Differentiable Structural EM for Explicit-Latent Linear–Gaussian DAGs

Mohammad Ali Javidian JAVIDIANMA@APPSTATE.EDU
R. Mitchell Parry PARRYRM@APPSTATE.EDU
Department of Computer Science
Appalachian State University
Boone, USA

Editors: Gustau Camps-Valls, Manuele Leonelli and Gherardo Varando

Abstract

We bridge two largely separate lines of work on directed graphical models: Classical Structural EM, which handles hidden variables and is commonly instantiated through search over discrete graph structures; and Continuous DAG learning methods, which avoid combinatorial search through smooth optimization, but are usually designed for fully observed data. We combine these ideas for linear–Gaussian models with explicit hidden variables. Our proposed method treats the graph weights as parameters of a generative Gaussian model over both observed and hidden variables. This makes the E-step exact: instead of filling in hidden variables with point estimates, it computes the needed posterior first and second moments. The M-step then uses smooth sparsity and acyclicity penalties to guide the graph search, but accepts an update only when it does not decrease the unpenalized expected complete-data objective. This preserves the usual monotone improvement of EM while still benefiting from continuous DAG optimization. Experiments show that both ingredients are necessary: exact second-moment inference in the E-step and the unpenalized acceptance check in the M-step. On hidden-confounding synthetic models and a linear–Gaussian version of Friedman's hidden-variable benchmark, a spectral warm start helps the explicit-latent model recover hidden structure, while observed-only and random-initialized alternatives perform substantially worse.

Keywords: Structural EM, latent variables, differentiable DAG learning, linear–Gaussian SEM, causal discovery

1. Introduction

Two largely parallel lines of work address structural learning in directed graphical models. Friedman (Friedman, 1997) introduced *Structural Expectation-Maximization (EM)* as a model-selection EM framework for learning from incomplete data, in which candidate models are evaluated using expected complete-data scores under the current model. Friedman (Friedman, 1998) subsequently developed the Bayesian Structural EM formulation for Bayesian-network structure learning, with expectations taken over missing or latent completions under the current network and parameters. In either formulation, the generalized M-step need not maximize the expected complete-data objective: it is sufficient to produce a model-and-parameter update that does not decrease it (Dempster et al., 1977). For Bayesian-network structure learning, this step is commonly implemented by searching over the discrete DAG space using decomposable scores and local edge operations (Friedman,

© 2026 M.A. Javidian & R.M. Parry.

1997, 1998; Chickering, 2002), although discrete local search is not required by the Structural EM principle itself. The central computational insight is that expectations computed under the current model can be reused to evaluate multiple candidate structures, avoiding a separate inference computation for every candidate. A practical limitation of conventional implementations is therefore not a requirement to use local search, but the combinatorial nature of optimization over a discrete DAG space, which can become brittle in higher dimensions and is difficult to integrate with continuous-optimization machinery.

A second line, *continuous directed acyclic graph (DAG) learning* reparameterizes structure learning as smooth optimization over a real-valued weighted adjacency matrix (Zheng et al., 2018; Bello et al., 2022; Ng et al., 2020; Yu et al., 2019; Zheng et al., 2020; Lachapelle et al., 2020). The unknowns become the entries of a real-valued weighted adjacency matrix $W \in \mathbb{R}^{p\times p}$, optimized end-to-end by gradient descent under a sparsity term and an acyclicity penalty $h(W)$ that vanishes only on DAGs. The recovered W is thresholded post hoc to extract a DAG. These methods have substantially expanded the algorithmic toolbox for structure discovery, but they are developed for the fully observed setting, where the score or likelihood is a function of observed samples alone.

When some variables are unobserved, a naive marriage of the two approaches runs into a precise interface problem. An EM procedure needs the E-step to condition on a probabilistic object representing the current model. In classical Structural EM that object is the current discrete graph. In continuous DAG learning the iterate is a real-valued W; its support pattern is a weighted directed graph, but during smooth optimization that support is not necessarily a DAG. Conditioning on a thresholded version of W (i.e., producing a DAG) would make the latent-variable posterior depend on an external post-processing hyperparameter rather than on the smooth objective being optimized. This is problematic because threshold specification in continuous DAG learning has been shown to substantially affect the recovered graph, with fixed thresholds sometimes removing true edges and optimal thresholds varying across settings (Ng et al., 2024). Moreover, in NOTEARS-style methods thresholding is typically a final post-processing or rounding step after optimization, not part of the model used during optimization (Zheng et al., 2018) Therefore, we do not threshold W during optimization.

A separate strand of work attacks latent confounding without an explicit latent inference step. Constraint-based methods such as FCI and RFCI (Spirtes et al., 2000; Zhang, 2008) return a partial ancestral graph. Acyclic Directed Mixed Graph (ADMG)-based score methods promote bidirected confounding edges to continuous parameters jointly optimized with the directed edges (Bhattacharya et al., 2021; Ashman et al., 2023; Wang and Drton, 2023). Low-rank-plus-sparse decompositions of the precision matrix (Chandrasekaran et al., 2012; Frot et al., 2019; Bellot and van der Schaar, 2021) absorb pervasive unobserved confounding into a low-rank component before a DAG learner. Kaltenpoth and Vreeken (2023) embed continuous latents in a non-linear differentiable framework but do not connect to the GEM monotonicity argument. Each modifies either the inference object or the loss landscape; none formulates the problem as a GEM with an *exact* latent E-step and a differentiable M-step over an explicit-latent structural parameter.

