Structure Learning on Clustered Data

Ryan Thompson RYAN.THOMPSON-1@UTS.EDU.AU
School of Mathematical and Physical Sciences
University of Technology Sydney
Ultimo, NSW 2007, Australia

Matt P. Wand MATT.WAND@UTS.EDU.AU
School of Mathematical and Physical Sciences
University of Technology Sydney
Ultimo, NSW 2007, Australia

Veerabhadran Baladandayuthapani VEERAB@UMICH.EDU
Department of Biostatistics
University of Michigan
Ann Arbor, MI 48105, USA

Abstract

Recent algorithmic advances have made directed acyclic graph (DAG) structure learning scalable for causal discovery. Yet, the currently available techniques assume a completely homogeneous population, precluding their application to clustered data where cluster-specific variations (e.g., patient-specific effects) are common. We address this issue by introducing a new approach that estimates a global structure while accounting for local cluster-level effects. The key idea is to extend the fixed- and random-effects framework of classical mixed models to the structure learning setting. Towards this end, we present a differentiable graph coupling mechanism that guarantees the union of the fixed- and random-effects graphs remains acyclic. Computationally, we provide a provably convergent first-order method and leverage efficient batched updates across clusters. Statistically, we establish identifiability of the model and show that our approach recovers the true structure asymptotically. In experiments on real and synthetic data, our proposal detects dependencies missed by alternative estimators, underscoring its value for structure learning in clustered settings.

Keywords: causal discovery, continuous acyclicity, directed acyclic graphs, heterogeneous data, mixed effects

1 Introduction

Directed acyclic graphs (DAGs)—graphs with directed edges and no cycles—are a fundamental tool for probabilistic modeling and causal inference (Spirtes et al., 2000; Pearl, 2009). Their capacity to compactly encode conditional independence relations makes them useful in a wide range of domains, e.g., psychology (Foster, 2010), economics (Imbens, 2020), and epidemiology (Tennant et al., 2021). The first formal treatments in statistics and machine learning appeared in the late 1980s (Lauritzen and Spiegelhalter, 1988; Pearl, 1988). Three decades later, Zheng et al. (2018) made a pivotal advance that transformed structure learning from a notoriously difficult combinatorial optimization problem to a tractable continuous optimization problem, opening the way to scalably learn graphs with hundreds of nodes using first-order algorithms. Surveys by Vowels et al. (2022) and Kitson et al. (2023) provide comprehensive overviews of this shift and its impact on modern structure learning.

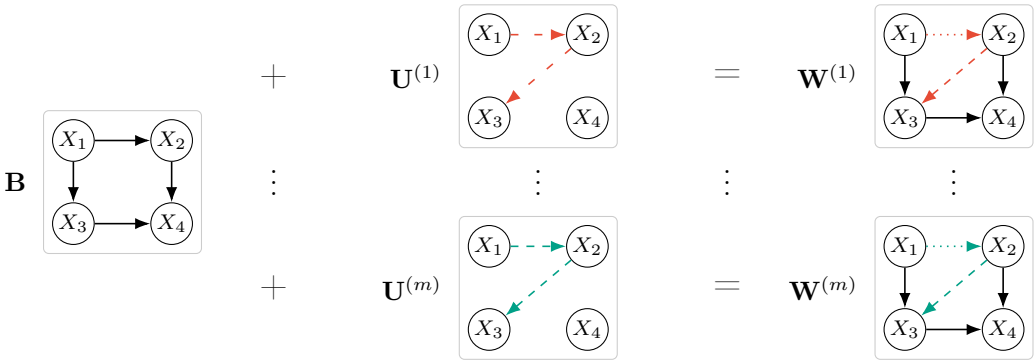


Figure 1: Schematic of mixed DAGs. The fixed-effects graph $\mathbf{B}$, shown with solid black arrows, is added to cluster-specific random-effect deviations $\mathbf{U}^{(i)}$, shown with colored dashed arrows, to form $\mathbf{W}^{(i)} = \mathbf{B} + \mathbf{U}^{(i)}$. Dotted colored arrows indicate edges with both fixed- and random-effect contributions. Different colors correspond to different clusters.

While recent advances have made structure learning scalable, most methods are confined to the iid setting, where the population is assumed to follow a single homogeneous causal model. This assumption is convenient but fails when causal effects vary within the population. Several strands of work have begun to relax homogeneity, including mixtures of DAGs (Saeed et al., 2020), DAGs adapted to distributional shifts (Huang et al., 2020), and DAGs that vary with modifying covariates (Thompson et al., 2024). However, one of the most typical sources of heterogeneity addressed by non-graphical models, clustered data, remains largely unexplored for DAGs. In clustered data, observations come with known cluster memberships, and while clusters may share a common causal structure, the strength or polarity of effects can differ across them. For example, in proteomic studies, repeated measures on the same patients form clusters. These clusters may exhibit protein interactions that broadly align across the population yet have effects that differ in detail across individuals or groups.

Drawing on the machinery of mixed-effects models (e.g., Jiang and Nguyen, 2021), we develop a scalable framework for learning DAGs on clustered data. Formally, let $\mathbf{X}^{(i)} \in \mathbb{R}^{n_i \times p}$ be n_i observations on p variables belonging to cluster $i = 1, \ldots, m$. A DAG on these variables can be represented by a *sparse* matrix $\mathbf{W}^{(i)} \in \mathbb{R}^{p \times p}$, known as a weighted adjacency matrix, with $w_{jk}^{(i)} \neq 0$ if and only if a directed edge exists from node j to node k. Hence, learning the DAG reduces to estimating an acyclic $\mathbf{W}^{(i)}$, which we decompose into population-level and cluster-specific components. This decomposition leads to a structural equation model:

$$\mathbf{X}^{(i)} = \mathbf{X}^{(i)}\mathbf{W}^{(i)} + \boldsymbol{\varepsilon}^{(i)}, \quad \mathbf{W}^{(i)} = \mathbf{B} + \mathbf{U}^{(i)}, \quad i = 1, \ldots, m, \tag{1}$$

where $\boldsymbol{\varepsilon}^{(i)} \in \mathbb{R}^{n_i \times p}$ is stochastic noise. Here, the matrix $\mathbf{B} \in \mathbb{R}^{p \times p}$ represents *fixed effects*, traditionally understood in linear mixed models as coefficients common to all clusters. Meanwhile, the matrices $\mathbf{U}^{(i)} \in \mathbb{R}^{p \times p}$ contain *random effects*, which capture cluster-specific deviations. In analogy with linear mixed models, the random effects associated with node k may be taken to follow $\mathbf{u}_k^{(i)} \sim \mathrm{N}(\mathbf{0}, \mathrm{diag}(\boldsymbol{\gamma}_k))$, where $\boldsymbol{\gamma}_k$ is the kth column of a matrix $\boldsymbol{\Gamma} \in \mathbb{R}_+^{p \times p}$ collecting all random-effect variances. As Figure 1 illustrates, unlike standard (purely fixed-effects) DAGs, where a single $\mathbf{W}$ applies uniformly across the population, mixed DAGs allow specific effects to vary with cluster membership under a global structure.

Under model (1), the overall DAG is determined by the union of two sparse matrices: the fixed-effects matrix $\mathbf{B}$, whose nonzero entries capture population-level effects, and the random-effects variance matrix $\mathbf{\Gamma}$, whose nonzero entries indicate which cluster-specific deviations are active across clusters. Learning the model therefore amounts to estimating $\mathbf{B}$ and $\mathbf{\Gamma}$ subject to the constraint that their union remains acyclic. We enforce this constraint through a graph coupling mechanism based on existing differentiable characterizations of acyclicity, guaranteeing that the combined fixed- and random-effects graphs are cycle-free. On the computational side, we develop a first-order optimization method with provable convergence and leverage efficient batched updates across clusters, making it feasible to learn graphs with up to hundreds of nodes. We further derive statistical results establishing identifiability of the model and correct recovery of the true structure asymptotically. Synthetic experiments and an application to proteomics confirm that our proposed approach recovers both population-level and cluster-specific effects that alternative approaches fail to detect.

The remainder of the paper is organized as follows. Section 2 situates our proposal in the context of the broader literature. Section 3 introduces our framework for learning mixed DAGs. Section 4 describes algorithms and strategies for scalable computation. Section 5 establishes an asymptotic statistical guarantee. Section 6 reports experiments on synthetic, semi-synthetic, and real data. Section 7 closes the paper with some final remarks.

2 Related Work

The breakthrough work by Zheng et al. (2018) reformulates the combinatorial acyclicity constraint as a continuous, differentiable constraint by exploiting the correspondence between cycles in a graph and traces of matrix powers. Subsequent work by Zhang et al. (2022) and Bello et al. (2022), among others, develops alternative continuous acyclicity characterizations with improved numerical behavior, including polynomial and log-determinant formulations. Zhang et al. (2025) provide a unified analytic perspective of these characterizations. Zheng et al. (2020) devise extensions to non-parametric DAGs, Thompson et al. (2025a) extend the framework to variational inference, while Deng et al. (2023b) and Deng et al. (2024) establish theoretical guarantees. While these approaches are nonconvex, Deng et al. (2023a) show how discrete combinatorics can be used to escape local minima if desired.

Though the above differentiable approaches assume iid observations, a few classical methods can handle clustered data. Bae et al. (2016) incorporate node-level random intercepts into DAGs, capable of capturing limited forms of cluster-level variation, while Li et al. (2018) allow edge strengths to consist of both fixed and random components. Both, however, rely on traditional discrete search and hence remain fundamentally combinatorial. More recent work considers DAGs that vary as a function of observed covariates rather than across clusters: Ni et al. (2019) use spline-based mapping and Thompson et al. (2024) employ a neural mapping with differentiable acyclicity. Oates et al. (2016) take a different route, jointly estimating observation-specific DAGs that are linked via a dependency network. These methods capture certain forms of heterogeneity, but none are designed specifically for clustered data.

Our paper also contributes to the growing literature on mixed-effects models adapted for machine learning. Xiong et al. (2019) and Simchoni and Rosset (2021, 2023, 2025) integrate random effects into neural networks—including feedforward, convolutional, and recurrent variants—to handle clustered data, with high-cardinality categorical features as an

important special case. Simchoni and Rosset (2024) extend this perspective to variational autoencoders, introducing random effects directly into the latent representation. Sholokhov et al. (2024) and Thompson et al. (2025b) develop algorithms for learning linear mixed models on high-dimensional data, enabling the selection of nonzero fixed and random effects across thousands of candidate predictors. These works, however, are confined to non-graphical models and so do not address multivariate relationships, let alone DAG structure learning.

3 Mixed DAGs

3.1 Objective Function

To formulate our estimator, we begin by narrowing our focus to a single node k within a single cluster i. The mixed-effects structural equation model (1) at this level takes the form

$$\mathbf{x}_k^{(i)} = \mathbf{X}^{(i)}\boldsymbol{\beta}_k + \mathbf{X}^{(i)}\mathbf{u}_k^{(i)} + \boldsymbol{\varepsilon}_k^{(i)},$$

where $\mathbf{x}_k^{(i)} \in \mathbb{R}^{n_i}$ denotes the kth column of $\mathbf{X}^{(i)}$, the vectors $\boldsymbol{\beta}_k \in \mathbb{R}^p$ and $\mathbf{u}_k^{(i)} \in \mathbb{R}^p$ represent the kth columns of the fixed-effects matrix $\mathbf{B}$ and random-effects matrix $\mathbf{U}^{(i)}$, respectively, and $\boldsymbol{\varepsilon}_k^{(i)} \in \mathbb{R}^{n_i}$ represents the kth column of the noise matrix $\boldsymbol{\varepsilon}^{(i)}$. It is common in the linear mixed model literature to assume Gaussian noise and random effects, so that $\boldsymbol{\varepsilon}_k^{(i)} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{u}_k^{(i)} \sim \mathrm{N}(\mathbf{0}, \mathrm{diag}(\boldsymbol{\gamma}_k))$, where $\boldsymbol{\gamma}_k \in \mathbb{R}_+^p$ is the kth column of $\boldsymbol{\Gamma}$. More generally, node-specific noise variances can be accommodated by taking $\boldsymbol{\varepsilon}_k^{(i)} \sim \mathrm{N}(\mathbf{0}, \sigma_k^2\mathbf{I})$ and replacing $\mathbf{I}$ below by $\sigma_k^2\mathbf{I}$, while here we set $\sigma_k^2 = 1$ to simplify notation. The diagonal covariance for the random effects is a common simplification in high-dimensional settings, where allowing unrestricted covariance is impractical. Throughout, we impose $\mathrm{diag}(\mathbf{B}) = \mathrm{diag}(\boldsymbol{\Gamma}) = \mathbf{0}$, excluding self-loops. Conditional on the parent variables of node k, the vector $\mathbf{x}_k^{(i)}$ follows

$$\mathbf{x}_k^{(i)} \sim \mathrm{N}(\mathbf{X}^{(i)}\boldsymbol{\beta}_k, \mathbf{V}^{(i)}(\boldsymbol{\gamma}_k)),$$

where the matrix $\mathbf{V}^{(i)}(\boldsymbol{\gamma}_k)$ captures correlations between observations within the ith cluster:

$$\mathbf{V}^{(i)}(\boldsymbol{\gamma}_k) := \mathbf{I} + \mathbf{X}^{(i)}\,\mathrm{diag}(\boldsymbol{\gamma}_k)\mathbf{X}^{(i)\top}.$$

We now move from this single-node description to the full model by applying the same construction across nodes $k = 1, \ldots, p$ and clusters $i = 1, \ldots, m$. On the feasible DAG set, after integrating out the random effects, the resulting nodewise conditional densities factorize over nodes and independent clusters, yielding a negative log-likelihood of the form

$$l(\mathbf{B}, \boldsymbol{\Gamma}) = \sum_{k=1}^{p}\sum_{i=1}^{m}\left\{\log\det\{\mathbf{V}^{(i)}(\boldsymbol{\gamma}_k)\} + (\mathbf{x}_k^{(i)} - \mathbf{X}^{(i)}\boldsymbol{\beta}_k)^\top\mathbf{V}^{-(i)}(\boldsymbol{\gamma}_k)(\mathbf{x}_k^{(i)} - \mathbf{X}^{(i)}\boldsymbol{\beta}_k)\right\}, \quad (2)$$

where $\mathbf{V}^{-(i)}(\boldsymbol{\gamma}_k) := \{\mathbf{V}^{(i)}(\boldsymbol{\gamma}_k)\}^{-1}$, and additive and multiplicative constants are omitted to simplify exposition. The negative log-likelihood (2) has explicit gradients with respect to the fixed-effects $\mathbf{B}$ and the random-effects variances $\boldsymbol{\Gamma}$. For the fixed-effect β_{jk}, the gradient is

$$\frac{\partial}{\partial\beta_{jk}}l(\mathbf{B}, \boldsymbol{\Gamma}) = -2\sum_{i=1}^{m}\mathbf{x}_j^{(i)\top}\mathbf{V}^{-(i)}(\boldsymbol{\gamma}_k)(\mathbf{x}_k^{(i)} - \mathbf{X}^{(i)}\boldsymbol{\beta}_k).$$

Likewise, the gradient with respect to the random-effect variance γ_{jk} is of the form

$$\frac{\partial}{\partial \gamma_{jk}} l(\mathbf{B}, \mathbf{\Gamma}) = \sum_{i=1}^{m} \left[\mathbf{x}_j^{(i)\top} \mathbf{V}^{-(i)}(\boldsymbol{\gamma}_k) \mathbf{x}_j^{(i)} - \left\{ \mathbf{x}_j^{(i)\top} \mathbf{V}^{-(i)}(\boldsymbol{\gamma}_k)(\mathbf{x}_k^{(i)} - \mathbf{X}^{(i)} \boldsymbol{\beta}_k) \right\}^2 \right].$$

These gradient expressions are useful for the algorithmic developments that follow later.

We take our estimator as a minimizer of the negative log-likelihood (2) over the fixed- and random-effects parameters while enforcing two key structural constraints. First, the graphs $\mathcal{G}(\mathbf{B})$ and $\mathcal{G}(\mathbf{\Gamma})$ induced by $\mathbf{B}$ and $\mathbf{\Gamma}$ must jointly remain acyclic. Here, $\mathcal{G}(\mathbf{W})$ denotes the directed graph on $\{1, \ldots, p\}$ having an edge $j \to k$ if and only if $w_{jk} \neq 0$. In other words, the graph $\mathcal{G}(\mathbf{B}) \cup \mathcal{G}(\mathbf{\Gamma})$ formed by the union of the individual edge sets must be a DAG. Second, sparsity is imposed to encourage parsimonious graphs with fewer edges and improve structure recovery. Together, these requirements lead to the constrained optimization problem

$$\min_{\mathbf{B} \in \mathbb{R}^{p \times p},\, \mathbf{\Gamma} \in \mathbb{R}_+^{p \times p}} l(\mathbf{B}, \mathbf{\Gamma}) + \lambda_1 \|\mathbf{B}\|_1 + \lambda_2 \|\mathbf{\Gamma}\|_1 \quad \text{s.t. } \mathcal{G}(\mathbf{B}) \cup \mathcal{G}(\mathbf{\Gamma}) \in \mathrm{DAG}_p, \tag{3}$$

where DAG_p is the set of all DAGs on nodes $\{1, \ldots, p\}$. Here, $\lambda_1, \lambda_2 \geq 0$ are sparsity penalty parameters, and $\|\cdot\|_1$ is the sum of the absolute values of the entries of its matrix argument. The optimization problem (3) is nonconvex, owing both to the form of the negative log-likelihood and to the DAG constraint. We denote its minimizers by $\hat{\mathbf{B}}$ and $\hat{\mathbf{\Gamma}}$.

To report cluster-specific graphs, we recover the random-effects $\mathbf{U}^{(i)}$ via their so-called best linear unbiased predictors (BLUPs) $\hat{\mathbf{U}}^{(i)}$. The BLUP for the kth column $\mathbf{u}_k^{(i)}$ of $\mathbf{U}^{(i)}$ is

$$\hat{\mathbf{u}}_k^{(i)} = \mathrm{diag}(\hat{\boldsymbol{\gamma}}_k) \mathbf{X}^{(i)\top} \mathbf{V}^{-(i)}(\hat{\boldsymbol{\gamma}}_k)(\mathbf{x}_k^{(i)} - \mathbf{X}^{(i)} \hat{\boldsymbol{\beta}}_k).$$

Combining $\hat{\mathbf{U}}^{(i)}$ with the estimated fixed-effects $\hat{\mathbf{B}}$ produces a DAG for cluster i as

$$\hat{\mathbf{W}}^{(i)} = \hat{\mathbf{B}} + \hat{\mathbf{U}}^{(i)}.$$

3.2 Differentiable Reformulation

As written, the DAG constraint in (3) constitutes a non-differentiable combinatorial restriction that precludes the use of scalable first-order optimization methods like gradient descent. Following the differentiable structure learning literature (e.g., Zheng et al., 2018; Bello et al., 2022), we reformulate the combinatorial DAG constraint as an *equivalent* differentiable DAG constraint. This line of work established the existence of smooth functions $h(\mathbf{W})$ such that

$$h(\mathbf{W}) = 0 \iff \mathcal{G}(\mathbf{W}) \in \mathrm{DAG}_p.$$

In other words, the differentiable function $h(\mathbf{W})$ vanishes exactly on the set of DAGs, so enforcing $h(\mathbf{W}) = 0$ is lossless relative to the original DAG constraint. We adopt the log-determinant form of h first proposed by Bello et al. (2022). For matrices $\mathbf{W} \in \mathbb{W} := \{\mathbf{W} \in \mathbb{R}_+^{p \times p} : \rho(\mathbf{W}) < 1\}$, where $\rho(\mathbf{W})$ denotes the spectral radius of $\mathbf{W}$, define the function

$$h(\mathbf{W}) := -\log\det(\mathbf{I} - \mathbf{W}).$$

Under the stated nonnegativity and spectral radius conditions, Bello et al. (2022) showed that $h(\mathbf{W}) = 0$ if and only if $\mathbf{W}$ is acyclic. In the real-valued case, $\mathbf{W}$ can be mapped to a

nonnegative surrogate via the Hadamard product $\mathbf{W} \odot \mathbf{W}$ before using h, preserving the zeros and hence the graph. Crucially, h is continuously differentiable on $\mathbb{W}$ with gradient

$$\nabla h(\mathbf{W}) = (\mathbf{I} - \mathbf{W})^{-\top}.$$

Our setting requires a stronger constraint than above: the *union* of the graphs induced by the fixed-effects matrix $\mathbf{B}$ and random-effects matrix $\mathbf{\Gamma}$ must be acyclic. It is therefore insufficient for $\mathbf{B}$ and $\mathbf{\Gamma}$ to be individually acyclic, as their union can still contain cycles, leading to cluster-specific matrices $\hat{\mathbf{W}}^{(i)}$ that are not valid DAGs. Since $\mathbf{\Gamma}$ has nonnegative entries, while $\mathbf{B}$ has real-valued entries, we enforce acyclicity of the union by applying the log-determinant characterization to the combined nonnegative matrix $\mathbf{B} \odot \mathbf{B} + \mathbf{\Gamma}$:

$$h(\mathbf{B}, \mathbf{\Gamma}) := -\log\det(\mathbf{I} - \mathbf{B} \odot \mathbf{B} - \mathbf{\Gamma}). \tag{4}$$

The following proposition formalizes the equivalence between acyclicity of the union graph induced by the fixed and random effects and the level-set of the log-determinant (4) at zero.

Proposition 1 *Let $h(\mathbf{B}, \mathbf{\Gamma})$ be given in (4) and define the domain*

$$\mathbb{D} := \{(\mathbf{B}, \mathbf{\Gamma}) \in \mathbb{R}^{p\times p} \times \mathbb{R}_+^{p\times p} : \rho(\mathbf{B} \odot \mathbf{B} + \mathbf{\Gamma}) < 1\}. \tag{5}$$

Then, for any $(\mathbf{B}, \mathbf{\Gamma}) \in \mathbb{D}$, it holds

$$\mathcal{G}(\mathbf{B}) \cup \mathcal{G}(\mathbf{\Gamma}) \in \mathrm{DAG}_p \iff h(\mathbf{B}, \mathbf{\Gamma}) = 0.$$

Proof See Appendix A. ■

Proposition 1 shows that, on the domain $\mathbb{D}$ defined in (5), enforcing $h(\mathbf{B}, \mathbf{\Gamma}) = 0$ is exactly equivalent to requiring that the union graph $\mathcal{G}(\mathbf{B}) \cup \mathcal{G}(\mathbf{\Gamma})$ be acyclic. Consequently, the differentiable constraint in (4) is lossless relative to the original combinatorial DAG restriction. The proof carries the argument of Bello et al. (2022) over to the mixed-effects setting.

