for $1,500$ iterations. We plot the trajectories for 30 runs with random initialization in the panels (a), (b), (c) of Fig. 1, from which we see that all three proposals work well. We have also tried $n = 100$ and observed good mixing performance, probably because with a smaller sample size the posterior distribution tends to be flatter [Agrawal et al., 2018]; we display the result in Section C.1 of the supplementary material. Given that adjacent transposition appears to yield the best mixing, it will be used for all the remaining numerical studies.

To compare our method with a score equivalent procedure, we consider the following posterior score, which is decomposable and yields the same value for Markov equivalent DAGs,

$$\phi_{\text{eq}}(G) = -|G| c_0 \log p - \frac{|G|}{2} \log[(1 + \alpha/\gamma)] - \frac{\alpha n + \kappa}{2} \sum_{j=1}^{p} \log\left(\text{RSS}_j(G)\right). \tag{16}$$

This score can be derived by a slight modification of our model: instead of assuming equal error variances, use an error variance parameter ω_j for each e_j in (2) and put an inverse-gamma prior on ω_j [Zhou and Chang, 2021]. To sample from the corresponding posterior distribution, we use the minimal I-MAP MCMC sampler of Agrawal et al. [2018], which is also a Metropolis-Hastings algorithm defined on the order space and proposes moves from the adjacent transposition neighborhood $\mathcal{N}_{\text{adj}}(\cdot)$; compared with our method, the main difference is that the minimal I-MAP MCMC uses conditional independence tests to find $\hat{G}_\sigma$. We run the minimal I-MAP MCMC for 10,000 iterations, and plot 30 trajectories with random



(a) (b) (c) (d)

Figure 1: Log posterior probability $\times 10^{-4}$ versus the effective number of iterations in 30 MCMC runs with random initialization. The red line gives the log posterior probability of the true ordering σ^*. Panel (d) is for the minimal I-MAP MCMC with decomposable score. Panels (a), (b), (c) correspond to our method with three types of proposals: (a) adjacent transposition, (b) random transposition, (c) random-to-random shuffle. We have checked that, for our method, all $30 \times 3 = 90$ runs have successfully reached the red line.

initialization in Fig. 1(d). Comparing it with Fig. 1(a), we see that our sampler with non-decomposable score mixes better in the sense that all 30 trajectories are able to visit some $\sigma \in [\sigma^*]$, while the minimal I-MAP MCMC may get stuck at local modes depending on the initialization. As we have explained in Section 1, score equivalence is likely to make the posterior distribution on the order space (or the DAG space) difficult to explore due to the existence of large equivalence classes. This simple numerical study verifies that the use of identifiability conditions does simplify the posterior distribution so that MCMC samplers tend to mix faster. In Section C.1 of the supplementary material, we show that the same observation can still be made if we simulate X using unequal error variances.

4.2 Performance evaluation

We conduct simulation studies to empirically evaluate the performance of the proposed order MCMC sampler. We still use G^* to denote the true p-node DAG that governs the data generating process described in (2) and let Γ^* be its adjacency matrix. Let $\hat{G}$ and $\hat{\Gamma}$ denote the corresponding estimators, and for our method, we always use $\hat{\Gamma} = \hat{\Gamma}^{\mathrm{RB}}$ where $\hat{\Gamma}^{\mathrm{RB}}$ is defined in Section 3.3. Entries of $\hat{\Gamma}$ are edge posterior inclusion probability estimates and thus take value in $[0, 1]$, while $\Gamma^* \in \{0, 1\}^{p \times p}$. We use four performance metrics to evaluate an estimator. The structural Hamming distance (HD) between G^* and $\hat{G}$ is the number of different edges between G^* and $\hat{G}$, which equals $\sum_{i,j} |\Gamma^*_{ij} - \hat{\Gamma}_{ij}|$. False negative rate (FNR) and false discovery rate (FDR) are defined as $(\sum_{i,j} \Gamma^*_{ij}(1-\hat{\Gamma}_{ij}))/|G^*| \times 100\%$ and $(\sum_{i,j}(1-\Gamma^*_{ij})\hat{\Gamma}_{ij})/|\hat{G}| \times 100\%$, respectively. The fourth metric, percentage of flipped edges, is calculated as $(\sum_{i,j} \Gamma^*_{ji}\hat{\Gamma}_{ij})/|G^*| \times 100\%$. We compare our method with two competing algorithms, the top-down method [Chen et al., 2019] and the algorithm of Ghoshal and Honorio [2018], and we follow the suggestions given in the two papers to choose the tuning parameters. These two algorithms are reported to have better performance than others. For our method, we fix $\alpha = 0.99, \gamma = 0.01, \kappa = 0$, $c_0 = 3$ and run MCMC for $3,000$ iterations for each simulated data set and discard the first 1,500 samples as burn-in. We always use

Method	Signal	Uniform($[-1,-0.3]\cup[0.3,1]$)			Uniform($[-1,-0.1]\cup[0.1,1]$)		
	n	100	500	1000	100	500	1000
Proposed	HD	10.0±0.5	0.8±0.2	0.1±0.1	13.9±0.7	5.2±0.3	3.0±0.3
	FNR	33.3±1.5	1.6±0.4	0.2±0.1	47.6±1.7	16.4±1.1	8.4±1.0
	FDR	3.2±0.8	1.4±0.4	0.2±0.1	2.4±0.6	2.6±0.5	2.5±0.4
	Flip	1.9±0.5	1.2±0.3	0.2±0.1	1.1±0.3	2.3±0.5	2.2±0.4
	Time	13.3±0.2	13.6±0.2	13.3±0.2	12.3±0.2	13.2±0.2	13.4±0.2
TD	HD	11.9±0.8	1.5±0.4	0.3±0.2	16.0±0.9	5.9±0.6	4.1±0.6
	FNR	37.8±1.8	2.3±0.5	0.3±0.2	52.5±1.7	15.0±1.3	8.2±0.9
	FDR	6.4±1.3	2.8±0.7	0.7±0.4	7.0±1.4	5.9±1.0	5.8±1.1
	Flip	3.6±0.8	1.8±0.5	0.3±0.2	2.9±0.7	4.1±0.7	4.1±0.6
	Time	0.6±0.0	0.5±0.0	0.5±0.0	0.5±0.0	0.6±0.0	0.5±0.0
LISTEN	HD	12.6±0.7	2.1±0.5	0.9±0.4	16.3±0.9	6.5±0.6	4.2±0.6
	FNR	39.5±1.7	3.1±0.7	1.0±0.4	52.2±1.7	15.9±1.1	8.9±1.0
	FDR	7.6±1.5	3.9±1.1	1.8±0.8	8.7±1.7	7.0±1.1	5.5±1.0
	Flip	3.6±0.7	2.6±0.7	1.0±0.4	3.3±0.7	4.6±0.7	4.3±0.7
	Time	0.5±0.0	0.5±0.0	0.6±0.0	0.6±0.0	0.6±0.0	0.5±0.0

Table 1: Uniform signal case with $p = 40$. TD and LISTEN refer to the top-down algorithm and the algorithm of Ghoshal and Honorio [2018], respectively. Each entry gives mean ± 1 standard error. Time is measured in seconds.

the following procedure to generate the true DAG G^*. We fix the true ordering to be $\sigma^* = (1, \ldots, p)$, and for each pair (i, j) such that $i < j$, we add edge $i \to j$ to G^* with probability $p_{\text{edge}} = 3/(2p-2)$. Hence, the expected number of edges of G^* is $3p/4$. The DAG G^* is resampled for each simulated data set.

We first generate the data from the structural equation models given in (2). We fix $p = 40$, set $\omega^* = 1$, and draw the edge weight B^*_{ij} for each edge $i \to j$ in G^* independently from some distribution F. We let sample size n be $100, 500$ or $1,000$, and repeat 30 times for each choice. In Table 1, we present the result for F being the uniform distribution on $[-1,-0.3] \cup [0.3,1]$ and that for F being the uniform distribution on $[-1,-0.1] \cup [0.1,1]$. The result for F being the standard Gaussian distribution is displayed in Section C.2 in the supplementary material. Table 1 shows that our method outperforms the other two methods in all settings by any of the four performance metrics, and in most cases, our method is better by a margin of at least one standard error.

p	n	d	FNR	FDR	Flip	Time
7	60	1.549	22.9±3.8	8.9±2.5	4.8±1.4	2.3±0.1
14	90	1.897	11.3±1.8	2.7±0.8	2.2±0.7	3.7±0.1
28	120	2.191	4.6±0.7	0.6±0.2	0.4±0.2	6.2±0.1
56	150	2.449	2.7±0.3	0.5±0.2	0.4±0.2	15.1±0.2
112	180	2.683	1.2±0.2	0.2±0.1	0.1±0.0	64.4±1.2
224	210	2.898	0.8±0.1	0.2±0.1	0.1±0.0	377.2±5.5
448	240	3.098	0.5±0.1	0.1±0.1	0.1±0.0	2896.4±48.1

Table 2: Simulation under a high-dimensional regime. Each entry gives mean ± 1 standard error. Time is measured in seconds.

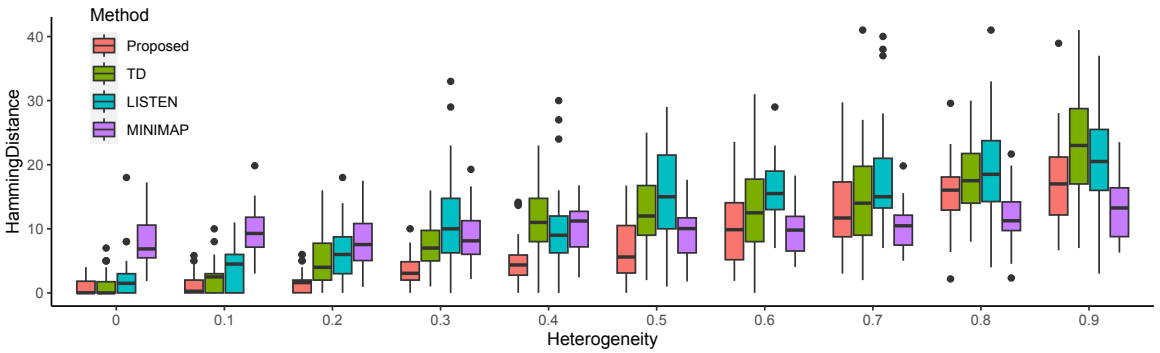


Figure 2: Boxplots for heterogeneous error variance case with $n = 500, p = 40$. We sample error variances from Uniform($[1-b, 1+b]$) for $b = 0, 0.1, \ldots, 0.9$ and nonzero edge weights from Uniform($[-1, -0.3] \cup [0.3, 1]$). The x-axis indicates the heterogeneity parameter b, and the y-axis represents the Hamming distance between the estimated DAG and G^*. MINIMAP is the minimal I-MAP MCMC that uses score (16), and thus it is score equivalent.

Next, we examine the performance of our method with varying n and p. To emulate a high-dimensional asymptotic regime where n grows linearly and p increases exponentially, we consider 7 settings where $n = 30(k+1)$ and $p = 7 \cdot 2^{k-1}$ in the k-th setting. When $k = 6$ or 7, we have $p > n$. We generate G^* with $p_{\text{edge}} = d/(p-1)$, where $d = 0.2\sqrt{n}$ is the expected number of neighbors for each node. We sample the edge weight B^*_{ij} for each $i \to j$ in G^* uniformly from $[-1, -0.5] \cup [0.5, 1]$ and set the error variance $\omega^* = 1$. We use the same values for $\alpha, \gamma, \kappa, c_0$ and run 3,000 MCMC iterations with 1,500 discarded samples as burn-in. The result of 30 replicates is summarized in Table 2, from which we see that FNR, FDR and flip rates all decrease as p increases. Further, the method is considerably scalable as it completes 3,000 iterations within an hour even when $p = 448$.

Lastly, we generate X by assuming each e_j in the structural equation models (2) has variance ω_j; thus, the equal variance assumption is violated. We repeat the simulation study presented in the left column of Table 1 by sampling ω_j independently from the uniform distribution on $[0.7, 1.3]$ for each j, and we observe that the advantage of the proposed method is more significant; see Section C.2 in the supplementary material for the result. To further examine how the heterogeneity of error variances affects the performance of our method, we fix $n = 500$ and $p = 40$, and sample ω_j from Uniform($[1-b, 1+b]$) for $b = 0, 0.1, \ldots, 0.9$. We plot the distribution of the HD metric over 30 replicates against b in Fig. 2. The proposed order MCMC sampler again performs uniformly better than competing algorithms. Besides, our method appears to be more robust, especially when b is not too large, which is probably due to the use of model averaging in Bayesian posterior inference.