The framework in this paper takes the explicit-latent route on a model class that resolves the interface problem cleanly: linear–Gaussian SEMs $Z = W^\top Z + \varepsilon$ with $Z = (X, H)$, X observed and H latent, and full joint graph acyclic. We interpret W as a parameter of

a generative Gaussian model. The only condition the E-step needs of W is that $(I - W^\top)$ be invertible (Appendix B), strictly weaker than exact acyclicity, and maintained in practice well before the acyclicity penalty $h(W)$ reaches zero. The E-step is then closed-form Gaussian conditioning, returning posterior first *and second moments* of H given X: the Gaussian analogue of Friedman's expected counts $\mathbb{E}[N_{ijk}]$. The M-step runs an inner Adam optimizer on a penalized likelihood, and at the outer-iteration boundary accepts the candidate only if the *unpenalized* expected complete-data objective Q_t has improved; the penalties shape the inner trajectory but do not enter the acceptance check. Section 2 keeps three notions of validity distinct: probabilistic validity along the trajectory, DAG semantics in the limit, and statistical identifiability.

Contributions. (i) A continuous explicit-latent realization of Structural EM for linear–Gaussian models, combining exact Gaussian inference, smooth graph optimization, and an unpenalized acceptance safeguard. (ii) A derivation of the required posterior first and second moments, showing why posterior covariance distinguishes proper Gaussian EM from posterior-mean imputation. (iii) A generalized-EM monotonicity theorem, kept distinct from probabilistic validity, DAG semantics, and identifiability. (iv) A 2×2 ablation of the key algorithmic ingredients and a linear–Gaussian replication of Friedman's hidden-variable experiment (Friedman, 1998, Sec. 5.3) on mediator and confounder topologies.

2. Method

Notation and model class. Let $\{x^{(i)}\}_{i=1}^n \subset \mathbb{R}^{p_x}$ denote n centered observations of the observed variables $X \in \mathbb{R}^{p_x}$, let $H \in \mathbb{R}^{p_h}$ denote the latent (never-observed) variables, and write $Z = (X, H) \in \mathbb{R}^p$ with $p = p_x + p_h$. We consider the linear–Gaussian structural equation model (SEM) in which each Z_i is a linear combination of its parents plus Gaussian noise. Stacking the coefficients into $W \in \mathbb{R}^{p \times p}$ with the convention:

$$W_{ij} \;=\; \text{weight of the edge } i \to j, \qquad W_{ii} = 0, \tag{1}$$

the structural equation for Z_i is $Z_i = \sum_j W_{ji} Z_j + \varepsilon_i$, or in matrix form:

$$Z \;=\; W^\top Z + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \Sigma_\varepsilon), \tag{2}$$

with Σ_ε diagonal. The outgoing edges $H_k \to X_j$ of the k-th latent are stored at $W_{p_x+k,\,j}$, i.e. in *row* $p_x + k$ of W; we call these the *latent rows*, and they encode the latent-to-observed loadings, which under the SEM convention play the role of factor loadings in classical factor analysis (Anderson, 2003, ch. 14). This matters for initializing the latent rows of W later.

Joint covariance and three notions of validity. The SEM (2) induces a Gaussian joint law whenever $(I - W^\top)$ is invertible, with covariance (Drton et al., 2011):

$$\Sigma_Z(W) \;=\; (I - W^\top)^{-1}\, \Sigma_\varepsilon\, (I - W)^{-1}. \tag{3}$$

We keep three notions of validity for W distinct. *Probabilistic validity:* Equation (3) requires only that $I - W^\top$ be invertible, equivalently that $1 \notin \operatorname{spec}(W)$, since W and $W^\top$ have the same spectrum. This condition is strictly weaker than acyclicity: a cyclic weighted graph may still satisfy $1 \notin \operatorname{spec}(W)$. Acyclicity is sufficient because an acyclic weighted adjacency matrix is nilpotent and therefore $I - W^\top$ is invertible; see Appendix B. *DAG semantics:* acyclicity ($h(W) = 0$) permits interpreting W as a directed acyclic graphical model. *Optimization trajectory:* during smooth acyclicity-penalized optimization, intermediate iterates can satisfy probabilistic validity without exact acyclicity; the penalty drives

$h(W) \to 0$ in the limit. The E-step algebra below relies on probabilistic validity only; DAG semantics apply to the final accepted iterate. Thresholding is never used inside the E-step or the optimization loop — only to extract a discrete graph from $\widehat{W}$ for evaluation after fitting completes.