The gradients of $h(\mathbf{B}, \mathbf{\Gamma})$ are

$$\nabla_{\mathbf{B}} h(\mathbf{B}, \mathbf{\Gamma}) = 2(\mathbf{I} - \mathbf{B} \odot \mathbf{B} - \mathbf{\Gamma})^{-\top} \odot \mathbf{B}$$

and

$$\nabla_{\mathbf{\Gamma}} h(\mathbf{B}, \mathbf{\Gamma}) = (\mathbf{I} - \mathbf{B} \odot \mathbf{B} - \mathbf{\Gamma})^{-\top}.$$

With this characterization, we can rewrite (3) without the combinatorial DAG constraint:

$$\min_{(\mathbf{B},\mathbf{\Gamma})\in\mathbb{D}} l(\mathbf{B}, \mathbf{\Gamma}) + \lambda_1\|\mathbf{B}\|_1 + \lambda_2\|\mathbf{\Gamma}\|_1 \quad \text{s.t. } h(\mathbf{B}, \mathbf{\Gamma}) = 0.$$

As discussed in the next section, it is possible to solve this optimization problem using first-order (gradient) methods, paving the way for scalable mixed-effects structure learning.

4 Scalable Computation

4.1 Path-Following Algorithm

Standard first-order algorithms do not readily accommodate hard constraints like $h(\mathbf{B}, \mathbf{\Gamma}) = 0$ directly. We therefore adopt a path-following strategy in the spirit of Bello et al. (2022). The key idea is to relax the constraint by incorporating it into the objective via a smooth barrier:

$$\min_{(\mathbf{B},\mathbf{\Gamma})\in\mathbb{D}} f_\mu(\mathbf{B}, \mathbf{\Gamma}; \lambda_1, \lambda_2) := \mu \{l(\mathbf{B}, \mathbf{\Gamma}) + \lambda_1 \|\mathbf{B}\|_1 + \lambda_2 \|\mathbf{\Gamma}\|_1\} + h(\mathbf{B}, \mathbf{\Gamma}). \tag{6}$$

Here, $\mu \geq 0$ is a continuation parameter that we decrease along a sequence $\mu^{(1)} > \mu^{(2)} > \cdots > \mu^{(T)} \geq 0$. At iteration $t+1$ of our strategy, we solve (6) with $\mu = \mu^{(t+1)}$ using the solution obtained from $\mu = \mu^{(t)}$ as an initialization point. Along this path, the log-determinant term $h(\mathbf{B}, \mathbf{\Gamma})$ acts as a smooth barrier over $\mathbb{D}$, keeping $\rho(\mathbf{B} \odot \mathbf{B} + \mathbf{\Gamma}) < 1$ and progressively steering iterates toward an acyclic union as μ decreases. In the limit as $\mu \to 0$, the barrier dominates and the resulting solution satisfies the acyclicity constraint $h(\mathbf{B}, \mathbf{\Gamma}) = 0$. Algorithm 1 summarizes the full procedure. In practice, we set the initial continuation

Algorithm 1 Path-following algorithm

Input: Initializer $(\mathbf{B}^{(0)}, \mathbf{\Gamma}^{(0)}) \in \mathbb{D}$ and parameters $\lambda_1, \lambda_2 > 0$ and $\mu^{(1)} > \cdots > \mu^{(T)} \geq 0$
for $t = 0, 1, \ldots, T-1$ **do**
  Initialize $(\mathbf{B}, \mathbf{\Gamma})$ at $(\mathbf{B}^{(t)}, \mathbf{\Gamma}^{(t)})$ and set

$$(\mathbf{B}^{(t+1)}, \mathbf{\Gamma}^{(t+1)}) \leftarrow \arg\min_{(\mathbf{B},\mathbf{\Gamma})\in\mathbb{D}} f_{\mu^{(t+1)}}(\mathbf{B}, \mathbf{\Gamma}; \lambda_1, \lambda_2)$$

end for
Output: Solution $(\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}}) = (\mathbf{B}^{(T)}, \mathbf{\Gamma}^{(T)})$

parameter $\mu^{(1)} = 1$ and take $\mu^{(t+1)} = 0.1\mu^{(t)}$ for $T = 5$ iterations. The initialization point $(\mathbf{B}^{(0)}, \mathbf{\Gamma}^{(0)})$ is set to matrices of zeros, i.e., $(\mathbf{B}^{(0)}, \mathbf{\Gamma}^{(0)}) = (\mathbf{0}, \mathbf{0})$, which trivially lies in $\mathbb{D}$.

4.2 Proximal Gradient Algorithm

To solve the inner problem (6) at each iteration of Algorithm 1, we design a proximal gradient algorithm (see, e.g., Polson et al., 2015). Towards this end, we use the fact that the objective function in (6) naturally decomposes into the sum of smooth and nonsmooth components as

$$f_\mu(\mathbf{B}, \mathbf{\Gamma}) := g_\mu(\mathbf{B}, \mathbf{\Gamma}) + r_\mu(\mathbf{B}, \mathbf{\Gamma}),$$

where the smooth component and nonsmooth component are defined respectively as

$$g_\mu(\mathbf{B}, \mathbf{\Gamma}) := \mu\{l(\mathbf{B}, \mathbf{\Gamma})\} + h(\mathbf{B}, \mathbf{\Gamma}), \qquad r_\mu(\mathbf{B}, \mathbf{\Gamma}) := \mu\{\lambda_1 \|\mathbf{B}\|_1 + \lambda_2 \|\mathbf{\Gamma}\|_1\} + \iota_{\mathbb{R}_+^{p\times p}}(\mathbf{\Gamma}). \tag{7}$$

We absorb the nonnegativity constraint on $\mathbf{\Gamma}$ into r_μ via the indicator $\iota_{\mathbb{R}_+^{p\times p}}(\mathbf{\Gamma})$, which equals 0 if $\mathbf{\Gamma} \in \mathbb{R}_+^{p\times p}$ and $+\infty$ otherwise. The smooth component g_μ admits gradients with respect to $\mathbf{B}$ and $\mathbf{\Gamma}$, while the nonsmooth term r_μ is amenable to proximal operators. This structure

leads to simple updates: given current iterates $(\mathbf{B}^{(k)}, \mathbf{\Gamma}^{(k)})$ we take a gradient step with step size $\alpha > 0$ on the smooth component and then apply proximal operators to the result, thus handling the nonsmooth penalties. The corresponding elementwise operators are

$$\text{soft}_\lambda(z) := \text{sign}(z)\max(|z| - \lambda, 0), \qquad \text{soft}^+_\lambda(z) := \max(z - \lambda, 0).$$

The first operator $\text{soft}_\lambda(z)$ is the well-known proximal operator for the ℓ_1-norm, applied to induce sparsity in the fixed-effects $\mathbf{B}$. The second operator $\text{soft}^+_\lambda(z)$ is the proximal operator for an ℓ_1-norm combined with a projection onto the nonnegative reals, used to produce sparse random-effects variances in $\mathbf{\Gamma}$. Algorithm 2 summarizes this proximal gradient procedure. In

Algorithm 2 Proximal gradient algorithm

Input: Initializer $(\mathbf{B}^{(0)}, \mathbf{\Gamma}^{(0)}) \in \mathbb{D}$, parameters $\lambda_1, \lambda_2 > 0$ and $\mu \geq 0$, and step size $\alpha > 0$
Initialize iterator $k \leftarrow 0$
while Not converged **do**
Perform proximal gradient updates

$$\mathbf{B}^{(k+1)} \leftarrow \text{soft}_{\alpha\mu\lambda_1}\left(\mathbf{B}^{(k)} - \alpha\nabla_{\mathbf{B}} g_\mu(\mathbf{B}^{(k)}, \mathbf{\Gamma}^{(k)})\right)$$

$$\mathbf{\Gamma}^{(k+1)} \leftarrow \text{soft}^+_{\alpha\mu\lambda_2}\left(\mathbf{\Gamma}^{(k)} - \alpha\nabla_{\mathbf{\Gamma}} g_\mu(\mathbf{B}^{(k)}, \mathbf{\Gamma}^{(k)})\right)$$

Increment iterator $k \leftarrow k + 1$
end while
Output: Solution $(\mathbf{B}^\star, \mathbf{\Gamma}^\star) = (\mathbf{B}^{(k)}, \mathbf{\Gamma}^{(k)})$

practice, we declare convergence when the relative change in g_μ falls below 10^{-6}.

To determine a suitable step size α in Algorithm 2, we derive a quadratic upper bound on the local change of the smooth component g_μ. This upper bound ensures that updates yield a controlled decrease in g_μ, thereby providing a principled basis for step size selection. The following proposition establishes this bound on any compact subset $\Omega \subset \mathbb{D}$.

Proposition 2 *Let $g_\mu(\mathbf{B}, \mathbf{\Gamma})$ be the smooth component (7) of the objective function, defined on the domain $\mathbb{D}$ given in (5). Fix any compact set $\Omega \subset \mathbb{D}$. Then there exists a constant $L < \infty$ such that for all $(\mathbf{B}, \mathbf{\Gamma}), (\mathbf{B}^+, \mathbf{\Gamma}^+) \in \Omega$ it holds*

$$g_\mu(\mathbf{B}^+, \mathbf{\Gamma}^+) \leq g_\mu(\mathbf{B}, \mathbf{\Gamma}) + \langle \nabla g_\mu(\mathbf{B}, \mathbf{\Gamma}), (\Delta\mathbf{B}, \Delta\mathbf{\Gamma})\rangle + \frac{L}{2}\|(\Delta\mathbf{B}, \Delta\mathbf{\Gamma})\|_F^2,$$

where $\Delta\mathbf{B} := \mathbf{B}^+ - \mathbf{B}$, $\Delta\mathbf{\Gamma} := \mathbf{\Gamma}^+ - \mathbf{\Gamma}$, and $\|(\Delta\mathbf{B}, \Delta\mathbf{\Gamma})\|_F^2 := \|\Delta\mathbf{B}\|_F^2 + \|\Delta\mathbf{\Gamma}\|_F^2$. Here, $\|\cdot\|_F$ denotes the Frobenius norm and $\langle\cdot,\cdot\rangle$ denotes the Frobenius inner product.

Proof See Appendix B. ■

Proposition 2 motivates the step size choice $\alpha \leq 1/L$, under which each proximal gradient descent update is guaranteed to be a descent step. In practice, we do not compute L explicitly when choosing the step size. Instead, we start from an initial step size and apply

backtracking line search until the inequality in Proposition 2 is satisfied by the trial update. This backtracking procedure can also be used to check that the trial update remains in $\mathbb{D}$.

With the step size condition above in hand, we now characterize the convergence properties of Algorithm 2 for fixed $\mu \geq 0$. The following theorem presents these properties.

Theorem 3 *Let $\mu \geq 0$. Fix any compact set $\Omega \subset \mathbb{D}$. Then there exists a constant $L < \infty$ such that, if Algorithm 2 is run with step size $\alpha < 1/L$ and the iterates $(\mathbf{B}^{(k)}, \mathbf{\Gamma}^{(k)})$ remain in Ω, the sequence of objective values $\{f_\mu(\mathbf{B}^{(k)}, \mathbf{\Gamma}^{(k)})\}$ is decreasing and converges to a finite limit $f_\mu^\star$. Moreover, as $k \to \infty$, the iterates satisfy*

$$\|(\mathbf{B}^{(k+1)}, \mathbf{\Gamma}^{(k+1)}) - (\mathbf{B}^{(k)}, \mathbf{\Gamma}^{(k)})\|_F \to 0.$$

Proof See Appendix C. ■

Theorem 3 shows that the proximal gradient descent iterates are well-behaved. In particular, the objective values decrease monotonically and the successive updates vanish asymptotically. Thus, under the stated step size condition, the objective sequence converges to a finite limit and the change from one iterate to the next becomes arbitrarily small as the algorithm progresses. The requirement that the iterates remain in a compact set is a technical condition used in the proof. Although this condition can be enforced explicitly if desired, our numerical experience is that it is not necessary to impose in practice for stable convergence.

4.3 Batched Updates

Direct evaluation of the negative log-likelihood in Algorithm 2 requires computing, for each node k and cluster i, the covariance matrix $\mathbf{V}^{(i)}(\boldsymbol{\gamma}_k)$ together with its log-determinant and the quadratic form involving its inverse. Both of these quantities depend on a Cholesky factorization of $\mathbf{V}^{(i)}(\boldsymbol{\gamma}_k)$. Looping over all m clusters and p nodes, therefore, leads to mp separate factorizations at every gradient step, each costing $O(n_i^3)$, which is computationally prohibitive when performed serially. To avoid this bottleneck, we perform the entire likelihood evaluation in parallel using batched linear algebra routines. Prior to running Algorithm 2, the cluster matrices $\mathbf{X}^{(i)}$ are padded into a block-aligned tensor $\mathbf{X} \in \mathbb{R}^{m \times \max_i n_i \times p}$ with a row mask to drop padded entries. Then, within each iteration of Algorithm 2, the random-effects variances $\boldsymbol{\gamma}_k$ are broadcast across clusters, producing the collection $\{\mathbf{V}^{(i)}(\boldsymbol{\gamma}_k)\}_{ik}$ as a four-dimensional tensor. Next, batched routines compute the factorizations concurrently, and the resulting triangular factors are used (again in batch) to extract the log-determinant terms from their diagonals and perform triangular solves for the quadratic terms.

In the setting $n_i \geq p$, the log-determinant and quadratic terms can be evaluated via Sylvester's determinant identity and the Woodbury identity using $p \times p$ batched factorizations, reducing the per-factorization cost from $O(n_i^3)$ to $O(p^3)$ (see, e.g., Harville, 1997).

4.4 Computational Complexity

Each iteration of Algorithm 2 evaluates the negative log-likelihood term $l(\mathbf{B}, \mathbf{\Gamma})$ and acyclicity term $h(\mathbf{B}, \mathbf{\Gamma})$, along with their gradients. The acyclicity term, which involves the log-determinant of a $p \times p$ matrix, requires $O(p^3)$ operations. The negative log-likelihood term requires $O(p \sum_{i=1}^m \min\{n_i^2 p + n_i^3, p^3\})$ operations. When the number of clusters m grows with the total sample size $n_\bullet := \sum_{i=1}^m n_i$ and the cluster sizes n_i remain uniformly bounded by a

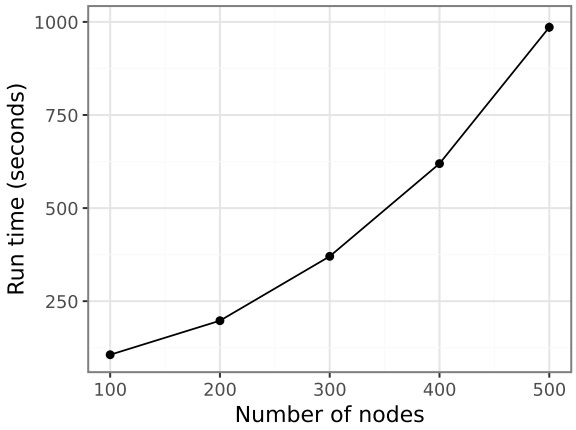


Figure 2: Run time in seconds as a function of the number of nodes p with $n_\bullet = 1{,}000$ observations and $m = 20$ clusters, measured on an NVIDIA RTX 4090. Averages (solid points) and standard errors (error bars) are measured over 30 datasets. All runs use regularization parameters $\lambda_1 = \lambda_2 = 0.001$.

constant n, the negative log-likelihood term reduces to $O(p^2 n_\bullet)$ operations. The gradients of both terms incur the same asymptotic costs. Hence, the cost per iteration of Algorithm 2 is $O(p\sum_{i=1}^m \min\{n_i^2 p + n_i^3, p^3\} + p^3)$ in general and $O(p^2 n_\bullet + p^3)$ in the bounded-cluster regime.

To examine how the run time scales with the number of nodes, we time Algorithm 1 as p varies. Figure 2 reports the results. As expected, the run time grows superlinearly in p, though not quite at a cubic rate (cubic complexity in p is worst-case complexity for computing the log-determinant). Overall, the times remain practical for many applications, with training taking about a minute for $p = 100$ and about 15 minutes for $p = 500$.

5 Statistical Properties

5.1 Setup

We now study some statistical properties of mixed DAGs. In particular, we first establish an identifiability result for the model and then present a statistical guarantee for our estimator in the asymptotic regime where the number of clusters $m \to \infty$. The guarantee addresses both (i) parameter estimation consistency and (ii) structure recovery consistency.

We assume the data-generating process is a structural equation model of the form (1):

$$\mathbf{X}^{(i)} = \mathbf{X}^{(i)}\mathbf{W}_0^{(i)} + \boldsymbol{\varepsilon}^{(i)}, \qquad \mathbf{W}_0^{(i)} = \mathbf{B}_0 + \mathbf{U}_0^{(i)}, \qquad i = 1, \ldots, m,$$

where the noise $\boldsymbol{\varepsilon}^{(i)} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$ and the random-effects matrix $\mathbf{U}_0^{(i)} \in \mathbb{R}^{p\times p}$ has columns satisfying $\mathbf{u}_{k,0}^{(i)} \sim \mathrm{N}(\mathbf{0}, \mathrm{diag}(\boldsymbol{\gamma}_{k,0}))$, with all random effects mutually independent and independent of the noise. The model parameters $\mathbf{B}_0 \in \mathbb{R}^{p\times p}$ and $\boldsymbol{\Gamma}_0 \in \mathbb{R}_+^{p\times p}$, representing the true fixed- and random-effects parameters, respectively, constitute a causal model satisfying

$$\mathcal{G}(\mathbf{B}_0) \cup \mathcal{G}(\boldsymbol{\Gamma}_0) \in \mathrm{DAG}_p.$$

We denote the edge sets (i.e., nonzero components) of the matrices $\mathbf{B}_0$ and $\mathbf{\Gamma}_0$ by

$$\mathcal{S}_{\mathbf{B}} := \{(j,k) : \beta_{jk,0} \neq 0\} \quad \text{and} \quad \mathcal{S}_{\mathbf{\Gamma}} := \{(j,k) : \gamma_{jk,0} \neq 0\}.$$

We assume $\mathcal{S}_{\mathbf{B}} \neq \emptyset$ and $\mathcal{S}_{\mathbf{\Gamma}} \neq \emptyset$ for simplicity, i.e., there is at least one edge in each graph.

To estimate $(\mathbf{B}_0, \mathbf{\Gamma}_0)$, we study the minimizer of the following optimization problem:

$$(\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}}) \in \operatorname*{arg\,min}_{(\mathbf{B},\mathbf{\Gamma})\in\mathcal{F}} F_m(\mathbf{B}, \mathbf{\Gamma}).$$

Here, the objective function F_m is the same penalized objective function in (3) but with the negative log-likelihood scaled by the number of clusters m to ensure a nondegenerate limit:

$$F_m(\mathbf{B}, \mathbf{\Gamma}) := \ell_m(\mathbf{B}, \mathbf{\Gamma}) + \lambda_1\|\mathbf{B}\|_1 + \lambda_2\|\mathbf{\Gamma}\|_1, \qquad \ell_m(\mathbf{B}, \mathbf{\Gamma}) = \frac{1}{m} l(\mathbf{B}, \mathbf{\Gamma}).$$

The feasible set $\mathcal{F}$ is the same as the constraint set in (3) but with added box constraints:

$$\mathcal{F} := \{(\mathbf{B}, \mathbf{\Gamma}) \in \mathbb{R}^{p\times p} \times \mathbb{R}_+^{p\times p} : \mathcal{G}(\mathbf{B}) \cup \mathcal{G}(\mathbf{\Gamma}) \in \mathrm{DAG}_p,\ \|\mathbf{B}\|_\infty \leq M_{\mathbf{B}},\ \|\mathbf{\Gamma}\|_\infty \leq M_{\mathbf{\Gamma}}\}. \quad (8)$$

The box constraints $\|\mathbf{B}\|_\infty \leq M_{\mathbf{B}}$ and $\|\mathbf{\Gamma}\|_\infty \leq M_{\mathbf{\Gamma}}$ are included in the constraint set as a technical convenience to guarantee compactness of $\mathcal{F}$. The constants $M_{\mathbf{B}} > 0$ and $M_{\mathbf{\Gamma}} > 0$ may be set arbitrarily large so long as they satisfy $M_{\mathbf{B}} > \|\mathbf{B}_0\|_\infty$ and $M_{\mathbf{\Gamma}} > \|\mathbf{\Gamma}_0\|_\infty$.

5.2 Assumptions

The consistency results for the estimator rely on two assumptions which we set forth below. Because some true components $\gamma_{jk,0}$ may equal zero, derivatives with respect to γ_{jk} at $\gamma_{jk} = 0$ are interpreted as the corresponding one-sided derivatives from within $\mathbb{R}_+^{p\times p}$.

Assumption 1 *As $m \to \infty$, the regularization parameters λ_1 and λ_2 satisfy $\lambda_1 \to 0$, $\lambda_2 \to 0$, $\sqrt{m}\lambda_1 \to \infty$, $\sqrt{m}\lambda_2 \to \infty$, and $\lambda_2/\lambda_1 \to \kappa \in (0, \infty)$.*

Assumption 1 specifies the scaling of the regularization parameters. The conditions $\lambda_1, \lambda_2 \to 0$ allow consistent parameter estimation by making the penalties asymptotically negligible. The additional conditions $\sqrt{m}\lambda_1, \sqrt{m}\lambda_2 \to \infty$ are needed for consistent structure recovery by ensuring that the penalties dominate the sampling fluctuations. The ratio condition $\lambda_2/\lambda_1 \to \kappa \in (0, \infty)$ implies that the two penalties shrink at the same asymptotic order.