4.3 Quantification of the bias caused by the equal variance assumption

When the true data generating process does not satisfy the equal variance assumption, our method is expected to have some bias. This is confirmed in Fig. 2, from which we see that HD increases with the heterogeneity of error variances. For comparison, we have also included in Fig. 2 the score-equivalent minimal I-MAP MCMC with score given by (16). Since this score does not encode the equal variance assumption, the minimal I-MAP MCMC sampler cannot determine the direction of an edge if reversing it yields another Markov equivalent

Method		$b = 0$	$b = 0.3$	$b = 0.5$	$b = 0.7$	$b = 0.9$	IG(3, 2)
Proposed	HD	0.1±0.0	0.5±0.2	1.6±0.4	2.1±0.5	2.6±0.5	3.3±0.8
	SHD	0.0±0.0	0.1±0.0	0.3±0.1	0.4±0.1	0.4±0.1	0.5±0.2
	Flip	1.1±0.7	4.0±1.5	10.0±2.4	13.4±3.0	18.5±3.9	21.1±4.1
MINIMAP	HD	3.0±0.3	2.5±0.2	2.6±0.3	2.6±0.2	2.7±0.2	2.6±0.2
	SHD	0.5±0.1	0.3±0.1	0.4±0.1	0.4±0.1	0.4±0.1	0.3±0.1
	Flip	23.0±2.9	22.3±3.1	23.4±3.2	23.7±3.2	24.7±3.1	23.7±3.0

Table 3: Analysis of the posterior distributions for $p = 7$. MINIMAP uses score (16), and thus it is score equivalent. The posterior inclusion probabilities of all edges are calculated exactly for both methods. The error variances are sampled from Uniform($[1 - b, 1 + b]$) or inverse-Gamma(3, 2). Each entry gives mean ± 1 standard error.

DAG. This can be clearly seen from Fig. 2: the performance of the minimal I-MAP MCMC does not change significantly with the heterogeneity level b, and it always has HD away from zero. When the heterogeneity level $b = 0.6$, which implies that the ratio between the maximum and minimum error variances can be as large as 4, the minimal I-MAP MCMC has a comparable performance to our method, and when $b \geq 0.7$, the minimal I-MAP MCMC performs better.

In order to better quantify the bias of our method, we exactly calculate the matrix Γ whose (i, j)-th element gives the posterior inclusion probability of the edge $i \to j$. We fix $p = 7$ so that we can enumerate all possible orderings, and the exact posterior inclusion probabilities corresponding to scores (7) and (16) can be calculated as

$$\Gamma_{ij} = \sum_{\sigma \in \mathbb{S}^p} \frac{e^{\phi(\hat{G}_\sigma)}}{\sum_{\sigma \in \mathbb{S}^p} e^{\phi(\hat{G}_\sigma)}} \mathbb{1}(\{i \to j\} \in \hat{G}_\sigma), \quad \Gamma^{\mathrm{eq}}_{ij} = \sum_{\sigma \in \mathbb{S}^p} \frac{e^{\phi_{\mathrm{eq}}(\hat{G}^{\mathrm{M}}_\sigma)}}{\sum_{\sigma \in \mathbb{S}^p} e^{\phi_{\mathrm{eq}}(\hat{G}^{\mathrm{M}}_\sigma)}} \mathbb{1}(\{i \to j\} \in \hat{G}^{\mathrm{M}}_\sigma),$$

where $\hat{G}_\sigma$ and $\hat{G}^{\mathrm{M}}_\sigma$ are the estimated DAGs given an ordering σ by our method and the minimal I-MAP method, respectively. We set $n = 100\,p$ and $p_{\mathrm{edge}} = 3/(2p - 2)$, sample nonzero edge weights from Uniform($[-1, -0.3] \cup [0.3, 1]$), and sample error variances from Uniform($[1 - b, 1 + b]$) and the inverse gamma distribution IG(a_1, a_2). We set $a_1 = 3$, which is the smallest integer that yields a finite variance, and set $a_2 = 2$ so that the expected value equals 1. We generate 30 replicates for each simulation setting. In Table 3, we report three metrics, HD, Flip, and the Hamming distance for skeletons (SHD); recall that the skeleton of a DAG is the undirected graph obtained by undirecting all edges. SHD is consistently close to zero throughout the simulation settings, which implies that the true skeleton is correctly identified by both methods regardless of the heterogeneity level b. Notably, in all the settings considered, even when $b = 0.9$ or in the inverse-gamma case, our method has a smaller flip rate than the minimal I-MAP method. That is, imposing the equal variance assumption does not increase the flip rate compared to a score-equivalent approach, which suggests that the computational gain resulting from this assumption is essentially obtained for free in this example.

5 Single-cell real data analysis

We use a real data set from the single-cell RNA database for Alzheimer's disease, known as scREAD [Jiang et al., 2020], to illustrate the advantages of the proposed algorithm. We

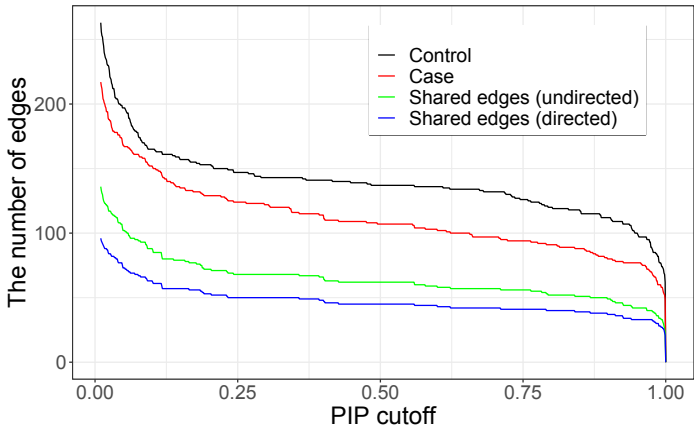


Figure 3: Result of the proposed method for the real data analysis. Given $\hat{\Gamma}_{ij}^{\mathrm{RB}}$, we infer the edge $i \to j$ exists in the DAG if $\hat{\Gamma}_{ij}^{\mathrm{RB}} > c$ where c is the cutoff of posterior inclusion probability. For each c, we count the number of edges occurring in the DAG for control samples (black), the number of edges in the DAG for case samples (red), the number of edges with edge direction ignored in both DAGs (green), and the number of directed edges in both DAGs (blue).

only consider genes involved in the brain-derived neurotrophic factor signaling pathway and expressed in the layer 2–3 glutamatergic neurons. The goal is to learn two DAG models, one from case samples and the other from control samples, and then inspect how different the two DAGs are. To mitigate potential batch effects, we only use samples that are generated at similar sequencing depths by checking the total and median expression level across all genes for each sample cell, which results in $n_0 = 2300$ control samples and $n_1 = 1666$ case samples. Next, we select the genes in this pathway expressed in at least half of the samples in both data sets, which yields $p = 73$. The data matrices for both case and control samples are obtained by performing normalization of log-transformed expression levels [Lee, 2007, Chapter 6].

For each of the two data sets, we run the proposed order MCMC sampler with iterative top-down initialization for 2×10^5 MCMC iterations, and then discard the first 10^5 iterations as burn-in. It only takes about 480 seconds for each data set. To infer the edge posterior inclusion probabilities, we use the conditioning scheme described in Section 3.3, and the result is presented in Fig. 3. The two DAGs learned from the data share a significant proportion of undirected edges, and more importantly, most of these edges have the same direction in both data sets: the gap between the blue and green lines in Fig. 3 is narrow. In other words, the orderings of the variables learned from the two data sets are very similar. The true ordering of the variables is hard to determine as there may even exist feedback loops among the selected genes, and we do not know to what extent the true model satisfies the equal variance assumption. But Fig. 3 suggests that the use of this score is very reasonable from a pragmatic perspective. For comparison, we have also tried the minimal I-MAP MCMC with the decomposable score given in (16), which represents a state-of-the-art score equivalent Bayesian structure learning procedure, and the result is shown in Section C.3 of the supplementary material. Given the same initialization and same number of MCMC and burn-in iterations, our method yields a higher proportion of shared directed edges than the minimal I-MAP MCMC. For example, with the posterior inclusion probability cutoff being

0.5, for our method 41% of the edges in the inferred DAG for case samples also occur in the same direction in the DAG for control samples, while this ratio drops to 26% for the minimal I-MAP MCMC.

To provide further evidence for the advantage of the proposed structure learning method, we repeat the above analysis 30 times using both our sampler with non-decomposable score and the minimal I-MAP MCMC with decomposable score. Then, for each pair (i, j) with $i \neq j$, we calculate the Gelman-Rubin scale factor [Gelman and Rubin, 1992] using Γ_{ij}, which is equal to 1 if $i \to j$ is in the sampled DAG and 0 otherwise. Thus, we get $p(p-1)$ Gelman-Rubin statistics for each data set, one for each directed edge. We find that 99.7% of the directed edges in the two DAGs have Gelman-Rubin statistics lower than 1.1 for our method, and 93.7% for the minimal I-MAP MCMC; we use the threshold 1.1 since this is the most common choice according to Vats and Knudson [2021]. Moreover, for the minimal I-MAP MCMC, Gelman-Rubin statistics of 90 directed edges yield infinity, which means that the within-chain variance of Γ_{ij} is zero for all 30 runs, but the between-chain variance is nonzero; that is, in some runs the edge $i \to j$ is selected in every iteration excluding burn-in, while in the other runs the edge $i \to j$ is never selected. This observation again illustrates that for a score equivalent procedure, traversing equivalence classes can sometimes be very difficult and cause slow mixing of MCMC samplers. In contrast, the maximum Gelman-Rubin statistic for our method is 2.56 for the control data set and 1.26 for the case data set.

Acknowledgement

The authors would like to thank the anonymous reviewers for their comments which helped improve the paper, and thank Prof. Mohsen Pourahmadi and Yongjian Yang for helpful discussions. HC and QZ were supported in part by NSF grant DMS-2245591. All authors were supported by the Triads for Transformation Grant of Texas A&M University.

References

Raj Agrawal, Caroline Uhler, and Tamara Broderick. Minimal I-MAP MCMC for scalable structure discovery in causal DAG models. In *International Conference on Machine Learning*, pages 89–98, 2018.

Hongzhi An, Da Huang, Qiwei Yao, and Cun-Hui Zhang. Stepwise searching for feature variables in high-dimensional linear regression. *Technical report*, 2008.

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

Bryon Aragam, Arash Amini, and Qing Zhou. Globally optimal score-based learning of directed acyclic graphs in high-dimensions. *Advances in Neural Information Processing Systems*, 32:4450–4462, 2019.

Emanuel Ben-David, Tianxi Li, Hélene Massam, and Bala Rajaratnam. High dimensional Bayesian inference for Gaussian directed acyclic graph models. *arXiv preprint arXiv:1109.4371*, 2011.

Megan Bernstein and Evita Nestoridi. Cutoff for random to random card shuffle. *The Annals of Probability*, 47(5):3303–3320, 2019.

Xuan Cao, Kshitij Khare, and Malay Ghosh. Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models. *The Annals of Statistics*, 47(1):319–348, 2019.

Carlos M Carvalho and James G Scott. Objective Bayesian model selection in Gaussian graphical models. *Biometrika*, 96(3):497–512, 2009.

Federico Castelletti and Guido Consonni. Bayesian inference of causal effects from observational data in Gaussian graphical models. *Biometrics*, 77(1):136–149, 2021.

Federico Castelletti, Guido Consonni, Marco L Della Vedova, and Stefano Peluso. Learning Markov equivalence classes of directed acyclic graphs: an objective Bayes approach. *Bayesian Analysis*, 13(4):1235–1260, 2018.

Hyunwoong Chang, Changwoo Lee, Zhao Tang Luo, Huiyan Sang, and Quan Zhou. Rapidly mixing multiple-try Metropolis algorithms for model selection problems. *Advances in Neural Information Processing Systems*, 35:25842–25855, 2022.

Wenyu Chen, Mathias Drton, and Y Samuel Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 2019.

David Maxwell Chickering. Learning equivalence classes of Bayesian-network structures. *Journal of machine learning research*, 2(Feb):445–498, 2002.

Mathias Drton and Marloes H Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393, 2017.

Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! In *Conference on learning theory*, pages 793–797. PMLR, 2018.

Nir Friedman and Daphne Koller. Being bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine learning*, 50(1):95–125, 2003.

Dan Geiger and David Heckerman. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *The Annals of Statistics*, 30 (5):1412–1440, 2002.

Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

Asish Ghoshal and Jean Honorio. Learning linear structural equation models in polynomial time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pages 1466–1475. PMLR, 2018.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Marco Grzegorczyk and Dirk Husmeier. Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Machine Learning*, 71(2-3):265, 2008.

Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *NIPS*, volume 21, pages 689–696. Citeseer, 2008.

Jing Jiang, Cankun Wang, Ren Qi, Hongjun Fu, and Qin Ma. scread: A single-cell RNA-Seq database for Alzheimer's disease. *Iscience*, 23(11):101769, 2020.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Jack Kuipers, Polina Suter, and Giusi Moffa. Efficient sampling and structure learning of Bayesian networks. *Journal of Computational and Graphical Statistics*, 31(3):639–650, 2022.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Steffen L Lauritzen. Propagation of probabilities, means, and variances in mixed graphical association models. *Journal of the American Statistical Association*, 87(420):1098–1108, 1992.

Kyoungjae Lee, Jaeyong Lee, and Lizhen Lin. Minimax posterior convergence rates and model selection consistency in high-dimensional DAG models based on sparse Cholesky factors. *The Annals of Statistics*, 47(6):3413–3437, 2019.

Mei-Ling Ting Lee. *Analysis of microarray gene expression data*. Springer Science & Business Media, 2007.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

David Madigan, Jeremy York, and Denis Allard. Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique*, pages 215–232, 1995.

Oren Mangoubi and Aaron Smith. Rapid mixing of Hamiltonian Monte Carlo on strongly log-concave distributions. *arXiv preprint arXiv:1708.07114*, 2017.