Exact Gaussian E-step. At iteration t, given the current parameters $\phi_t = (W_t, \Sigma_{\varepsilon,t})$ with $(I - W_t^\top)$ invertible, form $\Sigma_Z(\phi_t)$ from (3) and partition it into observed and latent blocks,

$$\Sigma_Z(\phi_t) \;=\; \begin{pmatrix} \Sigma_{XX,t} & \Sigma_{XH,t} \\ \Sigma_{HX,t} & \Sigma_{HH,t} \end{pmatrix}, \quad \Sigma_{XX,t} \in \mathbb{R}^{p_x \times p_x},\;\; \Sigma_{HH,t} \in \mathbb{R}^{p_h \times p_h}.$$

Because Z is zero-mean Gaussian under (2), the posterior $q_t(H \mid X = x^{(i)}) = p(H \mid X = x^{(i)}; \phi_t)$ is Gaussian with moments

$$m_t^{(i)} \;=\; \Sigma_{HX,t}\,\Sigma_{XX,t}^{-1}\,x^{(i)} \;\in \mathbb{R}^{p_h}, \qquad V_t \;=\; \Sigma_{HH,t} - \Sigma_{HX,t}\,\Sigma_{XX,t}^{-1}\,\Sigma_{XH,t} \;\in \mathbb{R}^{p_h \times p_h}. \tag{4}$$

This is the standard Gaussian conditional formula (Bishop and Nasrabadi, 2006; Murphy, 2012)— the same identity used in Gaussian process regression (Williams and Rasmussen, 2006)— applied to a finite-dimensional joint (X, H) under the SEM-induced covariance. Note that V_t does not depend on the sample index i; only the conditional mean does.

The E-step does not estimate hidden values; it estimates the sufficient statistics needed by the complete-data Gaussian likelihood. Writing $\tilde{z}^{(i)} = (x^{(i)}, h^{(i)}) \in \mathbb{R}^{p_x+p_h}$, the expected sufficient statistics passed to the M-step are a $(p_x + p_h)$-vector and a $(p_x + p_h) \times (p_x + p_h)$ matrix (where $m_t^{(i)} = \mathbb{E}_{q_t}[h^{(i)}]$, $\quad V_t = \mathrm{Var}_{q_t}(h^{(i)})$),

$$\underbrace{\mathbb{E}_{q_t}\big[\tilde{z}^{(i)}\big]}_{(p_x+p_h)\times 1} = \begin{pmatrix} x^{(i)} \\ m_t^{(i)} \end{pmatrix}, \qquad \underbrace{\mathbb{E}_{q_t}\big[\tilde{z}^{(i)}\tilde{z}^{(i)\top}\big]}_{(p_x+p_h)\times(p_x+p_h)} = \begin{pmatrix} \underbrace{x^{(i)}x^{(i)\top}}_{p_x\times p_x} & \underbrace{x^{(i)}m_t^{(i)\top}}_{p_x\times p_h} \\ \underbrace{m_t^{(i)}x^{(i)\top}}_{p_h\times p_x} & \underbrace{V_t + m_t^{(i)}m_t^{(i)\top}}_{p_h\times p_h} \end{pmatrix}. \tag{5}$$

Each block of the second-moment matrix is a matrix-valued outer product, not a scalar; the dimension annotations are explicit to avoid that reading. The V_t contribution to the bottom-right block is the Gaussian analogue of using expected counts $\mathbb{E}[N_{ijk}]$ rather than completed counts in discrete Structural EM (Friedman, 1998; Chickering, 2002). In fact, V_t accounts for posterior uncertainty in the latent variables. Thus the E-step passes second moments, not point imputations: $\mathbb{E}[HH^\top \mid X = x^{(i)}] = m_t^{(i)}m_t^{(i)\top} + V_t$. Omitting V_t would treat the latent variables as known after conditioning and would underestimate the quadratic sufficient statistics used by the Gaussian M-step. In another words, replacing V_t by zero — treating the posterior mean $m_t^{(i)}$ as observed data — yields posterior-mean imputation, which we show in Section 4 breaks observed-data monotonicity in practice even when the M-step acceptance check below is active.

M-step: penalized proposal, unpenalized acceptance. Given the posterior q_t from the E-step, define the expected complete-data log-likelihood $Q_t(\phi) = \mathbb{E}_{q_t}\big[\log p_\phi(X, H)\big]$. The M-step proceeds in two stages. First, an inner Adam optimizer (Kingma and Ba, 2015) runs for a fixed number of iterations on the penalized objective

$$\mathcal{L}_t(\phi) \;=\; -Q_t(\phi) \;+\; \lambda\,\big\|W_{\mathrm{obs}}\big\|_1 \;+\; \rho\,h(W), \tag{6}$$

where $\|W_{\mathrm{obs}}\|_1$ is the ℓ_1 norm over observed rows of W (latent rows unregularized; see Section 4), and $h(W)$ is a differentiable acyclicity penalty (Zheng et al., 2018 or Bello et al., 2022). The optimizer produces a candidate $\phi_t^{\mathrm{cand}} = (W_t^{\mathrm{cand}}, \Sigma_{\varepsilon,t}^{\mathrm{cand}})$. Second, the *acceptance*

check uses the *unpenalized* Q_t:

$$\phi_{t+1} \;=\; \begin{cases} \phi_t^{\text{cand}} & \text{if } Q_t(\phi_t^{\text{cand}}) \;\geq\; Q_t(\phi_t), \\ \phi_t & \text{otherwise.} \end{cases} \tag{7}$$