Assumption 2 *The population Hessian $\mathbf{H} := \nabla^2 L(\mathbf{B}_0, \mathbf{\Gamma}_0)$ is positive definite. In addition, let $\mathcal{S} := (\mathcal{S}_{\mathbf{B}}, \mathcal{S}_{\mathbf{\Gamma}})$ and write $\mathbf{H}_{\mathcal{S}\mathcal{S}}$ for the principal submatrix of $\mathbf{H}$ indexed by $\mathcal{S}$. For (j,k) indexing a component of $\mathbf{B}$ or $\mathbf{\Gamma}$, let $\mathbf{h}^{(\mathbf{B})}_{jk,\mathcal{S}}$ and $\mathbf{h}^{(\mathbf{\Gamma})}_{jk,\mathcal{S}}$ denote the row vectors of second derivatives between β_{jk} or γ_{jk} and the active components $(\mathbf{B}_{\mathcal{S}_{\mathbf{B}}}, \mathbf{\Gamma}_{\mathcal{S}_{\mathbf{\Gamma}}})$, respectively. Then there exists an $\eta \in (0,1)$ such that*

$$\left|\mathbf{h}^{(\mathbf{B})}_{jk,\mathcal{S}}\mathbf{H}^{-1}_{\mathcal{S}\mathcal{S}}\begin{pmatrix}\mathbf{s}_{\mathbf{B}}\\ \kappa\mathbf{1}\end{pmatrix}\right| \leq 1-\eta \quad \textit{for all} \quad (j,k) \in \mathcal{S}^c_{\mathbf{B}},$$

and

$$\mathbf{h}^{(\mathbf{\Gamma})}_{jk,\mathcal{S}}\mathbf{H}^{-1}_{\mathcal{S}\mathcal{S}}\begin{pmatrix}\mathbf{s}_{\mathbf{B}}\\ \kappa\mathbf{1}\end{pmatrix} \leq \kappa(1-\eta) \quad \textit{for all} \quad (j,k) \in \mathcal{S}^c_{\mathbf{\Gamma}},$$

where $\mathbf{s}_{\mathbf{B}} := (\mathrm{sign}(\beta_{jk,0}))_{(j,k)\in\mathcal{S}_{\mathbf{B}}}$ contains the signs of all active components of $\mathbf{B}$.

Assumption 2 imposes regularity conditions on the population loss. Positive definiteness of the Hessian at the true parameters makes the population loss locally strongly convex around $(\mathbf{B}_0, \mathbf{\Gamma}_0)$. The two inequalities constitute an irrepresentability condition, analogous to those used to analyze ℓ_1-penalized linear models (Zhao and Yu, 2006), that limits the influence of inactive components through the active parameters, needed for consistent structure recovery.

5.3 Results

The following theorem establishes identifiability of the model described in the setup.

Theorem 4 *Under the data-generating process described above, if $\mathcal{G}(\mathbf{B}_0) \cup \mathcal{G}(\mathbf{\Gamma}_0)$ is a DAG, then the parameters $(\mathbf{B}_0, \mathbf{\Gamma}_0)$ are identifiable from the distribution of the observed clustered data. Consequently, the graphs $\mathcal{G}(\mathbf{B}_0)$, $\mathcal{G}(\mathbf{\Gamma}_0)$, and their union are identifiable.*

Proof See Appendix D. ■

Theorem 4 implies that the data-generating parameters $(\mathbf{B}_0, \mathbf{\Gamma}_0)$ are uniquely determined by the distribution of the observed clustered data. Since $\ell_m(\mathbf{B}, \mathbf{\Gamma})$ is the negative log-likelihood under our structural equation model, the population loss $L(\mathbf{B}, \mathbf{\Gamma}) := \mathrm{E}[\ell_m(\mathbf{B}, \mathbf{\Gamma})]$ coincides with the Kullback–Leibler divergence between the true distribution and the model up to an additive constant. Consequently, $L(\mathbf{B}, \mathbf{\Gamma})$ is uniquely minimized at $(\mathbf{B}_0, \mathbf{\Gamma}_0)$. In the non-clustered setting, Peters and Bühlmann (2014) establish an analogous identifiability result for linear Gaussian structural equation models with equal noise variances.

The following theorem establishes asymptotic consistency of parameter estimation and structure recovery of the estimator $(\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}})$ with respect to the ground truth $(\mathbf{B}_0, \mathbf{\Gamma}_0)$.

Theorem 5 *Suppose the clusters $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(m)}$ are iid and have a common (deterministic) cluster size $n_i \equiv n$ with $1 \leq n < \infty$, and Assumptions 1 and 2 hold. Then the estimator $(\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}})$ is parameter estimation consistent as $m \to \infty$:*

$$\|(\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}}) - (\mathbf{B}_0, \mathbf{\Gamma}_0)\|_F \xrightarrow{p} 0.$$

Moreover, $(\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}})$ is structure recovery consistent as $m \to \infty$:

$$\Pr\left(\mathcal{G}(\hat{\mathbf{B}}) = \mathcal{G}(\mathbf{B}_0) \text{ and } \mathcal{G}(\hat{\mathbf{\Gamma}}) = \mathcal{G}(\mathbf{\Gamma}_0)\right) \to 1.$$

Proof See Appendix E. ■

Theorem 5 is stated for equal cluster sizes for simplicity. The result extends to unequal but uniformly bounded cluster sizes $1 \leq n_i \leq n < \infty$ with minor additional technicalities.

The proof of Theorem 5 involves four steps. We first establish consistency of the parameter estimates. Next, we show that this result rules out false negatives in the recovered structure. We then employ a primal–dual witness argument (Wainwright, 2009) to establish the absence of false positives. Finally, we combine the results to get structure recovery consistency.

To the best of our knowledge, no prior work has established structure recovery consistency for mixed DAGs. Related selection consistency results are available for non-graphical mixed-effects models. For example, Fan and Li (2012) establish model selection consistency for penalized procedures that select both fixed and random effects in linear mixed models.

Moreover, even in the classical homogeneous setting, structure recovery guarantees for DAG estimators are limited. One notable exception is the ordered-variable regime, where the DAG reduces to a lower-triangular adjacency and Shojaie and Michailidis (2010) show ℓ_1-penalized likelihood methods are structure recovery consistent under suitable conditions.

6 Experiments

6.1 Baselines and Metrics

We experimentally evaluate five estimators, using the following labels in figures and tables.

- Mixed DAG: our proposed mixed-effects DAG estimator, which jointly learns fixed- and random-effects structures while simultaneously enforcing acyclicity of their union.
- Fixed DAG: a fixed-effects-only DAG estimator following Bello et al. (2022), which does not include random effects and therefore ignores cluster-level heterogeneity.
- Mixed DAG (FM): a computationally simpler masked alternative to the mixed DAG that replaces the explicit acyclicity constraint with a fixed mask (FM) obtained from the fixed DAG estimate. Specifically, the fixed DAG estimate is converted into a topological ordering, and the resulting mask is applied to the fixed- and random-effects matrices, setting all entries that violate the ordering to zero. Thus, only edges from earlier to later variables in the ordering can be selected, thereby guaranteeing acyclicity.
- Mixed DAG (OM): the same masked mixed-effects DAG estimator as outlined above, but using an oracle mask (OM) obtained from a topological ordering of the true DAG.
- Mixed DG: a mixed-effects directed graph (DG) estimator that fits the same model as the mixed DAG but does not enforce acyclicity, and hence need not produce a DAG.

The oracle variant serves as a purely theoretical baseline, providing an idealized bound on performance. Among these estimators, those with an explicit acyclicity constraint use the log-determinant acyclicity characterization of Bello et al. (2022), providing a uniform basis for comparison. Sparsity parameters are tuned on a validation set generated independently and identically to the training set. Appendix F provides implementation details.

We consider several metrics that characterize different attributes of the estimators. We report the estimation error, defined as the average of squared differences between the estimated and true weighted adjacency matrices over all clusters. We report the structural Hamming distance (SHD), defined as the number of edge additions, deletions, and reversals needed to transform the estimated structure into the true structure. We also report the F1 score, defined as the harmonic mean of precision and recall in edge selection. Finally, we report the sparsity, defined as the total number of edges in the estimated graph.

6.2 Synthetic Data

We generate synthetic datasets from the structural equation model (1) as follows. First, we sample an Erdős–Rényi graph (Erdős and Rényi, 1959) with p nodes and s edges, draw a random ordering of the nodes, and orient each edge from the earlier node to the later node in that ordering, thereby obtaining a DAG. Each edge in this graph is then assigned a

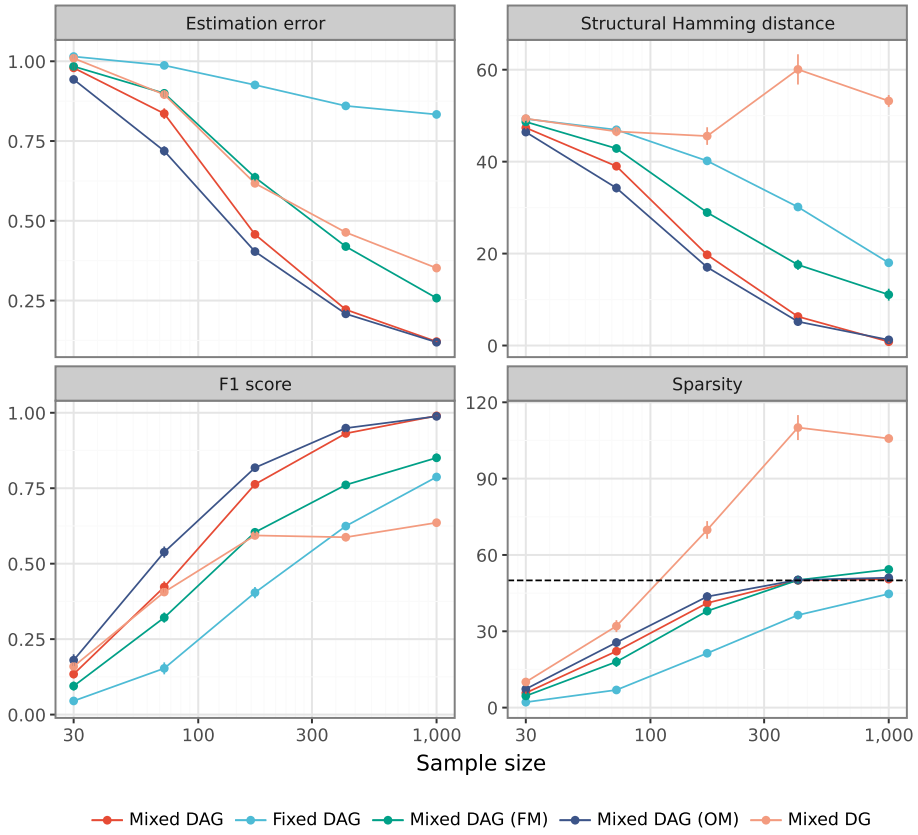


Figure 3: Performance on synthetic data generated from Erdős–Rényi DAGs with $p = 50$ nodes, $s = 50$ edges, and $m = \lceil \sqrt{n_\bullet} \rceil$ clusters. Averages (solid points) and standard errors (error bars) are measured over 30 datasets. The dashed horizontal line in the bottom right panel indicates the number of edges in the true DAG.

weight independently from $\text{Unif}([-0.3, -0.1] \cup [0.1, 0.3])$, producing the fixed-effects matrix $\mathbf{B}$. We then draw the random-effects variances independently from $\text{Unif}([0.1, 0.3])$ to form the matrix $\mathbf{\Gamma}$, and use these to generate the random-effects edge weights $\mathbf{U}^{(i)} \sim \text{N}(\mathbf{0}, \mathbf{\Gamma})$ for $i = 1, \ldots, m$. Independently, the noise terms are sampled for each cluster as $\boldsymbol{\varepsilon}^{(i)} \sim \text{N}(\mathbf{0}, \mathbf{I})$ for $i = 1, \ldots, m$. Finally, the total sample size $n_\bullet$ is varied over five values logarithmically spaced between 30 and 1,000. We set the number of clusters to $m = \lceil \sqrt{n_\bullet} \rceil$ and assign observations randomly to clusters, resulting in clusters that vary in size with a mean of $n_\bullet / m$.

Figures 3 and 4 present results for graphs with $p = 50$ and $p = 200$ nodes, respectively. In both settings, the mixed DAG exhibits excellent performance in both estimation and structure recovery. At very small sample sizes, where faithful recovery of the true DAG is intrinsically difficult, its results are comparable to those of the fixed DAG. As the sample size increases, however, a clear gap emerges, with the mixed DAG delivering substantially superior performance. Notably, the fixed DAG fails to recover the true graph structure even when $n_\bullet = 1{,}000$. The FM variant improves on the fixed DAG by incorporating cluster-specific edge weights, but its performance is limited by the quality of the ordering estimated from

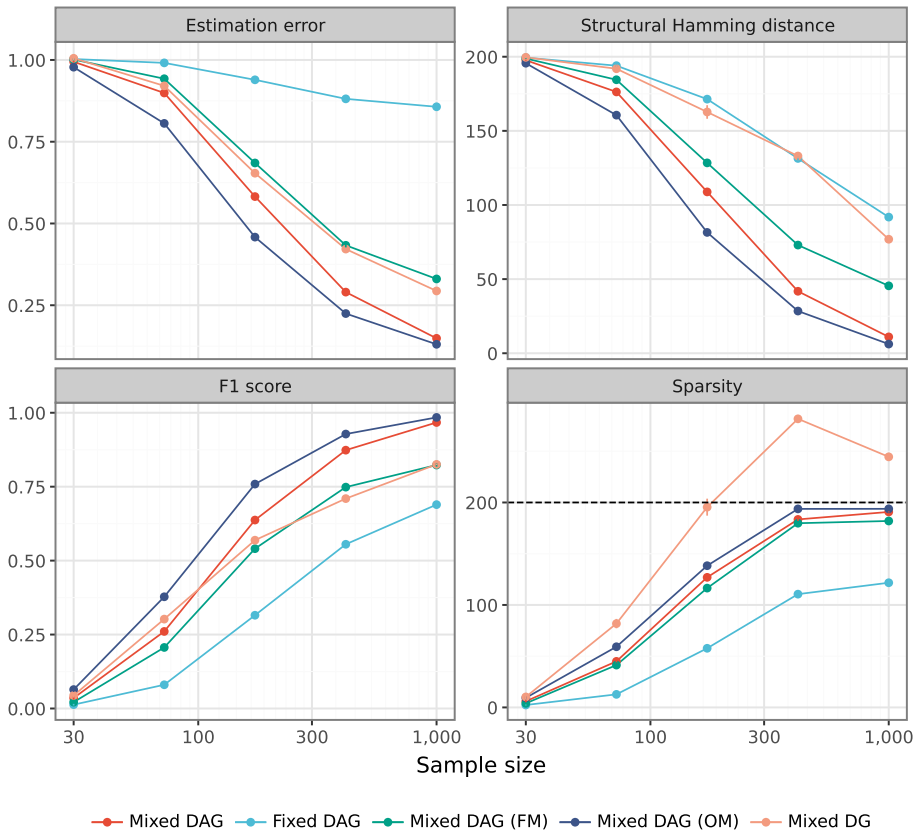


Figure 4: Performance on synthetic data generated from Erdős–Rényi DAGs with $p = 200$ nodes, $s = 200$ edges, and $m = \lceil \sqrt{n_\bullet} \rceil$ clusters. Averages (solid points) and standard errors (error bars) are measured over 30 datasets. The dashed horizontal line in the bottom right panel indicates the number of edges in the true DAG.

the fixed DAG. In contrast, the mixed DAG reliably recovers sparse, accurate structures, making it a practical method for learning DAGs from clustered data at scale. Moreover, its performance approaches that of the OM variant, which uses a topological ordering of the true DAG as an idealized benchmark. The mixed DG performs worst overall because it does not respect the acyclic nature of the graphs, highlighting the importance of the DAG constraint.

Further experiments under alternative simulation designs are reported in Appendix G. These additional results show that the mixed DAG continues to perform excellently under (i) alternative sparsity levels, (ii) alternative graph types, and (iii) alternative cluster regimes.

6.3 Semi-synthetic Data

To complement the fully synthetic graph designs in the previous experiments, we generate semi-synthetic datasets from the ANDES network, a well-known DAG from physics (Conati et al., 1997). It is publicly available from the Bayesian Network Repository (`https://www.bnlearn.com/bnrepository`). The network contains $p = 223$ nodes and $s = 338$ edges, representing a large-scale graph with realistic topologies and sparsity patterns. We generate

the datasets by treating the network adjacency matrix as the ground-truth DAG and draw datasets by sampling fixed effects, random effects, and noise in the same fashion as before.

Table 1 reports results for total sample size $n_\bullet = 1{,}000$, with the number of clusters $m = \lceil \sqrt{n_\bullet} \rceil = 32$ as in the main synthetic experiments. On this large graph, the mixed DAG maintains strong performance, achieving the best metrics among all practical competitors. The improvement over the fixed DAG is particularly pronounced, with a sevenfold decrease in estimation error and a sizable gain in F1 score. Importantly, the gap between the mixed DAG and the oracle mask variant is relatively modest, indicating that jointly learning the ordering and mixed effects does not incur a large statistical penalty at this scale. In contrast, methods that rely on a fixed ordering or omit acyclicity constraints exhibit degraded performance.

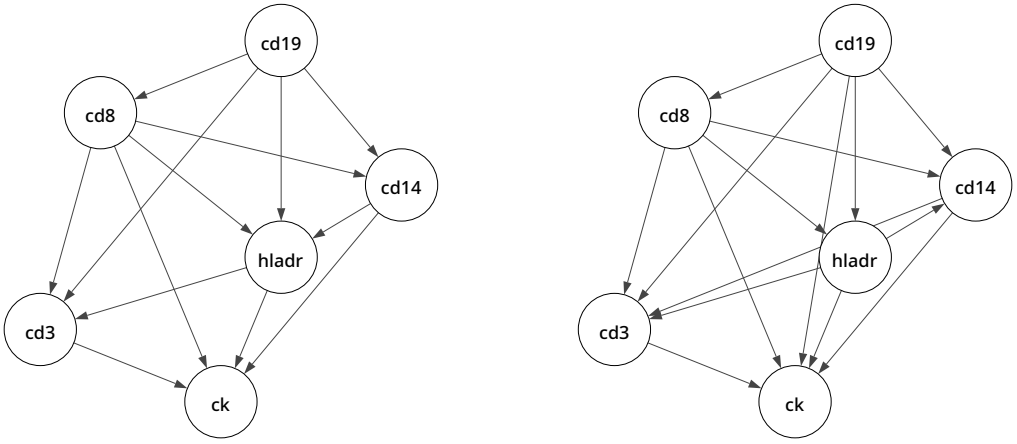
Table 1: Performance on semi-synthetic data generated from the ANDES network from the Bayesian Network Repository with $p = 223$ nodes, $s = 338$ edges, and $m = \lceil \sqrt{n_\bullet} \rceil = 32$ clusters. Averages (standard errors) are measured over 30 datasets.

	Est. error	SHD	F1 score	Sparsity
Mixed DAG	0.13 (0.00)	17.23 (0.67)	0.97 (0.00)	321.57 (0.64)
Fixed DAG	0.90 (0.00)	187.13 (2.12)	0.60 (0.01)	209.23 (2.39)
Mixed DAG (FM)	0.38 (0.01)	87.23 (1.13)	0.80 (0.00)	306.97 (1.29)
Mixed DAG (OM)	0.12 (0.00)	10.43 (0.50)	0.98 (0.00)	327.57 (0.50)
Mixed DG	0.29 (0.00)	118.97 (1.29)	0.83 (0.00)	364.20 (1.20)

6.4 Real Data

As an illustrative real-world example, we consider the lung cancer multiplex immunofluorescence dataset from the `R` package `VectraPolarisData`. The dataset contains cell-level protein measurements from $m = 153$ patients. The observational units are individual cells, and for each cell we observe the expression levels of $p = 6$ protein markers of tumor and immune activity: `cd19`, `cd3`, `cd14`, `cd8`, `hladr`, and `ck`. This low-dimensional setting enables direct visualization of learned graphs that might be indicative of tumor-immune proteomic interactions. We treat cells as clustered by patient to allow marker interactions to vary across individuals and to incorporate patient-level tumor heterogeneity, which is common in cancers and especially in lung cancer. To balance patient contributions in this illustrative analysis, we randomly select 300 cells per patient and split them equally into training, validation, and testing sets, yielding $n_\bullet = 153 \times 100 = 15{,}300$ cell-level observations in each split.

Figure 5 shows that both methods recover a coherent ordering in which immune-associated markers `cd19`, `cd8`, `hladr`, `cd3`, and `cd14` feed into `ck`, consistent with `ck` as an epithelial/tumor marker downstream of immune-associated features. Relative to the fixed DAG, which selects 13 edges, the mixed DAG learns a slightly denser graph containing 15 edges. Compared with the fixed DAG, the mixed DAG reverses the direction of the edge between `hladr` and `cd14`, yielding `hladr`→`cd14` rather than `cd14`→`hladr`, and also adds the dependencies `cd14`→`cd3` and `cd19`→`ck`. The remaining structure is consistent across the two graphs, including edges such as `cd8`→`cd14`, `cd3`→`ck`, and `cd14`→`ck`. These structural differences are reflected in a sizable gap in the mean squared reconstruction error on the testing set, which drops from 0.83 under the fixed DAG to 0.67 under the mixed DAG.



(a) Fixed DAG (b) Mixed DAG

Figure 5: Graphs learned on the `VectraPolarisData`. All edges in the graph produced by the mixed DAG contain both fixed and random effects.

7 Final Remarks

Many fields, including medicine and biology, give rise to clustered data with heterogeneous effects, making them unsuitable for existing approaches to causal discovery that assume a homogeneous population. Our approach addresses this issue by extending scalable structure learning techniques to clustered data through the integration of fixed- and random-effects principles within the framework of DAGs. We use a differentiable acyclicity constraint that guarantees the union of the fixed- and random-effects graphs remains acyclic, and develop a provably convergent first-order method that leverages efficient batched updates across clusters, making it feasible to learn graphs with hundreds of nodes. We further establish identifiability and show that our estimator recovers the true structure asymptotically. Synthetic, semi-synthetic, and real proteomics experiments demonstrate that the new estimator can recover both population-level and cluster-specific effects missed by alternative estimators. These results highlight its value as a principled tool for structure learning in clustered settings.

