Probabilistic Edge Modulation for High-Dimensional Causal Discovery

Seyong Hwang[1] **Kyoungjae Lee**[2] **Sunmin Oh**[3,4] **Gunwoong Park**[*,3,4,5]

[1]Department of Statistics, University of Michigan - Ann Arbor, U.S.A.,
[2]Department of Statistics, Sungkyunkwan University, South Korea
[3]Department of Statistics, Seoul National University, South Korea
[4]Institute for Data Innovation in Science, Seoul National University, South Korea
[5]Interdisciplinary Program in Artificial Intelligence, Seoul National University, South Korea
[*]Correspondence to: gw.park23@gmail.com ,
[*]*All authors contributed equally to this work.*

Abstract

Causal discovery in high-dimensional linear Bayesian networks is challenging, even with partial structural knowledge. Such information is often edge-specific and noisy, and naively enforcing uniform shrinkage or hard constraints can induce incorrect or unstable edge selection. We propose Probabilistic Edge Modulation (PEM), a principled probabilistic framework that replaces hard structural constraints with soft, edge-specific modulation via a spike-and-slab formulation. PEM integrates heterogeneous priors into ordering recovery and parent selection through a unified MAP formulation that remains computationally tractable in polynomial time. We establish high-dimensional consistency under both sub-Gaussian and heavy-tailed errors with bounded moments, and demonstrate robustness to prior misspecification. Experiments on synthetic and real retail data demonstrate improved structural stability and graph recovery in sparse and data-limited regimes; in a real e-commerce dataset, probabilistic transfer from a data-rich group stabilizes smaller groups and yields nontrivial graphs that do not collapse to near-empty graphs.

1 INTRODUCTION

Causal discovery in high-dimensional linear Bayesian networks (LBNs) remains fundamentally challenging, even when partial structural knowledge is available. Such information may arise from domain expertise, semantic or business rules, transfer across related datasets, or large language model-assisted knowledge extraction. It is commonly used to restrict the space of admissible graphs Vashishtha et al. [2023], Darvariu et al. [2024], Ghassami et al. [2018], Rodrıguez-López and Sucar [2022], Constantinou et al. [2023]. However, in learning LBNs, prior knowledge is often incorporated through heuristic rules, hard constraints, or algorithm-specific modifications, or omitted altogether [Shimizu et al., 2006, 2011, Ghoshal and Honorio, 2018, Chen et al., 2019, Park et al., 2021, Gao et al., 2022]. These approaches lack a unified probabilistic framework for handling uncertainty in prior information and typically lack high-dimensional consistency guarantees under heterogeneous or misspecified priors.

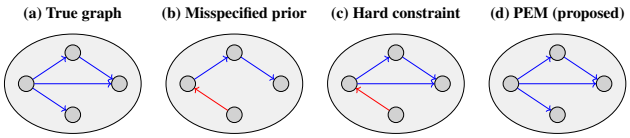


Figure 1: Conceptual illustration of prior-guided causal discovery under imperfect prior information. (a) True graph. (b) External prior containing partial error (red edge). (c) Hard-constraint-based estimation, which restricts the search space and propagates prior error. (d) Soft, edge-specific modulation (PEM), which allows prior information to guide estimation probabilistically and reduces sensitivity to prior misspecification.

Integrating uncertain prior information while preserving high-dimensional consistency is particularly challenging in settings with sparse and zero-inflated data—such as online retail and recommender systems—where limited samples provide weak signals relative to the underlying dependencies. In such regimes, existing structure learning methods face two key limitations. First, prior knowledge is often

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:2314–2335.

Table 1: Design space of structure learning methods for high-dimensional DAG models.

Domain	Method	DAG	Prior			Poly-Time	High-Dimensional Guarantees		
			Edge	Soft	Hard		Gaussian	Heavy-tail	No indeg.
Linear Bayesian networks	**This work (PEM-LBN)**	✓	✓	✓	✓	✓	✓	✓	✓
	Loh and Bühlmann [2014]	✓	✗	✗	✗	✓	✓	△	✓
	Ghoshal and Honorio [2018] *(LISTEN)*	✓	✗	✗	✗	✓	✓	✓	✓
	Chen et al. [2019] *(Top-Down)*	✓	✗	✗	✗	✗	✓	△	✗
	Park et al. [2021] *(HLSM)*	✓	✗	✗	✗	✓	✓	✓	✓
	Gao et al. [2022]	✓	✗	✗	✗	✗	✓	✗	✗
	Shimizu et al. [2011] (DirectLiNGAM)	✓	✗	✗	✗	✓	✗	✗	✗
	Tashiro et al. [2014] (ParcelLiNGAM)	✓	✓	✗	✓	✓	✗	✗	✗
Continuous optimization	Zheng et al. [2018] *(NOTEARS)*	✓	✗	✗	✗	✗	✗	✗	-
	Ng et al. [2020] *(GOLEM)*	✓	✗	✗	✗	✗	✓	✗	✗
Bayesian DAG	Friedman and Koller [2003]	✓	✗	✓	✗	✗	✗	✗	-
	Kuipers et al. [2022]	✓	✗	✓	✗	✗	✗	✗	-
	Zhou and Chang [2023]	✓	✗	✓	✗	✗	✓	✗	✗
MEC recovery	Perković et al. [2017]	✗	✗	✗	✓	✗	✗	✗	-
	Ejaz and Bareinboim [2025] *(LGES)*	✗	✓	✓	✓	✓	✓	✗	✗

Legend. ✓ (*yes*) indicates that the property is explicitly supported by the method; ✗ (*no*) indicates that it is not supported; △ indicates a conditional guarantee holding only under additional assumptions (e.g., identifiability conditions, sparsity, or bounded indegree); "–" denotes not applicable.
DAG: recovery of a single directed acyclic graph rather than an equivalence class. **Edge**: use of edge-specific prior information. **Soft**: incorporation of prior information through probabilistic weighting or regularization. **Hard**: imposition of deterministic constraints (e.g., forbidden or required edges). **Poly-Time**: existence of a polynomial-time algorithmic guarantee. **Gaussian / Heavy-tail**: high-dimensional consistency guarantees under Gaussian or non-Gaussian (heavy-tailed) noise. **No indeg.**: guarantees do not require a known maximum indegree bound.

incorporated as hard constraints that deterministically fix or exclude edges, reducing robustness when prior information is uncertain or noisy (e.g., Anjum et al., 2009, Amirkhani et al., 2016). Second, many existing methods do not provide an explicit mechanism to encode edge-specific prior inclusion strengths; instead, shrinkage is introduced through global [Friedman et al., 2008, Banerjee and Ghosal, 2015] or edge-wise adaptive penalty parameters [Zhou et al., 2009]. As a result, heterogeneous uncertainty across edges cannot be directly expressed, and uniform shrinkage may overshrink weak but meaningful edges or fail to adequately suppress spurious ones in finite samples. These contrasting mechanisms are illustrated in Figure 1.

We propose Probabilistic Edge Modulation (PEM), a precision-based framework for a GH-type identifiable class of LBNs, in which a topological ordering is recovered through recursive comparisons of diagonal precision entries Ghoshal and Honorio [2018]. This regime is plausible when variables are measured on comparable scales or arise from a similar domain Chen et al. [2019], Park et al. [2021], Gao et al. [2022]. Within this regime, PEM integrates heterogeneous edge-specific prior uncertainty directly into ordering-based high-dimensional LBN learning while preserving polynomial-time estimation and high-dimensional consistency guarantees. Specifically, PEM incorporates edge-specific inclusion strengths into a unified MAP estimator through a spike-and-slab formulation, where inclusion probabilities act as adaptive regularization weights within a single precision-driven optimization problem. Unlike prior approaches that treat inclusion probabilities primarily as sparsity-control mechanisms, PEM embeds them into the estimation procedure that jointly determines ordering and parent selection. As a result, prior information affects not only local edge inclusion but also the diagonal precision comparisons that govern ordering recovery. Table 1 situates PEM within the broader design space of structure learning methods along the axes of prior informativeness and flexibility.

Our contributions are summarized as follows:

- **Soft, edge-specific probabilistic modulation.** We encode uncertain prior information through heterogeneous inclusion strengths, enabling selective shrinkage without deterministic structural constraints.
- **Polynomial-time estimation.** The resulting MAP estimator admits a polynomial-time implementation.
- **Prior-adaptive high-dimensional consistency.** We establish structural consistency under sub-Gaussian and heavy-tailed noise without requiring a known maximum indegree bound, and show that informative edge-specific inclusion strengths relax the finite-sample burden of distinguishing the diagonal precision gaps for

ordering and off-diagonal precision signals for parent recovery.

2 BACKGROUND

Linear Bayesian Network. An LBN is defined by the structural equation

$$X = BX + \epsilon, \tag{1}$$

where $X = (X_1, \ldots, X_p)^\top$, $B \in \mathbb{R}^{p \times p}$ is a weighted adjacency matrix of a directed acyclic graph (DAG), i.e., $B_{j,k} \neq 0$ only if there is a directed edge from k to j. The noise vector $\epsilon = (\epsilon_1, \ldots, \epsilon_p)^\top$ consists of mutually independent variables with $\mathbb{E}[\epsilon_j] = 0$ and $\mathrm{Var}(\epsilon_j) = \sigma_j^2 > 0$. We assume causal sufficiency, i.e., there are no latent confounders among the observed variables.

Let Σ denote the population covariance matrix of X. The corresponding precision matrix $\Omega = \Sigma^{-1}$ admits the factorization

$$\Omega = (I_p - B)^\top (\Sigma^\epsilon)^{-1} (I_p - B), \tag{2}$$

where $\Sigma^\epsilon = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. For a matrix A, we write $A_{S,T}$ for the submatrix indexed by node sets $S, T \subseteq [p]$ and $A_{j,k}$ for the entry indexed by nodes $j, k \in [p]$. This representation motivates a class of structure learning methods that infer the DAG through estimation of the precision matrix (e.g., Peters et al., 2012, Ghoshal and Honorio, 2018, Chen et al., 2019, Park and Kim, 2020, Park et al., 2021, Gao et al., 2022).

Identifiability. The precision matrix alone is generally insufficient to identify the underlying DAG, and additional structural conditions are required. One well-known sufficient condition is the equal error variance assumption [Peters et al., 2012], which provides strong guarantees but restricts noise heterogeneity. Identifiability is therefore naturally characterized through ordering-based criteria driven by conditional precision (uncertainty) levels. The following lemma states an identifiability condition proposed in Ghoshal and Honorio [2018].

Lemma 2.1 (Identifiability) *Consider a linear Bayesian network* (1) *with DAG G and a topological ordering π. For $r = 1, \ldots, p$, let $T_r = \{\pi_1, \ldots, \pi_r\}$. Let $\mathrm{An}(v)$ denote the set of ancestors of node v in G. If for every $\ell \in \mathrm{An}(\pi_r)$,*

$$[\Omega_{T_r,T_r}]_{\pi_r,\pi_r} < [\Omega_{T_r,T_r}]_{\ell,\ell},$$

then G is uniquely identifiable.

Ordering-based LBN Learning. In ordering-based LBN learning, ordering is driven by comparisons of diagonal precision entries, while parent selection relies on the magnitudes of off-diagonal precision estimates. This asymmetry condition reflects the sequential elimination structure underlying ordering recovery. At each elimination step, the conditioning set T_r may include not only parents but also descendants and other variables. Under equal or comparable error variances—an assumption commonly encountered in homogeneous time-series (e.g., VAR) models, spatial autoregressive systems, or standardized Gaussian graphical models—conditioning on additional descendants reduces conditional variance. In contrast, a true source variable, whose parents are not yet included in the conditioning set, retains relatively larger conditional uncertainty. This strict separation in conditional variance translates into a strict ordering in conditional precision.

Although several methods achieve high-dimensional guarantees [Ghoshal and Honorio, 2018, Chen et al., 2019, Park et al., 2021, Gao et al., 2022], their finite-sample behavior can be unstable. Ordering recovery relies on comparisons of diagonal entries of sub-precision matrices, and small covariance estimation errors propagated through matrix inversion can reverse ordering decisions.