Because the entropy of the exact posterior $q_t = p(\cdot \mid X; \phi_t)$ does not depend on ϕ, improving Q_t is equivalent to improving the variational free energy $\mathcal{F}(q_t, \phi) = \mathbb{E}_{q_t}[\log p_\phi(X, H)] + \mathcal{H}(q_t)$ by the same amount, so (7) ensures $\mathcal{F}(q_t, \phi_{t+1}) \geq \mathcal{F}(q_t, \phi_t)$, the GEM condition Theorem 1 needs. The inner Adam routine serves only to generate a candidate; the complete generalized M-step consists of this proposal followed by the unpenalized acceptance check in (7). We call the combined inner-loop-plus-acceptance rule the *improving-step safeguard*; the 2×2 ablation in Section 4 verifies that disabling it breaks monotonicity at the observed-data likelihood level, consistent with the theory.

Why two stages with separate criteria? Penalized gradient steps sometimes decrease the loss but Q_t gets worse, so a "good" Adam step under (6) can still move *away* from the EM auxiliary function. In early experiments this produced occasional but significant decreases in observed-data likelihood across outer iterations, which is exactly the failure mode the standard EM monotonicity argument is supposed to rule out. The fix is to keep the penalties as a proposal mechanism but check acceptance on Q_t, as the theorem requires. This is principled because the ℓ_1 and acyclicity penalties are optimization heuristics, not part of the probabilistic model: they shape the search but they are not terms in $\log p_\phi(X, H)$, and therefore not part of Q_t or of the variational free energy $\mathcal{F}$. Excluding them from the acceptance check is what makes the GEM monotonicity argument apply unchanged.

Post-processing and evaluation. After the outer loop terminates, $\widehat{W}$ is real-valued. We binarize observed-row entries at a fixed magnitude threshold and use a separate, lower threshold for latent-to-observed entries (fitted latent loadings have smaller magnitudes, an identifiability artifact of linear–Gaussian latent SEMs). Both thresholds are applied only to $\widehat{W}$ at the end of fitting; the optimization itself never sees a thresholded matrix. For full-graph metrics on synthetic data where ground-truth latent identities are known, fitted latent indices are arbitrary and must be aligned to the truth; we use a greedy bipartite match on the overlap-minus-symmetric-difference of (thresholded) observed-children sets. This alignment uses ground-truth information and is strictly an *evaluation* step — it is not used by the optimizer, not used to choose thresholds, and not used to select among runs. The headline metric *latent-parent recovery* (Section 4) is alignment-free: it asks only whether *some* common fitted latent parent exists for each confounded pair.

Initialization (factor warm start). We initialize the latent rows of W from the leading principal directions of the empirical observed covariance. Let $\frac{1}{n} X^\top X = \sum_{k=1}^{p_x} \mu_k \, u_k u_k^\top$ with $\mu_1 \geq \mu_2 \geq \cdots$, and let $\epsilon_k \in \{-1, +1\}$ be a random sign. The latent row corresponding to H_k is initialized as

$$W^{(0)}_{p_x+k,\,1:p_x} \;=\; \epsilon_k \sqrt{\mu_k}\, u_k^\top, \qquad k = 1, \ldots, p_h, \tag{8}$$

so that each latent contributes a rank-one component of variance μ_k along direction u_k to the observed marginal. This is the classical factor-analysis loading initialization, placed in rows rather than columns per convention (1). Observed-row entries of $W^{(0)}$ are small Gaussian; $\Sigma_{\varepsilon,0}$ is set to the mean per-variable observed variance (see Appendix O for a formal justification).

Algorithm 1: Differentiable Structural EM for explicit-latent linear–Gaussian DAGs.

Input: Centered observations $\{x^{(i)}\}_{i=1}^{n}$; latent dimension p_h; penalty weights λ (sparsity) and ρ (acyclicity); outer iterations T, inner Adam iterations K, learning rate η; acyclicity penalty $h \in \{\text{NOTEARS}, \text{DAGMA}\}$.

Output: Fitted $\widehat{W}$ over (X, H) and noise estimate $\widehat{\Sigma}_\varepsilon$.

Initialize W_0 by factor warm start (8) on latent rows and small Gaussian noise on observed rows; initialize $\Sigma_{\varepsilon,0}$

for $t = 0, 1, \ldots, T-1$ **do**

  E-step: compute $\Sigma_Z(\phi_t)$ via (3), the posterior moments $(m_t^{(i)}, V_t)$ via (4), and the expected sufficient statistics (5)

  M-step (proposal): run K inner Adam steps on the penalized objective $\mathcal{L}_t$ from (6), producing a candidate ϕ_t^{cand}

  M-step (acceptance safeguard): evaluate the *unpenalized* Q_t at ϕ_t^{cand} and at ϕ_t; set ϕ_{t+1} according to (7)

end

return ϕ_T

Algorithm Description. Algorithm 1 summarizes the complete procedure, alternating the E-step (4)–(5) with the M-step (6)–(7) until the observed-data log-likelihood (computed from (3) after marginalizing H; derivation in Appendix C) plateaus.