Ryan Martin, Raymond Mess, and Stephen G Walker. Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli*, 23(3):1822–1847, 2017.

Gunwoong Park. Identifiability of additive noise models using conditional variances. *The Journal of Machine Learning Research*, 21(1):2896–2929, 2020.

J Peters, J Mooij, D Janzing, and B Schölkopf. Identifiability of causal graphs using functional models. In *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 589–598. AUAI Press, 2011.

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

Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing ℓ^1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Christian P Robert and Gareth O Roberts. Rao-Blackwellization in the MCMC era. *arXiv preprint arXiv:2101.01011*, 2021.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Ali Shojaie and George Michailidis. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538, 2010.

Alistair Sinclair. Improved bounds for mixing rates of Markov chains and multicommodity flow. *Combinatorics, probability and Computing*, 1(4):351–370, 1992.

David Strieder, Tobias Freidling, Stefan Haffner, and Mathias Drton. Confidence in causal discovery with linear causal models. In *Uncertainty in Artificial Intelligence*, pages 1217–1226. PMLR, 2021.

Chengwei Su and Mark E Borsuk. Improving structure MCMC for Bayesian networks through Markov blanket resampling. *The Journal of Machine Learning Research*, 17(1): 4042–4061, 2016.

Seth Sullivant, Kelli Talaska, and Jan Draisma. Trek separation for Gaussian graphical models. *The Annals of Statistics*, 38(3):1665–1685, 2010.

Mahlet G Tadesse and Marina Vannucci. Handbook of Bayesian variable selection. 2021.

Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, pages 436–463, 2013.

Sara Van de Geer and Peter Bühlmann. ℓ^0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, 41(2):536–567, 2013.

Dootika Vats and Christina Knudson. Revisiting the Gelman–Rubin diagnostic. *Statistical Science*, 36(4):518–529, 2021.

Yun Yang, Martin J Wainwright, and Michael I Jordan. On the computational complexity of high-dimensional Bayesian variable selection. *The Annals of Statistics*, 44(6):2497–2532, 2016.

Guo Yu and Jacob Bien. Learning local dependence in ordered data. *The Journal of Machine Learning Research*, 18(1):1354–1413, 2017.

Quan Zhou and Hyunwoong Chang. Complexity analysis of Bayesian learning of high-dimensional DAG models and their equivalence classes. *arXiv preprint arXiv:2101.04084*, 2021.

Shuheng Zhou. Thresholded Lasso for high dimensional variable selection and statistical estimation. *arXiv preprint arXiv:1002.1583*, 2010.

Supplementary material

A Algorithms

A.1 Overview of the proposed method

We outline the proposed order MCMC algorithm in Algorithm 3. For all displayed algorithms, we assume the data matrix X and model parameters $(c_0, \gamma, \alpha, \kappa, d_{\text{in}})$ are given. The `R` code for the proposed method and simulation studies can be found at https://github.com/hwchang1201/bayes.eqvar.

Algorithm 3: Bayesian order-based structure learning

Input: Number of MCMC iterations T, neighborhood function $\mathcal{N} = \mathcal{N}_{\text{adj}}, \mathcal{N}_{\text{rtr}}$ or $\mathcal{N}_{\text{rrs}}$, a DAG selection procedure $\hat{G}\colon \mathbb{S}^p \to \mathcal{G}_p$ (e.g. Algorithm 5)

$\sigma^{(0)} \leftarrow \hat{\sigma}^{\text{ITD}}$ // $\hat{\sigma}^{\text{ITD}}$ is the output of Algorithm 2

$G^{(0)} \leftarrow \hat{G}(\sigma^{(0)})$

for $t = 1, \ldots, T$ **do**

  Draw σ uniformly from $\mathcal{N}(\sigma^{(t-1)})$)

  Draw $u \sim \text{Uniform}(0, 1)$

  $a \leftarrow \min(\pi_n(\sigma)/\pi_n(\sigma^{(t-1)}), 1)$

  if $u \le a$ **then**

    $\sigma^{(t)} \leftarrow \sigma$

    $G^{(t)} \leftarrow \hat{G}(\sigma)$

  else

    $\sigma^{(t)} \leftarrow \sigma^{(t-1)}$

    $G^{(t)} \leftarrow G^{(t-1)}$

  $\hat{\Gamma}^{(t)} = \hat{\Gamma}(\sigma^{(t)}, G^{(t)})$ // See (15) for the definition of $\hat{\Gamma}$

Output: "Rao-Blackwellized" adjacency matrices $\{\hat{\Gamma}^{(t)}\}_{t=1}^{T}$

A.2 Forward-backward algorithms with non-decomposable scores

Recall the posterior score of a DAG given in (7). Define the nodewise score at node j by

$$\phi_j(S, \text{RSS}_{\cdot j}) = -|S| \log \left\{ p^{c_0} \sqrt{(1 + \alpha/\gamma)} \right\} - \frac{\alpha p n + \kappa}{2} \log \left(\text{RSS}_{\cdot j} + X_j^{\mathrm{T}} \Phi_S^{\perp} X_j \right), \tag{17}$$

for $S \subseteq P_j$, where $\text{RSS}_{\cdot j}$ denotes the total residual sum of squares of nodes other than j, and P_j is the potential parent set defined in (1). Hence, given $\text{RSS}_{\cdot j}$, we can use the standard forward-backward stepwise algorithm to select the parent set of node j; this is described in Algorithm 4. We allow using two different estimates for $\text{RSS}_{\cdot j}$, one for the forward phase and the other for the backward phase; the reason will become clear in the next subsection.

A.3 Implementation of order MCMC with non-decomposable scores

For our model, the main computational challenge is that a local change to the ordering σ can cause some global changes to the maximum a posteriori DAG estimator $\hat{G}_{\sigma}^{\text{MAP}}$, due to

Algorithm 4: Nodewise forward-backward selection

Input: Node index $j \in [p]$, a set of potential parent nodes $P_j \subset [p]$, two estimates for the total residual of sum of squares of other nodes $\mathrm{RSS}_{\text{-}j}$, $\mathrm{RSS}'_{\text{-}j}$

Forward phase: $S_{\mathrm{f}} \leftarrow \emptyset$
for $k = 1, \ldots, |P_j|$ **do**
  $\ell_0 \leftarrow \arg\max_{\ell \in P_j \setminus S_{\mathrm{f}}} \phi_j(S_{\mathrm{f}} \cup \{\ell\}, \mathrm{RSS}_{\text{-}j})$
  $\tilde{S}_{\mathrm{f}} \leftarrow S_{\mathrm{f}} \cup \{\ell_0\}$
  if $\phi_j(\tilde{S}_{\mathrm{f}}, \mathrm{RSS}_{\text{-}j}) \geq \phi_j(S_{\mathrm{f}}, \mathrm{RSS}_{\text{-}j})$ **then**
    $S_{\mathrm{f}} \leftarrow \tilde{S}_{\mathrm{f}}$
  else
    break
Backward phase: $S_{\mathrm{b}} \leftarrow S_{\mathrm{f}}$
for $k = 1, \ldots, |S_{\mathrm{f}}|$ **do**
  $\ell_1 \leftarrow \arg\max_{\ell \in S_{\mathrm{b}}} \phi_j(S_{\mathrm{b}} \setminus \{\ell\}, \mathrm{RSS}'_{\text{-}j})$
  $\tilde{S}_{\mathrm{b}} \leftarrow S_{\mathrm{b}} \setminus \{\ell_1\}$
  if $\phi_j(\tilde{S}_{\mathrm{b}}, \mathrm{RSS}'_{\text{-}j}) \geq \phi_j(S_{\mathrm{b}}, \mathrm{RSS}'_{\text{-}j})$ **then**
    $S_{\mathrm{b}} \leftarrow \tilde{S}_{\mathrm{b}}$
  else
    break
Output: A parent set S_{b} of node j

the use of the non-decomposable posterior score. Were the posterior score decomposable, whenever we use an adjacent transposition to move from σ to $\sigma' = \sigma \circ (i, i+1)_{\mathrm{c}}$, we know that $\mathrm{Pa}_j(\hat{G}^{\mathrm{MAP}}_{\sigma}) = \mathrm{Pa}_j(\hat{G}^{\mathrm{MAP}}_{\sigma'})$ for any $j \notin \{\sigma(i), \sigma(i+1)\}$, since maximizing the score of the entire DAG is equivalent to maximizing the local score at each node separately.

We describe a strategy for implementing local moves on $\mathbb{S}^p$ for our model, which is as efficient as with a decomposable posterior score. We start by proving two monotone properties of the nodewise score defined in (17).

Lemma 1. Let ϕ_j be as given in (17), $S \subset [p] \setminus \{j\}$, $k \notin S \cup \{j\}$ and $a > 0$.

(i) If $\phi_j(S \cup \{k\}, a) > \phi_j(S, a)$, then $\phi_j(S \cup \{k\}, b) > \phi_j(S, b)$ for any $0 < b < a$.

(ii) If $\phi_j(S \cup \{k\}, a) < \phi_j(S, a)$, then $\phi_j(S \cup \{k\}, b) < \phi_j(S, b)$ for any $b > a$.

Proof. To simplify the notation, let $K_0 = \log\{p^{c_0}\sqrt{(1+\alpha/\gamma)}\}$ and $K_1 = (\alpha p n + \kappa)/2$. A routine calculation shows that $\phi_j(S \cup \{k\}, a) > \phi_j(S, a)$ if and only if

$$\log \frac{a + X_j^{\mathrm{T}} \Phi_S^{\perp} X_j}{a + X_j^{\mathrm{T}} \Phi_{S \cup \{k\}}^{\perp} X_j} > \frac{K_0}{K_1}.$$

The claim follows by observing that the left-hand side is monotonically decreasing in a. □

Motivated by Lemma 1, we use the following procedure to find $\hat{G}^{\mathrm{MAP}}_{\sigma}$ for a given $\sigma \in \mathbb{S}^p$. First, for $j = 1, \ldots, p$, we find a lower bound and an upper bound on RSS_j such that

Algorithm 5: Forward-backward DAG selection

Input: $\sigma \in \mathbb{S}^p$

$G \leftarrow$ empty DAG

`// Forward phase`

while 1 **do**

  $(i_0, j_0) \leftarrow \arg\max_{i,j:\, \sigma^{-1}(i)<\sigma^{-1}(j), \{i\to j\}\notin G} \phi(G \cup \{i \to j\})$

  $\tilde{G} \leftarrow G \cup \{i_0 \to j_0\}$

  if $\phi(\tilde{G}) \geq \phi(G)$ **then**

    $G \leftarrow \tilde{G}$

  else

    break

`// Backward phase`

while 1 **do**

  $(i_1, j_1) \leftarrow \arg\max_{i,j:\, \{i\to j\}\in G} \phi(G \setminus \{i \to j\})$

  $\tilde{G} \leftarrow G \setminus \{i_1 \to j_1\}$

  if $\phi(\tilde{G}) \geq \phi(G)$ **then**

    $G \leftarrow \tilde{G}$

  else

    break

Output: DAG G

both bounds do not depend on σ. An obvious choice for the upper bound on RSS_j is given by $\overline{\mu}_j = X_j^{\mathrm{T}} X_j$, and if $p < n$, a lower bound is given by $\underline{\mu}_j = X_j^{\mathrm{T}} \Phi^{\perp}_{[p]\setminus\{j\}} X_j$ (we assume $\underline{\mu}_j$ is strictly positive). Next, for $j = 1, \ldots, p$, we apply Algorithm 4 with input $(j, P_j, \sum_{k\neq j} \underline{\mu}_k, \sum_{k\neq j} \overline{\mu}_k)$; that is, in the forward stage, we let the algorithm select as many parent nodes as possible by using minimum estimates for the residual sum of squares of other nodes, and in the backward stage, we let the algorithm remove as many nodes as possible. For all nodes, save the search paths of Algorithm 4, including the changes in residual sum of squares in each step, in the internal memory, and let $\overline{S}_j^{\sigma}$ denote the parent set of node j at the end of the forward stage. Denote by $\overline{G}_{\sigma}$ the DAG such that $\mathrm{Pa}_j(\overline{G}_{\sigma}) = \overline{S}_j^{\sigma}$ for each j. Now to find $\hat{G}_{\sigma}^{\mathrm{MAP}}$, we simply apply the backward stage of Algorithm 5 by initializing the DAG to $\overline{G}_{\sigma}$. This can be done very efficiently by using the search paths of Algorithm 4; no calculation of residual sum of squares is needed.

The above procedure enables an efficient updating algorithm for finding $\hat{G}_{\sigma}^{\mathrm{MAP}}$ when we move locally on the ordering space $\mathbb{S}^p$. For example, consider moving from σ to $\sigma' = \sigma \circ (i, i+1)_{\mathrm{c}}$. We only need to apply Algorithm 4 at nodes $\sigma(i)$ and $\sigma(i+1)$, and then perform backward DAG selection using the saved search paths of nodewise forward-backward selection. The computational time of the DAG selection step is negligible compared to that of Algorithm 4. Note that the parent sets of nodes other than $\sigma(i)$ and $\sigma(i+1)$ may change.

A.4 Three random walk proposals

Figure 4 describes (1) adjacent transposition, (2) random transposition, and (3) random-to-random shuffle, given the current topological ordering σ. The random transposition $\sigma \circ (i,j)_{\mathrm{c}}$ interchanges the i-th and the j-th elements of σ while keeping the others unchanged. The adjacent transposition is a special case of random transposition where i and j are adjacent, i.e., $|i-j|=1$. The random-to-random shuffle $\sigma \circ \xi(i,j)$ inserts the i-th element of σ to the j-th position.