Prior Incorporation Approaches. Prior incorporation in structure learning has been addressed through several paradigms. Hard structural constraints enforce must-have or forbidden edges directly [MEEK, 1995, Perković et al., 2017]. While effective when correct, they do not represent uncertainty and may exclude the true DAG under misspecification. Regularization-based methods introduce global shrinkage through a shared penalty parameter [Friedman et al., 2008, Banerjee and Ghosal, 2015]. Such shrinkage can stabilize estimation and promote sparsity, but it applies uniformly across edges and therefore does not differentiate edge-specific prior reliability. When empirical precision signals are close due to estimation noise, a single global penalty may provide limited flexibility for separating weak-but-true edges from spurious ones. Continuous optimization approaches [Zheng et al., 2018, Ng et al., 2020] modify objective functions but do not directly couple prior inclusion probabilities with precision-based ordering recovery. Fully Bayesian methods [Friedman and Koller, 2003, Kuipers et al., 2022, Chaudhuri et al., 2025] provide probabilistic interpretations but rely on combinatorial search or MCMC and are not naturally aligned with scalable MAP-based estimation.

3 PRIOR INFORMATION MODELING: PROBABILISTIC EDGE MODULATION

We introduce probabilistic edge modulation (PEM), in which prior information enters precision estimation through edge-specific inclusion strengths within a single MAP objective. This edge-level modulation directly guides parent selection via off-diagonal shrinkage and, through the coupled LogDet optimization, can also affect the diagonal ordering

scores used for ordering recovery.

Edge-specific Inclusion Strengths. We consider edge-specific prior information about potential DAG adjacencies, which may be directional, noisy, or incomplete. Such information influences structure learning only through precision estimation, which provides the ordering signal required by Lemma 2.1. Directional information is encoded via plausibility scores $w_{k\to j} \in [0,1]$, which are translated into undirected inclusion strengths by

$$\eta_{j,k} = \max\{w_{j\to k},\, w_{k\to j}\}, \qquad j < k.$$

If the directional prior is highly accurate, some orientation may indeed be lost. However, when the prior is noisy, using either direction as soft evidence of pairwise relevance can be more stable than enforcing a possible incorrect orientation, while the final direction is still through the ordering step.

Spike-and-slab Precision Modeling. We introduce PEM by adapting a spike-and-slab Laplace prior [Ročková and George, 2018, Gan et al., 2019] for precision matrix estimation to allow edge-specific inclusion strengths $\eta_{j,k}$ $(j < k)$:

$$\begin{aligned} \Omega_{j,k} \mid r_{j,k} = 0 &\sim \text{Laplace}(\nu_0), \\ \Omega_{j,k} \mid r_{j,k} = 1 &\sim \text{Laplace}(\nu_1), \quad \nu_1 > \nu_0 > 0, \\ r_{j,k} &\sim \text{Bern}(\eta_{j,k}), \quad 0 \le \eta_{j,k} \le 1, \end{aligned} \tag{3}$$

where the diagonal entries are assigned an exponential prior $\pi(\Omega_{j,j}) \propto \exp(-\tau\Omega_{j,j})$, and Ω is restricted to be positive definite.

Under this formulation, the inclusion strength $\eta_{j,k}$ directly modulates shrinkage. Edges with larger $\eta_{j,k}$ are more likely to be drawn from the slab component (weaker shrinkage), whereas smaller $\eta_{j,k}$ favor the spike component (stronger shrinkage). After marginalizing over $r_{j,k}$, this construction induces edge-specific regularization, yielding heterogeneous and probabilistic modulation of precision entries.

The parameters (ν_0, ν_1) determine the separation between spike and slab regimes, with $\nu_1 > \nu_0$ ensuring differential attenuation of noise-like versus signal-like coefficients. The diagonal penalty parameter τ controls regularization on conditional variance components and is taken small so that diagonal precision entries primarily reflect structural variance rather than shrinkage. Note that deterministic hard constraints correspond to the boundary cases $\eta_{j,k} \in \{0,1\}$, in which an edge is forced to be excluded or included with probability one. Hence, PEM generalizes hard-constraint methods within a continuous probabilistic framework.

PEM-MAP Precision Estimator. Given n independent samples $X^{1:n} := (X^{(i)})_{i=1}^n$, let $\widehat{\Sigma}$ denote the full sample covariance matrix. At elimination step r, we restrict attention to the active set $S(r)$ and define $\widehat{\Sigma}^{(r)} = \widehat{\Sigma}_{S^{(r)},S^{(r)}}$, the submatrix of the full sample covariance matrix restricted to this active set. Under the Gaussian likelihood in (1) and the prior in (3), we define the PEM-MAP precision estimator on the active set as

$$\widehat{\Omega}^{(r)} = \underset{\substack{\Theta \succ 0,\ \|\Theta\|_2 \le B_0 \\ \Theta \in \mathbb{R}^{(p+1-r)\times(p+1-r)}}}{\arg\min} \left\{ \frac{n}{2} D_{\text{LD}}(\widehat{\Sigma}^{(r)} \| \Theta^{-1}) + \mathcal{P}_\eta(\Theta) \right\}. \tag{4}$$

Here $D_{\text{LD}}(A\|B) = \text{tr}(AB^{-1}) - \log\det(AB^{-1}) - p$ denotes the LogDet Bregman divergence, and $\mathcal{P}_\eta(\Theta)$ is the negative log-prior induced by (3).

Impact on Diagonal Entries. Although PEM modulates shrinkage at the edge level, its effect is not confined to off-diagonal entries. The PEM-MAP estimator (4) is obtained by minimizing a coupled LogDet objective over $\Theta \succ 0$, where off-diagonal shrinkage and diagonal magnitude must jointly adjust to satisfy positive definiteness and best fit the data. Consequently, changing the inclusion strengths $\eta_{j,k}$ alters the optimal pattern of off-diagonal attenuation, which in turn induces systematic shifts in the diagonals through the coupled optimization. This coupling is essential in ordering-based LBN learning: since each elimination step compares diagonal entries of $\widehat{\Omega}^{(r)}$, edge-specific modulation provides a principled way to influence the relative ranking of these ordering scores when competing nodes have close conditional-precision values. In particular, heterogeneous $\eta_{j,k}$ can stabilize ordering decisions in finite samples by selectively suppressing spurious off-diagonal interactions that would otherwise distort diagonal precision comparisons.

From Sparsity to Identifiability. The PEM-MAP estimator builds on the Bayesian precision modeling framework of Gan et al. [2019], but shifts the theoretical focus of precision estimation. Rather than establishing support recovery for off-diagonal entries, we analyze the full precision matrix, including the diagonal elements that govern variable ordering in linear Bayesian networks. The theoretical objective therefore moves from sparsity consistency to ordering-based identifiability. Edge-specific inclusion strengths operate within this framework to modulate precision entries in a way that directly affects ordering recovery.

4 PEM-AUGMENTED LBN LEARNING

We perform LBN structure recovery via a backward-selection procedure in which precision estimation is carried out using the PEM-MAP formulation (4). Given edge-specific inclusion strengths $\{\eta_{j,k}\}$ and n independent samples $X^{1:n}$, the procedure proceeds iteratively. At iteration r, the PEM-MAP precision estimator is computed on the active set $S(r)$. The next variable in the ordering is selected according to the diagonal precision comparison underlying Lemma 2.1. Its parent set is then determined using posterior inclusion probabilities induced by the spike-and-slab prior. The full procedure is summarized in Algorithm 1.

Algorithm 1 PEM-LBN Learning Algorithm

Require: i.i.d. samples $X^{1:n}$, threshold $0 < T < 1$, edge-specific inclusion strengths $\{\eta_{j,k}\}$.

Ensure: Estimated DAG $\widehat{G} = (V, \widehat{E})$.

1: Initialize $\widehat{\pi}_{p+1} \leftarrow \emptyset$.
2: **for** $r = 1, \ldots, p-1$ **do**
3: Set $S(r) \leftarrow V \setminus \{\widehat{\pi}_{p+2-r}, \ldots, \widehat{\pi}_{p+1}\}$.
4: Estimate precision $\widehat{\Omega}^{(r)}$ for $X_{S(r)}$ by solving (4).
5: Select ordering element

$$\widehat{\pi}_{p+1-r} \leftarrow \arg\min_{j \in S(r)} [\widehat{\Omega}^{(r)}]_{j,j}.$$

6: Estimate parent set

$$\widehat{\mathrm{Pa}}(\widehat{\pi}_{p+1-r}) \leftarrow \left\{k \in S(r) \backslash \{\widehat{\pi}_{p+1-r}\} : p^{(r)}_{\widehat{\pi}_{p+1-r},k} \geq T\right\},$$

where $p^{(r)}_{j,k}$ denotes the posterior inclusion probability.
7: **end for**
8: **return** $\widehat{E} = \bigcup_{r=1}^{p-1} \left\{(k, \widehat{\pi}_{p+1-r}) : k \in \widehat{\mathrm{Pa}}(\widehat{\pi}_{p+1-r})\right\}$.

Specifically, the posterior inclusion probability of edge (j,k) at iteration r admits the closed-form expression

$$p^{(r)}_{j,k} = \frac{\frac{\eta_{j,k}}{2\nu_1} \exp\left(-\frac{|\widehat{\Omega}^{(r)}_{j,k}|}{\nu_1}\right)}{\frac{\eta_{j,k}}{2\nu_1} \exp\left(-\frac{|\widehat{\Omega}^{(r)}_{j,k}|}{\nu_1}\right) + \frac{1-\eta_{j,k}}{2\nu_0} \exp\left(-\frac{|\widehat{\Omega}^{(r)}_{j,k}|}{\nu_0}\right)}.$$

This quantity is evaluated at the MAP estimate $\widehat{\Omega}^{(r)}$ under the spike-and-slab model (3) (see Appendix B for details).

Given a threshold $T \in (0,1)$ for parent selection, edges are selected whenever $p^{(r)}_{j,k} \geq T$. Unless otherwise specified, $T = 0.5$ is used, corresponding to the median probability model [Barbieri and Berger, 2004]; similar thresholding rules are common in Bayesian graphical modeling (e.g., Lee and Cao, 2021, Banerjee and Ghosal, 2015).

Computational Complexity. The computational bottleneck of the r-th iteration lies in solving (4) to estimate the precision matrix on the active set $S(r)$, whose size is $|S(r)| = p + 1 - r$. Following the complexity discussion of the BAGUS estimator [Gan et al., 2019], the per-iteration cost is dominated by matrix operations on a $|S(r)| \times |S(r)|$ matrix, resulting in $O(|S(r)|^3)$ time per iteration (up to solver-dependent constants). Summing over $r = 1, \ldots, p-1$ yields $\sum_{j=2}^{p} O(j^3) = O(p^4)$.

This $O(p^4)$ complexity is not a limitation unique to PEM-LBN, but rather reflects a structural property of ordering-based LBN learning, in which precision matrix estimation must be repeatedly performed on shrinking variable sets. In this respect, PEM-LBN is computationally comparable to existing polynomial-time methods such as LISTEN and BHLSM, while remaining substantially more scalable than MCMC-based Bayesian DAG learning or best-subset-selection approaches, which are infeasible beyond relatively small graph sizes.

5 HIGH-DIMENSIONAL CONSISTENCY

We formalize the key probabilistic assumptions underlying our model to establish high-dimensional consistency of the proposed PEM-LBN algorithm 1. High-dimensional LBN recovery requires two complementary ingredients: (i) structural separation for identifiability, and (ii) stability of precision estimation under sparsity constraints. These conditions ensure that the structure estimation procedure remains well-posed under edge-level uncertainty and sparse graph regimes, and are standard in the analysis of high-dimensional LBNs with sub-Gaussian and $4m$-th bounded-moment error distributions (e.g., Ghoshal and Honorio, 2018, Chen et al., 2019, Park et al., 2021, Gao et al., 2022) (see Appendix F for the exact definition of error distributions). We formalize these requirements through the following assumptions.