Acknowledgments and Disclosure of Funding

Thompson and Wand acknowledge financial support from the Australian Research Council under Discovery Project DP230101179. Baladandayuthapani was partially supported by National Institutes of Health grants R01CA244845- 01A1 and P30 CA46592 for this work.

Appendix A. Proof of Proposition 1

Proof

To simplify exposition, define the matrix

$$\mathbf{S} := \mathbf{B} \odot \mathbf{B} + \mathbf{\Gamma}.$$

Since $(\mathbf{B}, \mathbf{\Gamma}) \in \mathbb{D}$, the matrix $\mathbf{S}$ is entrywise nonnegative and satisfies $\rho(\mathbf{S}) < 1$. Hence, Theorem 1 of (Bello et al., 2022) applies to $\mathbf{S}$ and gives

$$-\log\det(\mathbf{I} - \mathbf{S}) = 0 \iff \mathcal{G}(\mathbf{S}) \in \mathrm{DAG}_p.$$

Since the left-hand side is exactly $h(\mathbf{B}, \mathbf{\Gamma})$, it follows

$$h(\mathbf{B}, \mathbf{\Gamma}) = 0 \iff \mathcal{G}(\mathbf{S}) \in \mathrm{DAG}_p.$$

For each pair (j, k), the definition of $\mathbf{S}$ and the nonnegativity of β_{jk}^2 and γ_{jk} imply

$$s_{jk} = 0 \iff \beta_{jk}^2 = 0 \text{ and } \gamma_{jk} = 0 \iff \beta_{jk} = 0 \text{ and } \gamma_{jk} = 0.$$

Thus, $s_{jk} \neq 0$ when $\beta_{jk} \neq 0$ or $\gamma_{jk} \neq 0$, and hence

$$\mathcal{G}(\mathbf{S}) = \mathcal{G}(\mathbf{B} \odot \mathbf{B} + \mathbf{\Gamma}) = \mathcal{G}(\mathbf{B}) \cup \mathcal{G}(\mathbf{\Gamma}).$$

Combining these equivalences gives

$$h(\mathbf{B}, \mathbf{\Gamma}) = 0 \iff \mathcal{G}(\mathbf{B}) \cup \mathcal{G}(\mathbf{\Gamma}) \in \mathrm{DAG}_p,$$

thereby proving the claim of the proposition.

■

Appendix B. Proof of Proposition 2

The proof of the proposition requires two technical lemmas, which establish Lipschitz continuity of the gradients of the smooth parts of the objective function on a compact set. Lemma 6 first shows this property for the negative log-likelihood.

Lemma 6 *Let $l(\mathbf{B}, \mathbf{\Gamma})$ be the negative log-likelihood (2) defined on the domain $\mathbb{R}^{p\times p} \times \mathbb{R}_+^{p\times p}$. Fix any compact set $\Omega \subset \mathbb{R}^{p\times p} \times \mathbb{R}_+^{p\times p}$. Then there exists a constant $L_l < \infty$ such that*

$$\|\nabla l(\mathbf{B}_1, \mathbf{\Gamma}_1) - \nabla l(\mathbf{B}_2, \mathbf{\Gamma}_2)\|_F \leq L_l \|(\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2)\|_F,$$

for all $(\mathbf{B}_1, \mathbf{\Gamma}_1), (\mathbf{B}_2, \mathbf{\Gamma}_2) \in \Omega$. In particular, $\nabla l(\mathbf{B}, \mathbf{\Gamma})$ is L_l-Lipschitz continuous on Ω.

Proof

For each cluster $i = 1, \ldots, m$ and node $k = 1, \ldots, p$, define

$$\mathbf{r}_k^{(i)}(\boldsymbol{\beta}) := \mathbf{x}_k^{(i)} - \mathbf{X}^{(i)}\boldsymbol{\beta},$$

and

$$\phi_{ik}(\boldsymbol{\beta}, \boldsymbol{\gamma}) := \log\det\{\mathbf{V}^{(i)}(\boldsymbol{\gamma})\} + \mathbf{r}_k^{(i)}(\boldsymbol{\beta})^\top \mathbf{V}^{-(i)}(\boldsymbol{\gamma}) \mathbf{r}_k^{(i)}(\boldsymbol{\beta}).$$

Thus, writing $\boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}_k$ for the kth column of $\mathbf{B}$ and $\mathbf{\Gamma}$, respectively, we have

$$l(\mathbf{B}, \mathbf{\Gamma}) = \sum_{k=1}^{p} \sum_{i=1}^{m} \phi_{ik}(\boldsymbol{\beta}_k, \boldsymbol{\gamma}_k).$$

Now, for any $\boldsymbol{\gamma} \in \mathbb{R}^p_+$, it holds

$$\mathbf{V}^{(i)}(\boldsymbol{\gamma}) = \mathbf{I} + \mathbf{X}^{(i)} \operatorname{diag}(\boldsymbol{\gamma}) \mathbf{X}^{(i)\top} \succeq \mathbf{I},$$

where $\succeq$ denotes positive semidefinite ordering, and hence $\mathbf{V}^{(i)}(\boldsymbol{\gamma})$ is positive definite with $\lambda_{\min}\{\mathbf{V}^{(i)}(\boldsymbol{\gamma})\} \geq 1$. In particular, $\log\det\{\mathbf{V}^{(i)}(\boldsymbol{\gamma})\}$ and $\mathbf{V}^{-(i)}(\boldsymbol{\gamma})$ are well-defined, and these mappings are C^∞ on the open set

$$\mathcal{U} := \left\{ (\mathbf{B}, \boldsymbol{\Gamma}) \in \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p} : \mathbf{V}^{(i)}(\boldsymbol{\gamma}_k) \succ \mathbf{0} \, \forall \, i = 1, \ldots, m, \, k = 1, \ldots, p \right\}.$$

Since $\mathbf{V}^{(i)}(\boldsymbol{\gamma}_k) \succeq \mathbf{I}$ whenever $\boldsymbol{\Gamma} \geq \mathbf{0}$, we have $\mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p}_+ \subset \mathcal{U}$. Consequently, $l(\mathbf{B}, \boldsymbol{\Gamma})$ is twice continuously differentiable on $\mathcal{U}$. Now, since $\mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p}_+$ is convex and Ω is compact, its convex hull $\operatorname{conv}(\Omega)$ is compact and satisfies

$$\operatorname{conv}(\Omega) \subset \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p}_+ \subset \mathcal{U}.$$

Therefore, by continuity of the Hessian and compactness of $\operatorname{conv}(\Omega)$, there exists a finite constant

$$L_l := \sup_{(\mathbf{B},\boldsymbol{\Gamma}) \in \operatorname{conv}(\Omega)} \|\nabla^2 l(\mathbf{B}, \boldsymbol{\Gamma})\|_{\text{op}} < \infty,$$

where $\|\cdot\|_{\text{op}}$ denotes the operator norm. Next, take any $(\mathbf{B}_1, \boldsymbol{\Gamma}_1), (\mathbf{B}_2, \boldsymbol{\Gamma}_2) \in \Omega$ and for $t \in [0, 1]$ define the line segment

$$(\mathbf{B}(t), \boldsymbol{\Gamma}(t)) := (\mathbf{B}_2, \boldsymbol{\Gamma}_2) + t\{(\mathbf{B}_1, \boldsymbol{\Gamma}_1) - (\mathbf{B}_2, \boldsymbol{\Gamma}_2)\}.$$

By convexity of $\operatorname{conv}(\Omega)$ it holds $(\mathbf{B}(t), \boldsymbol{\Gamma}(t)) \in \operatorname{conv}(\Omega)$ for all $t \in [0, 1]$ and therefore

$$\nabla l(\mathbf{B}_1, \boldsymbol{\Gamma}_1) - \nabla l(\mathbf{B}_2, \boldsymbol{\Gamma}_2) = \int_0^1 \nabla^2 l(\mathbf{B}(t), \boldsymbol{\Gamma}(t)) \{(\mathbf{B}_1, \boldsymbol{\Gamma}_1) - (\mathbf{B}_2, \boldsymbol{\Gamma}_2)\} \, dt.$$

Taking norms, applying the triangle inequality, and then using $\|\nabla^2 l(\mathbf{B}(t), \boldsymbol{\Gamma}(t))\|_{\text{op}} \leq L_l$ for all $t \in [0, 1]$, we have

$$\begin{aligned}
\|\nabla l(\mathbf{B}_1, \boldsymbol{\Gamma}_1) - \nabla l(\mathbf{B}_2, \boldsymbol{\Gamma}_2)\|_F &\leq \left\| \int_0^1 \nabla^2 l(\mathbf{B}(t), \boldsymbol{\Gamma}(t)) \{(\mathbf{B}_1, \boldsymbol{\Gamma}_1) - (\mathbf{B}_2, \boldsymbol{\Gamma}_2)\} \, dt \right\|_F \\
&\leq \int_0^1 \|\nabla^2 l(\mathbf{B}(t), \boldsymbol{\Gamma}(t))\|_{\text{op}} \, dt \, \|(\mathbf{B}_1, \boldsymbol{\Gamma}_1) - (\mathbf{B}_2, \boldsymbol{\Gamma}_2)\|_F \\
&\leq L_l \|(\mathbf{B}_1, \boldsymbol{\Gamma}_1) - (\mathbf{B}_2, \boldsymbol{\Gamma}_2)\|_F.
\end{aligned}$$

We thus conclude that $\nabla l(\mathbf{B}, \boldsymbol{\Gamma})$ is L_l-Lipschitz on Ω.

■

Lemma 7 shows the same property for the log-determinant function.

Lemma 7 *Let $h(\mathbf{B}, \boldsymbol{\Gamma})$ be the log-determinant function (4) defined on the domain $\mathbb{D}$ given in (5). Fix any compact set $\Omega \subset \mathbb{D}$. Then there exists a constant $L_h < \infty$ such that*

$$\|\nabla h(\mathbf{B}_1, \boldsymbol{\Gamma}_1) - \nabla h(\mathbf{B}_2, \boldsymbol{\Gamma}_2)\|_F \leq L_h \|(\mathbf{B}_1, \boldsymbol{\Gamma}_1) - (\mathbf{B}_2, \boldsymbol{\Gamma}_2)\|_F,$$

for all $(\mathbf{B}_1, \boldsymbol{\Gamma}_1), (\mathbf{B}_2, \boldsymbol{\Gamma}_2) \in \Omega$. In particular, $\nabla h(\mathbf{B}, \boldsymbol{\Gamma})$ is L_h-Lipschitz continuous on Ω.

Proof
Define the quantities

$$\mathbf{S}(\mathbf{B},\mathbf{\Gamma}) := \mathbf{B} \odot \mathbf{B} + \mathbf{\Gamma}, \qquad \mathbf{M}(\mathbf{B},\mathbf{\Gamma}) := \mathbf{I} - \mathbf{S}(\mathbf{B},\mathbf{\Gamma}),$$

so

$$h(\mathbf{B},\mathbf{\Gamma}) = -\log\det\{\mathbf{M}(\mathbf{B},\mathbf{\Gamma})\}.$$

Now, for any $(\mathbf{B},\mathbf{\Gamma}) \in \mathbb{D}$, the matrix $\mathbf{S}(\mathbf{B},\mathbf{\Gamma})$ is entrywise nonnegative and satisfies $\rho\{\mathbf{S}(\mathbf{B},\mathbf{\Gamma})\} < 1$, meaning $\det\{\mathbf{M}(\mathbf{B},\mathbf{\Gamma})\} > 0$ (since it is a nonsingular M-matrix; see Meyer, 2000). In particular, these properties mean $h(\mathbf{B},\mathbf{\Gamma})$ is C^∞ on the open set

$$\mathcal{U} := \left\{(\mathbf{B},\mathbf{\Gamma}) \in \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p} : \det\{\mathbf{M}(\mathbf{B},\mathbf{\Gamma})\} > 0\right\}.$$

Since $\det\{\mathbf{M}(\mathbf{B},\mathbf{\Gamma})\} > 0$ for all $(\mathbf{B},\mathbf{\Gamma}) \in \mathbb{D}$, we have $\Omega \subset \mathcal{U}$. Consequently, $h(\mathbf{B},\mathbf{\Gamma})$ is twice continuously differentiable on $\mathcal{U}$, so the Hessian $\nabla^2 h(\mathbf{B},\mathbf{\Gamma})$ is continuous on $\mathcal{U}$. Now, since Ω is compact and $\mathcal{U}$ is open, there exists $\delta > 0$ such that the closed δ-neighborhood of Ω is contained in $\mathcal{U}$. Hence, by continuity of the Hessian and compactness of the closed δ-neighborhood of Ω, there exists a finite constant

$$M_h := \sup_{\inf_{(\tilde{\mathbf{B}},\tilde{\mathbf{\Gamma}})\in\Omega} \|(\mathbf{B},\mathbf{\Gamma})-(\tilde{\mathbf{B}},\tilde{\mathbf{\Gamma}})\|_F \le \delta} \|\nabla^2 h(\mathbf{B},\mathbf{\Gamma})\|_{\text{op}} < \infty.$$

Also, by continuity of the gradient and compactness of Ω, it holds

$$G_h := \sup_{(\mathbf{B},\mathbf{\Gamma})\in\Omega} \|\nabla h(\mathbf{B},\mathbf{\Gamma})\|_F < \infty.$$

Define the quantity

$$L_h := \max\left\{M_h, \frac{2G_h}{\delta}\right\}.$$

Next, take any $(\mathbf{B}_1,\mathbf{\Gamma}_1), (\mathbf{B}_2,\mathbf{\Gamma}_2) \in \Omega$. First, suppose case (i) is true:

$$\|(\mathbf{B}_1,\mathbf{\Gamma}_1) - (\mathbf{B}_2,\mathbf{\Gamma}_2)\|_F \le \delta.$$

For $t \in [0,1]$ define the line segment

$$(\mathbf{B}(t),\mathbf{\Gamma}(t)) := (\mathbf{B}_2,\mathbf{\Gamma}_2) + t\{(\mathbf{B}_1,\mathbf{\Gamma}_1) - (\mathbf{B}_2,\mathbf{\Gamma}_2)\}.$$

It then follows

$$\inf_{(\tilde{\mathbf{B}},\tilde{\mathbf{\Gamma}})\in\Omega} \|(\mathbf{B}(t),\mathbf{\Gamma}(t)) - (\tilde{\mathbf{B}},\tilde{\mathbf{\Gamma}})\|_F \le \|(\mathbf{B}(t),\mathbf{\Gamma}(t)) - (\mathbf{B}_2,\mathbf{\Gamma}_2)\|_F = t\|(\mathbf{B}_1,\mathbf{\Gamma}_1) - (\mathbf{B}_2,\mathbf{\Gamma}_2)\|_F \le \delta$$

for all $t \in [0,1]$, and hence $(\mathbf{B}(t),\mathbf{\Gamma}(t)) \in \mathcal{U}$ for all $t \in [0,1]$. Therefore, we have

$$\nabla h(\mathbf{B}_1,\mathbf{\Gamma}_1) - \nabla h(\mathbf{B}_2,\mathbf{\Gamma}_2) = \int_0^1 \nabla^2 h(\mathbf{B}(t),\mathbf{\Gamma}(t))\{(\mathbf{B}_1,\mathbf{\Gamma}_1) - (\mathbf{B}_2,\mathbf{\Gamma}_2)\}\, dt.$$

Taking norms, applying the triangle inequality, and then using

$$\|\nabla^2 h(\mathbf{B}(t),\mathbf{\Gamma}(t))\|_{\text{op}} \le M_h \le L_h$$

for all $t \in [0,1]$ gives

$$\begin{aligned}\|\nabla h(\mathbf{B}_1,\mathbf{\Gamma}_1) - \nabla h(\mathbf{B}_2,\mathbf{\Gamma}_2)\|_F &\leq \int_0^1 \|\nabla^2 h(\mathbf{B}(t),\mathbf{\Gamma}(t))\|_{\text{op}}\, dt\, \|(\mathbf{B}_1,\mathbf{\Gamma}_1) - (\mathbf{B}_2,\mathbf{\Gamma}_2)\|_F \\ &\leq L_h \|(\mathbf{B}_1,\mathbf{\Gamma}_1) - (\mathbf{B}_2,\mathbf{\Gamma}_2)\|_F.\end{aligned}$$

Next, suppose case (ii) is true:

$$\|(\mathbf{B}_1,\mathbf{\Gamma}_1) - (\mathbf{B}_2,\mathbf{\Gamma}_2)\|_F > \delta.$$

Then, using the triangle inequality, the definition of G_h, and the definition of L_h, it holds

$$\begin{aligned}\|\nabla h(\mathbf{B}_1,\mathbf{\Gamma}_1) - \nabla h(\mathbf{B}_2,\mathbf{\Gamma}_2)\|_F &\leq \|\nabla h(\mathbf{B}_1,\mathbf{\Gamma}_1)\|_F + \|\nabla h(\mathbf{B}_2,\mathbf{\Gamma}_2)\|_F \\ &\leq 2G_h \\ &\leq \frac{2G_h}{\delta}\|(\mathbf{B}_1,\mathbf{\Gamma}_1) - (\mathbf{B}_2,\mathbf{\Gamma}_2)\|_F \\ &\leq L_h \|(\mathbf{B}_1,\mathbf{\Gamma}_1) - (\mathbf{B}_2,\mathbf{\Gamma}_2)\|_F.\end{aligned}$$

Combining cases (i) and (ii), we conclude that $\nabla h(\mathbf{B},\mathbf{\Gamma})$ is L_h-Lipschitz on Ω.

■

We are now ready to prove Proposition 2.

Proof

Since $\mathbb{D}$ is open relative to $\mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p}_+$ and Ω is compact, there exists $\delta > 0$ such that the compact set

$$\Omega_\delta := \left\{ (\mathbf{B},\mathbf{\Gamma}) \in \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p}_+ : \inf_{(\tilde{\mathbf{B}},\tilde{\mathbf{\Gamma}})\in\Omega} \|(\mathbf{B},\mathbf{\Gamma}) - (\tilde{\mathbf{B}},\tilde{\mathbf{\Gamma}})\|_F \leq \delta \right\} \subset \mathbb{D}.$$

Applying Lemmas 6 and 7 to the compact set Ω_δ and then using the triangle inequality, there exists a finite constant $L_\delta < \infty$ such that

$$\|\nabla g_\mu(\mathbf{B}_1,\mathbf{\Gamma}_1) - \nabla g_\mu(\mathbf{B}_2,\mathbf{\Gamma}_2)\|_F \leq L_\delta \|(\mathbf{B}_1,\mathbf{\Gamma}_1) - (\mathbf{B}_2,\mathbf{\Gamma}_2)\|_F$$

for all $(\mathbf{B}_1,\mathbf{\Gamma}_1), (\mathbf{B}_2,\mathbf{\Gamma}_2) \in \Omega_\delta$. Also, by continuity of g_μ and ∇g_μ and compactness of $\Omega \times \Omega$, it holds

$$C := \sup_{(\mathbf{B}_1,\mathbf{\Gamma}_1),(\mathbf{B}_2,\mathbf{\Gamma}_2)\in\Omega} |g_\mu(\mathbf{B}_1,\mathbf{\Gamma}_1) - g_\mu(\mathbf{B}_2,\mathbf{\Gamma}_2) - \langle \nabla g_\mu(\mathbf{B}_2,\mathbf{\Gamma}_2), (\mathbf{B}_1,\mathbf{\Gamma}_1) - (\mathbf{B}_2,\mathbf{\Gamma}_2)\rangle| < \infty.$$

Define the quantity

$$L := \max\left\{ L_\delta, \frac{2C}{\delta^2} \right\}.$$

Next, take any $(\mathbf{B}_1,\mathbf{\Gamma}_1), (\mathbf{B}_2,\mathbf{\Gamma}_2) \in \Omega$. First, suppose case (i) is true:

$$\|(\mathbf{B}_1,\mathbf{\Gamma}_1) - (\mathbf{B}_2,\mathbf{\Gamma}_2)\|_F \leq \delta.$$

For $t \in [0,1]$ define the line segment

$$(\mathbf{B}(t), \mathbf{\Gamma}(t)) := (\mathbf{B}_2, \mathbf{\Gamma}_2) + t\{(\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2)\}.$$

It then follows

$$\inf_{(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}) \in \Omega} \|(\mathbf{B}(t), \mathbf{\Gamma}(t)) - (\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}})\|_F \leq \|(\mathbf{B}(t), \mathbf{\Gamma}(t)) - (\mathbf{B}_2, \mathbf{\Gamma}_2)\|_F = t\|(\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2)\|_F \leq \delta$$

for all $t \in [0,1]$, and hence $(\mathbf{B}(t), \mathbf{\Gamma}(t)) \in \Omega_\delta$ for all $t \in [0,1]$. Therefore, it holds

$$\begin{aligned}
g_\mu(\mathbf{B}_1, \mathbf{\Gamma}_1) - g_\mu(\mathbf{B}_2, \mathbf{\Gamma}_2) - \langle \nabla g_\mu(\mathbf{B}_2, \mathbf{\Gamma}_2), (\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2) \rangle & \\
&= \int_0^1 \langle \nabla g_\mu(\mathbf{B}(t), \mathbf{\Gamma}(t)) - \nabla g_\mu(\mathbf{B}_2, \mathbf{\Gamma}_2), (\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2) \rangle \, dt \\
&\leq \int_0^1 \|\nabla g_\mu(\mathbf{B}(t), \mathbf{\Gamma}(t)) - \nabla g_\mu(\mathbf{B}_2, \mathbf{\Gamma}_2)\|_F \, dt \, \|(\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2)\|_F \\
&\leq \int_0^1 L_\delta t \, dt \, \|(\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2)\|_F^2 \\
&= \frac{L_\delta}{2} \|(\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2)\|_F^2 \\
&\leq \frac{L}{2} \|(\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2)\|_F^2.
\end{aligned}$$

Rearranging the above inequality gives

$$g_\mu(\mathbf{B}_1, \mathbf{\Gamma}_1) \leq g_\mu(\mathbf{B}_2, \mathbf{\Gamma}_2) + \langle \nabla g_\mu(\mathbf{B}_2, \mathbf{\Gamma}_2), (\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2) \rangle + \frac{L}{2} \|(\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2)\|_F^2. \tag{9}$$

Next, suppose case (ii) is true:

$$\|(\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2)\|_F > \delta.$$

Then, by the definition of C and the definition of L, it holds

$$\begin{aligned}
g_\mu(\mathbf{B}_1, \mathbf{\Gamma}_1) - g_\mu(\mathbf{B}_2, \mathbf{\Gamma}_2) - \langle \nabla g_\mu(\mathbf{B}_2, \mathbf{\Gamma}_2), (\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2) \rangle &\leq C \\
&\leq \frac{C}{\delta^2} \|(\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2)\|_F^2 \\
&\leq \frac{L}{2} \|(\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2)\|_F^2.
\end{aligned}$$

Again, rearranging the above inequality gives

$$g_\mu(\mathbf{B}_1, \mathbf{\Gamma}_1) \leq g_\mu(\mathbf{B}_2, \mathbf{\Gamma}_2) + \langle \nabla g_\mu(\mathbf{B}_2, \mathbf{\Gamma}_2), (\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2) \rangle + \frac{L}{2} \|(\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2)\|_F^2. \tag{10}$$

Combining (9) and (10) and taking $(\mathbf{B}_1, \mathbf{\Gamma}_1) = (\mathbf{B}^+, \mathbf{\Gamma}^+)$ and $(\mathbf{B}_2, \mathbf{\Gamma}_2) = (\mathbf{B}, \mathbf{\Gamma})$ completes the proof. ■

Appendix C. Proof of Theorem 3

Proof

We break the proof into two parts, addressing the two claims of the theorem in turn.