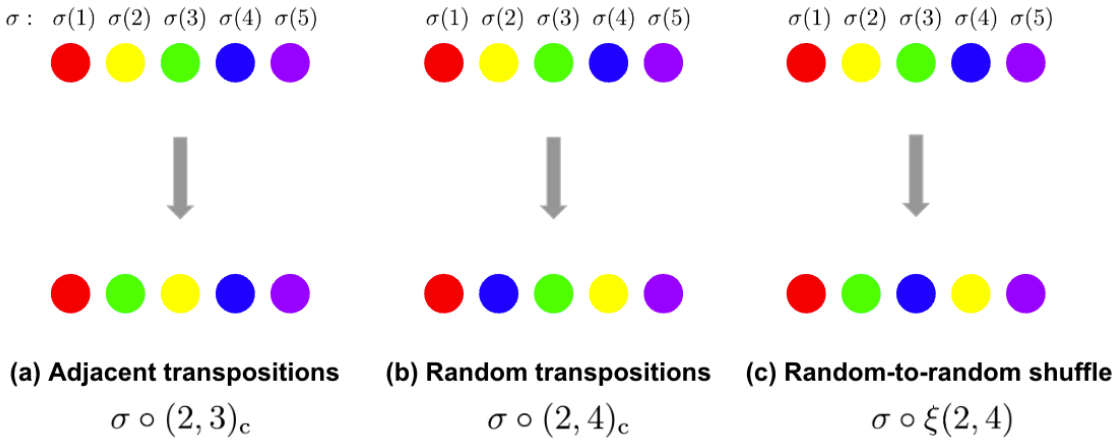


Figure 4: Illustration of the three proposals introduced in Section 3.1: adjacent transposition, the random transposition and the random-to-random shuffle.

B Proofs

B.1 High-probability events

Recall that we assume the data is generated according to the linear structural equation model (SEM) given in (10). Since the rows of X are assumed to be i.i.d. copies of X, we have

$$X_j = \sum_{i=1}^{p} (B^*)_{ij} X_i + \epsilon_j, \text{ where } \epsilon_j \sim N_n(0, \omega_j^* I), \text{ for all } j \in [p]. \tag{18}$$

By Remark 3, for each $\sigma \in \mathbb{S}^p$, we can derive a linear SEM equivalent to (18), which is given by

$$X_j = \sum_{i=1}^{p} (B_\sigma^*)_{ij} X_i + \epsilon_j^\sigma, \text{ where } \epsilon_j^\sigma \sim N_n(0, \omega_j^\sigma I), \text{ for all } j \in [p]. \tag{19}$$

We define the normalized error vectors by

$$z_j = (\omega_j^*)^{-\frac{1}{2}} \epsilon_j \text{ for } j \in [p], \qquad z_j^\sigma = (\omega_j^\sigma)^{-\frac{1}{2}} \epsilon_j^\sigma \text{ for } \sigma \in \mathbb{S}^p, j \in [p],$$

where z_j and z_j^σ are associated with the true model given in (18) and the linear SEM in (19), respectively. The sets of the corresponding normalized errors are defined by $\mathcal{Z}_0 = \{z_j : j \in$

$[p]\}$ and $\mathcal{Z}_1 = \{z_j^\sigma : \sigma \in \mathbb{S}^p, j \in [p]\}$. Clearly, $\mathcal{Z}_0 \subseteq \mathcal{Z}_1$ and $|\mathcal{Z}_0| = p$. Further, one can show that

$$|\mathcal{Z}_1| \leq p \cdot \binom{p}{d^*},$$

where d^* is defined in (12).

Before we prove the results given in the main text, we first define some event sets on which the random components of our generating SEM behaves as desired, and use concentration inequalities to show that they happen with high probability. We will then prove the main results of the paper by conditioning on these high-probability events. Recall P_j^σ defined in (1) and let $\mathcal{M}_p(d, P) = \{S \subseteq P : |S| \leq d\}$. Define

$$\mathcal{A} = \left\{ n\underline{\nu} \leq \min_{S \subseteq \mathcal{M}_p(2d_{\text{in}}, [p])} \lambda_{\min}(X_S^{\mathrm{T}} X_S) \leq \max_{S \subseteq \mathcal{M}_p(2d_{\text{in}}, [p])} \lambda_{\max}(X_S^{\mathrm{T}} X_S) \leq n\overline{\nu} \right\},$$

$$\mathcal{B} = \left\{ \min_{j \in [p]} \min_{S \subseteq \mathcal{M}_p(2d_{\text{in}}, P_j^{\sigma^*})} (z_j)^{\mathrm{T}} \Phi_S^{\perp} z_j \geq \frac{1}{2} n \right\},$$

$$\mathcal{B}' = \left\{ \min_{j \in [p], \sigma \in \mathbb{S}^p} \min_{S \subseteq \mathcal{M}_p(2d_{\text{in}}, P_j^{\sigma})} (z_j^\sigma)^{\mathrm{T}} \Phi_S^{\perp} z_j^\sigma \geq \frac{1}{2} n \right\},$$

$$\mathcal{C} = \left\{ \max_{j \in [p]} \max_{\substack{k \notin S \\ S \cup \{k\} \subseteq \mathcal{M}_p(2d_{\text{in}}, P_j^{\sigma^*})}} z_j^{\mathrm{T}} (\Phi_{S \cup \{k\}} - \Phi_S) z_j \leq \rho \log p \right\},$$

$$\mathcal{D} = \left\{ \min_{j \in [p], \sigma \in \mathbb{S}^p} \min_{S \subseteq \mathcal{M}_p(2d_{\text{in}}, P_j^{\sigma})} (z_j^\sigma)^{\mathrm{T}} \Phi_S^{\perp} z_j^\sigma > (1 - \frac{1}{2\eta}) n \right\},$$

$$\mathcal{E} = \left\{ \max_{j \in [p]} \max_{S \subseteq \mathcal{M}_p(2d_{\text{in}}, P_j^{\sigma^*})} z_j^{\mathrm{T}} \Phi_S^{\perp} z_j < (1 + \frac{1}{4\eta}) n \right\},$$

$$\mathcal{J} = \bigcap_{i,j \in [p]} \left\{ \left| \frac{X_i^{\mathrm{T}} X_j}{n} - \Sigma_{ij}^* \right| \leq 160\overline{\nu} \sqrt{\frac{\log p}{n}} \right\},$$

where $\eta, \rho > 0$ are universal constants.

Lemma 2. Under the conditions of Proposition 1, we have $\mathbb{P}^*(\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}) \geq 1 - 4p^{-1}$ for sufficiently large n.

Proof. From Lemma F1 of Zhou and Chang [2021], we have $\mathbb{P}^*(\mathcal{A}) \geq 1 - p^{-1}$ for sufficiently large n. The proof for the bounds of $\mathbb{P}^*(\mathcal{B})$ and $\mathbb{P}^*(\mathcal{C})$ is analogous to that of Lemma F2 of Zhou and Chang [2021]. A standard calculation using the tail bounds for chi-squared distributions [Laurent and Massart, 2000][Lemma 1] yields

$$\mathbb{P}^* \left\{ z_j^{\mathrm{T}} \Phi_S^{\perp} z_j \leq \frac{1}{2} n \right\} \leq e^{-n/48},$$

$$\mathbb{P}^* \left\{ z_j^{\mathrm{T}} (\Phi_{T \cup \{k\}} - \Phi_T) z_j \geq \rho \log p \right\} \leq 2e^{-\rho \log p / 2},$$

for any $j \in [p]$, $S \subseteq \mathcal{M}_p(2d_{\text{in}}, P_j^{\sigma^*})$ and $T \cup \{k\} \subseteq \mathcal{M}_p(2d_{\text{in}}, P_j^{\sigma^*})$. To conclude the proof, apply union bounds with the observations $|Z_0| \leq p$ and $|\mathcal{M}_p(2d_{\text{in}}, P_j^{\sigma^*})\}| \leq p^{2d_{\text{in}}+1}$ and the assumptions $d_{\text{in}} \log p = o(n)$ and $\rho > 4d_{\text{in}} + 6$. □

Lemma 3. Assume $d_{\text{in}} \log p = o(n)$ and $d^* \le d_{\text{in}}$. There exists some universal constant $c' = c'(\eta) > 0$ such that $\mathbb{P}^*(\mathcal{D} \cap \mathcal{E}) \ge 1 - 2e^{-c'n}$ for all sufficiently large n.

Proof. By Lemma 1 of Laurent and Massart [2000],

$$\mathbb{P}^*\left\{\frac{\chi_d^2}{d} \le 1 - a\right\} \le e^{-a^2 d/4}, \quad \mathbb{P}^*\left\{\frac{\chi_d^2}{d} \ge 1 + a + \frac{a^2}{2}\right\} \le e^{-a^2 d/4}, \tag{20}$$

where χ_d^2 denotes a chi-squared random variable with d degrees of freedom and $a > 0$ is arbitrary. Consider $\mathbb{P}^*(\mathcal{D})$ first. For any $j \in [p], \sigma \in \mathbb{S}^p$, and $S \in \mathcal{M}_p(2d_{\text{in}}, P_j^\sigma)$, by (20),

$$\mathbb{P}^*\left\{\frac{(z_j^\sigma)^{\mathrm{T}} \Phi_S^\perp z_j^\sigma}{n - |S|} \le 1 - \frac{1}{4\eta}\right\} \le \exp\left(-\frac{n - |S|}{64\eta^2}\right).$$

Since $|S| \le 2d_{\text{in}} = o(n/\log p)$, $n(n-|S|)^{-1}(1-(2\eta)^{-1}) \le 1 - (4\eta)^{-1}$ for sufficiently large n. Applying the union bound with $|\mathcal{Z}_1| \le p^{d_{\text{in}}+1}$ and $|\mathcal{M}_p(2d_{\text{in}}, P_j^\sigma)| \le p^{2d_{\text{in}}+1}$, we obtain

$$\mathbb{P}^*(\mathcal{D}^{\mathrm{c}}) \le p^{3d_{\text{in}}+2} \exp\left(-\frac{n}{128\eta^2}\right) \le e^{-c'n},$$

for sufficiently large n. Next, consider $\mathbb{P}^*(\mathcal{E})$. For any $j \in [p]$ and $S \in \mathcal{M}_p(2d_{\text{in}}, P_j^{\sigma^*})$, we have

$$\mathbb{P}^*\left\{\frac{z_j^{\mathrm{T}} \Phi_S^\perp z_j}{n - |S|} \ge 1 + \frac{1}{8\eta} + \frac{1}{128\eta^2}\right\} \le \exp\left(-\frac{n - |S|}{256\eta^2}\right),$$

by (20). Since $|\mathcal{Z}_0| = p$ and $|\mathcal{M}_p(2d_{\text{in}}, P_j^{\sigma^*})| \le p^{2d_{\text{in}}+1}$, the union bound gives

$$\mathbb{P}^*(\mathcal{E}^{\mathrm{c}}) \le p^{2d_{\text{in}}+2} \exp\left(-\frac{n}{512\eta^2}\right) \le e^{-c'n}.$$

Another application of the union bound yields the conclusion. □

Lemma 4. Under the conditions of Proposition 2, we have $\mathbb{P}^*(\mathcal{A} \cap \mathcal{B}' \cap \mathcal{J}) \ge 1 - 6p^{-1}$ for all sufficiently large n.

Proof. We have obtained the bound $\mathbb{P}^*(\mathcal{A}) \ge 1 - p^{-1}$ from Lemma 2, and the bound on $\mathbb{P}^*(\mathcal{B}')$ is proved in Lemma F2 of Zhou and Chang [2021]. Consider $\mathbb{P}^*(\mathcal{J}^{\mathrm{c}})$. Let

$$\mathcal{J}_{ij}^{\mathrm{c}} = \left\{\left|\frac{X_i^{\mathrm{T}} X_j}{n} - \Sigma_{ij}^*\right| > 160\overline{\nu}\sqrt{\frac{\log p}{n}}\right\}.$$

By Ravikumar et al. [2011, Lemma 1],

$$\mathbb{P}^*(\mathcal{J}_{ij}^{\mathrm{c}}) \le 4\exp(-3\overline{\nu}^2 \log p/(\max_i \Sigma_{ii}^*)^2) \le 4p^{-3},$$

from which we obtain $\mathbb{P}^*(\mathcal{J}^{\mathrm{c}}) = \mathbb{P}^*(\cup_{i,j\in[p]} \mathcal{J}_{ij}^{\mathrm{c}}) \le 4p^{-1}$ by the union bound. □

B.2 Proof of Proposition 1

We consider the proof of consistency for the estimator $\hat{G}^{\mathrm{MAP}}_{\sigma}$ defined in (9); that is, we show that the scoring criterion ϕ is consistent when the ordering σ is known. We first prove a technical lemma, which bounds the residual sum of squares $\mathrm{RSS}_j(G)$ when the node j is underfitted (i.e., $\mathrm{Pa}_j(G^*) \not\subseteq \mathrm{Pa}_j(G)$).