Assumption 1 (Structural Separation)

(i) There exist constants $k_1, k_2 > 0$ such that

$$\Lambda_{\min}(\Sigma) \geq k_1, \quad \textit{and} \quad \max_{1 \leq j \leq p} \Sigma_{j,j} < k_2.$$

(ii) (Minimum gap) For any $r \in \{2, \ldots, p\}$, and $\ell \in An(\pi_r)$, there exists a constant $\tau_{\min} > 0$ such that

$$\Omega^{(r)}_{\ell,\ell} - \Omega^{(r)}_{\pi_r,\pi_r} > \tau_{\min}.$$

(iii) (Minimum signal) For any $r \in \{1, \ldots, p-1\}$, there exists a constant $\theta_{\min} > 0$ such that

$$\min_{j \in Pa(\pi_{p+1-r})} \left|\Omega^{(r)}_{j,\pi_{p+1-r}}\right| > \theta_{\min},$$

where $\Lambda_{min}(\cdot)$ denotes the minimum eigenvalue and $\Omega^{(r)}$ is the precision matrix of $\{X_{\pi_1}, \ldots, X_{\pi_{p+1-r}}\}$.

Assumption 1 collectively enforces structural separation and identifiability. In particular, condition (i) ensures invertibility of the covariance matrix, condition (ii) guarantees sufficient diagonal separation for correct ordering recovery, and condition (iii) imposes a minimum signal requirement for reliable parent identification. While Assumption 1 addresses identifiability, high-dimensional consistency further requires uniform control of precision estimation errors. To this end, we impose the following stability condition. For ease of notation, let $S = \{(i,j) : \Omega_{i,j} \neq 0\}$ and $S^{(r)} = \{(i,j) : \Omega^{(r)}_{i,j} \neq 0\}$.

Assumption 2 (Estimation Stability) *There exist positive constants M_1 and M_2, independent of p, such that*

$$\|\Sigma\|_\infty \le M_1, \quad \max_{1\le r\le p-1} \|(\Omega^{(r)}\otimes\Omega^{(r)})_{S^{(r)},S^{(r)}}\|_\infty \le M_2,$$

where $\|\cdot\|_\infty$ denotes the ℓ_∞/ℓ_∞ operator norm.

Assumption 2 addresses estimation stability by controlling the ℓ_∞/ℓ_∞ operator norm of the restricted precision Hessian uniformly over elimination steps, thereby ensuring stable support recovery of the penalized precision estimator. Examples demonstrating that Assumptions 1 and 2 are satisfied under representative graph structures are provided in Appendix E.

Theorem 5.1 (High-dimensional Consistency) *Consider an LBN* (1) *with sub-Gaussian or $4m$-th bounded-moment errors. Suppose that Assumptions 1 and 2 hold. Then, for some hyper-parameters $(\nu_0, \nu_1, \{\eta_{j,k}\}, \tau)$ and threshold T (specified in Appendix C), Algorithm 1 recovers the true graph G with high probability. In particular,*

- *for sub-Gaussian errors,*

$$\Pr(\widehat{G} = G) \ge 1 - 4p^2 \exp\left(-C_{sub}\frac{n}{(d_M+1)^2}\right),$$

- *for $4m$-th bounded-moment errors,*

$$\Pr(\widehat{G} = G) \ge 1 - C_{bm}\, p^2 \frac{(d_M+1)^{2m}}{n^m},$$

where d_M denotes the maximum degree of the moralized graph of G, obtained by connecting nodes that share a common child and then dropping edge directions, and C_{sub} and C_{bm} are positive constants.

Theorem 5.1 establishes consistent structure recovery in high-dimensional regimes, with sample complexities $n = \Omega(d_M^2 \log p)$ under sub-Gaussian errors and $n = \Omega(d_M^2 p^{2/m})$ under $4m$-th bounded-moment errors. These rates are of the same order as those of existing high-dimensional LBN learning methods (see Appendix A for comparison and Appendix D for the full proof).

Although the scaling in n is comparable across methods, the structural requirements differ. As d_M increases, diagonal separation may shrink and stability conditions tighten, making dense graphs intrinsically harder to recover. Many high-dimensional procedures rely on incoherence- or compatibility-type constraints, whose constants may deteriorate rapidly as d_M increases. In contrast, PEM-LBN relies on stepwise separation and restricted operator norm control, without imposing irrepresentable-type incoherence conditions. Although stability must hold uniformly over iteration-specific subproblems, the required control is restricted to the active supports rather than the entire precision matrix, which makes its dependence on d_M comparatively milder.

Benefit of Informative η. More informative edge-specific inclusion strengths tend to strengthen the separation conditions required for ordering and parent recovery. In particular, accurate prior modulation reduces the effective noise level in off-diagonal precision entries, thereby stabilizing diagonal comparisons and facilitating satisfaction of Assumption 1. In the extreme case of fully informative inclusion strengths, recovery becomes deterministic.

Proposition 5.2 (Oracle inclusion strengths) *Suppose that the inclusion strengths satisfy*

$$\eta_{j,k} = \begin{cases} 1, & (j,k) \in E, \\ 0, & (j,k) \notin E, \end{cases}$$

where E is an edge set. Then Algorithm 1 returns the true graph G.

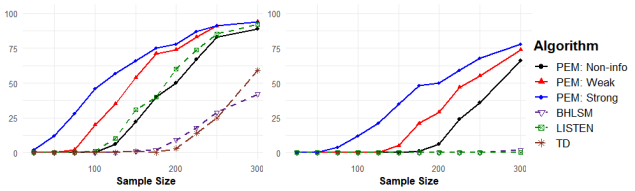
6 NUMERICAL EXPERIMENTS

This section evaluates the empirical performance of Algorithm 1 for learning linear Bayesian networks under both Gaussian and heavy-tailed noise. The experiments are designed to assess two aspects: (i) the practical benefit of incorporating edge-wise prior information in high-dimensional regimes, especially as graph density increases and the structural separation conditions in Section 5 become harder to satisfy, and (ii) the robustness of the method to misspecification in the prior inclusion probabilities $\eta_{j,k}$.

6.1 EXPERIMENTAL SETUP

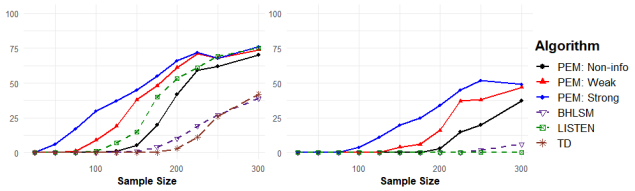
We report results over 100 realizations of p-node LBNs. DAGs were generated as Erdös–Rényi graphs with edge probability $q = \min(1, 3d_M/p)$ under maximum degree $d_M \in \{3, 6\}$ [Ghoshal and Honorio, 2017, Park et al., 2021]. Non-zero coefficients were sampled from $\beta_{k,j} \in (-1.0, -0.5) \cup (0.5, 1.0)$, and Gaussian noise follows $\epsilon_j \sim \mathcal{N}(0, 2^2)$. For robustness evaluation, Gaussian errors were replaced by Student-t errors with degrees of freedom $\{10, 15, 20\}$ assigned cyclically across nodes.

Algorithm 1 was compared with representative high-dimensional methods, including LISTEN [Ghoshal and Honorio, 2018], BHLSM [Park et al., 2021], and TD [Chen et al., 2019], using average graph recovery rates. Continuous optimization approaches such as NOTEARS and GOLEM are widely used benchmarks; however, in the high-dimensional regimes considered here, their implementations did not achieve meaningful recovery. Fully Bayesian DAG methods, while flexible in incorporating prior information, lack scalable polynomial-time guarantees and become computationally prohibitive for $p \ge 100$, and are therefore excluded from the main comparisons.



(a) $d_M = 3$ (b) $d_M = 6$

Figure 2: Average graph recovery rates for 200-node Gaussian LBNs with maximum degree $d_M \in \{3, 6\}$.



(a) $d_M = 3$ (b) $d_M = 6$

Figure 3: Average graph recovery rates for 200-node LBNs with Student-t errors and maximum degree $d_M \in \{3, 6\}$.

Prior information enters through the edge-wise inclusion probabilities $\eta_{j,k}$ in (3). To evaluate (i), we construct an oracle η based on the true graph and consider three regimes: (PEM: Weak), where $\eta_{j,k} = 0.7$ for true edges and 0.3 otherwise; (PEM: Strong), where $\eta_{j,k} = 0.95$ for true edges and 0.05 otherwise; and (PEM: Non-info), where $\eta_{j,k} = 0.5$ for all pairs (j, k), corresponding to a homogeneous inclusion prior without edge-wise heterogeneity.

To evaluate (ii), we inject noise into η in the weak setting by flipping a fraction $\rho \in \{0.1, 0.2, 0.3, 0.4\}$ of true-edge and true non-edge entries ($0.7 \leftrightarrow 0.3$), yielding an overall corruption level 2ρ. Hence, some true edges receive low inclusion strengths and some true non-edges receive high inclusion strengths. All hyper-parameters were fixed across experiments (see Appendix G.1).

6.2 GAUSSIAN LINEAR BAYESIAN NETWORKS

Figure 2 reports average graph recovery rates for 200-node Gaussian LBNs with maximum degree $d_M \in \{3, 6\}$ as the sample size varies over $n \in \{25, 50, \ldots, 300\}$. We compare three variants of Algorithm 1 (strong, weak, and non-info) with LISTEN, BHLSM, and TD. The benchmark methods exhibit performance comparable to the no-prior variant of the proposed algorithm. TD is omitted for $d_M = 6$ due to prohibitive computational cost.

First, as the sample size increases, all methods converge toward similar recovery rates, consistent with the high-dimensional consistency guarantees in Theorem 5.1. Additionally, incorporating edge-wise prior information consistently improves recovery in high-dimensional finite-sample regimes. Both strong and weak prior variants outperform fully data-driven approaches, particularly when n is small relative to p. This confirms the benefit of prior modulation suggested by the structural separation and stability conditions in Section 5.

The benefit of prior modulation becomes more pronounced as graph density increases. For $d_M = 6$, substantial performance gaps emerge, whereas for $d_M = 3$ differences are modest once n is moderate. As d_M grows, stronger local dependence reduces precision separation, making Assumption 1(ii)–(iii) harder to satisfy and tightening the ℓ_∞/ℓ_∞ stability requirement in Assumption 2. In this regime, PEM-LBN remains more robust, suggesting that its separation and stability conditions deteriorate more gradually with density than fully data-driven alternatives.

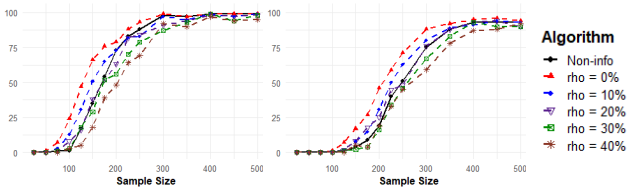
6.3 HEAVY-TAILED LINEAR BAYESIAN NETWORKS

This section examines robustness to heavy-tailed noise by replacing Gaussian errors with Student-t distributions. Samples were generated as in Section 6.2, with the sole modification to the error distribution. Figure 3 reports the corresponding graph recovery rates. The qualitative behavior closely mirrors the Gaussian setting. As the sample size increases, all methods converge toward comparable recovery rates, consistent with the high-dimensional consistency guarantees established in Section 5. Incorporating prior information remains beneficial in small-sample regimes, particularly for denser graphs. Overall, these results indicate that the structural separation and stability conditions degrade gracefully under moderate heavy-tailed deviations.

6.4 ROBUSTNESS TO NOISY INFORMATION

This section examines the sensitivity of Algorithm 1 to misspecification in edge-wise prior information. Starting from the weak prior setting, we randomly flipped a fraction $\rho \in \{0.1, 0.2, 0.3, 0.4\}$ of the prior inclusion probabilities. Hence, some true edges were assigned low prior weights while some non-edges received high prior weights. Larger values of ρ correspond to increasing inconsistency between the prior specification and the true underlying graph. For the non-informative prior baseline (Non-info), all inclusion probabilities were fixed at $\eta_{j,k} = 0.5$ for all (j, k), corresponding to the red curves in Figure 2.

Figure 4 reports the resulting graph recovery rates across different misspecification levels. As the misspecification rate ρ increases, graph recovery performance degrades gradually. For small to moderate levels of misspecification ($\rho \leq 0.2$), incorporating prior information continues to improve recovery accuracy relative to the no-prior setting, particularly in high-dimensional regimes with limited sample sizes. As the



(a) $d_M = 3$ (b) $d_M = 6$

Figure 4: Average graph recovery rates for 100-node Gaussian LBNs with $d_M \in \{3, 6\}$ as the edge-wise prior misspecification rate ρ increases from 0% to 40%.

corruption level approaches 0.5 ($\rho \approx 0.2$–0.3), recovery becomes nearly indistinguishable from the non-informative baseline, indicating that highly noisy priors no longer contribute informative signal for structure recovery. This behavior reflects the robustness of soft prior integration via posterior inclusion probabilities, as opposed to hard structural enforcement.