Part 1: Convergence of objective values We begin by proving the first claim of the theorem that the objective values are decreasing and convergent. Our proof works by upper bounding g_μ and r_μ to attain a sufficient decrease property for f_μ. First, recall that Proposition 2 provides an upper bound for g_μ as

$$g_\mu(\mathbf{B}^{(k+1)}, \mathbf{\Gamma}^{(k+1)}) \leq g_\mu(\mathbf{B}^{(k)}, \mathbf{\Gamma}^{(k)}) + \left\langle \nabla g_\mu(\mathbf{B}^{(k)}, \mathbf{\Gamma}^{(k)}), (\Delta\mathbf{B}^{(k)}, \Delta\mathbf{\Gamma}^{(k)}) \right\rangle + \frac{L}{2}\|(\Delta\mathbf{B}^{(k)}, \Delta\mathbf{\Gamma}^{(k)})\|_F^2, \tag{11}$$

where

$$\Delta\mathbf{B}^{(k)} := \mathbf{B}^{(k+1)} - \mathbf{B}^{(k)}, \qquad \Delta\mathbf{\Gamma}^{(k)} := \mathbf{\Gamma}^{(k+1)} - \mathbf{\Gamma}^{(k)}.$$

To bound r_μ, recall that the proximal gradient updates

$$\mathbf{B}^{(k+1)} \leftarrow \text{soft}_{\alpha\mu\lambda_1}\left(\mathbf{B}^{(k)} - \alpha\nabla_{\mathbf{B}} g_\mu(\mathbf{B}^{(k)}, \mathbf{\Gamma}^{(k)})\right)$$

and

$$\mathbf{\Gamma}^{(k+1)} \leftarrow \text{soft}^+_{\alpha\mu\lambda_2}\left(\mathbf{\Gamma}^{(k)} - \alpha\nabla_{\mathbf{\Gamma}} g_\mu(\mathbf{B}^{(k)}, \mathbf{\Gamma}^{(k)})\right)$$

are the minimizers of the proximal quadratic model

$$Q(\mathbf{B}, \mathbf{\Gamma}) := \left\langle \nabla g_\mu(\mathbf{B}^{(k)}, \mathbf{\Gamma}^{(k)}), (\mathbf{B}, \mathbf{\Gamma}) - (\mathbf{B}^{(k)}, \mathbf{\Gamma}^{(k)}) \right\rangle + \frac{1}{2\alpha}\|(\mathbf{B}, \mathbf{\Gamma}) - (\mathbf{B}^{(k)}, \mathbf{\Gamma}^{(k)})\|_F^2 + r_\mu(\mathbf{B}, \mathbf{\Gamma}).$$

Hence, by virtue of the fact that $\mathbf{B}^{(k+1)}$ and $\mathbf{\Gamma}^{(k+1)}$ minimize $Q(\mathbf{B}, \mathbf{\Gamma})$, it holds

$$Q(\mathbf{B}^{(k+1)}, \mathbf{\Gamma}^{(k+1)}) \leq Q(\mathbf{B}^{(k)}, \mathbf{\Gamma}^{(k)}). \tag{12}$$

Directly evaluating the expression for $Q(\mathbf{B}^{(k)}, \mathbf{\Gamma}^{(k)})$ on the right-hand side of (12) yields

$$Q(\mathbf{B}^{(k)}, \mathbf{\Gamma}^{(k)}) = r_\mu(\mathbf{B}^{(k)}, \mathbf{\Gamma}^{(k)}).$$

Likewise, evaluating the expression for $Q(\mathbf{B}^{(k+1)}, \mathbf{\Gamma}^{(k+1)})$ on the left-hand side of (12) and subsequently rearranging terms yields

$$r_\mu(\mathbf{B}^{(k+1)}, \mathbf{\Gamma}^{(k+1)}) \leq r_\mu(\mathbf{B}^{(k)}, \mathbf{\Gamma}^{(k)}) - \left\langle \nabla g_\mu(\mathbf{B}^{(k)}, \mathbf{\Gamma}^{(k)}), (\Delta\mathbf{B}^{(k)}, \Delta\mathbf{\Gamma}^{(k)}) \right\rangle - \frac{1}{2\alpha}\|(\Delta\mathbf{B}^{(k)}, \Delta\mathbf{\Gamma}^{(k)})\|_F^2. \tag{13}$$

Finally, to bound the objective function f_μ we add (11) and (13) and cancel the inner product terms to get

$$f_\mu(\mathbf{B}^{(k+1)}, \mathbf{\Gamma}^{(k+1)}) \leq f_\mu(\mathbf{B}^{(k)}, \mathbf{\Gamma}^{(k)}) - \left(\frac{1}{2\alpha} - \frac{L}{2}\right)\|(\Delta\mathbf{B}^{(k)}, \Delta\mathbf{\Gamma}^{(k)})\|_F^2. \tag{14}$$

By assumption, the step size $\alpha < 1/L$, so $1/(2\alpha) - L/2 > 0$. It immediately follows

$$f_\mu(\mathbf{B}^{(k+1)}, \mathbf{\Gamma}^{(k+1)}) \leq f_\mu(\mathbf{B}^{(k)}, \mathbf{\Gamma}^{(k)}).$$

Finally, since a continuous function on a compact set attains its minimum, f_μ is bounded below on Ω. Moreover, because a monotone decreasing sequence bounded below converges, the sequence of objective values must converge, proving the first claim of the theorem.

Part 2: Vanishing difference in iterates We now prove the second claim of the theorem that the difference in iterates vanishes. Define the constant $c := 1/(2\alpha) - L/2$. As before, since $\alpha < 1/L$, we have $c > 0$. Hence, for every k, it holds from (14)

$$c\|(\Delta\mathbf{B}^{(k)}, \Delta\mathbf{\Gamma}^{(k)})\|_F^2 \leq f_\mu(\mathbf{B}^{(k)}, \mathbf{\Gamma}^{(k)}) - f_\mu(\mathbf{B}^{(k+1)}, \mathbf{\Gamma}^{(k+1)}).$$

Summing both sides from $k = 0$ to $K - 1$ yields the bound

$$c \sum_{k=0}^{K-1} \|(\Delta\mathbf{B}^{(k)}, \Delta\mathbf{\Gamma}^{(k)})\|_F^2 \leq f_\mu(\mathbf{B}^{(0)}, \mathbf{\Gamma}^{(0)}) - f_\mu(\mathbf{B}^{(K)}, \mathbf{\Gamma}^{(K)}).$$

Now, define the quantity

$$f_\mu^{\inf} := \inf_{(\mathbf{B},\mathbf{\Gamma})\in\Omega} f_\mu(\mathbf{B}, \mathbf{\Gamma}) > -\infty.$$

From the fact that $f_\mu(\mathbf{B}^{(K)}, \mathbf{\Gamma}^{(K)}) \geq f_\mu^{\inf}$, we get

$$c \sum_{k=0}^{K-1} \|(\Delta\mathbf{B}^{(k)}, \Delta\mathbf{\Gamma}^{(k)})\|_F^2 \leq f_\mu(\mathbf{B}^{(0)}, \mathbf{\Gamma}^{(0)}) - f_\mu^{\inf}.$$

Thus, the sums on the left-hand side are uniformly bounded in K. Taking $K \to \infty$, we obtain

$$\sum_{k=0}^{\infty} \|(\Delta\mathbf{B}^{(k)}, \Delta\mathbf{\Gamma}^{(k)})\|_F^2 < \infty.$$

Since this series is convergent and has nonnegative terms, its terms must converge to zero. Hence, it must hold

$$\|(\Delta\mathbf{B}^{(k)}, \Delta\mathbf{\Gamma}^{(k)})\|_F^2 \to 0,$$

from which the second claim of the theorem immediately follows.

■

Appendix D. Proof of Theorem 4

Proof Fix one observation row from one cluster and suppress the cluster and row indices, writing the resulting random vector as $(x_1, \ldots, x_p)^\top$. Since the joint distribution of the clustered data determines the marginal distribution of this vector, it suffices to recover $(\mathbf{B}_0, \mathbf{\Gamma}_0)$ from that marginal distribution.

Because $\mathcal{G}(\mathbf{B}_0) \cup \mathcal{G}(\mathbf{\Gamma}_0)$ is a DAG, there exists a topological ordering of the nodes. Fix one such ordering for notational convenience and index the variables accordingly. Then we may write

$$x_k = \sum_{j<k} (\beta_{jk,0} + u_{jk}) x_j + \varepsilon_k, \qquad k = 1, \ldots, p,$$

where the random-effects $u_{jk} \sim \mathrm{N}(0, \gamma_{jk,0})$ are independent, the noise terms $\varepsilon_k \sim \mathrm{N}(0,1)$ are independent, and all random effects are independent of all noise terms. Conditional on the random effects, $(x_1, \ldots, x_p)$ is Gaussian with positive definite covariance, so its marginal distribution has a strictly positive density on $\mathbb{R}^p$.

A node is a source (i.e., it has no parents) if and only if its variance equals one. Indeed, if node k is a source, then $x_k = \varepsilon_k$, so $\mathrm{Var}(x_k) = 1$. Conversely, if node k is not a source, then at least one coefficient $\beta_{jk,0} + u_{jk}$ with $j < k$ is nonzero almost surely. Since the random effects entering node k do not appear in the structural equations for $x_1, \ldots, x_{k-1}$, they are independent of $(x_1, \ldots, x_{k-1})$. Therefore, the variance of x_k satisfies

$$\mathrm{Var}(x_k) = \mathrm{Var}\left(\sum_{j<k} (\beta_{jk,0} + u_{jk}) x_j \right) + 1 > 1,$$

because $(x_1, \ldots, x_{k-1})$ has positive definite covariance and the coefficients are not zero almost surely. Thus, the source set is determined by the observed distribution.

Choose any source and relabel it as node one. Then $x_1 = \varepsilon_1$, so x_1 is independent of all remaining noise terms and all random effects. For $k = 2, \ldots, p$, we have

$$x_k = (\beta_{1k,0} + u_{1k}) x_1 + \sum_{1<j<k} (\beta_{jk,0} + u_{jk}) x_j + \varepsilon_k.$$

Thus, conditional on $x_1 = 0$, the vector $(x_2, \ldots, x_p)$ again follows a mixed-effects DAG model of the same form, with node one removed. Applying the previous variance characterization recursively to these conditional distributions identifies one source after another, and therefore allows us to construct a topological ordering of $\mathcal{G}(\mathbf{B}_0) \cup \mathcal{G}(\mathbf{\Gamma}_0)$.

After constructing a topological ordering in the manner above, the parameters are recoverable node by node. For each $k = 1, \ldots, p$ and any $x_1, \ldots, x_{k-1} \in \mathbb{R}$, the conditional distribution of x_k given $(x_1, \ldots, x_{k-1})$ satisfies

$$x_k \mid (x_1, \ldots, x_{k-1}) \sim \mathrm{N}\left(\sum_{j<k} \beta_{jk,0} x_j, \sum_{j<k} \gamma_{jk,0} x_j^2 + 1 \right),$$

because $u_{1k}, \ldots, u_{k-1,k}$ and ε_k are independent mean-zero Gaussians and the u_{jk} are independent of $(x_1, \ldots, x_{k-1})$. Since $(x_1, \ldots, x_{k-1})$ has strictly positive density on $\mathbb{R}^{k-1}$, the conditional mean uniquely determines the fixed effects $\beta_{jk,0}$ for $j < k$, and the conditional variance uniquely determines the random-effect variances $\gamma_{jk,0}$ for $j < k$. Applying this argument for $k = 1, \ldots, p$ shows that every entry of $\mathbf{B}_0$ and $\mathbf{\Gamma}_0$ is uniquely determined by the observed distribution. Therefore, $(\mathbf{B}_0, \mathbf{\Gamma}_0)$ is identifiable, and so are $\mathcal{G}(\mathbf{B}_0)$, $\mathcal{G}(\mathbf{\Gamma}_0)$, and their union. ■

Appendix E. Proof of Theorem 5

The proof requires three technical lemmas. All lemmas are stated and proved under the assumptions of Theorem 5. Lemma 8 first shows that the population loss is uniquely minimized at the true parameters.

Lemma 8 *For all* $(\mathbf{B}, \mathbf{\Gamma}) \in \mathcal{F}$*, where* $\mathcal{F}$ *is defined in* (8)*, the population loss satisfies the inequality*

$$L(\mathbf{B}, \mathbf{\Gamma}) \geq L(\mathbf{B}_0, \mathbf{\Gamma}_0),$$

holding with equality if and only if $(\mathbf{B}, \mathbf{\Gamma}) = (\mathbf{B}_0, \mathbf{\Gamma}_0)$.

Proof Let $p_{\mathbf{B},\mathbf{\Gamma}}(\mathbf{X}^{(1)})$ denote the cluster-level density under $(\mathbf{B}, \mathbf{\Gamma})$. Since $\ell_m(\mathbf{B}, \mathbf{\Gamma})$ is exactly the average negative log-likelihood (after restoring additive and multiplicative constants), taking expectation gives

$$L(\mathbf{B}, \mathbf{\Gamma}) = -\mathrm{E}\left[\log\left\{p_{\mathbf{B},\mathbf{\Gamma}}(\mathbf{X}^{(1)})\right\}\right], \qquad L(\mathbf{B}_0, \mathbf{\Gamma}_0) = -\mathrm{E}\left[\log\left\{p_{\mathbf{B}_0,\mathbf{\Gamma}_0}(\mathbf{X}^{(1)})\right\}\right].$$

Taking the difference of the two terms above yields

$$L(\mathbf{B}, \mathbf{\Gamma}) - L(\mathbf{B}_0, \mathbf{\Gamma}_0) = \mathrm{E}\left[\log\left\{\frac{p_{\mathbf{B}_0,\mathbf{\Gamma}_0}(\mathbf{X}^{(1)})}{p_{\mathbf{B},\mathbf{\Gamma}}(\mathbf{X}^{(1)})}\right\}\right].$$

The right-hand side is the Kullback–Leibler divergence from $p_{\mathbf{B}_0,\mathbf{\Gamma}_0}$ to $p_{\mathbf{B},\mathbf{\Gamma}}$ and is therefore nonnegative. Equality holds if and only if $p_{\mathbf{B},\mathbf{\Gamma}}(\mathbf{X}^{(1)}) = p_{\mathbf{B}_0,\mathbf{\Gamma}_0}(\mathbf{X}^{(1)})$ almost surely, which by Theorem 4, implies $(\mathbf{B}, \mathbf{\Gamma}) = (\mathbf{B}_0, \mathbf{\Gamma}_0)$. ■

Lemma 9 provides a uniform convergence result for the empirical loss and its Hessian over compact parameter sets.

Lemma 9 *Let* $K \subset \mathbb{R}^{p\times p} \times \mathbb{R}_+^{p\times p}$ *be compact. Then it holds*

$$\sup_{(\mathbf{B},\mathbf{\Gamma})\in K} |\ell_m(\mathbf{B}, \mathbf{\Gamma}) - L(\mathbf{B}, \mathbf{\Gamma})| \xrightarrow{p} 0$$

and

$$\sup_{(\mathbf{B},\mathbf{\Gamma})\in K} \left\|\nabla^2 \ell_m(\mathbf{B}, \mathbf{\Gamma}) - \nabla^2 L(\mathbf{B}, \mathbf{\Gamma})\right\|_{\mathrm{op}} \xrightarrow{p} 0.$$

Proof For each cluster i, write $a(\mathbf{X}^{(i)}, \mathbf{B}, \mathbf{\Gamma})$ for the contribution of cluster i to the negative log-likelihood, so that

$$\ell_m(\mathbf{B}, \mathbf{\Gamma}) = \frac{1}{m}\sum_{i=1}^{m} a(\mathbf{X}^{(i)}, \mathbf{B}, \mathbf{\Gamma}).$$

Because $\mathbf{\Gamma} \geq \mathbf{0}$, each covariance matrix

$$\mathbf{V}^{(i)}(\boldsymbol{\gamma}_k) = \mathbf{I} + \mathbf{X}^{(i)} \operatorname{diag}(\boldsymbol{\gamma}_k)\mathbf{X}^{(i)\top} \succeq \mathbf{I},$$

so the log-determinant and inverse terms appearing in $a(\mathbf{X}^{(i)}, \mathbf{B}, \mathbf{\Gamma})$ are well-defined for every $(\mathbf{B}, \mathbf{\Gamma}) \in K$. Consequently, for almost every $\mathbf{X}^{(i)}$, the maps

$$(\mathbf{B}, \mathbf{\Gamma}) \mapsto a(\mathbf{X}^{(i)}, \mathbf{B}, \mathbf{\Gamma}), \qquad (\mathbf{B}, \mathbf{\Gamma}) \mapsto \nabla^2 a(\mathbf{X}^{(i)}, \mathbf{B}, \mathbf{\Gamma}),$$

are continuous on K. Next, since K is compact and $n_i \equiv n$, $a(\mathbf{X}^{(i)}, \mathbf{B}, \mathbf{\Gamma})$ and every scalar component of $\nabla^2 a(\mathbf{X}^{(i)}, \mathbf{B}, \mathbf{\Gamma})$ is bounded uniformly over $(\mathbf{B}, \mathbf{\Gamma}) \in K$ by a polynomial in $\|\mathbf{X}^{(i)}\|_F$. Under the Gaussian structural equation model, $\mathbf{X}^{(i)}$ has finite moments of all orders, so these bounds have finite expectation. The same bounds justify differentiation under the expectation, and therefore

$$\nabla^2 L(\mathbf{B}, \mathbf{\Gamma}) = \mathrm{E}\left[\nabla^2 a(\mathbf{X}^{(1)}, \mathbf{B}, \mathbf{\Gamma})\right].$$

Thus, the uniform law of large numbers (e.g., Newey and McFadden, 1994, Lemma 2.4), applied on the compact set K using the continuity and polynomial bounds above, gives

$$\sup_{(\mathbf{B}, \mathbf{\Gamma}) \in K} |\ell_m(\mathbf{B}, \mathbf{\Gamma}) - L(\mathbf{B}, \mathbf{\Gamma})| \xrightarrow{p} 0.$$

The same result also gives entrywise uniform convergence of $\nabla^2 \ell_m(\mathbf{B}, \mathbf{\Gamma})$ to $\nabla^2 L(\mathbf{B}, \mathbf{\Gamma})$ on K. Since p is fixed, the Hessian has finitely many entries, and it follows

$$\sup_{(\mathbf{B}, \mathbf{\Gamma}) \in K} \left\|\nabla^2 \ell_m(\mathbf{B}, \mathbf{\Gamma}) - \nabla^2 L(\mathbf{B}, \mathbf{\Gamma})\right\|_{\mathrm{op}} \xrightarrow{p} 0.$$

■

Finally, Lemma 10 establishes local strong convexity of the empirical loss in a neighborhood of the true parameter values.