Lemma 5. Fix some $S \subseteq [p]$ such that $|S| \leq d_{\mathrm{in}}$ and $S \neq S^* = \mathrm{Pa}_j(G^*)$. Suppose we are on the event $\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}$ and the conditions of Proposition 1 hold. Then

$$X_j^{\mathrm{T}}(\Phi_{S\cup\{k_0\}} - \Phi_S)X_j \geq 9c_0\overline{\nu}\log p/\alpha,$$

for some $k_0 \in S^* \setminus S$.

Proof. We denote $X_j = Z_j + \epsilon_j$, $Z_j = X_{S^*}(B^*_j)_{S^*}$, where B^*_j is j-th column of the true weighted adjacency matrix B^*. Let $k_0 = \arg\max_{k\in S^*\setminus S} Z_j^{\mathrm{T}}(\Phi_{S\cup\{k\}} - \Phi_S)Z_j$. By the triangle inequality,

$$X_j^{\mathrm{T}}(\Phi_{S\cup\{k_0\}} - \Phi_S)X_j \ \geq (||(\Phi_{S\cup\{k_0\}} - \Phi_S)Z_j|| - ||(\Phi_{S\cup\{k_0\}} - \Phi_S)\epsilon_j||)^2. \tag{21}$$

On the event set $\mathcal{C}$, we can use $c_0 > \alpha\rho$ from condition (C2) to obtain that

$$||(\Phi_{S\cup\{k_0\}} - \Phi_S)\epsilon_j||^2 \leq \rho\omega^*_j \log p \leq \rho\overline{\nu}\log p < \frac{c_0}{\alpha}\overline{\nu}\log p,$$

and thus by Lemma E2 of Zhou and Chang [2021],

$$||(\Phi_{S\cup\{k_0\}} - \Phi_S)Z_j||^2 \geq \frac{||B^*_{S^*\setminus S}||^2}{|S^*\setminus S|}\frac{n\underline{\nu}^2}{\overline{\nu}} \geq 16c_0\frac{\overline{\nu}^2\log p}{\alpha\underline{\nu}^2 n}\frac{n\underline{\nu}^2}{\overline{\nu}} \geq \frac{16c_0}{\alpha}\overline{\nu}\log p.$$

The second inequality follows from condition (C3). Plugging the above two displayed bounds into (21), we obtain the asserted result. □

Proof of Proposition 1. On the event $\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}$ defined in Section B.1, we will show that all the three events stated in the proposition happen. For a non-negative integer d, define

$$\mathcal{G}^*_p(d) = \bigcup_{\sigma\in[\sigma^*]} \mathcal{G}^{\sigma}_p(d).$$

Event (i). Fix an arbitrary $G \in \mathcal{G}^*_p(2d_{\mathrm{in}})$ such that $\mathrm{Pa}_j(G^*) \subset \mathrm{Pa}_j(G)$ for some $j \in [p]$. We prove that we can remove all the redundant parents of node j. This is slightly stronger than the asserted result, but it will be useful later for proving the claim for event (iii). Pick an arbitrary $k \in \mathrm{Pa}_j(G) \setminus \mathrm{Pa}_j(G^*)$ and define $G' = G \setminus \{k \to j\}$. On the event $\mathcal{B} \cap \mathcal{C}$, we have

$$X_j^{\mathrm{T}}(\Phi^{\perp}_{\mathrm{Pa}_j(G')} - \Phi^{\perp}_{\mathrm{Pa}_j(G)})X_j = \epsilon_j^{\mathrm{T}}(\Phi_{\mathrm{Pa}_j(G)} - \Phi_{\mathrm{Pa}_j(G')})\epsilon_j \leq \omega^*_j\rho\log p,$$
$$\mathrm{RSS}_i(G) = X_i^{\mathrm{T}}\Phi^{\perp}_{\mathrm{Pa}_i(G)}X_i \geq \epsilon_i^{\mathrm{T}}\Phi^{\perp}_{\mathrm{Pa}_i(G)}\epsilon_i \geq \frac{n\omega^*_i}{2} \text{ for } i \in [p].$$

Since $1 + x \leq \exp(x)$ for $x \in \mathbb{R}$ and $\sqrt{1+\alpha/\gamma} > 1$, we find that

$$\begin{aligned}
\frac{\exp(\phi(G))}{\exp(\phi(G'))} &= \left(p^{c_0}\sqrt{1+\alpha/\gamma}\right)^{-1} \left(\frac{\sum_{i\neq j}^{p} \mathrm{RSS}_i(G) + \mathrm{RSS}_j(G')}{\sum_{i=1}^{p} \mathrm{RSS}_i(G)}\right)^{\frac{\alpha p n+\kappa}{2}} \\
&< p^{-c_0} \left(1 + \frac{X_j^{\mathrm{T}}(\Phi^{\perp}_{\mathrm{Pa}_j(G')} - \Phi^{\perp}_{\mathrm{Pa}_j(G)})X_j}{\sum_{i=1}^{p} \mathrm{RSS}_i(G)}\right)^{\frac{\alpha p n+\kappa}{2}} \\
&\leq p^{-c_0} \exp\left(\frac{\alpha np+\kappa}{2} \frac{X_j^{\mathrm{T}}(\Phi^{\perp}_{\mathrm{Pa}_j(G')} - \Phi^{\perp}_{\mathrm{Pa}_j(G)})X_j}{\sum_{i=1}^{p} \mathrm{RSS}_i(G)}\right) \\
&\leq p^{-c_0} \exp\left\{\frac{(\alpha np+\kappa)\omega_j^* \rho \log p}{(\min_i \omega_i^*) np}\right\} \\
&\leq p^{\{\max_{i\neq j}(\omega_j^*/\omega_i^*)\}(\alpha+1)\rho - c_0} < 1.
\end{aligned}$$

In the last line, we have used $\kappa \leq np$ and $c_0 > \max_{i\neq j}(\omega_j^*/\omega_i^*)(\alpha+1)\rho$ from condition (C2). The same argument implies that if we define G_0 such that $\mathrm{Pa}_j(G_0) = \mathrm{Pa}_j(G^*)$ and $\mathrm{Pa}_i(G_0) = \mathrm{Pa}_i(G)$ for $i \neq j$, then we have

$$\frac{\exp(\phi(G))}{\exp(\phi(G_0))} < p^{(|\mathrm{Pa}_j(G)| - |\mathrm{Pa}_j(G^*)|)\{\max_{i\neq j}(\omega_j^*/\omega_i^*)(\alpha+1)\rho - c_0\}} < 1.$$

Event (ii). Fix an arbitrary $G \in \mathcal{G}_p^*(d_{\mathrm{in}})$ such that $\mathrm{Pa}_j(G^*) \not\subseteq \mathrm{Pa}_j(G)$ for some $j \in [p]$. Since there exists some $\sigma \in [\sigma^*]$ such that $G, G^* \in \mathcal{G}_p^{\sigma}(d_{\mathrm{in}})$, we can apply Lemma 5 to show that there exists some $k \in \mathrm{Pa}_j(G^*) \setminus \mathrm{Pa}_j(G)$ such that the DAG $G' = G \cup \{k \to j\}$ satisfies $X_j^{\mathrm{T}}(\Phi_{\mathrm{Pa}_j(G')} - \Phi_{\mathrm{Pa}_j(G)})X_j \geq 9c_0\overline{\nu}\log p/\alpha$. Further, on the event $\mathcal{A}$, we have $\mathrm{RSS}_i(G) \leq X_i^{\mathrm{T}} X_i \leq n\overline{\nu}$. Now using $\sqrt{1+\alpha/\gamma} \leq p$, which follows from condition (C2), we find that

$$\begin{aligned}
\frac{\exp(\phi(G))}{\exp(\phi(G'))} &= \left(p^{c_0}\sqrt{1+\alpha/\gamma}\right) \left(\frac{\sum_{i\neq j}^{p} \mathrm{RSS}_i(G) + \mathrm{RSS}_j(G')}{\sum_{i=1}^{p} \mathrm{RSS}_i(G)}\right)^{\frac{\alpha p n+\kappa}{2}} \\
&\leq p^{(c_0+1)} \left(1 - \frac{X_j^{\mathrm{T}}(\Phi^{\perp}_{\mathrm{Pa}_j(G)} - \Phi^{\perp}_{\mathrm{Pa}_j(G')})X_j}{\sum_{i=1}^{p} \mathrm{RSS}_i(G)}\right)^{\frac{\alpha p n+\kappa}{2}} \\
&\leq p^{(c_0+1)} \exp\left(-\frac{\alpha np+\kappa}{2} \frac{X_j^{\mathrm{T}}(\Phi_{\mathrm{Pa}_j(G')} - \Phi_{\mathrm{Pa}_j(G)})X_j}{\sum_{i=1}^{p} \mathrm{RSS}_i(G)}\right) \\
&\leq p^{(c_0+1)} \exp\left\{-\frac{\alpha np+\kappa}{2} \frac{9c_0\overline{\nu}\log p/\alpha}{np\overline{\nu}}\right\} \leq p^{(-7c_0/2+1)}.
\end{aligned}$$

This implies $\exp(\phi(G)) < \exp(\phi(G'))$ since $c_0 > 4d_{\mathrm{in}} + 6 > 2/7$. The same argument shows that if we define $G_1 \in \mathcal{G}_p^{\sigma}$ such that $\mathrm{Pa}_j(G_1) = \mathrm{Pa}_j(G^*) \cup \mathrm{Pa}_j(G)$ and $\mathrm{Pa}_i(G_1) = \mathrm{Pa}_i(G)$ for $i \neq j$, then we have

$$\frac{\exp(\phi(G))}{\exp(\phi(G_1))} \leq p^{|\mathrm{Pa}_j(G^*)\setminus\mathrm{Pa}_j(G)|(-7c_0/2+1)}. \tag{22}$$

Event (iii). Consider an arbitrary $G \in \mathcal{G}_p^*(d_{\mathrm{in}})$ such that $G \neq G^*$. Then, there exists some $j \in [p]$ such that $\mathrm{Pa}_j(G) \neq \mathrm{Pa}_j(G^*)$. If the node j is overfitted (i.e., $\mathrm{Pa}_j(G^*) \subset \mathrm{Pa}_j(G)$), event (i) shows that there exists some $G_0 \in \mathcal{G}_p^*(d_{\mathrm{in}})$ such that $\phi(G_0) > \phi(G)$. If the node j is underfitted, i.e., $\mathrm{Pa}_j(G^*) \not\subseteq \mathrm{Pa}_j(G)$, inequality (22) shows that there exists some $G_1 \in$

$\mathcal{G}_p^*(2d_{\text{in}})$ such that $\phi(G_1) > \phi(G)$ and node j is overfitted. But event (i) again implies that there exists some $G_2 \in \mathcal{G}_p^*(d_{\text{in}})$ such that $\phi(G_2) > \phi(G_1)$. Hence, G cannot be the maximizer of ϕ in $\mathcal{G}_p^\sigma(d_{\text{in}})$; that is, G^* is the unique DAG in $\mathcal{G}_p^*(d_{\text{in}})$ that maximizes ϕ, which completes the proof. □

B.3 Proof of Theorem 1

For $\tau \notin [\sigma^*]$, the ratio of $\exp(\phi(\hat{G}_\tau))$ to $\exp(\phi(G^*))$ is

$$\frac{\exp(\phi(\hat{G}_\tau))}{\exp(\phi(G^*))} = \left(p^{c_0}\sqrt{1+\alpha/\gamma}\right)^{|G^*|-|\hat{G}_\tau|} \left(\frac{\sum_{j=1}^p \text{RSS}_j(\hat{G}_\tau)}{\sum_{j=1}^p \text{RSS}_j(G^*)}\right)^{-\frac{\alpha pn+\kappa}{2}}. \tag{23}$$

On the event $\mathcal{D} \cap \mathcal{E}$ defined in Section B.1, we have

$$\begin{aligned}
\frac{\sum_{j=1}^p \text{RSS}_j\left(\hat{G}_\tau\right)}{\sum_{j=1}^p \text{RSS}_j\left(G^*\right)} &\geq \frac{\sum_{j=1}^p X_j^{\mathrm{T}} \Phi^{\perp}_{\text{Pa}_j(\hat{G}_\tau)\cup\text{Pa}_j(G^*_\tau)} X_j}{\sum_{j=1}^p X_j^{\mathrm{T}} \Phi^{\perp}_{\text{Pa}_j(G^*)} X_j} \\
&= \frac{\sum_{j=1}^p (\epsilon_j^\tau)^{\mathrm{T}} \Phi^{\perp}_{\text{Pa}_j(\hat{G}_\tau)\cup\text{Pa}_j(G^*_\tau)} \epsilon_j^\tau}{\sum_{j=1}^p \epsilon_j^{\mathrm{T}} \Phi^{\perp}_{\text{Pa}_j(G^*)} \epsilon_j} \\
&\geq \frac{\text{tr}(\Omega^*_\tau)}{\text{tr}(\Omega^*_{\sigma^*})} \cdot \frac{(1-1/(2\eta))}{1+1/(4\eta)}
\end{aligned}$$

where the error vectors $\epsilon_j, \epsilon_j^\tau$ are as defined in (18) and (19). Without loss of generality, we can assume $\eta > 3$ in Assumption A, from which we obtain that

$$\frac{\sum_{j=1}^p \text{RSS}_j(\hat{G}_\tau)}{\sum_{j=1}^p \text{RSS}_j(G^*)} \geq \frac{(1+1/\eta)(1-1/(2\eta))}{1+1/(4\eta)} > \frac{1+1/(3\eta)}{1+1/(4\eta)} > 1 + \frac{1}{\eta'},$$

for some universal $\eta' > 0$. Hence,

$$\frac{\exp(\phi(\hat{G}_\tau))}{\exp(\phi(G^*))} \leq p^{c_0|G^*|}\left(1+\frac{1}{\eta'}\right)^{-\frac{\alpha pn+\kappa}{2}} \leq p^{c_0 p d_{\text{in}}}\left(1+\frac{1}{\eta'}\right)^{-\frac{\alpha pn+\kappa}{2}}.$$