7 REAL DATA

The proposed algorithm is applied to a real-world order dataset from an online shopping mall, publicly available at **Kaggle E-commerce Dataset**. The dataset contains order histories of 38,997 customers over 42 products, along with payment method and product category information. Products fall into four categories, 'Auto & Accessories', 'Fashion', 'Electronic', and 'Home & Furniture'. Orders are grouped by payment method, and for each group, we construct an $n \times p$ customer–product count matrix. We then study product–product relationships within each group and learn a directed graph. We focus on payment groups with at least 1,000 customers: credit card ($n = 31{,}055$), e-wallet ($n = 2{,}746$), and money order ($n = 9{,}089$).

Small sample sizes within certain payment groups lead to unstable directed graph estimation. In particular, when the e-wallet and money order groups are fitted without prior information, data-driven structure learning methods tend to produce near-empty graphs. This is largely because order histories are highly sparse, and each user contributes only a handful of purchases, leaving too little signal to reliably identify an ordering and parent sets. As a result, high-dimensional procedures tend to shrink toward overly sparse structures. By contrast, the credit card group admits a non-trivial DAG due to its substantially larger sample size. We therefore learn a reference graph from the credit card data and transfer it as probabilistic prior structural information when fitting the smaller payment groups (Figure 1) (details are provided in Appendix H).

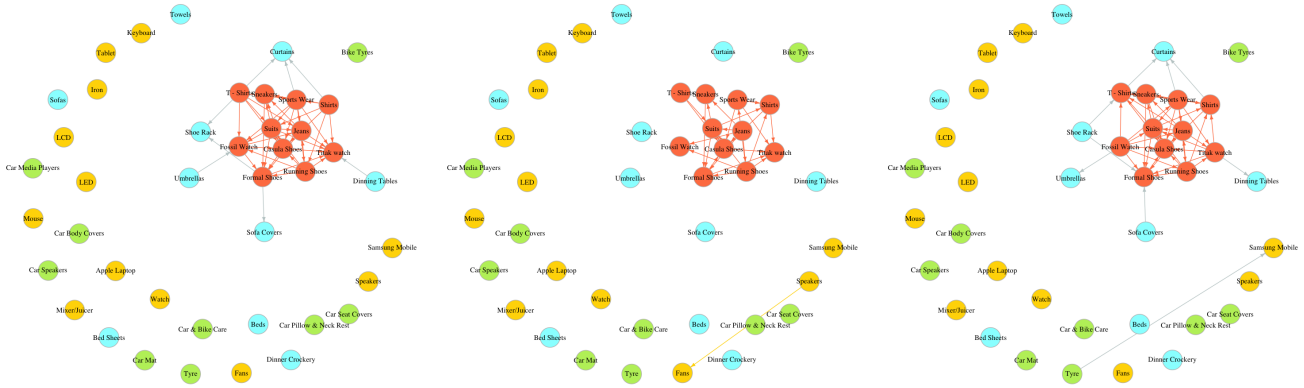
Figure 5 shows the estimated DAGs for each payment method. The underlying transaction data exhibit strong within-category co-purchasing patterns, a phenomenon widely documented in retail network analyses and market basket studies (e.g., Agrawal et al., 1993). The estimated graphs closely align with this empirical structure: in the credit card graph, 42 of 50 edges lie within 'Fashion', and in the money order graph 27 of 28 edges remain within the same category. The e-wallet graph retains a similar concentration (42 of 51 edges within 'Fashion'), indicating that the dominant category-level structure is consistently recovered. Differences across payment methods are also reflected in the learned graphs. The money order graph exhibits almost no cross-category connectivity, whereas the e-wallet graph introduces 9 cross-category links. Such variation suggests responsiveness to group-specific purchasing patterns, rather than mechanical replication of a reference structure.

Edge directions further align with established demand-driven network patterns. Products with higher order volumes tend to have larger outdegrees (Spearman correlations 0.802 for money order and 0.877 for e-wallet), and the highest-volume products serve as outgoing hubs. This demand–centrality relationship mirrors degree heterogeneity patterns commonly observed in product networks (e.g., Leskovec et al., 2007). Overall, the learned DAGs align with known structural properties of retail transaction networks while differentiating payment-specific interaction patterns, indicating that the proposed method recovers interpretable and data-consistent dependence structures.

8 CONCLUSION

We proposed probabilistic edge modulation (PEM) for learning linear Bayesian networks with uncertain edge-specific prior information. By replacing hard constraints with calibrated inclusion strengths, PEM integrates noisy priors while preserving ordering-based identifiability. We established high-dimensional consistency under sub-Gaussian and heavy-tailed noise with polynomial-time MAP estimation, achieving sample complexity $n = \Omega(d_M^2 \log p)$ and $n = \Omega(d_M^2 p^{2/m})$. Empirical results further show that informative priors improve structural stability, particularly in sample-limited regimes.

Future directions include extending uncertainty quantification beyond edge-level posterior inclusion probabilities to node orderings via scalable Bayesian approximations [Kuipers et al., 2022], using posterior edge probabilities to refine prior beliefs and prioritize follow-up experimental validation, improving computational efficiency using topological-layer representations [Gao et al., 2020, Zhou et al., 2021, Park, 2023], and generalizing edge-wise prior modulation to nonlinear and discrete identifiable DAG models [Zhang and Hyvärinen, 2009, Hoyer et al., 2009, Mooij et al., 2009, Peters et al., 2012, Park and Raskutti, 2018].



(a) credit card. (b) money order. (c) e-wallet.

Figure 5: Payment-specific DAG estimates using a credit-card-based graph as prior knowledge. The node color indicates product category: 'Auto & Accessories' (green), 'Fashion' (orange), 'Electronic' (yellow), and 'Home & Furniture' (blue).

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) Grant funded by the Korea government(MSIT) (RS-2026-25473737), Institute of Information & communications Technology Planning & Evaluation (IITP) Grant funded by the Korea government(MSIT) (N). RS-2021-II211343), and LAMP Program of the National Research Foundation of Korea(NRF) grant funded by the Ministry of Education(No. RS-2023-00301976) for Gunwoong Park. This work was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00211974 and RS-2025-02216235) for Kyoungjae Lee. Additionally, this research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (RS-2025-25419298) for Sunmin Oh.

References

Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, 1993.

Hossein Amirkhani, Mohammad Rahmati, Peter JF Lucas, and Arjen Hommersom. Exploiting experts' knowledge for structure learning of bayesian networks. *IEEE transactions on pattern analysis and machine intelligence*, 39 (11):2154–2170, 2016.

Shahzia Anjum, Arnaud Doucet, and Chris C Holmes. A boosting approach to structure learning of graphs with and without prior knowledge. *Bioinformatics*, 25(22): 2929–2936, 2009.

Sayantan Banerjee and Subhashis Ghosal. Bayesian structure learning in graphical models. *Journal of Multivariate Analysis*, 136:147–162, 2015.

Maria Maddalena Barbieri and James O Berger. Optimal predictive model selection. *Annals of Statistics*, 32(3): 870–897, 2004.

Anamitra Chaudhuri, Anirban Bhattacharya, and Yang Ni. Consistent dag selection for bayesian causal discovery under general error distributions. *arXiv preprint arXiv:2508.00993*, 2025.

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

Anthony C Constantinou, Zhigao Guo, and Neville K Kitson. The impact of prior knowledge on causal structure learning. *Knowledge and Information Systems*, 65(8): 3385–3434, 2023.

Victor-Alexandru Darvariu, Stephen Hailes, and Mirco Musolesi. Large language models are effective priors for causal graph discovery. *arXiv preprint arXiv:2405.13551*, 2024. URL `https://arxiv.org/abs/2405.13551`.

Adiba Ejaz and Elias Bareinboim. Less greedy equivalence search. *arXiv preprint arXiv:2506.22331*, 2025.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Nir Friedman and Daphne Koller. Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks. *Machine Learning*, 50: 95–125, 2003.

Lingrui Gan, Naveen N Narisetty, and Feng Liang. Bayesian regularization for graphical models with unequal shrinkage. *Journal of the American Statistical Association*, 114 (527):1218–1231, 2019.

Ming Gao, Yi Ding, and Bryon Aragam. A polynomial-time algorithm for learning nonparametric causal graphs. *arXiv preprint arXiv:2006.11970*, 2020.

Ming Gao, Wai Ming Tai, and Bryon Aragam. Optimal estimation of gaussian dag models. In *International Conference on Artificial Intelligence and Statistics*, pages 8738–8757. PMLR, 2022.

AmirEmad Ghassami, Negar Kiyavash, Biwei Huang, and Kun Zhang. Multi-domain causal structure learning in linear systems. *Advances in neural information processing systems*, 31, 2018.

Asish Ghoshal and Jean Honorio. Learning identifiable gaussian bayesian networks in polynomial time and sample complexity. In *Advances in Neural Information Processing Systems*, pages 6457–6466, 2017.

Asish Ghoshal and Jean Honorio. Learning linear structural equation models in polynomial time and sample complexity. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1466–1475, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.

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

Kyoungjae Lee and Xuan Cao. Bayesian inference for high-dimensional decomposable graphs. *Electronic Journal of Statistics*, 15(1):1549–1582, 2021.

Jure Leskovec, Lada A Adamic, and Bernardo A Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5–es, 2007.

Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *The Journal of Machine Learning Research*, 15 (1):3065–3105, 2014.

C MEEK. Causal inference and causal explanation with background knowledge. In *Proc. Conf. on Uncertainty in Artificial Intelligence (UAI-95)*, pages 403–410, 1995.

Joris Mooij, Dominik Janzing, Jonas Peters, and Bernhard Schölkopf. Regression by dependence minimization and its application to causal inference in additive noise models. In *Proceedings of the 26th annual international conference on machine learning*, pages 745–752. ACM, 2009.

Irene Y Ng, Dianbo Huang, Yichong Zhang, and Marzyeh Ghassemi. Golem: Guaranteeing structural constraints in high-dimensional causal discovery. In *Advances in Neural Information Processing Systems*, 2020.

Gunwoong Park. Identifiability of additive noise models using conditional variances. *Journal of Machine Learning Research*, 21(75):1–34, 2020.

Gunwoong Park. Computationally efficient learning of gaussian linear structural equation models with equal error variances. *Journal of Computational and Graphical Statistics*, pages 1–26, 2023.

Gunwoong Park and Youngwhan Kim. Identifiability of gaussian linear structural equation models with homogeneous and heterogeneous error variances. *Journal of the Korean Statistical Society*, 49(1):276–292, 2020.

Gunwoong Park and Garvesh Raskutti. Learning quadratic variance function (qvf) dag models via overdispersion scoring (ods). *Journal of Machine Learning Research*, 18 (224):1–44, 2018.

Gunwoong Park, Sang Jun Moon, Sion Park, and Jong-June Jeon. Learning a high-dimensional linear structural equation model via l1-regularized regression. *Journal of Machine Learning Research*, 22(102):1–41, 2021.

Emilija Perković, Markus Kalisch, and Maloes H Maathuis. Interpreting and using cpdags with background knowledge. 2017.

Jonas Peters, Joris Mooij, Dominik Janzing, and Bernhard Schölkopf. Identifiability of causal graphs using functional models. *arXiv preprint arXiv:1202.3757*, 2012.

Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, Bin Yu, et al. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Veronika Ročková and Edward I George. The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444, 2018.

Verónica Rodrıguez-López and Luis Enrique Sucar. Knowledge transfer for learning subject-specific causal models. In *International Conference on Probabilistic Graphical Models*, pages 385–396. PMLR, 2022.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *The Journal of Machine Learning Research*, 7:2003–2030, 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research*, 12 (Apr):1225–1248, 2011.

Yusuke Tashiro, Kawahara Takuya, and Shohei Shimizu. Parcel-based causal discovery from fmri data for large-scale brain network identification. In *ICASSP*. IEEE, 2014.

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N. Balasubramanian, and Amit Sharma. Causal inference using LLM-guided discovery. *arXiv preprint arXiv:2310.15117*, 2023. URL `https://arxiv.org/abs/2310.15117`.