Lemma 10 *There exist constants $c > 0$ and $\delta > 0$ such that, with probability tending to one, the empirical loss $\ell_m(\mathbf{B}, \mathbf{\Gamma})$ satisfies*

$$\ell_m(\mathbf{B}_1, \mathbf{\Gamma}_1) \geq \ell_m(\mathbf{B}_2, \mathbf{\Gamma}_2) + \langle \nabla \ell_m(\mathbf{B}_2, \mathbf{\Gamma}_2), (\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2) \rangle + \frac{c}{2} \|(\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2)\|_F^2$$

for all $(\mathbf{B}_1, \mathbf{\Gamma}_1), (\mathbf{B}_2, \mathbf{\Gamma}_2) \in \mathcal{F}$ such that

$$\|(\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_0, \mathbf{\Gamma}_0)\|_F \leq \delta \quad \textit{and} \quad \|(\mathbf{B}_2, \mathbf{\Gamma}_2) - (\mathbf{B}_0, \mathbf{\Gamma}_0)\|_F \leq \delta.$$

Proof By Assumption 2, $\nabla^2 L(\mathbf{B}_0, \mathbf{\Gamma}_0)$ is positive definite. Let

$$\lambda_0 := \lambda_{\min}\{\nabla^2 L(\mathbf{B}_0, \mathbf{\Gamma}_0)\} > 0.$$

Since $\nabla^2 L(\mathbf{B}, \mathbf{\Gamma})$ is continuous in a neighborhood of $(\mathbf{B}_0, \mathbf{\Gamma}_0)$ relative to $\mathbb{R}^{p \times p} \times \mathbb{R}_+^{p \times p}$, there exists $\delta > 0$ such that

$$\lambda_{\min}\{\nabla^2 L(\mathbf{B}, \mathbf{\Gamma})\} \geq \frac{\lambda_0}{2}$$

whenever

$$\|(\mathbf{B}, \mathbf{\Gamma}) - (\mathbf{B}_0, \mathbf{\Gamma}_0)\|_F \leq \delta.$$

Define the compact set

$$K_\delta := \left\{ (\mathbf{B}, \mathbf{\Gamma}) \in \mathbb{R}^{p \times p} \times \mathbb{R}_+^{p \times p} : \|(\mathbf{B}, \mathbf{\Gamma}) - (\mathbf{B}_0, \mathbf{\Gamma}_0)\|_F \leq \delta \right\},$$

and apply Lemma 9 to get

$$\sup_{(\mathbf{B}, \mathbf{\Gamma}) \in K_\delta} \left\| \nabla^2 \ell_m(\mathbf{B}, \mathbf{\Gamma}) - \nabla^2 L(\mathbf{B}, \mathbf{\Gamma}) \right\|_{\text{op}} \xrightarrow{p} 0.$$

Therefore, with probability tending to one, it holds

$$\sup_{\|(\mathbf{B}, \mathbf{\Gamma}) - (\mathbf{B}_0, \mathbf{\Gamma}_0)\|_F \leq \delta} \left\| \nabla^2 \ell_m(\mathbf{B}, \mathbf{\Gamma}) - \nabla^2 L(\mathbf{B}, \mathbf{\Gamma}) \right\|_{\text{op}} \leq \frac{\lambda_0}{4}.$$

On this event, we have

$$\lambda_{\min}\{\nabla^2 \ell_m(\mathbf{B}, \mathbf{\Gamma})\} \geq \lambda_{\min}\{\nabla^2 L(\mathbf{B}, \mathbf{\Gamma})\} - \left\| \nabla^2 \ell_m(\mathbf{B}, \mathbf{\Gamma}) - \nabla^2 L(\mathbf{B}, \mathbf{\Gamma}) \right\|_{\text{op}} \geq \frac{\lambda_0}{4}$$

for all $(\mathbf{B}, \mathbf{\Gamma})$ in the neighborhood. Define $c := \lambda_0/4$ and for any two points $(\mathbf{B}_1, \mathbf{\Gamma}_1)$ and $(\mathbf{B}_2, \mathbf{\Gamma}_2)$ in this neighborhood, define

$$(\mathbf{B}(t), \mathbf{\Gamma}(t)) := (\mathbf{B}_2, \mathbf{\Gamma}_2) + t\{(\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2)\}$$

for $t \in [0, 1]$. Since K_δ is convex, it holds $(\mathbf{B}(t), \mathbf{\Gamma}(t)) \in K_\delta$ for all $t \in [0, 1]$. Moreover, because $\mathbf{\Gamma}(t) \in \mathbb{R}_+^{p \times p}$, each covariance matrix appearing in $\ell_m(\mathbf{B}(t), \mathbf{\Gamma}(t))$ is positive definite, so $\ell_m(\mathbf{B}(t), \mathbf{\Gamma}(t))$ is twice continuously differentiable on $[0, 1]$. Therefore, Taylor's theorem with the integral form of the remainder yields

$$\ell_m(\mathbf{B}_1, \mathbf{\Gamma}_1) = \ell_m(\mathbf{B}_2, \mathbf{\Gamma}_2) + \langle \nabla \ell_m(\mathbf{B}_2, \mathbf{\Gamma}_2), (\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2) \rangle + \int_0^1 (1 - t) Q(t)\, dt,$$

where

$$Q(t) := \left\langle (\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2), \nabla^2 \ell_m((\mathbf{B}_2, \mathbf{\Gamma}_2) + t\{(\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2)\})\{(\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2)\} \right\rangle.$$

Using the uniform lower bound on the Hessian gives

$$Q(t) \geq c \|(\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2)\|_F^2$$

for all $t \in [0, 1]$. Therefore, it holds

$$\ell_m(\mathbf{B}_1, \mathbf{\Gamma}_1) \geq \ell_m(\mathbf{B}_2, \mathbf{\Gamma}_2) + \langle \nabla \ell_m(\mathbf{B}_2, \mathbf{\Gamma}_2), (\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2) \rangle + \frac{c}{2} \|(\mathbf{B}_1, \mathbf{\Gamma}_1) - (\mathbf{B}_2, \mathbf{\Gamma}_2)\|_F^2,$$

which completes the proof. ■

With these results in place, we are now ready to prove Theorem 5.

Proof

Part 1: Consistency of parameter estimation By the definition of F_m and the fact that $\|\mathbf{B}\|_\infty \le M_{\mathbf{B}}$, $\|\mathbf{\Gamma}\|_\infty \le M_{\mathbf{\Gamma}}$, and p is fixed, we have

$$\sup_{(\mathbf{B},\mathbf{\Gamma})\in\mathcal{F}} |F_m(\mathbf{B},\mathbf{\Gamma}) - L(\mathbf{B},\mathbf{\Gamma})| \le \sup_{(\mathbf{B},\mathbf{\Gamma})\in\mathcal{F}} |\ell_m(\mathbf{B},\mathbf{\Gamma}) - L(\mathbf{B},\mathbf{\Gamma})| + \lambda_1 p^2 M_{\mathbf{B}} + \lambda_2 p^2 M_{\mathbf{\Gamma}}.$$

We now show the two terms on the right-hand side converge to zero. Since $\mathcal{F}$ is compact, Lemma 9 implies

$$\sup_{(\mathbf{B},\mathbf{\Gamma})\in\mathcal{F}} |\ell_m(\mathbf{B},\mathbf{\Gamma}) - L(\mathbf{B},\mathbf{\Gamma})| \xrightarrow{p} 0.$$

Moreover, by Assumption 1 the penalty terms on the right-hand side vanish, and hence

$$z_m := \sup_{(\mathbf{B},\mathbf{\Gamma})\in\mathcal{F}} |F_m(\mathbf{B},\mathbf{\Gamma}) - L(\mathbf{B},\mathbf{\Gamma})| \xrightarrow{p} 0.$$

Now, fix any $\epsilon > 0$ and define the set

$$\mathcal{A}_\epsilon := \{(\mathbf{B},\mathbf{\Gamma}) \in \mathcal{F} : \|(\mathbf{B},\mathbf{\Gamma}) - (\mathbf{B}_0,\mathbf{\Gamma}_0)\|_F \ge \epsilon\}.$$

Since $\mathcal{F}$ is closed and bounded (since the set of acyclic matrices is closed), it is compact and hence $\mathcal{A}_\epsilon$ is also compact. Moreover, since L is continuous on $\mathcal{F}$ and, by Lemma 8, uniquely minimized at $(\mathbf{B}_0,\mathbf{\Gamma}_0)$, it follows

$$\Delta_\epsilon := \inf_{(\mathbf{B},\mathbf{\Gamma})\in\mathcal{A}_\epsilon} \{L(\mathbf{B},\mathbf{\Gamma}) - L(\mathbf{B}_0,\mathbf{\Gamma}_0)\} > 0.$$

Now, using that $(\hat{\mathbf{B}},\hat{\mathbf{\Gamma}})$ is the global minimizer of F_m, we have

$$L(\hat{\mathbf{B}},\hat{\mathbf{\Gamma}}) \le F_m(\hat{\mathbf{B}},\hat{\mathbf{\Gamma}}) + z_m \le F_m(\mathbf{B}_0,\mathbf{\Gamma}_0) + z_m \le L(\mathbf{B}_0,\mathbf{\Gamma}_0) + 2z_m.$$

Therefore, on the event $z_m < \Delta_\epsilon/2$, it holds

$$L(\hat{\mathbf{B}},\hat{\mathbf{\Gamma}}) < L(\mathbf{B}_0,\mathbf{\Gamma}_0) + \Delta_\epsilon. \tag{15}$$

Suppose for the sake of contradiction

$$\|(\hat{\mathbf{B}},\hat{\mathbf{\Gamma}}) - (\mathbf{B}_0,\mathbf{\Gamma}_0)\|_F \ge \epsilon,$$

then $(\hat{\mathbf{B}},\hat{\mathbf{\Gamma}}) \in \mathcal{A}_\epsilon$, implying

$$L(\hat{\mathbf{B}},\hat{\mathbf{\Gamma}}) \ge L(\mathbf{B}_0,\mathbf{\Gamma}_0) + \Delta_\epsilon,$$

which contradicts (15). Hence, whenever $z_m < \Delta_\epsilon/2$, it must hold

$$\|(\hat{\mathbf{B}},\hat{\mathbf{\Gamma}}) - (\mathbf{B}_0,\mathbf{\Gamma}_0)\|_F < \epsilon.$$

Since $z_m \xrightarrow{p} 0$, we have

$$\Pr\left(\|(\hat{\mathbf{B}},\hat{\mathbf{\Gamma}}) - (\mathbf{B}_0,\mathbf{\Gamma}_0)\|_F \ge \epsilon\right) \le \Pr(z_m \ge \Delta_\epsilon/2) \to 0,$$

and hence

$$(\hat{\mathbf{B}},\hat{\mathbf{\Gamma}}) \xrightarrow{p} (\mathbf{B}_0,\mathbf{\Gamma}_0),$$

thereby proving the first claim of the theorem.

Part 2: No false negatives Let $(j,k) \in \mathcal{S}_{\mathbf{B}}$. Since p is fixed, the set $\mathcal{S}_{\mathbf{B}}$ is finite. Define

$$c_{\mathbf{B}} := \min_{(a,b)\in\mathcal{S}_{\mathbf{B}}} |\beta_{ab,0}| > 0.$$

If $\hat{\beta}_{jk} = 0$, then $|\hat{\beta}_{jk} - \beta_{jk,0}| = |\beta_{jk,0}| \geq c_{\mathbf{B}}$. Hence, by Part 1, we have

$$\Pr(\hat{\beta}_{jk} = 0) \leq \Pr(|\hat{\beta}_{jk} - \beta_{jk,0}| \geq c_{\mathbf{B}}) \to 0.$$

Therefore, $\Pr(\hat{\beta}_{jk} \neq 0) \to 1$ for all $(j,k) \in \mathcal{S}_{\mathbf{B}}$. Since p is fixed, $\mathcal{S}_{\mathbf{B}}$ is finite and the union bound gives

$$\Pr\left(\exists\,(j,k) \in \mathcal{S}_{\mathbf{B}} : \hat{\beta}_{jk} = 0\right) \leq \sum_{(j,k)\in\mathcal{S}_{\mathbf{B}}} \Pr(\hat{\beta}_{jk} = 0) \to 0. \tag{16}$$

Similarly, let $(j,k) \in \mathcal{S}_{\mathbf{\Gamma}}$. Since p is fixed, the set $\mathcal{S}_{\mathbf{\Gamma}}$ is finite. Define

$$c_{\mathbf{\Gamma}} := \min_{(a,b)\in\mathcal{S}_{\mathbf{\Gamma}}} \gamma_{ab,0} > 0.$$

If $\hat{\gamma}_{jk} = 0$, then $|\hat{\gamma}_{jk} - \gamma_{jk,0}| = \gamma_{jk,0} \geq c_{\mathbf{\Gamma}}$. Hence, again by Part 1, we have

$$\Pr\left(\hat{\gamma}_{jk} = 0\right) \leq \Pr(|\hat{\gamma}_{jk} - \gamma_{jk,0}| \geq c_{\mathbf{\Gamma}}) \to 0,$$

which implies $\Pr(\hat{\gamma}_{jk} \neq 0) \to 1$ for all $(j,k) \in \mathcal{S}_{\mathbf{\Gamma}}$. The union bound therefore gives

$$\Pr(\exists\,(j,k) \in \mathcal{S}_{\mathbf{\Gamma}} : \hat{\gamma}_{jk} = 0) \leq \sum_{(j,k)\in\mathcal{S}_{\mathbf{\Gamma}}} \Pr(\hat{\gamma}_{jk} = 0) \to 0. \tag{17}$$

Combining (16) and (17) yields

$$\Pr\left(\hat{\beta}_{jk} \neq 0\,\forall\,(j,k) \in \mathcal{S}_{\mathbf{B}} \wedge \hat{\gamma}_{jk} \neq 0\,\forall\,(j,k) \in \mathcal{S}_{\mathbf{\Gamma}}\right) \to 1,$$

proving $(\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}})$ contains no false negatives.

Part 3: No false positives

Part 3(a) Define the structure-restricted feasible set

$$\mathcal{F}_{\mathcal{S}} := \{(\mathbf{B}, \mathbf{\Gamma}) \in \mathcal{F} : \beta_{jk} = 0\,\forall\,(j,k) \in \mathcal{S}_{\mathbf{B}}^{c},\ \gamma_{jk} = 0\,\forall\,(j,k) \in \mathcal{S}_{\mathbf{\Gamma}}^{c}\}.$$

This set is nonempty because $(\mathbf{B}_0, \mathbf{\Gamma}_0) \in \mathcal{F}_{\mathcal{S}}$. Since $\mathcal{G}(\mathbf{B}_0) \cup \mathcal{G}(\mathbf{\Gamma}_0)$ is a DAG, any spanning subgraph (i.e., a graph obtained by deleting edges) is also a DAG. Therefore, every element of $\mathcal{F}_{\mathcal{S}}$ is acyclic. In particular, for any $(\mathbf{B}, \mathbf{\Gamma}) \in \mathcal{F}_{\mathcal{S}}$, it holds

$$\mathcal{G}(\mathbf{B}) \cup \mathcal{G}(\mathbf{\Gamma}) \subseteq \mathcal{G}(\mathbf{B}_0) \cup \mathcal{G}(\mathbf{\Gamma}_0).$$

Let $(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}})$ be a global minimizer of F_m over $\mathcal{F}_{\mathcal{S}}$. We work on the event that $(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}})$ is close enough to $(\mathbf{B}_0, \mathbf{\Gamma}_0)$ such that:

- $\operatorname{sign}(\tilde{\beta}_{jk}) = \operatorname{sign}(\beta_{jk,0})$ and $|\tilde{\beta}_{jk}| < M_{\mathbf{B}}$ for all $(j,k) \in \mathcal{S}_{\mathbf{B}}$; and

- $0 < \tilde{\gamma}_{jk} < M_{\mathbf{\Gamma}}$ for all $(j,k) \in \mathcal{S}_{\mathbf{\Gamma}}$.

This event occurs with probability tending to one due to the same line of reasoning as in Part 1, but on $\mathcal{F}_{\mathcal{S}}$, together with the inequalities $\|\mathbf{B}_0\|_\infty < M_{\mathbf{B}}$ and $\|\mathbf{\Gamma}_0\|_\infty < M_{\mathbf{\Gamma}}$.

Define the gradients on the nonzero components as

$$\nabla_{\mathbf{B},\mathcal{S}_{\mathbf{B}}} \ell_m(\mathbf{B},\mathbf{\Gamma}) := \left(\frac{\partial}{\partial \beta_{jk}} \ell_m(\mathbf{B},\mathbf{\Gamma})\right)_{(j,k)\in\mathcal{S}_{\mathbf{B}}}$$

and

$$\nabla_{\mathbf{\Gamma},\mathcal{S}_{\mathbf{\Gamma}}} \ell_m(\mathbf{B},\mathbf{\Gamma}) := \left(\frac{\partial}{\partial \gamma_{jk}} \ell_m(\mathbf{B},\mathbf{\Gamma})\right)_{(j,k)\in\mathcal{S}_{\mathbf{\Gamma}}},$$

which can be stacked together as

$$\nabla_{\mathcal{S}} \ell_m(\mathbf{B},\mathbf{\Gamma}) := \begin{pmatrix} \nabla_{\mathbf{B},\mathcal{S}_{\mathbf{B}}} \ell_m(\mathbf{B},\mathbf{\Gamma}) \\ \nabla_{\mathbf{\Gamma},\mathcal{S}_{\mathbf{\Gamma}}} \ell_m(\mathbf{B},\mathbf{\Gamma}) \end{pmatrix}.$$

On the event above, the acyclicity constraint is automatic on $\mathcal{F}_{\mathcal{S}}$ (each such union graph is a spanning subgraph of the true union DAG), the box constraints are inactive on the nonzero components, and the nonnegative constraint on $\mathbf{\Gamma}$ is inactive on $\mathcal{S}_{\mathbf{\Gamma}}$. The first-order conditions for $(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}})$ on the nonzero components are

$$\nabla_{\mathbf{B},\mathcal{S}_{\mathbf{B}}} \ell_m(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}) + \lambda_1 \mathbf{s}_{\mathbf{B}} = \mathbf{0} \tag{18}$$

and, since $\gamma_{jk} \geq 0$ and nonzero $\tilde{\gamma}_{jk} > 0$,

$$\nabla_{\mathbf{\Gamma},\mathcal{S}_{\mathbf{\Gamma}}} \ell_m(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}) + \lambda_2 \mathbf{1} = \mathbf{0}. \tag{19}$$

Stacking these equations together gives

$$\nabla_{\mathcal{S}} \ell_m(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}) + \lambda_1 \begin{pmatrix} \mathbf{s}_{\mathbf{B}} \\ \lambda_2/\lambda_1 \mathbf{1} \end{pmatrix} = \mathbf{0}. \tag{20}$$

Part 3(b) Define the sets of acyclicity-preserving inactive indices

$$\mathcal{T}_{\mathbf{B}} := \{(j,k) \in \mathcal{S}_{\mathbf{B}}^c : \mathcal{G}(\mathbf{B}_0) \cup \mathcal{G}(\mathbf{\Gamma}_0) \cup \{j \to k\} \in \mathrm{DAG}_p\}$$

and

$$\mathcal{T}_{\mathbf{\Gamma}} := \{(j,k) \in \mathcal{S}_{\mathbf{\Gamma}}^c : \mathcal{G}(\mathbf{B}_0) \cup \mathcal{G}(\mathbf{\Gamma}_0) \cup \{j \to k\} \in \mathrm{DAG}_p\},$$

where $\{j \to k\}$ denotes the directed graph with node set $\{1,\ldots,p\}$ and edge set $\{(j,k)\}$. We show that the strict dual feasibility conditions hold with probability tending to one on the acyclicity-preserving inactive sets defined above. For indices $\mathcal{S}_{\mathbf{B}}^c \setminus \mathcal{T}_{\mathbf{B}}$ and $\mathcal{S}_{\mathbf{\Gamma}}^c \setminus \mathcal{T}_{\mathbf{\Gamma}}$, any nonzero value would create a cycle in the union graph and hence violate feasibility of $(\mathbf{B},\mathbf{\Gamma}) \in \mathcal{F}$ on the event that all true edges are present (i.e., the no false negatives event established in Part 2). It therefore suffices to establish dual feasibility on $\mathcal{T}_{\mathbf{B}}$ and $\mathcal{T}_{\mathbf{\Gamma}}$. Specifically, we show the following inequalities with probability tending to one:

- For every $(j,k) \in \mathcal{T}_{\mathbf{B}}$, it holds

$$\left| \frac{\partial}{\partial \beta_{jk}} \ell_m(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}) \right| \leq \left(1 - \frac{\eta}{2}\right) \lambda_1; \tag{21}$$

and

- For every $(j,k) \in \mathcal{T}_{\mathbf{\Gamma}}$, it holds

$$\frac{\partial}{\partial \gamma_{jk}} \ell_m(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}) > -\left(1 - \frac{\eta}{2}\right) \lambda_2. \tag{22}$$

These inequalities ensure strict dual feasibility for all inactive coordinates that admit feasible perturbations.