Using $d_{\text{in}} \log p = o(n)$ and Stirling's formula, we get

$$\frac{\sum_{\tau\notin[\sigma^*]} \exp(\phi(\hat{G}_\tau))}{\exp(\phi(G^*))} \leq p! \frac{\exp(\phi(\hat{G}_\tau))}{\exp(\phi(G^*))} \leq e^{-Cnp},$$

for some universal $C > 0$. For sufficiently large n, by Assumption B and Lemma 3, the event $\mathcal{D} \cap \mathcal{E} \cap \left(\cap_{\sigma\in[\sigma^*]}\{\hat{G}_\sigma = G^*\}\right)$ happens with probability at least $1 - \zeta(p) - 2e^{-c'n}$, on which we have

$$\pi_n(G^*) = \frac{\sum_{\sigma\in[\sigma^*]} e^{\phi(G^*)}}{\sum_{\tau\in\mathbb{S}^p} e^{\phi(\hat{G}_\tau)}} \geq 1 - \frac{\sum_{\tau\notin[\sigma^*]} e^{\phi(\hat{G}_\tau)}}{\sum_{\sigma\in[\sigma^*]} e^{\phi(G^*)}} \geq 1 - e^{-Cnp}.$$

That is, $\pi_n(G^*)$ converges to 1 in probability. □

B.4 Proof for the case of sub-Gaussian errors

Let X be an $n \times p$ random matrix, each of whose rows is an i.i.d. copy of p-dimensional sub-Gaussian random vector with mean zero and covariance matrix Σ^* with a sub-Gaussian parameter bounded by a universal constant C_{sub}. We define Σ^*_S as the submatrix of Σ^* with both rows and columns indexed by the set S. Let $\Sigma^*_{j|S} = \Sigma^*_{j,j} - \Sigma^*_{j,S}(\Sigma^*_S)^{-1}\Sigma^*_{S,j}$ denote the partial covariance and let $\hat{\Sigma}_{j|S} = n^{-1}X_j\Phi^{\perp}_S X_j$ be its estimator for $|S| \le d_{\text{in}}$ and $j \notin S$. Denote $\|\cdot\|_{\text{op}}$ as the operator norm.

In the sub-Gaussian case, zero correlation does not imply independence anymore, and thus we need more stringent assumptions. The first condition is that

$$\frac{\overline{\nu}^4 d_{\text{in}} \log p}{\underline{\nu}^6 n} \to 0, \tag{24}$$

as n goes to infinity. Second, we need $\text{Pa}_j(\hat{G}_\tau) \subseteq \text{Pa}_j(G^*_\tau)$ for $\tau \notin [\sigma^*]$, which means that the stepwise selection method should estimate the minimal I-map G^*_τ sparser and should not include an edge that is not in G^*_τ. For the consistency result, the ratio $\hat{\Sigma}_{j|S}/\Sigma^*_{j|S}$ need to be controlled. To this end, we need the following lemmas.

Lemma 6. Suppose $d_{\text{in}} \log p = o(n)$. There exists a constant K_0, which only depend on C_{sub}, satisfying for sufficiently large n,

$$\max_{S \in \mathcal{M}_p(2d_{\text{in}},[p])} \left\| n^{-1}X_S^\top X_S - \Sigma^*_S \right\|_{\text{op}} \le K_0 \sqrt{\frac{d_{\text{in}} \log p}{n}},$$

with probability at least $1 - 2p^{-d_{\text{in}}}$.

Proof. See Lemma F3 in Zhou and Chang [2021]. □

Lemma 7. Suppose $d_{\text{in}} \log p = o(n)$ and a set S and j satisfy $|S| \le d_{\text{in}}$ and $j \notin S$. Let K_0 be the constant in Lemma 6. Then, for sufficiently large n, we have

$$|\hat{\Sigma}_{j|S} - \Sigma^*_{j|S}| \le K_0 \frac{\overline{\nu}^2}{\underline{\nu}^2} \sqrt{\frac{d_{\text{in}} \log p}{n}},$$

with probability at least $1 - 2p^{-d_{\text{in}}}$.

Proof. Apply the proof of Lemma E4 of Zhou and Chang [2021] by setting $T = \{j\}$, where T is a set defined in Lemma E4 of Zhou and Chang [2021]. □

Now, we are ready to prove the sub-Gaussian case. It is sufficient to show

$$\frac{\sum_{j=1}^p \text{RSS}_j(\hat{G}_\tau)}{\sum_{j=1}^p \text{RSS}_j(G^*)} > 1 + \frac{1}{\eta'}.$$

For fixed $\eta > 0$, by the condition (24), a sufficiently large n satisfies $K_0(\overline{\nu}^2/\underline{\nu}^2)\sqrt{d_{\text{in}} \log p/n}$ $< \underline{\nu}/(4\eta)$. It follows that

$$\begin{aligned} \hat{\Sigma}_{j|S} &> \Sigma^*_{j|S} - K_0 \frac{\overline{\nu}^2}{\underline{\nu}^2} \sqrt{\frac{d_{\text{in}} \log p}{n}} \\ &> \Sigma^*_{j|S} - \frac{\underline{\nu}}{2\eta}, \end{aligned}$$

which implies that $\hat{\Sigma}_{j|S}/\Sigma^*_{j|S} > 1-(2\eta)^{-1}$ by the fact $\underline{\nu} \le \Sigma^*_{j|S}$. The other direction can be obtained by

$$\begin{aligned}\hat{\Sigma}_{j|S} &< \Sigma^*_{j|S} + K_0 \frac{\overline{\nu}^2}{\underline{\nu}^2}\sqrt{\frac{d_{\text{in}}\log p}{n}} \\ &< \Sigma^*_{j|S} + \frac{\underline{\nu}}{4\eta},\end{aligned}$$

which yields $\hat{\Sigma}_{j|S}/\Sigma^*_{j|S} < 1+(4\eta)^{-1}$. Therefore,

$$\begin{aligned}\frac{\sum_{j=1}^p \text{RSS}_j\left(\hat{G}_\tau\right)}{\sum_{j=1}^p \text{RSS}_j\left(G^*\right)} &\ge \frac{\sum_{j=1}^p X_j^{\mathrm{T}}\Phi^{\perp}_{\text{Pa}_j(G^*_\tau)}X_j}{\sum_{j=1}^p X_j^{\mathrm{T}}\Phi^{\perp}_{\text{Pa}_j(G^*)}X_j} \\ &= \frac{\sum_{j=1}^p \hat{\Sigma}_{j|\text{Pa}_j(G^*_\tau)}}{\sum_{j=1}^p \hat{\Sigma}_{j|\text{Pa}_j(G^*)}} \\ &\ge \frac{\text{tr}(\Omega^*_\tau)}{\text{tr}(\Omega^*_{\sigma^*})}\cdot\frac{(1-1/(2\eta))}{1+1/(4\eta)} \\ &\ge \frac{(1+1/\eta)(1-1/(2\eta))}{1+1/(4\eta)} > 1+\frac{1}{\eta'},\end{aligned}$$

for some universal constant $\eta' > 0$. The rest of the proof is identical to the Gaussian case. □

B.5 Proof of Proposition 2

By $(C1')$, we have $\omega^*_1 = \cdots = \omega^*_p = \omega^*$ in (18) for the true data generating model. Without loss of generality, assume that $\text{id} = (1,\ldots,p)$ is a true ordering. Define

$$\theta = d_{\text{in}}^2\frac{\overline{\nu}^2\log p}{\underline{\nu}^3 n}.$$

Lemma 8. Under the setting of Proposition 2,

$$\Sigma^*_{ii} = \omega^* + O(\theta/d_{\text{in}}), \qquad \Sigma^*_{ij} = O(\sqrt{\theta}/d_{\text{in}}),$$

for all $i,j \in [p]$ and $i \ne j$.

Proof. For ease of notation, in this proof we write $B = B^*$, and without loss of generality, we assume the true error variance ω^* equals 1. Since B is a strictly upper triangular matrix, its operator norm is zero and $B^p = 0$. So we can expand Σ using the Neumann series by

$$\begin{aligned}\Sigma = (I-B^{\mathrm{T}})^{-1}(I-B)^{-1} &= \sum_{k=0}^{\infty}(B^{\mathrm{T}})^k\sum_{k=0}^{\infty}B^k \\ &= \sum_{k=0}^{\infty}\sum_{r+s=k}(B^{\mathrm{T}})^r B^s = \sum_{k=0}^{2p-2}\sum_{\substack{r+s=k\\ r,s<p}}(B^{\mathrm{T}})^r B^s.\end{aligned}$$

We can calculate B^s and $(B^{\mathrm{T}})^r$ by treating B^* and $(B^*)^{\mathrm{T}}$ as weighted transition matrices for a random walk on the DAG with weighted adjacency matrix B. Explicitly, define the set of all paths from node i to node j with s steps by

$$\text{PATH}^s_{ij} = \{q = (q_0, q_1, \ldots, q_s)\colon B_{q_k q_{k+1}} \ne 0, \text{ for } k = 0,\ldots,s-1, q_0 = i, q_s = j\},$$

and the weight W_q of an s-length path $q = (q_0, \ldots, q_s)$ by $W_q = \prod_{k=1}^{s} B_{q_{k-1}q_k}$. We have $|\mathrm{W}_q| = O(\theta^{s/2}/d_{\mathrm{in}}^s)$, since $|B_{ij}| = O(\sqrt{\theta}/d_{\mathrm{in}})$ for any i, j by the condition (C1'). It follows that the (i,j)-th entry of $(B^{\mathrm{T}})^r B^s$ is given by

$$\begin{aligned}\left((B^{\mathrm{T}})^r B^s\right)_{ij} &= \sum_{k\in[p]} (B^{\mathrm{T}})^r_{ik} B^s_{kj} = \sum_{k\in[p]} \left(\sum_{q\in \mathrm{PATH}^s_{kj}} W_q\right)\left(\sum_{q\in \mathrm{PATH}^r_{ki}} W_q\right) \\ &= \sum_{k\in[p]} \sum_{q\in\mathrm{PATH}^s_{kj}, q'\in\mathrm{PATH}^r_{ki}} W_{q'} W_q = N^{r,s}(i,j) O(\theta^{(r+s)/2}/d_{\mathrm{in}}^{r+s}),\end{aligned}$$

where $N^{r,s}(i,j)$ denotes the number of possible "paths" that start from node i, move backwards for r steps, move forwards for s steps and arrive at node j; such paths are called treks [Uhler et al., 2013, Sullivant et al., 2010] and we denote them by $q = (q'_0, q'_1, \ldots, q'_{r-1}, q'_r = q_s, q_{s-1}, \ldots, q_1, q_0)$, where $q'_0 = i$, $q_0 = j$. Since d is the maximum number of parent nodes, given i, j, there are at most d_{in} different choices for q'_1 and q_1. Similarly, given q'_1 and q_1, there are at most d_{in} choices for q'_2 and q_2. Repeating this argument yields that $N^{r,s}(i,j) \le d_{\mathrm{in}}^{r+s-1}$, and it follows that $\left((B^{\mathrm{T}})^r B^s\right)_{ii} = O(\theta^{(r+s)/2}/d_{\mathrm{in}})$. Therefore, for sufficiently large n,

$$\begin{aligned}\Sigma_{ii} &= \sum_{k=0}^{2p-2} \sum_{\substack{r+s=k \\ r,s<p}} ((B^{\mathrm{T}})^r B^s)_{ii} \\ &= 1 + \sum_{k=2}^{p} \sum_{1\le r\le k-1} ((B^{\mathrm{T}})^r B^{k-r})_{ii} + \sum_{k=p+1}^{2p-2} \sum_{k-p+1\le r\le p-1} ((B^{\mathrm{T}})^r B^{k-r})_{ii} \\ &= 1 + \sum_{k=2}^{p} d_{\mathrm{in}}^{-1}(k-1) O(\theta^{k/2}) + \sum_{k=p+1}^{2p-2} d_{\mathrm{in}}^{-1}(2p-1-k) O(\theta^{k/2}) \\ &= 1 + \sum_{k=2}^{\infty} d_{\mathrm{in}}^{-1} O(2^{k-2}\theta^{k/2}) = 1 + O(\theta/d_{\mathrm{in}}).\end{aligned}$$

Similarly, for any $i < j$,

$$\begin{aligned}\Sigma_{ij} &= \sum_{k=0}^{2p-2} \sum_{\substack{r+s=k \\ r,s<p}} ((B^{\mathrm{T}})^r B^s)_{ij} \\ &= B_{ij} + \sum_{k=2}^{p} d_{\mathrm{in}}^{-1}(k-1) O(\theta^{k/2}) + \sum_{k=p+1}^{2p-2} d_{\mathrm{in}}^{-1}(2p-1-k) O(\theta^{k/2}),\end{aligned}$$

from which we obtain that $\Sigma_{ij} = O(\sqrt{\theta}/d_{\mathrm{in}}) + O(\theta/d_{\mathrm{in}}) = O(\sqrt{\theta}/d_{\mathrm{in}})$. □

Proof of Proposition 2. Define $\mathcal{G}_p(d_{\mathrm{in}}) = \cup_{\sigma\in\mathbb{S}^p} \mathcal{G}_p^\sigma(d_{\mathrm{in}})$. Let $G_1, G_2 \in \mathcal{G}_p(d_{\mathrm{in}})$ be such that $\{i \to j\} \in G_1$ and G_2 can be obtained from G_1 by reversing $i \to j$. Let $S = \mathrm{Pa}_i(G_1)$ and $T = \mathrm{Pa}_j(G_2)$; see Fig. 5. The sets S and T may not be disjoint.