Kun Zhang and Aapo Hyvärinen. Causality discovery with additive disturbances: An information-theoretical perspective. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 570–585. Springer, 2009.

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 (NeurIPS)*, 2018. URL `https://arxiv.org/abs/1803.01422`.

Quan Zhou and Hyunwoong Chang. Complexity analysis of bayesian learning of high-dimensional DAG models and their equivalence classes. *The Annals of Statistics*, 51 (3):1058–1085, 2023. doi: 10.1214/23-AOS2280.

Shuheng Zhou, Sara van de Geer, and Peter Bühlmann. Adaptive lasso for high dimensional regression and gaussian graphical modeling. *arXiv preprint arXiv:0903.2515*, 2009.

Wei Zhou, Xin He, Wei Zhong, and Junhui Wang. Efficient learning of quadratic variance function directed acyclic graphs via topological layers. *arXiv preprint arXiv:2111.01560*, 2021.

Probabilistic Edge Modulation for High-Dimensional Causal Discovery (Supplementary Material)

Seyong Hwang[1] **Kyoungjae Lee**[2] **Sunmin Oh**[3,4] **Gunwoong Park**[*,3,4,5]

[1]Department of Statistics, University of Michigan - Ann Arbor, U.S.A.,
[2]Department of Statistics, Sungkyunkwan University, South Korea
[3]Department of Statistics, Seoul National University, South Korea
[4]Institute for Data Innovation in Science, Seoul National University, South Korea
[5]Interdisciplinary Program in Artificial Intelligence, Seoul National University, South Korea
[*]Correspondence to: gw.park23@gmail.com ,
[*]*All authors contributed equally to this work.*

A EXISTING APPROACHES TO LEARNING LINEAR BAYESIAN NETWORKS

This section provides a review of existing methods for learning LBNs. One popular approach is based on regression. For example, Park and Kim [2020] applies standard regression and conditional independence tests to learn a low-dimensional Gaussian LBN. It estimates the ordering using the variance of residuals, and subsequently infers directed edges through conditional independence tests. Chen et al. [2019] develops the ℓ_0- and ℓ_1-regression-combined TD algorithm for sub-Gaussian LBNs. The algorithm estimates the ordering using the best-subset-regression, and then infers the edges using a ℓ_1-regularized approach. The TD algorithm requires a sample size of $n = \Omega(q^2 \log p)$ for accurate ordering estimation, where q is the predetermined upper bound of the maximum indegree. Similarly, Park et al. [2021] develops the ℓ_1-regression-based LBN learning (LSEM) algorithm with the sample complexities of $n = \Omega(d_M^4 \log p)$ and $n = \Omega(d_M^4 p^{2/m})$ for sub-Gaussian and $4m$-th bounded-moment LBNs, respectively. Finally, Gao et al. [2022] proposes the best-subset-selection-based optimal learning algorithm for Gaussian LBNs. Its sample complexity is optimal $n = \Omega(d_{in} \log \frac{p}{d_{in}})$ under the known maximum indegree d_{in}.

Another common approach relies on the precision matrix. Loh and Bühlmann [2014] estimates the precision matrix by applying graphical Lasso, and infers the moralized graph from the support of the estimated precision matrix. Loh and Bühlmann [2014] then selects the best-scoring graph amongst DAGs that are consistent with the moralized graph. Ghoshal and Honorio [2018] develops the LISTEN algorithm that applies CLIME to estimate the precision matrix and estimates the last element of the ordering using the diagonal entries of the estimated matrix. It then determines its parents based on the non-zero entries in its row of the precision matrix. After eliminating the last element of the ordering, this procedure is repeated until the graph is fully estimated. Ghoshal and Honorio [2018] proves that the LISTEN algorithm requires sample sizes of $n = \Omega(d_M^4 \log p)$ and $n = \Omega(d_M^4 p^{2/m})$, when the error distributions are sub-Gaussian and $4m$-th bounded-moment, respectively.

In summary, existing frequentist approaches to learning linear Bayesian networks primarily rely on tools such as OLS, Lasso, graphical Lasso, and CLIME. While these methods have achieved notable theoretical and empirical progress, they typically employ uniform shrinkage penalties, require strong structural assumptions such as incoherence conditions or known bounds on the maximum indegree, and may face computational challenges in less sparse regimes. Moreover, these approaches offer limited flexibility for incorporating partial prior information about the graph structure. These considerations motivate the Bayesian precision-matrix–based framework developed in this work.

B COMPUTATION OF POSTERIOR INCLUSION PROBABILITIES

The posterior inclusion probability under the proposed method is given by

$$\Pr(r_{j,k} = 1 \mid X^{1:n}) = \int \Pr(r_{j,k} = 1 \mid \Omega_{j,k}) \pi(\Omega_{j,k} \mid X^{1:n}) d\Omega_{j,k}.$$

Following the approach of Gan et al. [2019], we approximate the posterior inclusion probability by replacing the integral over $\Omega_{j,k}$ with a plug-in MAP estimator as follows

$$\begin{aligned}\widehat{\Pr}(r_{j,k} = 1 \mid X^{1:n}) &= \Pr(r_{j,k} = 1 \mid \widehat{\Omega}_{j,k}) \\ &= \frac{\frac{\eta_{j,k}}{2\nu_1} \exp\big(-\frac{|\widehat{\Omega}_{j,k}|}{\nu_1}\big)}{\frac{\eta_{j,k}}{2\nu_1} \exp\big(-\frac{|\widehat{\Omega}_{j,k}|}{\nu_1}\big) + \frac{1-\eta_{j,k}}{2\nu_0} \exp\big(-\frac{|\widehat{\Omega}_{j,k}|}{\nu_0}\big)}.\end{aligned}$$

C DETAILED CONDITION FOR THEOREM 5.1

This section outlines the specific constraints on the constants and hyper-parameters required to establish the consistency results presented in Theorem 5.1.

C.1 CONSTANTS AND ASSUMPTIONS

For ease of notation, we define $M_\Sigma = \|\Sigma\|_\infty$, $M_\Gamma = \|(\Omega \otimes \Omega)_{S,S}\|_\infty$ and $M_{\Gamma^{(r)}} = \|(\Omega^{(r)} \otimes \Omega^{(r)})_{S^{(r)},S^{(r)}}\|_\infty$ for any $r \in \{1, 2, ..., p-1\}$, where $\|\cdot\|_\infty$ and $\otimes$ denote the ℓ_∞/ℓ_∞ operator norm and Kronecker product, respectively. Let $d = \max_{r\in\{1,2,\cdots,p\}} \max_{i\in\{\pi_1,\cdots,\pi_r\}} card\{j : \Omega^{(r)}_{i,j} \neq 0\}$. Lastly, suppose that $M_{\Gamma_{\max}} = \max_{r\in\{1,2,\ldots,p-1\}} M_{\Gamma^{(r)}}$ and $M_{\Gamma_{\min}} = \min_{r\in\{1,2,\ldots,p-1\}} M_{\Gamma^{(r)}}$.

We assume there exist constants $C_1 > 0$, $C_2 > C_3 > 0$, and $\epsilon_1 > 0$ that satisfy the following inequalities:

$$\begin{aligned}&C_3\epsilon_1 \leq \frac{k_1^2 p}{2}, \text{ and} \\ &(C_1 + C_3) < \frac{1}{4M_{\Gamma_{max}}} \min\left\{\tau_{\min}, 2\theta_{\min} d, k_1^2, \frac{2}{3M_\Sigma}, \ \frac{2}{3M_{\Gamma_{max}} M_\Sigma^3}\right\}.\end{aligned}$$

Additionally, we define the auxiliary constant C_4 as:

$$C_4 = C_1 + M_\Sigma^2 2(C_1 + C_3) M_{\Gamma_{max}} + 6(C_1 + C_3)^2 M_{\Gamma_{max}}^2 M_\Sigma^3$$

C.2 HYPER-PARAMETER CONDITIONS

For the consistency of Algorithm 1, the prior hyper-parameters $(\nu_0, \nu_1, \{\eta_{j,k}\}, \tau)$, the threshold parameter T, and the spectral norm bound B_0 must be chosen to satisfy the following six conditions: for some $\epsilon_1 > 0$,

- Slab precision scaling:

$$\frac{1}{\nu_1} = \frac{C_3}{1 + \epsilon_1} \frac{n}{p}$$

- Spike precision scaling:

$$\frac{1}{\nu_0} > C_4 \frac{n}{p}$$

- Signal-to-noise ratio (prior separation):
  For all $(j, k) \in S$,

$$\frac{\nu_1^2(1 - \eta_{j,k})}{\nu_0^2 \eta_{j,k}} \leq \epsilon_1 \exp\left\{2(C_2 - C_3) M_{\Gamma_{min}} (C_4 - C_3) \frac{n}{p^2}\right\}$$

- Off-diagonal penalty:

$$\tau \leq C_3 \frac{n}{2p}$$

- Thresholding condition:
  For $\forall (j,k) \notin S$,

$$\eta_{jk} < \frac{T\nu_1}{(1-T)\nu_0 + T\nu_1}.$$

  For $\forall (j,k) \in S$,

$$\eta_{jk} > \frac{T\nu_1 e^{-\alpha\left(\frac{1}{\nu_0}-\frac{1}{\nu_1}\right)}}{(1-T)\nu_0 + T\nu_1 e^{-\alpha\left(\frac{1}{\nu_0}-\frac{1}{\nu_1}\right)}},$$

  where $\alpha = \left(\theta_{\min} - 2(C_1+C_3)M_\Gamma \frac{1}{d}\right)$.

- Spectral norm bound:

$$\frac{1}{k_1} + 2(C_1+C_3)M_{\Gamma_{max}} < B_0 < \sqrt{2n\nu_0}$$

D PROOF FOR THEOREM 5.1

This section proves that Algorithm 1 accurately recovers the underlying graph with high probability. The proof is built upon the prior works of LBN learning algorithms Park [2020], Ghoshal and Honorio [2018] and the theoretical result of BAGUS in Gan et al. [2019].

For convenience, we utilize notations $M_\Sigma, M_\Gamma, M_{\Gamma^{(r)}}, M_{\Gamma_{max}}, M_{\Gamma_{min}}$, and d defined in Section C.1, and denote $\|A\|_{max} = \max_{i,j} |A_{i,j}|$.

D.1 PRELIMINARIES

Lemma D.1 *Consider an LBN* (1) *with sub-Gaussian and* $4m$*-th bounded-moment errors. For any pre-defined constants* $C_1 > 0$ *and* $C_2 > C_3 > 0$*, assume*

i) the prior hyper-parameters $\nu_0, \nu_1, \{\eta_{j,k}\}$*, and* τ *satisfy*

$$\begin{cases} \frac{1}{n\nu_1} = C_3 \frac{1}{p} \frac{1}{1+\epsilon_1}, \frac{1}{n\nu_0} > C_4 \frac{1}{p}, \\ \frac{\nu_1^2(1-\eta_{j,k})}{\nu_0^2 \eta_{j,k}} \leq \epsilon_1 \exp\left[2(C_2-C_3)M_\Gamma(C_4-C_3)n/p^2\right], \quad \forall (j,k) \in S, \\ \tau \leq C_3 \frac{n}{2} \frac{1}{p} \end{cases}$$

for some constant $\epsilon_1 > 0$*, where* $C_4 = C_1 + M_\Sigma^2 2(C_1+C_3)M_\Gamma + 6(C_1+C_3)^2 M_\Gamma^2 M_\Sigma^3$*,*

ii) the spectral norm bound B_0 *satisfies* $\frac{1}{k_1} + 2(C_1+C_3)M_\Gamma < B_0 < (2n\nu_0)^{\frac{1}{2}}$*,*

iii) $\max\left\{2(C_1+C_3)M_\Gamma \max\left(3M_\Sigma, 3M_\Gamma M_\Sigma^3, \frac{2}{k_1^2}\right), \ \frac{2C_3\epsilon_1}{k_1^2 p}\right\} \leq 1$*,*

iv) $\Lambda_{\min}(\Sigma) \geq k_1$*.*

If $\|\widehat{\Sigma} - \Sigma\|_{max} < C_1 \frac{1}{d}$*, then*

$$\|\widehat{\Omega} - \Omega\|_m < 2(C_1+C_3)M_\Gamma \frac{1}{d}.$$

Proof 1 (Proof of Lemma D.1) *Before presenting the proof, we will define two penalizing functions $pen_{SS}(\theta), pen_1(\theta)$ as follows:*

$$pen_{SS}(\theta) = -\log\left[\left(\frac{\eta}{2\nu_1}\right)e^{-\frac{|\theta|}{\nu_1}} + \left(\frac{1-\eta}{2\nu_0}\right)e^{-\frac{|\theta|}{\nu_0}}\right],$$
$$pen_1(\theta) = \tau|\theta|.$$

By the definition, we can find bounds of the first and second derivatives of $pen_{SS}(\delta)$, respectively.