We begin by performing a Taylor expansion on the nonzero entries. Since $\mathbf{\Gamma} \in \mathbb{R}_+^{p \times p}$ on $\mathcal{F}$, each covariance matrix $\mathbf{V}^{(i)}(\boldsymbol{\gamma}_k)$ is positive definite, so $\log\det\{\mathbf{V}^{(i)}(\boldsymbol{\gamma}_k)\}$ and $\mathbf{V}^{-(i)}(\boldsymbol{\gamma}_k)$ are C^2 (indeed C^∞) on an open neighborhood of $\mathcal{F}$. Hence, $\ell_m(\mathbf{B}, \mathbf{\Gamma})$ is C^2 on an open neighborhood of $\mathcal{F}$, and we may expand $\nabla_{\mathcal{S}} \ell_m(\mathbf{B}, \mathbf{\Gamma})$ around $(\mathbf{B}_0, \mathbf{\Gamma}_0)$ as

$$\nabla_{\mathcal{S}} \ell_m(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}) = \nabla_{\mathcal{S}} \ell_m(\mathbf{B}_0, \mathbf{\Gamma}_0) + \mathbf{H}_{\mathcal{S}\mathcal{S}}^{(m)} \begin{pmatrix} \tilde{\mathbf{B}}_{\mathcal{S}_{\mathbf{B}}} - \mathbf{B}_{\mathcal{S}_{\mathbf{B}},0} \\ \tilde{\mathbf{\Gamma}}_{\mathcal{S}_{\mathbf{\Gamma}}} - \mathbf{\Gamma}_{\mathcal{S}_{\mathbf{\Gamma}},0} \end{pmatrix}, \tag{23}$$

where $\mathbf{H}_{\mathcal{S}\mathcal{S}}^{(m)}$ is the mean Hessian along the line segment between $(\mathbf{B}_0, \mathbf{\Gamma}_0)$ and $(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}})$ restricted to the nonzero components. Combining (20) and (23) gives

$$\mathbf{H}_{\mathcal{S}\mathcal{S}}^{(m)} \begin{pmatrix} \tilde{\mathbf{B}}_{\mathcal{S}_{\mathbf{B}}} - \mathbf{B}_{\mathcal{S}_{\mathbf{B}},0} \\ \tilde{\mathbf{\Gamma}}_{\mathcal{S}_{\mathbf{\Gamma}}} - \mathbf{\Gamma}_{\mathcal{S}_{\mathbf{\Gamma}},0} \end{pmatrix} = -\nabla_{\mathcal{S}} \ell_m(\mathbf{B}_0, \mathbf{\Gamma}_0) - \lambda_1 \begin{pmatrix} \mathbf{s}_{\mathbf{B}} \\ \lambda_2/\lambda_1 \mathbf{1} \end{pmatrix}.$$

Since $\mathcal{F}_{\mathcal{S}}$ is compact, Lemma 9 gives

$$\sup_{(\mathbf{B},\mathbf{\Gamma}) \in \mathcal{F}_{\mathcal{S}}} \left\| \nabla^2 \ell_m(\mathbf{B}, \mathbf{\Gamma}) - \nabla^2 L(\mathbf{B}, \mathbf{\Gamma}) \right\|_{\text{op}} \xrightarrow{p} 0.$$

Moreover, the same argument as in Part 1, but with minimization carried out over $\mathcal{F}_{\mathcal{S}}$, gives $(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}) \xrightarrow{p} (\mathbf{B}_0, \mathbf{\Gamma}_0)$. Together with continuity of $\nabla^2 L(\mathbf{B}, \mathbf{\Gamma})$, these two results imply that $\mathbf{H}_{\mathcal{S}\mathcal{S}}^{(m)} \xrightarrow{p} \mathbf{H}_{\mathcal{S}\mathcal{S}}$. Since $\mathbf{H}_{\mathcal{S}\mathcal{S}}$ is nonsingular (because $\mathbf{H}$ is positive definite by Assumption 2), $\mathbf{H}_{\mathcal{S}\mathcal{S}}^{(m)}$ is invertible with probability tending to one, and so

$$\begin{pmatrix} \tilde{\mathbf{B}}_{\mathcal{S}_{\mathbf{B}}} - \mathbf{B}_{\mathcal{S}_{\mathbf{B}},0} \\ \tilde{\mathbf{\Gamma}}_{\mathcal{S}_{\mathbf{\Gamma}}} - \mathbf{\Gamma}_{\mathcal{S}_{\mathbf{\Gamma}},0} \end{pmatrix} = -\mathbf{H}_{\mathcal{S}\mathcal{S}}^{-(m)} \left[\nabla_{\mathcal{S}} \ell_m(\mathbf{B}_0, \mathbf{\Gamma}_0) + \lambda_1 \begin{pmatrix} \mathbf{s}_{\mathbf{B}} \\ \lambda_2/\lambda_1 \mathbf{1} \end{pmatrix} \right].$$

Because the clusters are iid and satisfy $n_i \equiv n$, each coordinate of $\nabla \ell_m(\mathbf{B}_0, \mathbf{\Gamma}_0)$ can be written as an average of iid mean-zero clusterwise contributions with finite variance under the Gaussian structural equation model. Therefore, by Chebyshev's inequality, $\nabla_{\mathcal{S}} \ell_m(\mathbf{B}_0, \mathbf{\Gamma}_0) = O_p(m^{-1/2})$ coordinatewise. Further, $\mathbf{H}_{\mathcal{S}\mathcal{S}}^{-(m)} = O_p(1)$ and, by Assumption 1, $m^{-1/2} = o(\lambda_1)$. Hence, it holds

$$\begin{pmatrix} \tilde{\mathbf{B}}_{\mathcal{S}_{\mathbf{B}}} - \mathbf{B}_{\mathcal{S}_{\mathbf{B}},0} \\ \tilde{\mathbf{\Gamma}}_{\mathcal{S}_{\mathbf{\Gamma}}} - \mathbf{\Gamma}_{\mathcal{S}_{\mathbf{\Gamma}},0} \end{pmatrix} = -\lambda_1 \mathbf{H}_{\mathcal{S}\mathcal{S}}^{-(m)} \begin{pmatrix} \mathbf{s}_{\mathbf{B}} \\ \lambda_2/\lambda_1 \mathbf{1} \end{pmatrix} + o_p(\lambda_1). \tag{24}$$

We now perform a Taylor expansion on the zero entries. Take a zero fixed-effect entry $(j,k) \in \mathcal{T}_{\mathbf{B}}$. Expand its partial derivative around $(\mathbf{B}_0, \mathbf{\Gamma}_0)$ as

$$\frac{\partial}{\partial \beta_{jk}} \ell_m(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}) = \frac{\partial}{\partial \beta_{jk}} \ell_m(\mathbf{B}_0, \mathbf{\Gamma}_0) + \mathbf{h}_{jk,\mathcal{S}}^{(m,\mathbf{B})} \begin{pmatrix} \tilde{\mathbf{B}}_{\mathcal{S}_{\mathbf{B}}} - \mathbf{B}_{\mathcal{S}_{\mathbf{B}},0} \\ \tilde{\mathbf{\Gamma}}_{\mathcal{S}_{\mathbf{\Gamma}}} - \mathbf{\Gamma}_{\mathcal{S}_{\mathbf{\Gamma}},0} \end{pmatrix}, \tag{25}$$

where $\mathbf{h}_{jk,\mathcal{S}}^{(m,\mathbf{B})}$ is the mean cross-Hessian row between β_{jk} and the active entries. Substitute (24) into (25) to get

$$\frac{\partial}{\partial \beta_{jk}} \ell_m(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}) = \frac{\partial}{\partial \beta_{jk}} \ell_m(\mathbf{B}_0, \mathbf{\Gamma}_0) - \lambda_1 \mathbf{h}_{jk,\mathcal{S}}^{(m,\mathbf{B})} \mathbf{H}_{\mathcal{S}\mathcal{S}}^{-(m)} \begin{pmatrix} \mathbf{s}_{\mathbf{B}} \\ \lambda_2/\lambda_1 \mathbf{1} \end{pmatrix} + o_p(\lambda_1).$$

The first term on the right-hand side can be written as

$$\frac{\partial}{\partial \beta_{jk}} \ell_m(\mathbf{B}_0, \mathbf{\Gamma}_0) = \frac{\partial}{\partial \beta_{jk}} L(\mathbf{B}_0, \mathbf{\Gamma}_0) + \left\{ \frac{\partial}{\partial \beta_{jk}} \ell_m(\mathbf{B}_0, \mathbf{\Gamma}_0) - \frac{\partial}{\partial \beta_{jk}} L(\mathbf{B}_0, \mathbf{\Gamma}_0) \right\}.$$

Since $(\mathbf{B}_0, \mathbf{\Gamma}_0)$ uniquely minimizes L over $\mathcal{F}$ by Lemma 8, the population derivative vanishes along any direction that preserves acyclicity of the union graph. The term in brackets is an average of iid mean-zero clusterwise contributions with finite variance and is therefore $O_p(m^{-1/2})$ by Chebyshev's inequality. Hence, we have

$$\frac{\partial}{\partial \beta_{jk}} \ell_m(\mathbf{B}_0, \mathbf{\Gamma}_0) = O_p(m^{-1/2}) = o_p(\lambda_1),$$

where the final equality follows from Assumption 1. Also $\lambda_2/\lambda_1 \to \kappa$ by Assumption 1. Then, since $\mathbf{h}_{jk,\mathcal{S}}^{(m,\mathbf{B})} \xrightarrow{p} \mathbf{h}_{jk,\mathcal{S}}^{(\mathbf{B})}$, we have

$$\mathbf{h}_{jk,\mathcal{S}}^{(m,\mathbf{B})} \mathbf{H}_{\mathcal{S}\mathcal{S}}^{-(m)} \begin{pmatrix} \mathbf{s}_{\mathbf{B}} \\ \lambda_2/\lambda_1 \mathbf{1} \end{pmatrix} \xrightarrow{p} \mathbf{h}_{jk,\mathcal{S}}^{(\mathbf{B})} \mathbf{H}_{\mathcal{S}\mathcal{S}}^{-1} \begin{pmatrix} \mathbf{s}_{\mathbf{B}} \\ \kappa \mathbf{1} \end{pmatrix}.$$

By Assumption 2, the right-hand side has absolute value at most $1 - \eta$, therefore

$$\Pr\left(\left| \frac{\partial}{\partial \beta_{jk}} \ell_m(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}) \right| \leq \left(1 - \frac{\eta}{2}\right) \lambda_1 \right) \to 1$$

for each $(j,k) \in \mathcal{T}_{\mathbf{B}}$. Since p is fixed, $\mathcal{T}_{\mathbf{B}}$ is finite, so the above convergence holds uniformly over all such $(j,k) \in \mathcal{T}_{\mathbf{B}}$ by the union bound, proving (21).

The same line of argument as above applies to a zero random-effect variance entry $(j,k) \in \mathcal{T}_{\mathbf{\Gamma}}$, yielding

$$\frac{\partial}{\partial \gamma_{jk}} \ell_m(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}) = \frac{\partial}{\partial \gamma_{jk}} \ell_m(\mathbf{B}_0, \mathbf{\Gamma}_0) - \lambda_1 \mathbf{h}_{jk,\mathcal{S}}^{(m,\mathbf{\Gamma})} \mathbf{H}_{\mathcal{S}\mathcal{S}}^{-(m)} \begin{pmatrix} \mathbf{s}_{\mathbf{B}} \\ \lambda_2/\lambda_1 \mathbf{1} \end{pmatrix} + o_p(\lambda_1).$$

The first term on the right-hand side can be written as

$$\frac{\partial}{\partial \gamma_{jk}} \ell_m(\mathbf{B}_0, \mathbf{\Gamma}_0) = \frac{\partial}{\partial \gamma_{jk}} L(\mathbf{B}_0, \mathbf{\Gamma}_0) + \left\{ \frac{\partial}{\partial \gamma_{jk}} \ell_m(\mathbf{B}_0, \mathbf{\Gamma}_0) - \frac{\partial}{\partial \gamma_{jk}} L(\mathbf{B}_0, \mathbf{\Gamma}_0) \right\}.$$

Since $(\mathbf{B}_0, \mathbf{\Gamma}_0)$ uniquely minimizes L over $\mathcal{F}$ by Lemma 8, the right derivative of L is nonnegative along any direction that preserves acyclicity of the union graph. The term in brackets is an average of iid clusterwise contributions with finite variance and is therefore $O_p(m^{-1/2})$ by Chebyshev's inequality. Hence, we have

$$\frac{\partial}{\partial \gamma_{jk}} \ell_m(\mathbf{B}_0, \mathbf{\Gamma}_0) = \frac{\partial}{\partial \gamma_{jk}} L(\mathbf{B}_0, \mathbf{\Gamma}_0) + O_p(m^{-1/2}) \geq -o_p(\lambda_2),$$

where the final inequality follows by Assumption 1. Then, since $\mathbf{h}^{(m,\mathbf{\Gamma})}_{jk,\mathcal{S}} \xrightarrow{p} \mathbf{h}^{(\mathbf{\Gamma})}_{jk,\mathcal{S}}$, we have

$$\mathbf{h}^{(m,\mathbf{\Gamma})}_{jk,\mathcal{S}} \mathbf{H}^{-(m)}_{\mathcal{S}\mathcal{S}} \begin{pmatrix} \mathbf{s_B} \\ \lambda_2/\lambda_1 \mathbf{1} \end{pmatrix} \xrightarrow{p} \mathbf{h}^{(\mathbf{\Gamma})}_{jk,\mathcal{S}} \mathbf{H}^{-1}_{\mathcal{S}\mathcal{S}} \begin{pmatrix} \mathbf{s_B} \\ \kappa \mathbf{1} \end{pmatrix}.$$

Again, Assumption 2 upper bounds the right-hand side by $\kappa(1-\eta)$, therefore

$$\Pr\left(\frac{\partial}{\partial \gamma_{jk}} \ell_m(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}) > -\left(1 - \frac{\eta}{2}\right) \lambda_2 \right) \to 1$$

for each $(j,k) \in \mathcal{T}_{\mathbf{\Gamma}}$. As before, the above convergence holds uniformly over all $(j,k) \in \mathcal{T}_{\mathbf{\Gamma}}$ by the union bound, proving (22).

Part 3(c) By Part 1, $(\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}}) \xrightarrow{p} (\mathbf{B}_0, \mathbf{\Gamma}_0)$, and as explained in Part 3(b), $(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}) \xrightarrow{p} (\mathbf{B}_0, \mathbf{\Gamma}_0)$. Since $(\mathbf{B}_0, \mathbf{\Gamma}_0)$ is the unique minimizer of L over $\mathcal{F}$ by Lemma 8, it is also the unique minimizer of L over the subset $\mathcal{F}_{\mathcal{S}}$. Let $\delta > 0$ and $c > 0$ be the constants from Lemma 10. Hence, with probability tending to one, we have

$$\|(\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}}) - (\mathbf{B}_0, \mathbf{\Gamma}_0)\|_F \leq \delta \quad \text{and} \quad \|(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}) - (\mathbf{B}_0, \mathbf{\Gamma}_0)\|_F \leq \delta.$$

On this event, Lemma 10 with $(\mathbf{B}_1, \mathbf{\Gamma}_1) = (\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}})$ and $(\mathbf{B}_2, \mathbf{\Gamma}_2) = (\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}})$ yields

$$\ell_m(\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}}) - \ell_m(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}) \geq \left\langle \nabla \ell_m(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}), (\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}}) - (\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}) \right\rangle + \frac{c}{2} \|(\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}}) - (\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}})\|_F^2.$$

Adding the difference of ℓ_1 penalties to both sides gives

$$\begin{aligned} F_m(\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}}) - F_m(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}) \geq & \sum_{(j,k) \in \mathcal{S}_{\mathbf{B}}} \left[\frac{\partial}{\partial \beta_{jk}} \ell_m(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}})(\hat{\beta}_{jk} - \tilde{\beta}_{jk}) + \lambda_1(|\hat{\beta}_{jk}| - |\tilde{\beta}_{jk}|) \right] \\ & + \sum_{(j,k) \in \mathcal{S}^c_{\mathbf{B}}} \left[\frac{\partial}{\partial \beta_{jk}} \ell_m(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}) \hat{\beta}_{jk} + \lambda_1 |\hat{\beta}_{jk}| \right] \\ & + \sum_{(j,k) \in \mathcal{S}_{\mathbf{\Gamma}}} \left[\frac{\partial}{\partial \gamma_{jk}} \ell_m(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}})(\hat{\gamma}_{jk} - \tilde{\gamma}_{jk}) + \lambda_2(\hat{\gamma}_{jk} - \tilde{\gamma}_{jk}) \right] \\ & + \sum_{(j,k) \in \mathcal{S}^c_{\mathbf{\Gamma}}} \left[\frac{\partial}{\partial \gamma_{jk}} \ell_m(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}) \hat{\gamma}_{jk} + \lambda_2 \hat{\gamma}_{jk} \right] \\ & + \frac{c}{2} \|(\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}}) - (\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}})\|_F^2. \end{aligned}$$

By (18) and the inequality

$$|a| - |b| \geq \text{sign}(b)(a - b),$$

the sum over $\mathcal{S}_{\mathbf{B}}$ is nonnegative. By (19), the sum over $\mathcal{S}_{\mathbf{\Gamma}}$ is equal to zero. By (21), we have

$$\frac{\partial}{\partial \beta_{jk}} \ell_m(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}) \hat{\beta}_{jk} + \lambda_1 |\hat{\beta}_{jk}| \geq \frac{\eta}{2} \lambda_1 |\hat{\beta}_{jk}|$$

for all $(j,k) \in \mathcal{T}_{\mathbf{B}}$. Likewise, since $\hat{\gamma}_{jk} \geq 0$ and (22) holds, we have

$$\frac{\partial}{\partial \gamma_{jk}} \ell_m(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}) \hat{\gamma}_{jk} + \lambda_2 \hat{\gamma}_{jk} \geq \frac{\eta}{2} \lambda_2 \hat{\gamma}_{jk}$$

for all $(j,k) \in \mathcal{T}_{\mathbf{\Gamma}}$. On the event from Part 2 (no false negatives), feasibility of $(\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}}) \in \mathcal{F}$ implies

$$\hat{\beta}_{jk} = 0 \, \forall \, (j,k) \in \mathcal{S}_{\mathbf{B}}^c \setminus \mathcal{T}_{\mathbf{B}} \quad \text{and} \quad \hat{\gamma}_{jk} = 0 \, \forall \, (j,k) \in \mathcal{S}_{\mathbf{\Gamma}}^c \setminus \mathcal{T}_{\mathbf{\Gamma}},$$

since turning on any such coordinate would create a directed cycle in the union graph. It follows

$$F_m(\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}}) - F_m(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}) \geq \frac{\eta}{2} \lambda_1 \sum_{(j,k) \in \mathcal{T}_{\mathbf{B}}} |\hat{\beta}_{jk}| + \frac{\eta}{2} \lambda_2 \sum_{(j,k) \in \mathcal{T}_{\mathbf{\Gamma}}} \hat{\gamma}_{jk} + \frac{c}{2} \|(\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}}) - (\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}})\|_F^2.$$

Since $(\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}})$ is the global minimizer of F_m over $\mathcal{F}$ and $(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}) \in \mathcal{F}_{\mathcal{S}} \subseteq \mathcal{F}$, it holds

$$F_m(\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}}) \leq F_m(\tilde{\mathbf{B}}, \tilde{\mathbf{\Gamma}}).$$

Combining the previous inequalities, we conclude

$$\sum_{(j,k) \in \mathcal{T}_{\mathbf{B}}} |\hat{\beta}_{jk}| = 0 \quad \text{and} \quad \sum_{(j,k) \in \mathcal{T}_{\mathbf{\Gamma}}} \hat{\gamma}_{jk} = 0,$$

with probability tending to one. Together with the feasibility implications above, this result gives

$$\hat{\beta}_{jk} = 0 \, \forall \, (j,k) \in \mathcal{S}_{\mathbf{B}}^c \quad \text{and} \quad \hat{\gamma}_{jk} = 0 \, \forall \, (j,k) \in \mathcal{S}_{\mathbf{\Gamma}}^c,$$

proving $(\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}})$ contains no false positives.

Part 4: Consistency of structure recovery Part 2 guarantees $(\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}})$ contain no false negatives. Part 3 guarantees $(\hat{\mathbf{B}}, \hat{\mathbf{\Gamma}})$ contain no false positives. It immediately follows

$$\Pr\Big(\mathcal{G}(\hat{\mathbf{B}}) = \mathcal{G}(\mathbf{B}_0) \text{ and } \mathcal{G}(\hat{\mathbf{\Gamma}}) = \mathcal{G}(\mathbf{\Gamma}_0)\Big) \to 1,$$

thereby proving the second claim of the theorem.

■

Appendix F. Implementation Details

F.1 Sparsity parameters

The sparsity parameters λ_1 and λ_2 are taken equal and swept over a grid of 30 values equally spaced on a logarithmic grid between 1 and 10^{-3}. As with all differentiable approaches, the weighted adjacency matrices produced by `MixDAG` can contain small nonzero values that violate acyclicity. These violations arise because gradient descent does not produce exact zeros when finite stopping rules are used. We thus follow the common practice of hard thresholding (e.g., Zheng et al., 2018; Bello et al., 2022) and set any elements of $\mathbf{B}$ and $\mathbf{\Gamma}$ that are below 0.05 in absolute value to zero. All baselines are thresholded similarly.

F.2 Polishing

A drawback to ℓ_1-induced sparsity is that it biases the estimates of $\mathbf{B}$ and $\mathbf{\Gamma}$ towards zero. This bias is a consequence of the shrinkage inherent to the ℓ_1 norm and is a well-documented side effect of ℓ_1 penalties (Hastie et al., 2020). We correct for this bias in all baselines through a "polishing" step. Specifically, after fitting the model, we retain the selected edges and reoptimize over this fixed structure, removing the ℓ_1 penalties (and the DAG constraint since the solution is already acyclic). Let $\mathbf{M}^{\mathbf{B}}, \mathbf{M}^{\mathbf{\Gamma}} \in \{0,1\}^{p\times p}$ denote binary masks with entries

$$M^{\mathbf{B}}_{jk} := 1(\hat{\beta}_{jk} \neq 0), \qquad M^{\mathbf{\Gamma}}_{jk} := 1(\hat{\gamma}_{jk} \neq 0).$$

We then solve

$$\min_{\mathbf{B}\in\mathbb{R}^{p\times p}, \mathbf{\Gamma}\in\mathbb{R}^{p\times p}_{+}} l(\mathbf{M}^{\mathbf{B}} \odot \mathbf{B}, \mathbf{M}^{\mathbf{\Gamma}} \odot \mathbf{\Gamma}).$$

This polishing step is typically fast, as it optimizes over a reduced parameter space and involves the negative log-likelihood only and not the acyclicity constraint or ℓ_1 penalties.

F.3 Computational Environment

The experiments are run on a Linux machine with eight NVIDIA RTX 4090 graphics cards. Each experimental run used a single graphics card, with multiple runs performed in parallel. Our implementations are in Python and built on PyTorch (Paszke et al., 2019).

Appendix G. Additional Experiments

We assess the robustness of `MixDAG` to a range of alternative simulation designs. In particular, we vary the graph sparsity, the graph type, and the cluster regime. Figures 6 and 7 show results for denser graphs with $s = 2p$ and $s = 3p$ edges, respectively, rather than the $s = p$ edges used in the main experiments. Figures 8 and 9 show results for scale-free graphs (Barabási and Albert, 1999) with degree exponents 2 and 3, respectively, in place of the Erdős–Rényi graphs used in the main experiments. Finally, Figures 10 and 11 show results for alternative cluster regimes with fixed average cluster size ($m = \lceil n_\bullet/10 \rceil$) and fixed number of clusters ($m = 10$), respectively, instead of the regime in the main experiments where both the number of clusters and cluster sizes increase with $n_\bullet$. Across all designs, `MixDAG` continues to perform well relative to the competing methods, indicating the gains in the main experiments are not specific to a particular graph sparsity, graph type, or cluster regime.