3. Theoretical Results

The first result is an algorithmic monotonicity guarantee for the observed-data log-likelihood under the exact E-step of (4)–(5) and the unpenalized-Q acceptance rule (7). The second clarifies what the guarantee does *not* say, in particular about the recovered graph's identifiability.

For parameters $\phi = (W, \Sigma_\varepsilon)$ with $(I - W^\top)$ invertible (so (2) defines a valid Gaussian joint law) and any distribution q on H, define the variational free energy

$$\mathcal{F}(q, \phi) \;=\; \mathbb{E}_q\big[\log p_\phi(X, H)\big] \;+\; \mathcal{H}(q), \tag{9}$$

where $\mathcal{H}(q)$ is the entropy of q. The standard variational identity gives the decomposition

$$\log p(X; \phi) \;=\; \mathcal{F}(q, \phi) \;+\; \mathrm{KL}\big(q(H) \,\big\|\, p(H \mid X; \phi)\big), \tag{10}$$

which by non-negativity of KL implies $\log p(X; \phi) \geq \mathcal{F}(q, \phi)$, with equality iff $q(H) = p(H \mid X; \phi)$ almost everywhere. The derivation is reproduced in Appendix A.

Theorem 1 (Monotonicity of differentiable Structural EM) *Let $\{\phi_t\}_{t \geq 0}$ be the sequence produced by Algorithm 1, and assume $(I - W_t^\top)$ is invertible at every iterate. Suppose that at iteration t:*

(i) The posterior $q_t = p(\cdot \mid X; \phi_t)$ is the exact Gaussian conditional given by (4).

(ii) The M-step's acceptance step (7) produces ϕ_{t+1} satisfying $Q_t(\phi_{t+1}) \geq Q_t(\phi_t)$.

Then the observed-data log-likelihood is monotonically nondecreasing: $\log p(X; \phi_{t+1}) \geq \log p(X; \phi_t)$.

Proof The entropy $\mathcal{H}(q_t)$ does not depend on ϕ, so condition (ii) is equivalent to $\mathcal{F}(q_t, \phi_{t+1}) \geq \mathcal{F}(q_t, \phi_t)$. Applying (10) at ϕ_t with the exact posterior q_t from (i) yields $\log p(X; \phi_t) =$

$\mathcal{F}(q_t, \phi_t)$ (the KL term vanishes). Applying (10) at ϕ_{t+1} with the same q_t gives $\log p(X; \phi_{t+1}) \geq \mathcal{F}(q_t, \phi_{t+1})$. Chaining, $\log p(X; \phi_{t+1}) \geq \mathcal{F}(q_t, \phi_{t+1}) \geq \mathcal{F}(q_t, \phi_t) = \log p(X; \phi_t)$. ■

Corollary 2 (Role of the penalties) *The acyclicity penalty $h(W)$ and the sparsity penalty $\lambda\|W\|_1$ in (6) do not enter condition (ii) of Theorem 1. They shape the inner-Adam trajectory inside the M-step but are not used by the acceptance rule (7), which compares only $Q_t(\phi_t^{\text{cand}})$ to $Q_t(\phi_t)$. Consequently the inner optimizer can be replaced by any procedure that proposes candidates — gradient descent, second-order methods, or penalty schedules of any kind — without affecting the monotonicity guarantee, provided the outer acceptance check is preserved.*

What Theorem 1 does and does not say. The theorem is a statement about the optimizer alone. Its *algorithmic* validity rests only on probabilistic validity: $(I - W_t^\top)$ invertible along the trajectory, which holds well before $h(W_t)$ reaches zero. *DAG semantics* apply to the final accepted $\widehat{W}$ once $h(\widehat{W}) \approx 0$; they govern how the recovered W is interpreted as a graphical model but not how the optimization proceeds. *Statistical identifiability* is a third question on which the monotonicity guarantee is silent. Observational linear–Gaussian DAGs are generally not fully identifiable; one recovers at best a Markov equivalence class in the fully observed case (Chickering, 2002; Spirtes et al., 2000), and explicit latent variables enlarge the ambiguity. We read our experimental claims either as (a) holding within an identifiable subclass such as the equal-error-variance setting of Peters and Bühlmann (2014), in which the strong recovery below obtains; or (b) outside such subclasses, as likelihood-based optimization whose structural conclusions are interpreted up to the observability-pattern equivalence. Recovering ancestral graphs over the observables alone (Richardson and Spirtes, 2002) is a different problem we do not address.