- *Bound of the first derivative of $pen_{SS}(\delta)$:*
  Using equation (5) in the appendix of Gan et al. [2019], the following inequality holds:

$$\frac{1}{n}|pen'_{SS}(\delta)| < \frac{1}{n\nu_1}\left(1 + \frac{\frac{\nu_1^2(1-\eta)}{\nu_0^2\eta}}{e^{\frac{|\delta|}{\nu_0}-\frac{|\delta|}{\nu_1}}}\right).$$

  With the conditions in this lemma and letting $\nu_1^2(1-\eta)/(\nu_0^2\eta) = \xi\exp\left[\psi(C_4-C_3)n/p^2\right]$ where $\xi < \epsilon_1$,

$$\frac{\frac{\nu_1^2(1-\eta)}{\nu_0^2\eta}}{e^{\frac{|\delta|}{\nu_0}-\frac{|\delta|}{\nu_1}}} \le \frac{\xi\exp\left[\psi(C_4-C_3)n/p^2\right]}{\exp\left[\psi(C_4-C_3)n/p^2\right]} \le \xi,$$

  with $|\delta| \ge \psi/p$. So we can get the bound of $\frac{1}{n}|pen'_{SS}(\delta)|$ as

$$\frac{1}{n}|pen'_{SS}(\delta)| < \frac{1}{n\nu_1}\left(1 + \frac{\frac{\nu_1^2(1-\eta)}{\nu_0^2\eta}}{e^{\frac{|\delta|}{\nu_0}-\frac{|\delta|}{\nu_1}}}\right) \le C_3\frac{1}{p}\frac{1}{1+\epsilon_1}(1+\xi) < C_3\frac{1}{p} \le C_3\frac{1}{d}.$$

- *Bound of the second derivative of $pen_{SS}(\delta)$:*

  Using equation (7) in the appendix of Gan et al. [2019] and since the condition iii) implies $\frac{1}{p} \le \frac{k_1^2}{2C_3\epsilon_1}$, the following inequality holds:

$$\frac{1}{2n}|pen''_{SS}(\delta)| < \frac{\xi}{2n\nu_1} < \frac{C_3}{2}\xi\frac{1}{p} < \frac{C_3}{2}\epsilon_1\frac{1}{p} \le \frac{1}{4}k_1^2.$$

Using the bounds of the first and second derivatives of $pen_{SS}(\delta)$ derived above, this lemma can be proved in the same way as Theorem A in the appendix of Gan et al. [2019] changing $\sqrt{\frac{\log p}{n}}$ to $\frac{1}{d}$. Moreover, we can omit the sample size bound $n \ge M^2\log p$ in Theorem A in the appendix of Gan et al. [2019], so this lemma can be used more freely without considering the condition of n and p.

Lemma D.2 *Suppose that Assumptions 1 and all the conditions in Theorem 5.1 are satisfied. If $\|\widehat{\Sigma} - \Sigma\|_{max} < C_1/d$, for $r \in \{2, \ldots, p\}$,*

$$\widehat{Pa}(\pi_r) = Pa(\pi_r),$$

where $\widehat{Pa}(\pi_r) = \{i : p_{i,\pi_r}^{(r)} \ge T\}$.

Proof 2 (Proof of Lemma D.2) *Recall the thresholding condition in Section C.2:*

$$\begin{cases} \eta_{j,k} < \frac{T\nu_1}{(1-T)\nu_0 + T\nu_1}, & \forall (j,k) \notin S, \\ \eta_{j,k} > \frac{T\nu_1 e^{-\alpha\left(\frac{1}{\nu_0}-\frac{1}{\nu_1}\right)}}{(1-T)\nu_0 + T\nu_1 e^{-\alpha\left(\frac{1}{\nu_0}-\frac{1}{\nu_1}\right)}}, & \forall (j,k) \in S, \end{cases} \tag{5}$$

where $\alpha = \left(\theta_{\min} - 2(C_1+C_3)M_\Gamma\frac{1}{d}\right)$.

The posterior inclusion probabilities satisfy the following equation (see Appendix B):

$$\log\left(\frac{p^{(r)}_{i,\pi_r}}{1-p^{(r)}_{i,\pi_r}}\right) = -\log\left(\frac{\nu_1(1-\eta_{i,\pi_r})}{\nu_0\eta_{i,\pi_r}}\right) + |\widehat{\Omega}^{(r)}_{i,\pi_r}|\left(\frac{1}{\nu_0}-\frac{1}{\nu_1}\right)$$
$$= \log\left(\frac{\nu_0\eta_{i,\pi_r}}{\nu_1(1-\eta_{i,\pi_r})}\right) + |\widehat{\Omega}^{(r)}_{i,\pi_r}|\left(\frac{1}{\nu_0}-\frac{1}{\nu_1}\right).$$

To check whether a support of the estimated precision matrix is correctly estimated, we consider two cases which an element of the true precision matrix is zero or not, $\Omega^{(r)}_{i,\pi_r} = 0$ and $\Omega^{(r)}_{i,\pi_r} \neq 0$.

- *When $\Omega^{(r)}_{i,\pi_r} = 0$, by the construction in the proof of Lemma D.1, $\widehat{\Omega}^{(r)}_{i,\pi_r} = 0$. Hence, the following inequality holds using the bound of $\eta_{i,\pi_r} \notin S$ in Range* (5)*:*

$$\log\left(\frac{p^{(r)}_{i,\pi_r}}{1-p^{(r)}_{i,\pi_r}}\right) = \log\left(\frac{\nu_0\eta_{i,\pi_r}}{\nu_1(1-\eta_{i,\pi_r})}\right) < \log\left(\frac{T}{1-T}\right).$$

  This inequality implies $p^{(r)}_{i,\pi_r} < T$, so the zero entries are recovered correctly.

- *When $\Omega^{(r)}_{i,\pi_r} \neq 0$, applying Lemma D.1 with Assumption 1 (iii), the following inequality holds:*

$$|\widehat{\Omega}^{(r)}_{i,\pi_r}| > \theta_{\min} - 2(C_1+C_3)M_{\Gamma^{(r)}}\frac{1}{d} > 0.$$

  Then the following inequality holds using the bound of $\eta_{i,\pi_r} \in S$ in Range (5)*:*

$$\log\left(\frac{p^{(r)}_{i,\pi_r}}{1-p^{(r)}_{i,\pi_r}}\right) = \log\left(\frac{\nu_0\eta_{i,\pi_r}}{\nu_1(1-\eta_{i,\pi_r})}\right) + |\widehat{\Omega}^{(r)}_{i,j}|\left(\frac{1}{\nu_0}-\frac{1}{\nu_1}\right)$$
$$> \log\left(\frac{\nu_0\eta_{i,\pi_r}}{\nu_1(1-\eta_{i,\pi_r})}\right) + \left(\theta_{\min} - 2(C_1+C_3)M_{\Gamma^{(r)}}\frac{1}{d}\right)\left(\frac{1}{\nu_0}-\frac{1}{\nu_1}\right)$$
$$> \log\left(\frac{T}{1-T}\right).$$

  This inequality implies $p^{(r)}_{i,\pi_r} > T$, so the nonzero entries are recovered correctly.

Therefore, a parent set of π_r is correctly estimated, $\widehat{Pa}(\pi_r) = Pa(\pi_r)$.

Lemma D.3 *If k is a terminal vertex of the graph, then $\mathrm{Pa}(k) = supp(\Omega_{k,*}) \setminus \{k\}$, where $\Omega_{k,*}$ is the k-th row of Ω.*

Proof 3 (Proof of Lemma D.3) *By the factorization of the precision matrix*

$$\Omega = (I_p - B)^\top (\Sigma^\epsilon)^{-1}(I_p - B),$$

the entries $\Omega_{k,j}$ are calculated as follows:

$$\Omega_{k,j} = -\frac{1}{\sigma_j^2}\beta_{k,j} - \frac{1}{\sigma_k^2}\beta_{j,k} + \sum_{l\in \mathrm{Ch}(k)\cap\mathrm{Ch}(j)} \frac{\beta_{k,l}\beta_{j,l}}{\sigma_l^2}. \tag{6}$$

If k is a terminal vertex of the graph, $\beta_{k,j} = 0$ for all $j \in V \setminus \{k\}$. Then the following equation holds:

$$\Omega_{k,j} = -\frac{1}{\sigma_k^2}\beta_{j,k}.$$

Since $\mathrm{Pa}(k) = \{j \in V | \beta_{j,k} \neq 0\}$, the parent set of k can be found using the support set of the k-th row of the precision matrix as following:

$$\mathrm{Pa}(k) = supp(\Omega_{k,*}) \setminus \{k\}.$$

Lemma D.4 *For an LBN, the following inequality holds between the column sparsity of the precision matrix (d) and the maximum degree of the moralized graph (d_M):*

$$d \leq d_M + 1.$$

Proof 4 (Proof of Lemma D.4) *If two nodes j, k are not connected in the moralized graph, it means that $B_{j,k} = B_{k,j} = 0$ and there is no common child between nodes j and k. Then by (6), it can be shown that $\Omega_{k,j} = 0$. So if $\Omega_{j,k} \neq 0$, nodes j and k are connected in the moralized graph. Hence, the following inequality holds:*

$$\begin{aligned} d &= \max_{i=1,2,\cdots p} card\{j : \Omega_{i,j} \neq 0\} \\ &= 1 + \max_{i=1,2,\cdots p} card\{j \neq i : \Omega_{i,j} \neq 0\} \\ &\leq 1 + \max_{i=1,2,\cdots p} card\{j : (i,j) \in M(G)\} \\ &= d_M + 1, \end{aligned}$$

where $M(G)$ is the moralized graph of G.

Proposition D.5 (Error Bound for the Sample Covariance Matrix) *Consider a random vector $(X_j)_{j=1}^p$ whose covariance matrix is $\Sigma \in \mathbb{R}^{p \times p}$.*

- *Lemma 1 of Ravikumar et al. [2011]: When $(X_j)_{j=1}^p$ follows a sub-Gaussian distribution,*

$$\Pr\left(\left|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\right| \geq \zeta\right) \leq 4 \cdot \exp\left(\frac{-n\zeta^2}{128(1 + 4s_{\max}^2)\max_j(\Sigma_{j,j})^2}\right),$$

  for all $\zeta \in (0, \max_j \Sigma_{j,j} 8(1 + 4s_{\max}^2))$.
- *Lemma 2 of Ravikumar et al. [2011]: When $(X_j)_{j=1}^p$ follows a 4m-th bounded-moment distribution,*

$$\Pr\left(\left|\widehat{\Sigma}_{j,k} - \Sigma_{j,k}\right| \geq \zeta\right) \leq 4 \cdot \frac{2^{2m}\max_j(\Sigma_{j,j})^{2m} C_m (K_{\max} + 1)}{n^m \zeta^{2m}},$$

  where C_m is a constant depending only on m.

D.2 PROOF OF THEOREM 5.1

Without loss of generality, assume that the true ordering satisfying the backward selection condition in Lemma 2.1 is $\pi = (\pi_1, \pi_2, ..., \pi_p) = (p, p-1, ..., 1)$, and hence $\pi_{p+1-r} = r$. For ease of notation, let $\pi_{1:j} = (\pi_1, \pi_2, ..., \pi_j)$. Lastly, define $\sigma_{\max}^2 = \max_{1 \leq j \leq p} \sigma_j^2$ as the maximum error variance.

In this proof, we show that Algorithm 1 accurately recovers the graph structure if a sample covariance matrix error is sufficiently small using mathematical induction.