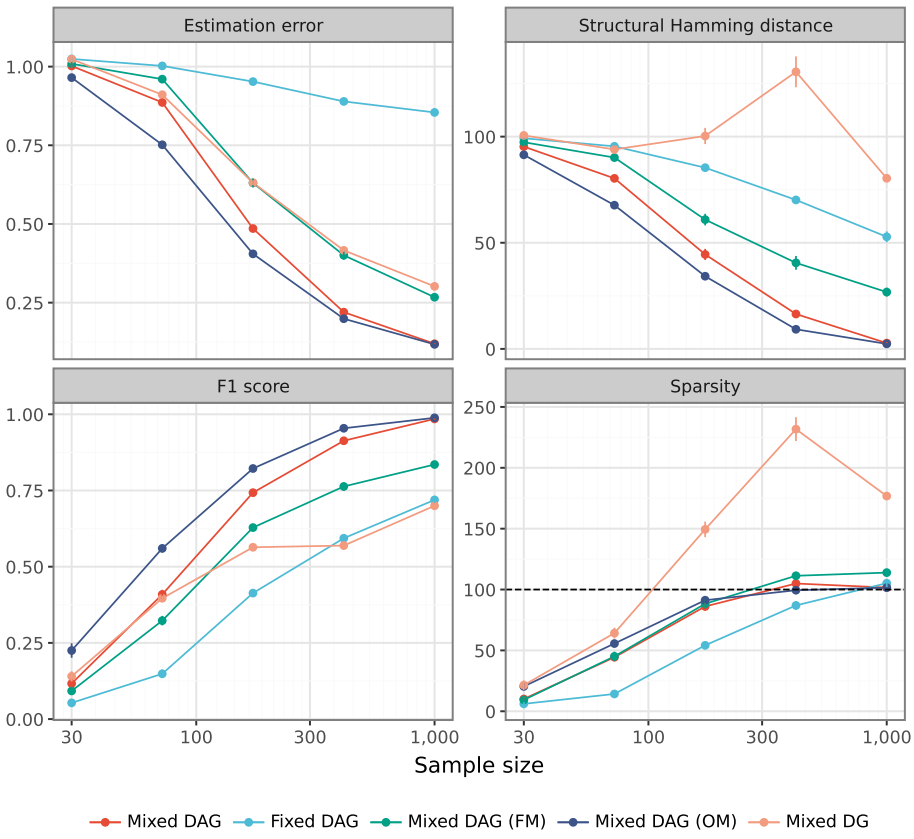


Figure 6: Performance on synthetic data generated from Erdős–Rényi DAGs with $p = 50$ nodes, $s = 100$ edges, and $m = \lceil \sqrt{n_\bullet} \rceil$ clusters. Averages (solid points) and standard errors (error bars) are measured over 30 datasets. The dashed horizontal line in the bottom right panel indicates the number of edges in the true DAG.

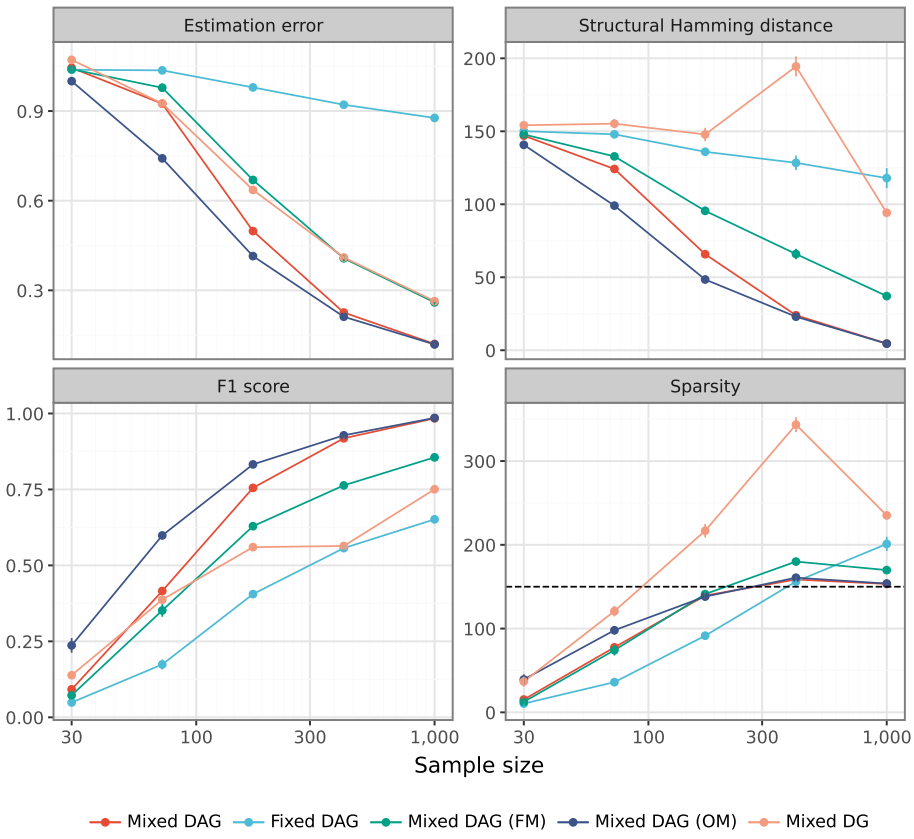


Figure 7: Performance on synthetic data generated from Erdős–Rényi DAGs with $p = 50$ nodes, $s = 150$ edges, and $m = \lceil \sqrt{n_\bullet} \rceil$ clusters. Averages (solid points) and standard errors (error bars) are measured over 30 datasets. The dashed horizontal line in the bottom right panel indicates the number of edges in the true DAG.

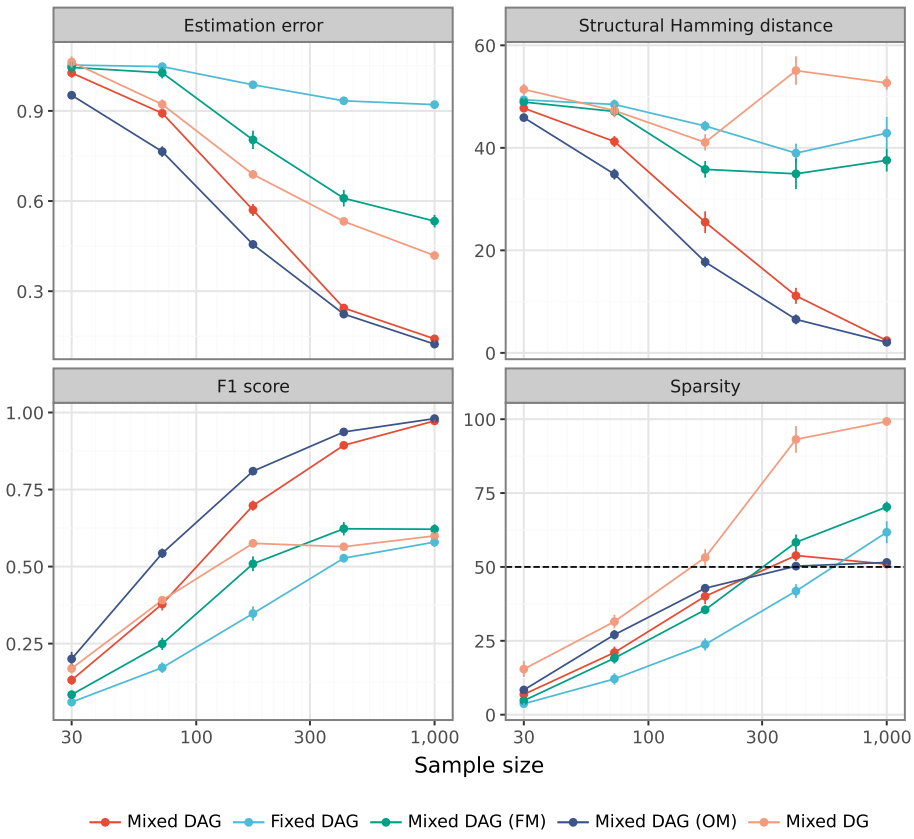


Figure 8: Performance on synthetic data generated from scale-free (degree exponent 2) DAGs with $p = 50$ nodes, $s = 50$ edges, and $m = \lceil \sqrt{n_\bullet} \rceil$ clusters. Averages (solid points) and standard errors (error bars) are measured over 30 datasets. The dashed horizontal line in the bottom right panel indicates the number of edges in the true DAG.

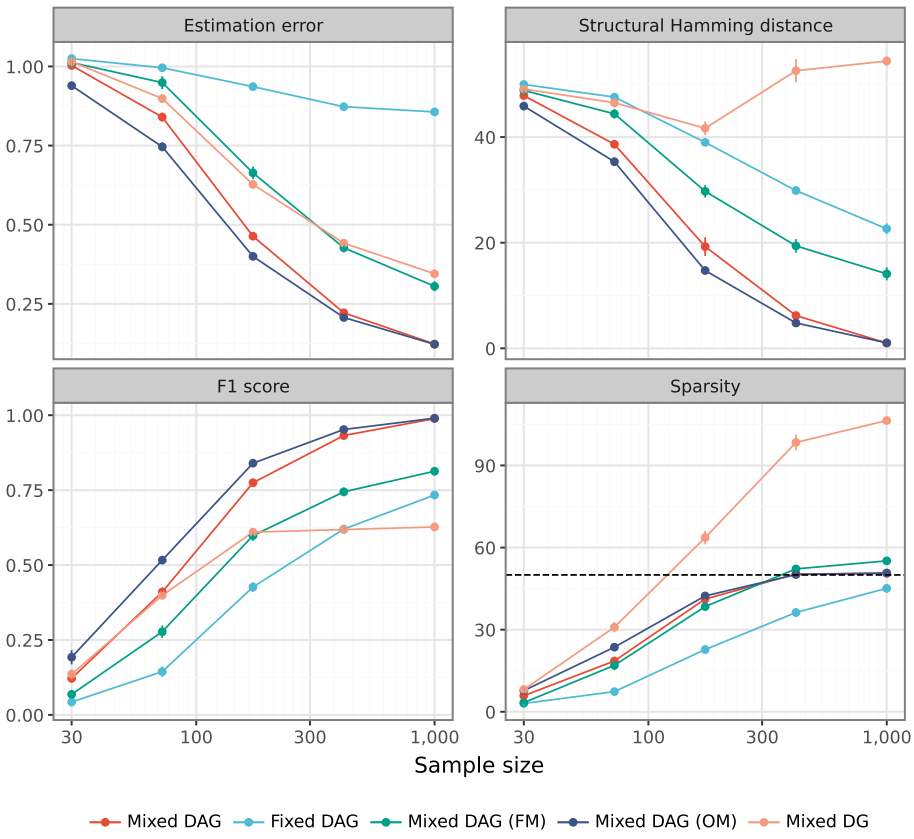


Figure 9: Performance on synthetic data generated from scale-free (degree exponent 3) DAGs with $p = 50$ nodes, $s = 50$ edges, and $m = \lceil \sqrt{n_\bullet} \rceil$ clusters. Averages (solid points) and standard errors (error bars) are measured over 30 datasets. The dashed horizontal line in the bottom right panel indicates the number of edges in the true DAG.

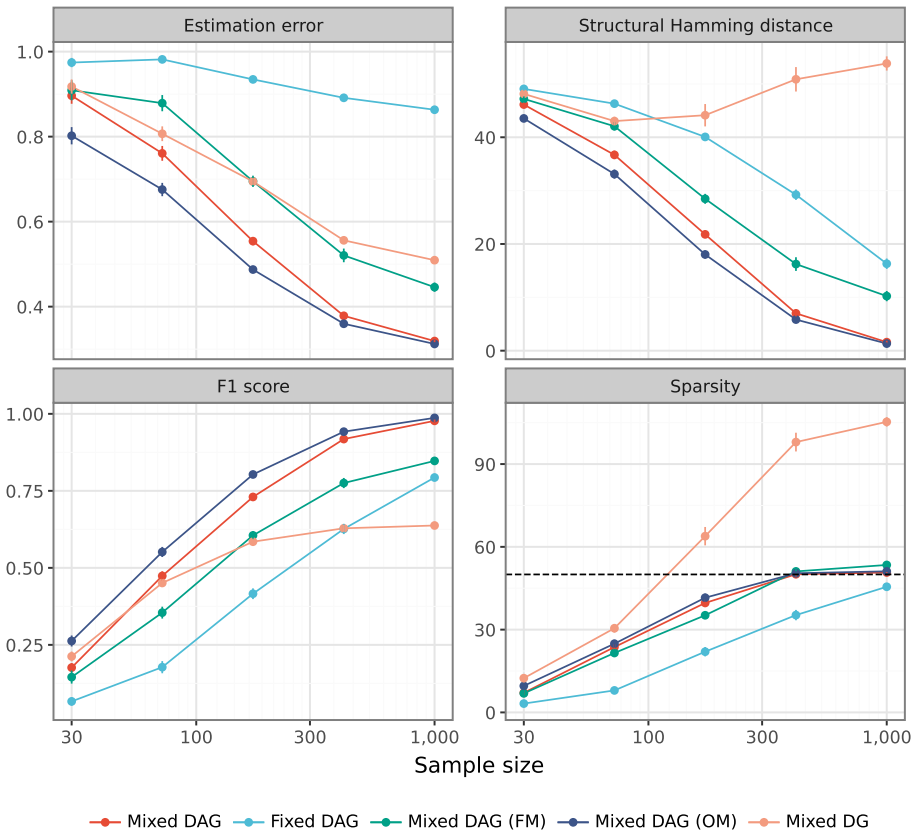


Figure 10: Performance on synthetic data generated from Erdős–Rényi DAGs with $p = 50$ nodes, $s = 50$ edges, and $m = \lceil n_\bullet/10 \rceil$ clusters. Averages (solid points) and standard errors (error bars) are measured over 30 datasets. The dashed horizontal line in the bottom right panel indicates the number of edges in the true DAG.

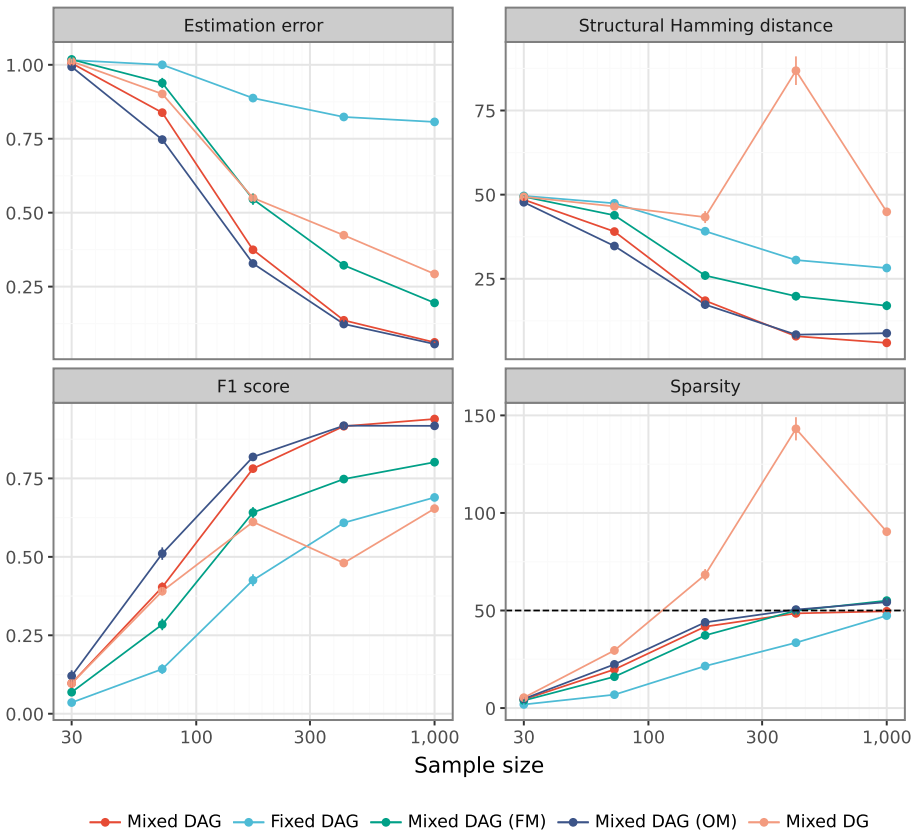


Figure 11: Performance on synthetic data generated from Erdős–Rényi DAGs with $p = 50$ nodes, $s = 50$ edges, and $m = 10$ clusters. Averages (solid points) and standard errors (error bars) are measured over 30 datasets. The dashed horizontal line in the bottom right panel indicates the number of edges in the true DAG.

References

Harold Bae, Stefano Monti, Monty Montano, Martin H Steinberg, Thomas T Perls, and Paola Sebastiani. Learning Bayesian networks from correlated data. *Scientific Reports*, 6, 2016.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. DAGMA: Learning DAGs via M-matrices and a log-determinant acyclicity characterization. In *Advances in Neural Information Processing Systems*, volume 35, pages 8226–8239, 2022.

Cristina Conati, Abigail S Gertner, Kurt Vanlehn, and Marek J Druzdzel. On-line student modeling for coached problem solving using bayesian networks. In *User Modeling: Proceedings of the Sixth International Conference*, pages 231–242, 1997.

Chang Deng, Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. Optimizing NOTEARS objectives via topological swaps. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 7563–7595, 2023a.

Chang Deng, Kevin Bello, Pradeep Ravikumar, and Bryon Aragam. Global optimality in bivariate gradient-based DAG learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 17929–17968, 2023b.

Chang Deng, Kevin Bello, Pradeep Ravikumar, and Bryon Aragam. Markov equivalence and consistency in differentiable structure learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 91756–91797, 2024.

P Erdős and A Rényi. On random graphs. I. *Publicationes Mathematicae Debrecen*, 6(3-4): 290–297, 1959.

Yingying Fan and Runze Li. Variable selection in linear mixed effects models. *Annals of Statistics*, 40(4):2043–2068, 2012.

E Michael Foster. Causal inference and developmental psychology. *Developmental Psychology*, 46(6):1454–1480, 2010.

David A Harville. *Matrix Algebra From a Statistician's Perspective*. Springer, New York, NY, USA, 1997.

Trevor Hastie, Robert Tibshirani, and Ryan Tibshirani. Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons. *Statistical Science*, 35(4):579–592, 2020.

Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21:1–53, 2020.

Guido W Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4): 1129–1179, 2020.

Jiming Jiang and Thuan Nguyen. *Linear and Generalized Linear Mixed Models and Their Applications*. Springer, New York, NY, USA, 2nd edition, 2021.

Neville Kenneth Kitson, Anthony C Constantinou, Zhigao Guo, Yang Liu, and Kiattikun Chobtham. A survey of Bayesian network structure learning. *Artificial Intelligence Review*, 56(8):8721–8814, 2023.

S L Lauritzen and D J Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–224, 1988.

Xiang Li, Shanghong Xie, Peter McColgan, Sarah J Tabrizi, Rachael I Scahill, Donglin Zeng, and Yuanjia Wang. Learning subject-specific directed acyclic graphs with mixed effects structural equation models from observational data. *Frontiers in Genetics*, 9, 2018.

Carl D Meyer. *Matrix Analysis and Applied Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2000.

Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier, 1994.

Yang Ni, Francesco C Stingo, and Veerabhadran Baladandayuthapani. Bayesian graphical regression. *Journal of the American Statistical Association*, 114(525):184–197, 2019.

Chris J Oates, Jim Q Smith, Sach Mukherjee, and James Cussens. Exact estimation of multiple directed acyclic graphs. *Statistics and Computing*, 26(4):797–811, 2016.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, CA, USA, 1988.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

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

Nicholas G Polson, James G Scott, and Brandon T Willard. Proximal algorithms in statistics and machine learning. *Statistical Science*, 30(4):559–581, 2015.

Basil Saeed, Snigdha Panigrahi, and Caroline Uhler. Causal structure discovery from distributions arising from mixtures of DAGs. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 8336–8345, 2020.

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

Aleksei Sholokhov, James V Burke, Damian F Santomauro, Peng Zheng, and Aleksandr Aravkin. A relaxation approach to feature selection for linear mixed effects models. *Journal of Computational and Graphical Statistics*, 33(1):261–275, 2024.

Giora Simchoni and Saharon Rosset. Using random effects to account for high-cardinality categorical features and repeated measures in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 25111–25122, 2021.

Giora Simchoni and Saharon Rosset. Integrating random effects in deep neural networks. *Journal of Machine Learning Research*, 24:1–57, 2023.

Giora Simchoni and Saharon Rosset. Integrating random effects in variational autoencoders for dimensionality reduction of correlated data, 2024. arXiv:2412.16899.

Giora Simchoni and Saharon Rosset. Flexible copula-based mixed models in deep learning: A scalable approach to arbitrary marginals. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pages 91–99, 2025.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, USA, 2nd edition, 2000.

Peter W G Tennant, Eleanor J Murray, Kellyn F Arnold, Laurie Berrie, Matthew P Fox, Sarah C Gadd, Wendy J Harrison, Claire Keeble, Lynsie R Ranker, Johannes Textor, Georgia D Tomova, Mark S Gilthorpe, and George T H Ellison. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: Review and recommendations. *International Journal of Epidemiology*, 50(2):620–632, 2021.

Ryan Thompson, Edwin V Bonilla, and Robert Kohn. Contextual directed acyclic graphs. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pages 2872–2880, 2024.

Ryan Thompson, Edwin V Bonilla, and Robert Kohn. ProDAG: Projected variational inference for directed acyclic graphs. In *Advances in Neural Information Processing Systems*, volume 38, pages 162714–162739, 2025a.

Ryan Thompson, Matt P Wand, and Joanna J J Wang. Scalable subset selection in linear mixed models, 2025b. arXiv:2506.20425.

Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like DAGs? A survey on structure learning and causal discovery. *ACM Computing Surveys*, 55(4):1–36, 2022.

Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

Yunyang Xiong, Hyunwoo J Kim, and Vikas Singh. Mixed effects neural networks (MeNets) with applications to gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7743–7752, 2019.

Zhen Zhang, Ignavier Ng, Dong Gong, Yuhang Liu, Ehsan M Abbasnejad, Mingming Gong, Kun Zhang, and Javen Qinfeng Shi. Truncated matrix power iteration for differentiable DAG learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 18390–18402, 2022.

Zhen Zhang, Ignavier Ng, Dong Gong, Yuhang Liu, Mingming Gong, Biwei Huang, Kun Zhang, Anton van den Hengel, and Javen Qinfeng Shi. Analytic DAG constraints for differentiable DAG learning. In *International Conference on Learning Representations*, 2025.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric P Xing. Learning sparse nonparametric DAGs. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108, pages 3414–3425, 2020.