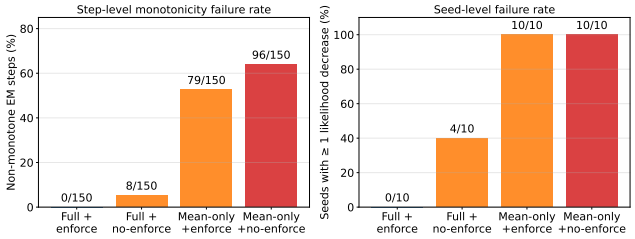
4. Experiment 1: Synthetic latent confounding

We evaluate the method on synthetic linear–Gaussian SEMs with explicit latent confounders. Two questions guide the experiments. First, does the full algorithm exhibit the monotone observed-data likelihood behavior predicted by Theorem 1, and which of its assumptions are necessary in practice rather than merely sufficient in principle? Second, under strong hidden confounding, does explicit latent modeling improve structure recovery relative to observed-only differentiable DAG learning, and what is responsible for any gain?

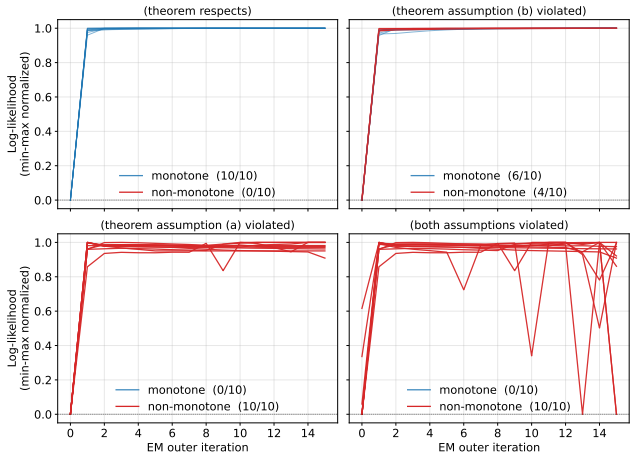
Setting. All experiments use linear–Gaussian SEMs with $p_x = 8$ observed and $p_h = 2$ latent variables, equal noise variances, and sample sizes $n \in \{200, 1000, 5000\}$. Results average over 10 random seeds. The proposed method runs for 15 outer EM iterations, each with 300 inner M-step Adam steps. Observed data are centered before fitting. The data-generating process puts most of $\Sigma_X = \text{Cov}(X)$ into hidden common causes rather than direct $X_i \to X_j$ edges: each latent is connected to four observed nodes with weights uniform in $[1.0, 2.0]$ (random sign), while observed–observed edges are sampled at density 0.10 along a random topological order with weights in $[0.3, 1.0]$. This is the regime in which an observed-only learner is structurally incentivized to fabricate spurious direct edges between confounded pairs. Code to reproduce the experiments is available at https://github.com/majavid/PGM26. More experimental settings are available at Appendicies L,M,N.

Methods. We compare twelve estimators along five orthogonal axes: acyclicity penalty (NOTEARS matrix-exponential vs DAGMA log-determinant); observability (observed-only vs explicit-latent wrapper); estimator (single fit vs stability selection); initialization (random vs warm start); and latent ℓ_1 regularization (asymmetric, latent rows unregularized, vs symmetric). The fully observed oracle that fits a DAG over (X, H) with H revealed serves as the upper bound. All baselines share the same ℓ_1 weight, the same ρ schedule for the acyclicity penalty, and the same Adam learning rate, so cross-method differences are attributable to the axis being varied. Two warm starts are included, both derived from the empirical second-moment structure of the observed data and matched in construction so their effects can be distinguished. The *factor warm start* (Section 2, Eq. (8)) initializes the latent rows of W from the top-p_h principal directions of $\frac{1}{n}X^\top X$, scaled by the square roots of the corresponding eigenvalues (Anderson, 2003; Bartholomew et al., 2011). The *OLS warm start* for the observed-only baselines initializes W from per-variable ridge OLS regressions of each X_j on the rest; this is the unconstrained Gaussian MLE that the M-step would converge to without the acyclicity penalty, and it is the analogue available to observed-only baselines without introducing a latent parameterization.

Metrics. The headline metric is the number of *spurious confounder-induced edges*: estimated directed edges between two observed variables that share a latent parent in the truth but no direct edge between them. We complement it with structural Hamming distance (SHD) on the observed subgraph, directed-edge precision, and recall. For latent-modeling methods we additionally report *latent-parent recovery*: the number of confounded observed pairs connected via a common fitted latent parent in the recovered model. This checks that improvements come from genuinely using the latent part of the model. Full-graph metrics on $(X, H) \times (X, H)$ are reported in Appendix F.



(a) Aggregate failure rates.



(b) Per-seed trajectories.

Figure 1: Monotonicity ablation across the 2×2 design (E-step variant × acceptance rule). **(a)** Fraction of EM steps with a log-likelihood decrease (left bars) and fraction of seeds with at least one decrease (right bars). **(b)** Per-seed log-likelihood trajectories (min–max normalized per seed). Top row: full Gaussian E-step; bottom row: mean imputation. Left column: unpenalized-Q acceptance check on; right column: check off.