Assume

$$\|\widehat{\Sigma} - \Sigma\|_{max} < C_1 \frac{1}{d}.$$

The last element of the ordering, say π_p, is achieved by comparing diagonal entries of the MAP estimator $\widehat{\Omega}$ in Equation (3). More specifically, it is estimated as $\widehat{\pi}_p = \arg\min_{k \in V} \widehat{\Omega}_{k,k}$. If the following inequality holds, Algorithm 1 estimates node 1 as $\widehat{\pi}_p$, so the terminal vertex of the graph is correctly recovered, $\widehat{\pi}_p = \pi_p$.

$$\min_{k \in \pi_{1:p-1}} \left(\widehat{\Omega}_{k,k} - \widehat{\Omega}_{1,1}\right) > 0. \tag{7}$$

Since

$$\begin{aligned} \min_{k \in \pi_{1:p-1}} \left(\widehat{\Omega}_{k,k} - \widehat{\Omega}_{1,1}\right) &= \min_{k \in \pi_{1:p-1}} \left\{\left(\Omega_{k,k} - \Omega_{1,1}\right) - \left(\widehat{\Omega}_{1,1} - \Omega_{1,1}\right) + \left(\widehat{\Omega}_{k,k} - \Omega_{k,k}\right)\right\} \\ &\geq \min_{k \in \pi_{1:p-1}} \left\{\left(\Omega_{k,k} - \Omega_{1,1}\right) - \left|\widehat{\Omega}_{1,1} - \Omega_{1,1}\right| - \left|\widehat{\Omega}_{k,k} - \Omega_{k,k}\right|\right\}, \end{aligned}$$

Inequality (7) holds if

$$\min_{k\in\pi_{1:p-1}} (\Omega_{k,k} - \Omega_{1,1}) > 4(C_1 + C_3)M_\Gamma \frac{1}{d} \tag{8}$$

$$\text{and} \quad \max_{k\in V} \left|\widehat{\Omega}_{k,k} - \Omega_{k,k}\right| < 2(C_1 + C_3)M_\Gamma \frac{1}{d}. \tag{9}$$

Inequality (8) holds due to Assumption 1 (ii) and the fact that $\tau_{min} > 4(C_1 + C_3)M_\Gamma/d$ by the condition in Theorem 5.1 (specified in Section C.1). Moreover, since the constants C_1, C_2, C_3 and hyper-parameters $\nu_0, \nu_1, \{\eta_{j,k}\}, \tau$ satisfy the conditions i), ii), iii) in Lemma D.1 and condition iv) holds by Assumption 1 (i), we can apply Lemma D.1. Then Inequality (9) holds. Hence, Inequality (7) holds, which implies that the last element of the ordering is correctly estimated, i.e., $\widehat{\pi}_p = \pi_p = 1$, if $\|\widehat{\Sigma} - \Sigma\|_{\max} < C_1/d$. Furthermore, by Lemmas D.2 and D.3, if $\|\widehat{\Sigma} - \Sigma\|_{max} < C_1/d$, parents of π_p are correctly estimated:

$$\widehat{\mathrm{Pa}}\left(\widehat{\pi}_p\right) = \mathrm{Pa}\left(\pi_p\right).$$

Now, suppose that we have correctly estimated $\{\pi_{p+2-r}, \ldots, \pi_p\}$ for some $r = 2, 3, \ldots, p-1$, i.e., $\{\widehat{\pi}_{p+2-r}, \ldots, \widehat{\pi}_p\} = \{\pi_{p+2-r}, \ldots, \pi_p\}$. Then if $\|\widehat{\Sigma} - \Sigma\|_{max} < C_1/d$, the following inequality holds:

$$\|\widehat{\Sigma}^{(r)} - \Sigma^{(r)}\|_{max} \leq \|\widehat{\Sigma} - \Sigma\|_{max} \leq C_1 \frac{1}{d},$$

where $\widehat{\Sigma}^{(r)}$ and $\Sigma^{(r)}$ are the sample and true covariance matrices of $(X_{\pi_1}, \ldots, X_{\pi_{p+1-r}})$. Moreover, we can derive the lower bound of a minimum eigenvalue of $\Sigma^{(r)}$ by the following inequality:

$$\Lambda_{\min}(\Sigma^{(r)}) = \min_{||x||_2=1, x\in\mathbb{R}^{p+1-r}} x^\top \Sigma^{(r)} x \geq \min_{||x||_2=1, x\in\mathbb{R}^p} x^\top \Sigma x = \Lambda_{\min}(\Sigma) \geq k_1 \tag{10}$$

The $(p+1-r)$-th element of the ordering, say π_{p+1-r}, is achieved by comparing diagonal entries of the estimated precision matrix of $(X_{\pi_1}, \ldots, X_{\pi_{p+1-r}})$, denoted as $\widehat{\Omega}^{(r)}$. It is correctly estimated if the following inequality holds:

$$\min_{k\in\pi_{1:p-r}} \left(\widehat{\Omega}^{(r)}_{k,k} - \widehat{\Omega}^{(r)}_{r,r}\right) > 0. \tag{11}$$

By similar arguments used to show Inequality (7), the above inequality holds if both following inequalities are satisfied.

$$\min_{k\in\pi_{1:p-r}} \left(\Omega^{(r)}_{k,k} - \Omega^{(r)}_{r,r}\right) > 4(C_1 + C_3)M_{\Gamma^{(r)}} \frac{1}{d},$$

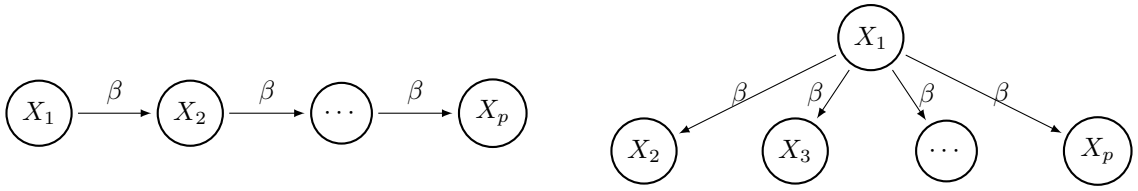
$$\max_{k\in\pi_{1:p+1-r}} \left(\widehat{\Omega}^{(r)}_{k,k} - \Omega^{(r)}_{k,k}\right) < 2(C_1 + C_3)M_{\Gamma^{(r)}} \frac{1}{d}.$$

The first inequality holds due to Assumption 1 (ii) and the fact that $\tau_{min} > 4(C_1 + C_3)M_{\Gamma^{(r)}}/d$ by the condition in Theorem 5.1 (Section C.1). Moreover, since the constants C_1, C_2, C_3 and hyper-parameters $\nu_0, \nu_1, \{\eta_{j,k}\}, \tau$ satisfy the conditions i), ii), iii) in Lemma D.1, condition iv) holds by (10) and $\|\widehat{\Sigma}^{(r)} - \Sigma^{(r)}\|_{max} \leq C_1 \frac{1}{d}$, we can apply Lemma D.1 to $\widehat{\Sigma}^{(r)}$ and $\widehat{\Omega}^{(r)}$. Then the second inequality holds. Thus, if $\|\widehat{\Sigma} - \Sigma\|_{max} < C_1/d$, Inequality (11) holds, which implies $\widehat{\pi}_{p+1-r} = \pi_{p+1-r} = r$. Again by Lemmas D.2 and D.3, if $\|\widehat{\Sigma} - \Sigma\|_{max} < C_1/d$, parents of π_{p+1-r} are correctly estimated:

$$\widehat{\mathrm{Pa}}\left(\widehat{\pi}_{p+1-r}\right) = \widehat{\mathrm{Pa}}\left(\pi_{p+1-r}\right) = \mathrm{Pa}\left(\pi_{p+1-r}\right).$$

Therefore, by mathematical induction, $\widehat{G} = G$ holds if $\|\widehat{\Sigma} - \Sigma\|_{max} < C_1/d$. Using this result, the following inequality holds:

$$\begin{aligned}
\Pr\left(\widehat{G} = G\right) &\geq \Pr\left(\|\widehat{\Sigma} - \Sigma\|_{max} < C_1 \frac{1}{d}\right) \\
&= 1 - \Pr\left(\|\widehat{\Sigma} - \Sigma\|_{max} \geq C_1 \frac{1}{d}\right) \\
&= 1 - \Pr\left(\cup_{i,j\in\{1,2,\ldots,p\}} \left[|\widehat{\Sigma}_{i,j} - \Sigma_{i,j}| \geq C_1 \frac{1}{d}\right]\right) \\
&\geq 1 - \sum_{i,j\in\{1,2,\ldots,p\}} \Pr\left(|\widehat{\Sigma}_{i,j} - \Sigma_{i,j}| \geq C_1 \frac{1}{d}\right).
\end{aligned}$$



G_1: Chain G_2: Star

Figure 6: p-node chain and star graphs

Applying Lemmas D.5 and D.4, for a sub-Gaussian LBN (12),

$$\begin{aligned}\Pr\left(\widehat{G}=G\right) &\geq 1-\sum_{i,j\in\{1,2,\ldots,p\}}\Pr\left(|\widehat{\Sigma}_{i,j}-\Sigma_{i,j}|\geq C_1\frac{1}{d}\right)\\ &\geq 1-4p^2\exp\left(-\frac{C_1^2}{A_3}\frac{n}{d^2}\right)\\ &\geq 1-4p^2\exp\left(-\frac{C_1^2}{A_3}\frac{n}{(d_M+1)^2}\right)\\ &= 1-4p^2\exp\left(-C_{sub}\frac{n}{(d_M+1)^2}\right),\end{aligned}$$

where $A_3=128(1+4s_{\max}^2)\max_j\left(\Sigma_{j,j}\right)^2$ and $C_{sub}=C_1^2/A_3$.

Applying Lemma D.5 and D.4, for a $4m$-th bounded-moment LBN (13),

$$\begin{aligned}\Pr\left(\widehat{G}=G\right) &\geq 1-\sum_{i,j\in\{1,2,\ldots,p\}}\Pr\left(|\widehat{\Sigma}_{i,j}-\Sigma_{i,j}|\geq C_1\frac{1}{d}\right)\\ &\geq 1-4p^2\frac{A_4d^{2m}}{n^mC_1^{2m}}\\ &\geq 1-4p^2\frac{A_4(d_M+1)^{2m}}{n^mC_1^{2m}}\\ &= 1-C_{bm}p^2\frac{(d_M+1)^{2m}}{n^m},\end{aligned}$$

where $A_4=2^{2m}\max_j\left(\Sigma_{j,j}\right)^{2m}C_m(K_{\max}+1)$ and $C_{bm}=4A_4/C_1^{2m}$.

■

E ILLUSTRATION OF THE ASSUMPTIONS AND HYPER-PARAMETERS

This section confirms the validity of Assumptions 1 and 2. Additionally, it provides appropriate hyper-parameters ν_0,ν_1,η,τ and the inclusion probability threshold T (under uniform shrinkage setting; $\eta_{j,k}=\eta$) when recovering chain and star LBNs illustrated in Figure 6. These are popular examples of sparse and dense models, respectively, because the chain graph has a small maximum degree of the moralized graph ($d_M=2$), whereas the star graph has a large maximum degree ($d_M=p-1$). Specifically, the considered chain and star LBNs are defined as follows: For all $j\in\{1,2,...,p-1\}$,

$$\begin{cases}\text{Chain:} & X_{j+1}=\beta X_j+\epsilon_{j+1},\\ \text{Star:} & X_{j+1}=\beta X_1+\epsilon_{j+1}\end{cases}$$

where $X_1=\epsilon_1$ in both LBNs and all the error variances are σ^2. Additionally, $\beta\in(-1,1)$ is a small edge weight.