Assume we are on the event $\mathcal{B}' \cap \mathcal{J}$ defined in Section B.1. Since G_1, G_2 have the same

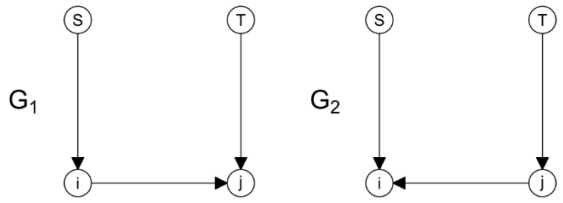


Figure 5: Local structure of G_1, G_2 in the proof of Proposition 2.

number of edges, the posterior ratio of G_1 to G_2 is

$$\begin{aligned}
\frac{\exp(\phi(G_1))}{\exp(\phi(G_2))} &= \left(\frac{\sum_{k=1}^p X_k^{\mathrm{T}}\Phi^{\perp}_{\mathrm{Pa}_k(G_2)}X_k}{\sum_{k=1}^p X_k^{\mathrm{T}}\Phi^{\perp}_{\mathrm{Pa}_k(G_1)}X_k}\right)^{\frac{\alpha pn+\kappa}{2}} \\
&= \left(1+\frac{X_j^{\mathrm{T}}(\Phi_{T\cup\{i\}}-\Phi_T)X_j - X_i^{\mathrm{T}}(\Phi_{S\cup\{j\}}-\Phi_S)X_i}{\sum_{k=1}^p X_k^{\mathrm{T}}\Phi^{\perp}_{\mathrm{Pa}_k(G_1)}X_k}\right)^{\frac{\alpha pn+\kappa}{2}} \\
&\le \exp\left(\frac{\alpha pn+\kappa}{2}\frac{X_j^{\mathrm{T}}(\Phi_{T\cup\{i\}}-\Phi_T)X_j - X_i^{\mathrm{T}}(\Phi_{S\cup\{j\}}-\Phi_S)X_i}{np\underline{\nu}/2}\right) \\
&\le \exp\left\{\frac{\alpha+1}{\underline{\nu}}[X_j^{\mathrm{T}}(\Phi_{T\cup\{i\}}-\Phi_T)X_j - X_i^{\mathrm{T}}(\Phi_{S\cup\{j\}}-\Phi_S)X_i]\right\},
\end{aligned}$$

where the first inequality follows from the inequality $1+x \le \exp(x)$ for all $x \in \mathbb{R}$ and the second follows from the observation that $X_k^{\mathrm{T}}\Phi^{\perp}_{\mathrm{Pa}_k(G_1)}X_k \ge n\underline{\nu}/2$ for any $k \in [p]$ on the event $\mathcal{B}'$. To conclude the proof, we need to show

$$X_j^{\mathrm{T}}(\Phi_{T\cup\{i\}}-\Phi_T)X_j - X_i^{\mathrm{T}}(\Phi_{S\cup\{j\}}-\Phi_S)X_i = o((\overline{\nu}^2/\underline{\nu}^2)\log p). \tag{25}$$

By Lemma 8 and condition (C2'), on the event $\mathcal{J}$, we have

$$\begin{aligned}
\frac{X_i^{\mathrm{T}}X_i}{n} &= \Sigma_{ii} + O(\underline{\nu}\sqrt{\theta}/d_{\mathrm{in}}) = \omega^* + O(\theta/d_{\mathrm{in}}) + O(\underline{\nu}\sqrt{\theta}/d_{\mathrm{in}}) = \omega^* + o(1), \\
\frac{X_i^{\mathrm{T}}X_j}{n} &= \Sigma_{ij} + O(\underline{\nu}\sqrt{\theta}/d_{\mathrm{in}}) = O(\sqrt{\theta}/d_{\mathrm{in}}) = o(1).
\end{aligned}$$

Hence, by Neumann series, for any $S \subseteq [p]$ such that $|S| \le d_{\mathrm{in}}$, we have $(n^{-1}X_S^{\mathrm{T}}X_S)^{-1} = (\omega^*)^{-1}I + R_S$ where R_S is a matrix with all entries being $O(\sqrt{\theta}/d_{\mathrm{in}})$. This yields, for all $i, j \in [p] \setminus S$,

$$\begin{aligned}
\frac{X_i^{\mathrm{T}}\Phi_S X_j}{n} &= \frac{X_i^{\mathrm{T}}X_S}{n}\left(\frac{X_S^{\mathrm{T}}X_S}{n}\right)^{-1}\frac{X_S^{\mathrm{T}}X_j}{n} \\
&= \begin{bmatrix} O(\sqrt{\theta}/d_{\mathrm{in}}) & \cdots & O(\sqrt{\theta}/d_{\mathrm{in}}) \end{bmatrix}((\omega^*)^{-1}I + R_S)\begin{bmatrix} O(\sqrt{\theta}/d_{\mathrm{in}}) \\ \vdots \\ O(\sqrt{\theta}/d_{\mathrm{in}})) \end{bmatrix} \\
&= d_{\mathrm{in}}O(\theta/d_{\mathrm{in}}^2) + d_{\mathrm{in}}^2 O(\theta^{3/2}/d_{\mathrm{in}}^3) = O(\theta/d_{\mathrm{in}}) = o(1).
\end{aligned}$$

It follows that

$$X_j^{\mathrm{T}}(\Phi_{T\cup\{i\}}-\Phi_T)X_j - X_i^{\mathrm{T}}(\Phi_{S\cup\{j\}}-\Phi_S)X_i = \frac{(X_j^{\mathrm{T}}\Phi_T^{\perp}X_i)^2}{X_i^{\mathrm{T}}\Phi_T^{\perp}X_i} - \frac{(X_j^{\mathrm{T}}\Phi_S^{\perp}X_i)^2}{X_j^{\mathrm{T}}\Phi_S^{\perp}X_j}$$

$$= n\frac{\left[\frac{X_j^{\mathrm{T}}X_i}{n} - \frac{X_j^{\mathrm{T}}\Phi_T X_i}{n}\right]^2}{\frac{X_i^{\mathrm{T}}X_i}{n} - \frac{X_i^{\mathrm{T}}\Phi_T X_i}{n}} - n\frac{\left[\frac{X_j^{\mathrm{T}}X_i}{n} - \frac{X_j^{\mathrm{T}}\Phi_S X_i}{n}\right]^2}{\frac{X_j^{\mathrm{T}}X_j}{n} - \frac{X_j^{\mathrm{T}}\Phi_S X_j}{n}}$$

$$= n(\omega^*)^{-1}\left\{(1+o(1))\left[\frac{X_j^{\mathrm{T}}X_i}{n} - \frac{X_j^{\mathrm{T}}\Phi_T X_i}{n}\right]^2 - (1+o(1))\left[\frac{X_j^{\mathrm{T}}X_i}{n} - \frac{X_j^{\mathrm{T}}\Phi_S X_i}{n}\right]^2\right\}$$

$$= n(\omega^*)^{-1}\left\{-\frac{2X_j^{\mathrm{T}}X_i}{n}\left[\frac{X_j^{\mathrm{T}}\Phi_T X_i}{n} - \frac{X_j^{\mathrm{T}}\Phi_S X_i}{n}\right] + \left(\frac{X_j^{\mathrm{T}}\Phi_T X_i}{n}\right)^2 - \left(\frac{X_j^{\mathrm{T}}\Phi_S X_i}{n}\right)^2 + o(\theta/d_{\mathrm{in}}^2)\right\}$$

$$= n\left\{O(\sqrt{\theta}/d_{\mathrm{in}})O(\theta/d_{\mathrm{in}}) + O(\theta^2/d_{\mathrm{in}}^2) + o(\theta/d_{\mathrm{in}}^2)\right\} = no(\theta/d_{\mathrm{in}}^2) = o((\overline{\nu}^2/\underline{\nu}^2)\log p),$$

which completes the proof of (25). □

B.6 Proof of Theorem 2

Let $\delta = \underline{\nu}^2 C_{\min}(d_{\mathrm{in}}+1)^{-1}(\underline{\nu}C_{\min} + 3\omega^*(1+C_{\min}))^{-1}$ and $\hat{\Sigma}_{ij} = X_i^{\mathrm{T}}X_j/n$ for each (i,j). Define $\mathcal{K} = \left\{\max_{i,j\in[p]}|\hat{\Sigma}_{ij} - \Sigma^*_{ij}| \le \delta\right\}$. For any $\epsilon > 0$, using Lemma 1 of Ravikumar et al. [2011] and our Lemma 2, we can show that $\mathbb{P}^*(\mathcal{A}\cap\mathcal{B}\cap\mathcal{C}\cap\mathcal{K}) \ge 1-\epsilon$ and

$$\mathbb{P}^*\{|\hat{\Sigma}_{ij} - \Sigma^*_{ij}| > \delta\} \le 4\exp\left\{-\frac{n\delta^2}{3200\max_k(\Sigma^*_{ij})^2}\right\} \le \frac{\epsilon}{p(p+1)}.$$

Further, from the proof of Proposition 1, we know that on the event $\mathcal{A}\cap\mathcal{B}\cap\mathcal{C}$, we have

$$\arg\max_{S\subset P_j:\,|S|\le d_{\mathrm{in}}} \phi_j\left(S, \sum_{i\neq j}\mathrm{RSS}_i(G)\right) = \mathrm{Pa}_j(G^*),$$

for any $j\in[p]$, $P_j \supseteq \mathrm{Pa}_j(G^*)$, and $G\in\mathcal{G}_p^*(2d_{\mathrm{in}})$. Observe that Theorem 2 holds if we can show that for any $G\in\mathcal{G}_p^*(d_{\mathrm{in}})$, Algorithm 1 with input $\mathrm{RSS} = (\mathrm{RSS}_1(G),\ldots,\mathrm{RSS}_p(G))$ returns some $\sigma\in[\sigma^*]$, but this follows by an argument completely analogous to the proof of Theorem 2 of Chen et al. [2019]. □

B.7 Derivation of the posterior distribution

Let $L(B,\omega)$ be the likelihood function in (2). The α-fractional posterior distribution of B,ω, given the prior distributions in (3) and (4), is

$$\pi_n(B,\omega \mid G,\sigma) \propto \pi_0(B,\omega\mid G,\sigma)L(B,\omega)^\alpha$$
$$= \frac{\pi_0(B,\omega\mid G,\sigma)}{L(B,\omega)^{1-\alpha}}L(B,\omega),$$

where the first term in the last equation can be regarded as the effective prior distribution for $(B,\omega)\mid(G,\sigma)$. By the normal-inverse-gamma conjugacy, the α-fractional marginal likelihood

of (G, σ) is given by

$$
\begin{aligned}
f_\alpha(G,\sigma) &\propto \int \pi_0(B,\omega \mid G,\sigma) L(B,\omega)^\alpha d(B,\omega) \\
&= \int \pi_0(B \mid \omega, G, \sigma)\pi_0(\omega \mid G,\sigma) L(B,\omega)^\alpha d(B,\omega) \\
&\propto \int \left(\frac{\omega}{\gamma}\right)^{-|G|/2} \prod_{j=1}^{p} \det\left(X_{\mathrm{Pa}_j}^{\mathrm{T}} X_{\mathrm{Pa}_j}\right)^{1/2} \exp\left\{-\frac{\gamma}{2\omega}\sum_{j=1}^{p}(B_{\mathrm{Pa}_j,j} - \hat{B}_{\mathrm{Pa}_j,j})^{\mathrm{T}}(X_{\mathrm{Pa}_j}^{\mathrm{T}} X_{\mathrm{Pa}_j})(B_{\mathrm{Pa}_j,j} - \hat{B}_{\mathrm{Pa}_j,j})\right\} \times \\
&(\omega^{-\frac{\kappa}{2}-1}) \left[\omega^{-\frac{\alpha n p}{2}} \exp\left\{-\frac{\alpha}{2\omega}\sum_{j=1}^{p}(X_j - B_{\mathrm{Pa}_j,j}^{\mathrm{T}} X_{\mathrm{Pa}_j})^{\mathrm{T}}(X_j - B_{\mathrm{Pa}_j,j}^{\mathrm{T}} X_{\mathrm{Pa}_j})\right\}\right] d(B,\omega) \\
&\propto \int \left(\frac{\omega}{\gamma}\right)^{-|G|/2} \omega^{-\frac{\alpha n p+\kappa}{2}-1} \exp\left\{-\frac{\alpha}{2\omega}\sum_{j=1}^{p} X_j^{\mathrm{T}} \Phi_{\mathrm{Pa}_j}^{\perp} X_j\right\} \left(\frac{\alpha+\gamma}{\omega}\right)^{-|G|/2} \times \\
&\int \left(\frac{\omega}{\alpha+\gamma}\right)^{-|G|/2} \prod_{j=1}^{p} \det\left(X_{\mathrm{Pa}_j}^{\mathrm{T}} X_{\mathrm{Pa}_j}\right)^{1/2} \times \\
&\exp\left\{-\frac{\alpha+\gamma}{2\omega}\sum_{j=1}^{p}(B_{\mathrm{Pa}_j,j} - \hat{B}_{\mathrm{Pa}_j,j})^{\mathrm{T}}(X_{\mathrm{Pa}_j}^{\mathrm{T}} X_{\mathrm{Pa}_j})(B_{\mathrm{Pa}_j,j} - \hat{B}_{\mathrm{Pa}_j,j})\right\} dB d\omega \\
&= \left(1+\frac{\alpha}{\gamma}\right)^{-|G|/2} \int \omega^{-\frac{\alpha n p+\kappa}{2}-1} \exp\left\{-\frac{\alpha}{2\omega}\sum_{j=1}^{p} X_j^{\mathrm{T}} \Phi_{\mathrm{Pa}_j}^{\perp} X_j\right\} d\omega \\
&\propto \left(1+\frac{\alpha}{\gamma}\right)^{-|G|/2} \left(\sum_{j=1}^{p} \mathrm{RSS}_j(G)\right)^{-\frac{\alpha n p+\kappa}{2}}.
\end{aligned}
$$