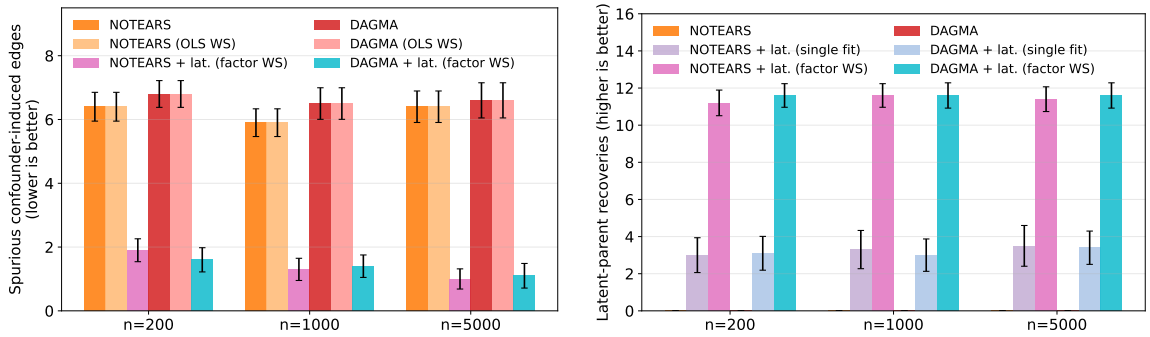
4.1. Monotonicity ablation: both ingredients are necessary

We first verify that the full procedure satisfies Theorem 1 in practice and isolate the contribution of each of its two conditions. Crossing the E-step variant (full posterior moments, i.e. condi-

tion (i), vs. posterior-mean imputation, i.e. V_t set to zero in (5)) with the acceptance rule (unpenalized-Q check enforced as in (7), i.e. condition (ii), vs. acceptance disabled so the candidate is always taken) gives four configurations. We refer to the conjunction of these two ingredients as the *improving-step safeguard*. As shown in Figure 1(a), the configuration that respects both Theorem 1 conditions produces 0/150 non-monotone steps and 0/10 seeds with any decrease. Disabling the acceptance check while keeping the exact E-step degrades to 8/150 steps and 4/10 seeds. With mean-only imputation, monotonicity breaks sharply — 79/150 steps and 10/10 seeds — even when the acceptance check is on. Both conditions of Theorem 1 are necessary in practice, and the posterior-second-moment assumption (i) dominates the acceptance-check assumption (ii) in the observed failure rates. The per-seed trajectories (Figure 1(b)) make the dichotomy along the E-step axis visible at the seed level: top trajectories are monotone ascents that converge within the first few EM iterations; bottom trajectories exhibit visible decreases, sometimes recovered later and sometimes not.

4.2. Structure recovery

The central structural finding compares, within each acyclicity family, three estimators: the observed-only baseline, its OLS warm-started variant (the most aggressive observation-side initialization we can give without a latent parameterization), and the explicit-latent model with factor warm start. If structure recovery were limited mainly by optimization quality, OLS warm start should already close most of the gap. It does not. For



(a) Spurious confounder-induced edges.

(b) Latent-parent recovery.

Figure 2: Structure-recovery headline. **(a)** OLS warm-starting the observed-only baselines does not change their behaviour under either acyclicity penalty; the latent + factor warm start closes most of the gap to the oracle. **(b)** Factor-warm-started latents approach the fully observed oracle, while random-init latents recover roughly a quarter of the structure. Latent capacity and a latent-aware initialization are both necessary; neither alone suffices.

DAGMA, the observed-only baseline and its OLS-warm-started variant produce $6.8, 6.5, 6.6$ and $6.8, 6.5, 6.6$ spurious edges at $n = 200, 1000, 5000$ respectively—identical to within standard error—and SHD is unchanged at roughly 8.0 throughout. The same null holds under NOTEARS. Optimization quality alone therefore does not explain any gain; even an aggressively warm-started observed-only model continues to fabricate spurious direct edges because its model class has no representation of a hidden common cause. The factor-warm-started latent variants tell the opposite story: under DAGMA, spurious edges drop to

1.6 ± 0.38, 1.4 ± 0.35, 1.1 ± 0.39, SHD drops to $2.7, 2.5, 2.1$, and precision rises from ~ 0.20 to 0.54–0.64. Recall is preserved at ~ 0.82. Under NOTEARS the pattern is the same: 1.9 ± 0.38, 1.3 ± 0.37, 1.0 ± 0.33 spurious edges with SHD $3.2, 2.7, 2.2$. The improvement is consistent across both acyclicity choices, supporting the claim that the framework is orthogonal to the underlying differentiable DAG mechanism (Figure 2).