In the chain LBN, we can determine fixed bounds for $M_{\Gamma_{\max}}$, $M_{\Gamma_{\min}}$, and M_Σ, which are as follows:

$$M_{\Gamma_{\max}} = \left(\beta^4 + 2|\beta|^3 + 4\beta^2 + 2|\beta| + 1\right)/\sigma^4,$$
$$M_{\Gamma_{\min}} = \left(\beta^4 + 2|\beta|^3 + 3\beta^2 + 2|\beta| + 1\right)/\sigma^4,$$
$$M_\Sigma = \frac{(1-|\beta|^p)(1-|\beta|^{p+1})}{(1-|\beta|)(1-|\beta|^2)}\sigma^2 \leq \frac{1}{(1-|\beta])(1-|\beta|^2)}\sigma^2.$$

Consequently, Assumption 2 is satisfied if:

$$M_1 > \frac{1}{(1-|\beta|)(1-|\beta|^2)}\sigma^2, \quad M_2 > \frac{\beta^4 + 2|\beta|^3 + 4\beta^2 + 2|\beta| + 1}{\sigma^4}$$

Moreover, simple algebra yields that Assumption 1 is satisfied if:

$$k_1 \leq \frac{\sigma^2}{(1+|\beta|)^2}, \quad \tau_{\min} \leq \frac{\beta^2}{\sigma^2}, \quad \theta_{\min} \leq \frac{|\beta|}{\sigma^2}.$$

In contrast, the star LBN exhibits diverging values of $M_{\Gamma_{max}}$ and M_Σ as the number of nodes increases:

$$M_{\Gamma_{max}} = \frac{A_2}{\sigma^4},$$
$$M_{\Gamma_{min}} = \frac{\left(2\beta^4 + 2|\beta|^3 + 3\beta^2 + 2|\beta| + 1\right)}{\sigma^4},$$
$$M_\Sigma = \max\left\{(p-1)|\beta| + 1, (p-1)\beta^2 + |\beta| + 1\right\}\sigma^2,$$

where $A_2 = 2(p-1)^2\beta^4 + 2(p-1)^2|\beta|^3 + 3(p-1)\beta^2 + 2(p-1)|\beta| + 1$. As a result, Assumption 2 rarely holds in large-scale LBNs due to diverging M_1 and M_2, providing heuristic support that this assumption is related to the sparsity of the moralized graph.

The success of the proposed approach depends on the existence of appropriate hyper-parameters. Hence, we now turn our attention to the existence of appropriate hyper-parameters ν_0, ν_1, η, τ, and T according to Theorem 5.1. By setting suitable constants C_1, C_2, C_3, and T, we can easily find proper values for these hyper-parameters. A possible choice is as follows: For any $\epsilon_1 > 0$, set

$$\nu_1 = \frac{p(1+\epsilon_1)}{nC_3},\ \nu_0 = \frac{p}{nC_4},\ \tau = \frac{nC_3}{2p},\ \text{and}$$
$$T = \frac{\nu_0\eta}{\nu_1(1-\eta) + \nu_0\eta},\ \text{where } \eta = \frac{\nu_1^2}{\nu_1^2 + \nu_0^2\epsilon_1}.$$

For both chain and star LBNs, we may also set:

$$C_1 = C_3/10, C_2 = d\theta_{\min}/(2M_{\Gamma_{max}}),$$
$$C_3 = \frac{1}{2}\min\left\{\frac{1}{6M_{\Gamma_{\max}}M_\Sigma}, \frac{1}{6M_{\Gamma_{\max}}^2 M_\Sigma^3}, \frac{k_1^2}{4M_{\Gamma_{\max}}}, \frac{\tau_{\min}}{4M_{\Gamma_{\max}}}, \frac{d\theta_{\min}}{2M_{\Gamma_{\max}}}, \frac{k_1^2 p}{2\epsilon_1}\right\}.$$

As discussed, the chain BN has fixed values for $M_{\Gamma_{\min}}$, $M_{\Gamma_{max}}$, and M_Σ, resulting in the fixed hyper-parameters regardless of the number of sample size and nodes. However, the star LBN has diverging values for $M_{\Gamma_{max}}$ and M_Σ, leading to unacceptable value of hyper-parameters, such as ν_0 and ν_1 diverge, $\tau \to 0$, and $T \to 1$, as p increases in this setting.

So far, we have demonstrated that appropriate hyper-parameters exist for chain BNs regardless of the number of nodes or samples. However, finding appropriate hyper-parameters for a large-scale star BN is difficult. Hence, this underscores the importance of Assumption 2 in achieving consistency, and highlights that the proposed method is well-suited for learning high-dimensional sparse LBNs, while it may not be the optimal choice for recovering dense LBNs.

Of course, the choice of required hyper-parameters depends on the true model quantities which are typically unknown in practice. Hence, in applications, we find ourselves in a similar position as for other graph learning algorithms (e.g., the PC, HGSM, BHLSM, and TD algorithms), where the output depends on test and tuning parameters. To select good hyper-parameter values for the BAGUS method, cross-validation can be applied as discussed in Gan et al. [2019]. However, due to the heavy computational cost, we used fixed hyper-parameters, $\nu_0 = \sqrt{1/(100n)}, \nu_1 = 1, \tau = 0.0001$, and $T = 0.5$, assuming unknown true model information in all our numerical experiments. The simulation results heuristically confirm that the proposed algorithm successfully recovers sparse graphs with high probability.

F TAIL CONDITIONS FOR LINEAR BAYESIAN NETWORKS

This section collects technical assumptions on error distributions used in the high-dimensional analysis.

F.1 ERROR PROPAGATION IN LBNS

In an LBN, each variable X_j admits the representation

$$X_j = \sum_{k \in \mathrm{Pa}(j)} \beta_{k,j} X_k + \epsilon_j = \sum_{k \in \mathrm{An}(j)} \beta_{k \to j} \epsilon_k + \epsilon_j,$$

where $\beta_{k \to j}$ denotes the sum of products of coefficients along all directed paths from k to j. Consequently, tail properties of the error variables propagate to the observed variables.

F.2 SUB-GAUSSIAN LINEAR BAYESIAN NETWORKS

An LBN is said to be sub-Gaussian if, for each $j \in V$, the standardized error $\epsilon_j / \sqrt{\mathrm{Var}(\epsilon_j)}$ is sub-Gaussian with parameter $s_{\max}^2$. In this case, X_j is also sub-Gaussian and satisfies

$$\mathbb{E}[\exp(tX_j)] \leq \exp\left(\frac{s_{\max}^2 \Sigma_{j,j} t^2}{2} \right), \qquad t \in \mathbb{R}. \tag{12}$$

F.3 $4m$-TH BOUNDED-MOMENT LINEAR BAYESIAN NETWORKS

An LBN is said to have bounded $4m$-th moments if there exists $K_{\max} > 0$ such that

$$\max_{j \in V} \mathbb{E}[(X_j)^{4m}] \leq K_{\max} \Sigma_{j,j}^{2m}. \tag{13}$$

This condition allows for heavier-tailed error distributions while retaining sufficient concentration for high-dimensional analysis.

G NUMERICAL EXPERIMENTS

G.1 HYPER-PARAMETER SETTINGS

The hyper-parameters of the proposed algorithm were fixed across experiments to isolate the effect of prior information and ensure reproducibility. Precisely, they were set to $\nu_0 = 0.5\sqrt{1/n}$, $\nu_1 = 1$, $\tau = 10^{-4}$, and $T = 0.5$. These choices follow the guidelines discussed in Section 4. In particular, a small value of ν_0 is used to enforce sparsity through the spike component, with its magnitude decreasing as the sample size increases, while a small tolerance τ ensures numerical stability of the optimization procedure. The threshold $T = 0.5$ corresponds to a symmetric decision rule on posterior inclusion probabilities.

For numerical stability, minor implementation-level modifications were applied to LISTEN and TD. In particular, LISTEN was implemented using CLIME to estimate each step of the ordering, rather than applying it only at the final step followed by a matrix update. This choice improves stability in high-dimensional settings by avoiding error propagation across successive updates. The regularized regression parameter was set to $0.5\sqrt{\log p/n}$, and the hard-thresholding parameter was chosen as half of the minimum magnitude of the true edge weights, $\min_{(j,k) \in E}(|\beta_{j,k}|/2)$. For TD, ℓ_1-regularized regression was used for parent estimation, consistent with the implementation of BHLSM. The regularization parameter was set to $4\sqrt{\log p/n}$. Finally, q was set to the maximum degree of the true moralized graph, d_M.

H BELIEF SWEEP AND PRIOR STRENGTH SELECTION

We examine the effect of prior strength on graph learning by sweeping a belief parameter $b \in [0, 1]$, which controls the edge-specific inclusion strengths $\eta_{j,k}$ in the PEM spike-and-slab prior; see the Bernoulli inclusion model in (3). Concretely,

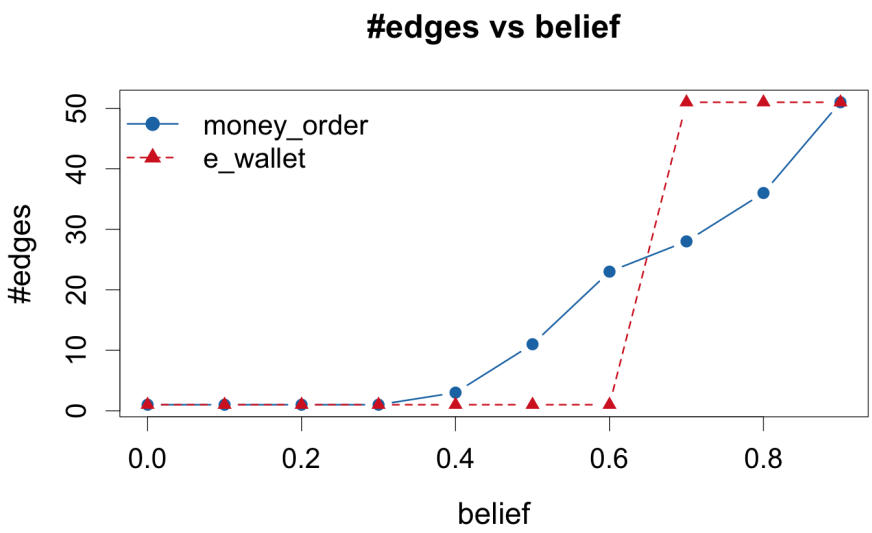


Figure 7: Number of directed edges in the learned DAG as a function of belief for money order and e-wallet.

let $\mathcal{E}_{\text{ref}}$ denote the set of directed edges in the reference DAG learned from the credit-card group. For each pair $j < k$, we set

$$\eta_{j,k}(b) = \begin{cases} \eta(b), & (j \to k) \in \mathcal{E}_{\text{ref}} \text{ or } (k \to j) \in \mathcal{E}_{\text{ref}}, \\ 1/2, & \text{otherwise}, \end{cases}$$

where the scalar mapping $b \mapsto \eta(b)$ is

$$\eta(b) = \{1 + \exp(-3b)\}^{-1}. \tag{14}$$

Thus $b = 0$ yields $\eta(b) = 0.5$ (non-informative in (3)), and larger b increasingly favors edges aligned with the reference structure. We sweep b from 0 to 1 in increments of 0.1.

Figure 7 shows the number of directed edges in the learned DAGs for the money order and e-wallet payment groups as a function of b. For both groups, the learned graphs are essentially empty under weak enforcement and become non-trivial only once the prior is sufficiently strong. In particular, for e-wallet, the edge count remains at 1 up to $b = 0.60$ and increases sharply to 51 at $b = 0.70$. For money order, the edge count grows more gradually, from 1 edge at $b = 0.30$ to 3 at 0.40, 11 at 0.50, 23 at 0.60, and 28 at 0.70.

Based on this sweep, we choose $b = 0.70$ as the smallest belief value at which the learned graphs enter a stable, non-degenerate regime in terms of edge count. As shown in Figure 7, e-wallet exhibits a sharp transition at $b = 0.70$ (from near-empty to ≈ 51 edges) and then remains essentially saturated for larger b, while money order increases more smoothly and shows substantially reduced sensitivity beyond $b \approx 0.70$. We therefore take $b = 0.70$ as a conservative choice that is strong enough to stabilize structure learning in the most sample-limited group without unnecessarily enforcing the prior further.

Importantly, using prior information does not reduce to copying the credit-card reference structure. At $b = 0.70$, the learned graphs only partially overlap with the reference DAG, matching 19 of 28 edges for money order and 28 of 51 edges for e-wallet, with deviations in both edge presence and direction. For instance, money order drops several cross-category links present in the credit-card graph and retains a smaller core within *Fashion*, whereas e-wallet introduces additional cross-category links such as (dining tables $\to$ titan watch) and (umbrellas $\to$ fossil watch), along with multiple *Fashion*–to–*Home & Furniture* edges (e.g., formal shoes $\to$ shoe rack).