Given the prior distribution (5), we obtain the posterior distribution of (G, σ) as

$$
\begin{aligned}
\pi_n(G,\sigma) &\propto f_\alpha(G,\sigma)\pi_0(G,\sigma) \\
&= \left(1+\frac{\alpha}{\gamma}\right)^{-|G|/2} \cdot \left(\sum_{j=1}^{p} \mathrm{RSS}_j(G)\right)^{-\frac{\alpha n p+\kappa}{2}} \cdot p^{-c_0 \log p} \cdot \mathbb{1}_{\{\hat{G}_\sigma\}}(G) \\
&= e^{\phi(G)} \mathbb{1}_{\{\hat{G}_\sigma\}}(G).
\end{aligned}
$$

C Simulation results

C.1 Mixing behavior

In Fig. 6 we examine the mixing behavior of the three types of proposals for a moderately small sample size. We repeat the simulation studies shown in panels (a), (b), and (c) of Fig. 1 in Section 4.1 by choosing $n = 100$ and keeping all the other simulation settings unchanged. We confirm that all 90 trajectories have reached the red line, which appears to be the global mode. Figure 7 shows the mixing behavior of our method and the minimal I-MAP MCMC for the heterogeneous case where, for each $j \in [p]$, we sample error variance ω_j for node j uniformly from $[0.5, 1.5]$. We still observe that some trajectories of the minimal

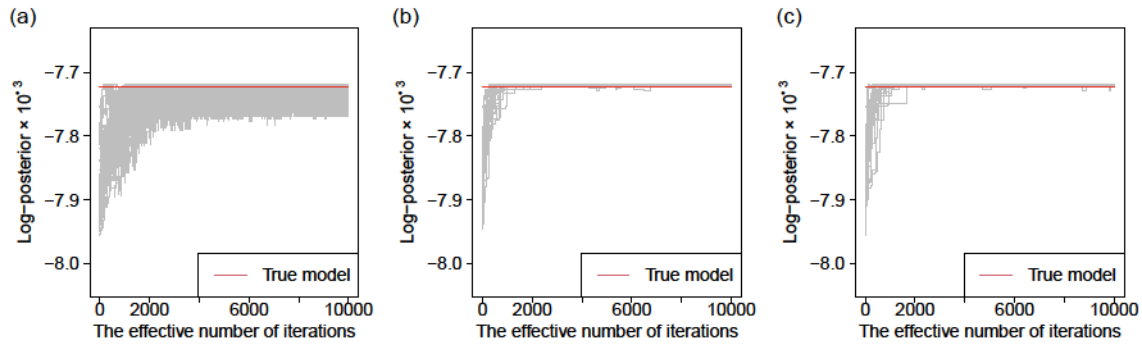


Figure 6: Log posterior probability times 10^{-3} versus the effective number of iterations of 30 MCMC runs for $p = 20$ and $n = 100$. The red line represents the true ordering σ^*.

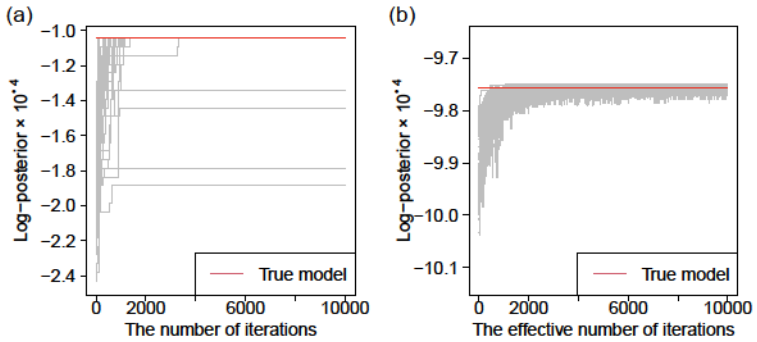


Figure 7: Log posterior probability $\times 10^{-4}$ versus the effective number of iterations of 30 MCMC runs with random initialization for the heterogeneous case with $p = 20$ and $n = 1000$: (a) minimal I-MAP MCMC, (b) the proposed method. The red line represents the true ordering σ^*.

I-MAP MCMC get stuck at local modes, while the mixing behavior of the proposed method is consistently good despite of the model misspecification.

C.2 Performance evaluation

We consider more scenarios for the simulation study described in Section 4.2. We always fix $p = 40$. In Table 4, we still generate X under the equal variance assumption but we sample each B^*_{ij} for each edge $i \to j$ in the DAG G^* from the standard Gaussian distribution. The advantage of the proposed method is as significant as in Table 1 presented in the main text. In Table 5, we sample the error variance ω_j for each j uniformly from $[0.7, 1.3]$ and sample each B^*_{ij} from the uniform distribution on $[-1, -0.3] \cup [0.3, 1]$. Comparing Table 5 with the left column of Table 1, we see that the advantage of our method over the competing ones becomes more substantial.

	Signal		N(0, 1)	
Method	n	100	500	1000
Proposed	HD	**10.4±0.8**	**5.2±0.5**	**4.2±0.4**
	FNR	**34.2±1.7**	17.0±1.7	13.8±1.2
	FDR	**2.3±0.5**	**1.7±0.5**	**1.6±0.4**
	Flip	0.8±0.3	**1.2±0.3**	**1.0±0.3**
	Time	12.8±0.2	13.2±0.2	13.2±0.2
TD	HD	12.0±0.8	6.3±0.6	6.4±0.6
	FNR	39.3±1.8	18.1±1.5	15.5±1.1
	FDR	3.7±0.8	4.8±1.2	6.8±1.2
	Filp	1.3±0.4	2.4±0.6	3.1±0.6
	Time	0.6±0.0	0.5±0.0	0.5±0.0
LISTEN	HD	12.5±0.8	6.5±0.6	5.9±0.6
	FNR	39.3±1.8	18.8±1.5	15.3±1.2
	FDR	6.6±1.1	4.8±1.1	5.8±1.1
	Flip	2.0±0.4	2.6±0.6	2.8±0.5
	Time	0.5±0.0	0.5±0.0	0.5±0.0

Table 4: Standard Gaussian signal case with $p = 40$. Each entry gives mean ± 1 standard error. The best performance with a margin of more than one se is highlighted in boldface. Time is measured in seconds.

We also conduct simulation studies on the proposed algorithm with weakly increasing error variances. We fix $n = 1,000$ and $p = 40$, and sample the error variance $\omega_j \sim \text{Uniform}([1-b, 1+b])$ for 6 different heterogeneity levels b. We set $\sigma^* = (1, \ldots, p)$ to be the true ordering and sort the error variances in ascending order to make them weakly increasing in σ^*. We generate G^* by adding $i \to j$ for $i < j$ with probability $p_{\text{edge}} = 3/(2p-2)$ and draw the edge weight B^*_{ij} independently from some distribution F. In Table 6, we present the results with 4 metrics: Hamming distance (HD), the false negative rate (FNR), false discover rate (FDR), and the percentage of flipped edges (Flip). The rows of Uniform and Gaussian indicate the result for F being $\text{Uniform}([-1, -0.3] \cup [0.3, 1])$ and that for F being the standard normal distribution, respectively. Notably, the Flip rate is always very low, which indicates that the algorithm can accurately identify the true ordering. When $b = 0.9$, FNR tends to be significantly larger. This is because some nodes may have very large error variances when $b = 0.9$, and thus the signal-to-noise ratio is low, making it challenging for the algorithm to detect edges.

C.3 Single-cell real data analysis

Figure 8 shows the result of the minimal I-MAP MCMC (with decomposable score) for the real data analysis. See Section 5 in the main text for details.

	Signal	Heterogeneity		
Method	n	100	500	1000
Proposed	HD	**10.3±0.6**	**3.2±0.5**	**4.4±0.8**
	FNR	**33.1±1.6**	**6.0±1.0**	**6.0±0.8**
	FDR	**4.4±0.7**	**6.1±1.2**	**8.9±1.5**
	Flip	**2.8±0.5**	**5.4±1.0**	**6.0±0.8**
	Time	12.0±0.2	11.6±0.2	12.3±0.2
TD	HD	15.8±1.0	6.8±0.8	8.0±1.2
	FNR	45.5±2.0	10.0±1.1	9.1±1.2
	FDR	14.8±1.6	13.4±1.6	16.3±2.3
	Filp	7.5±0.9	9.2±1.1	9.0±1.2
	Time	0.5±0.0	0.5±0.0	0.5±0.0
LISTEN	HD	16.0±1.0	8.4±1.0	8.9±1.2
	FNR	46.2±1.9	11.3±1.0	10.0±1.1
	FDR	15.2±1.8	16.4±1.8	17.9±2.3
	Flip	7.1±0.8	10.5±1.0	9.7±1.1
	Time	0.5±0.0	0.6±0.0	0.5±0.0

Table 5: Heterogeneous error variance case with $p = 40$. Each entry gives mean ± 1 standard error. The best performance with a margin of more than one se is highlighted in boldface. Time is measured in seconds.

Signal		$b = 0$	$b = 0.1$	$b = 0.3$	$b = 0.5$	$b = 0.7$	$b = 0.9$
Uniform	HD	0.2±0.1	0.1±0.1	0.1±0.1	0.1±0.1	0.2±0.1	1.2±0.2
	FNR	0.3±0.2	0.3±0.2	0.2±0.2	0.3±0.2	0.5±0.2	4.0±0.6
	FDR	0.3±0.2	0.2±0.1	0.1±0.1	0.2±0.1	0.2±0.1	0.3±0.2
	Flip	0.3±0.2	0.2±0.1	0.1±0.1	0.2±0.1	0.2±0.1	0.2±0.1
Gaussian	HD	4.9±0.5	4.3±0.4	4.5±0.4	4.7±0.4	5.1±0.4	6.0±0.4
	FNR	15.4±1.4	15.3±1.3	14.9±1.2	15.5±1.2	17.1±1.3	20.2±1.3
	FDR	2.2±0.6	0.4±0.2	0.5±0.2	0.3±0.2	0.3±0.2	0.4±0.2
	Flip	1.4±0.4	0.3±0.1	0.4±0.2	0.2±0.1	0.2±0.1	0.2±0.1

Table 6: A table for increasing error variances with heterogeneity level $b = 0, 0.1, \ldots, 0.9$ with $p = 40$. We sample error variance from Uniform($[1 - b, 1 + b]$) and sort in ascending order. Nonzero edge weights are from Uniform($[-1, -0.3] \cup [0.3, 1]$) in Uniform case and $N(0, 1)$ in Gaussian case. Each entry gives mean ± 1 standard error.

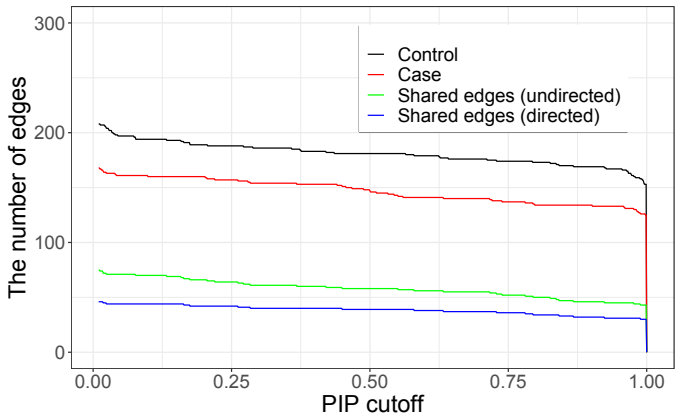


Figure 8: Result of the minimal I-MAP MCMC for the real case-control data analysis. Given an estimate $\hat{\Gamma}_{ij}$ from MCMC samples, we infer the edge $i \to j$ exists in the DAG if $\hat{\Gamma}_{ij} > c$ where c is the posterior inclusion probability cutoff. For each c, we count the number of edges occurring in the DAG for control samples (black), the number of edges in the DAG for case samples (red), the number of edges (edge direction ignored) in both DAGs (green), and the number of directed edges in both DAGs (blue).