The improvements above are not a coincidence of sparser fitted graphs: the factor-warm-started latents recover the underlying confounding structure at near-oracle fidelity. The oracle, given H directly, recovers 11.8–11.9 confounded pairs through a common latent parent (out of a per-seed total around 12); the DAGMA factor-warmstart variant recovers 11.6 ± 0.6 pairs at every n, statistically indistinguishable from the oracle, while the NOTEARS variant recovers 11.2–11.5 pairs. Random-init latent single-fit models recover only 3–3.5 pairs on average. Two further negative results sharpen the interpretation. Stability selection over random restarts does not help; in our experiments it is consistently worse than the warm-started single fit and sometimes worse than the random-init single fit (Appendix G). And regularising latent rows with the same ℓ_1 strength as observed rows (the symmetric-ℓ_1 ablation, Appendix H) improves only marginally over random-init and remains far from the factor-warmstart estimator. The asymmetric latent ℓ_1 is therefore not the source of the gains; it is a minor choice that prevents latent collapse but does not by itself solve the optimization problem. The complete twelve-method numerical table is in Appendix E. Two conclusions follow. Algorithmically, the procedure behaves as the GEM analysis predicts: monotone ascent obtains only when both ingredients of Theorem 1 are respected. Structurally, the bottleneck in the explicit-latent setting is not the probabilistic semantics (which are well-defined throughout optimization) but the nonconvex search induced by the latent parameterization; once the optimizer is guided into a latent-aware basin, the explicit-latent formulation recovers the confounding structure at near-oracle fidelity.

5. Experiment 2: Replication of Friedman's hidden-variable experiment

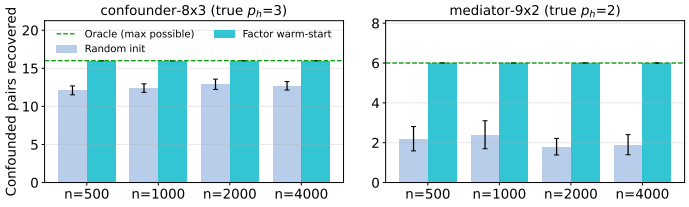
To position our method directly against Friedman's original Bayesian Structural EM (Friedman, 1998), we replicate the experimental design of his §5.3 ("Hidden Variables") on his two synthetic networks, adapted to our linear–Gaussian setting. *Network A (mediator-9×2)* has $p_x = 9$ observed and $p_h = 2$ hidden variables; the hiddens are not roots, but mediators H_1, H_2 whose parents are top observed variables X_1, X_2, X_3 and whose children are two disjoint groups of three bottom observed variables (Friedman's "3+2+{3+3}" topology). *Network B (confounder-8×3)* has $p_x = 8$ and $p_h = 3$; the three hidden roots each parent four consecutive observed variables with adjacent hiddens sharing two children (Friedman's "3×8" topology). See Figure 9 in Appendix K for diagrams.

Protocol. Following Friedman, we vary $n_{\text{train}} \in \{500, 1000, 2000, 4000\}$ and sweep $p_h^{\text{fit}} \in \{0, 1, 2, 3, 4\}$—the learner does not know the true number of latents. Held-out test sets of $n_{\text{test}} = 5000$ are drawn independently from the same SEM and results average over 10 seeds. We compare three estimators: observed-only DAGMA; our explicit-latent EM with random initialization; and our explicit-latent EM with the factor warm start of Section 4. Friedman's original networks are binary multinomial Bayesian networks; we replicate the *topologies* but parameterize them as linear–Gaussian SEMs (edge weights $|w| \sim \text{Uniform}[0.7, 1.3]$ with random sign, equal noise variance) because our method does not apply to discrete

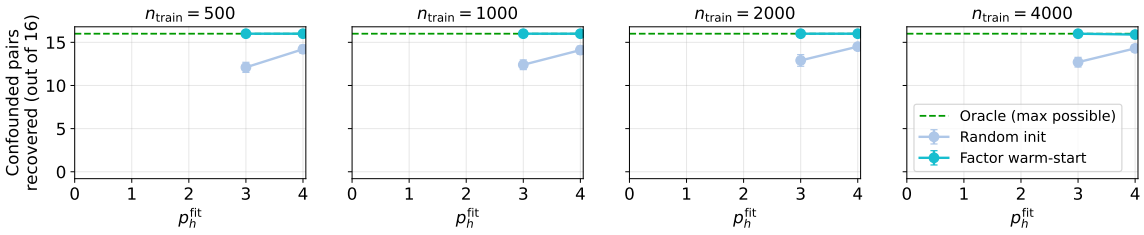
multinomials. The structural-recovery question—can the algorithm recover hidden-variable structure when p_h is unknown?—remains identical; only the distributional family differs.

Two metrics, two purposes. We report both metrics that Friedman did not have to distinguish, because in the Gaussian setting they tell different stories. *Test-set KL divergence to ground truth* (Friedman's metric) is computed using the closed-form Gaussian marginal $p(X \mid \widehat{W}, \widehat{\sigma}^2)$ obtained by integrating out the fitted latents analytically; the derivation is in Appendix C. This metric cleanly separates a no-latent fit from any-latent fit, but does not discriminate among $p_h^{\text{fit}} \in \{1, 2, 3, 4\}$ in our setting, since any latent model with sufficient rank matches the second-order moments of X to within sampling noise—a structural difference from Friedman's discrete setting, where misspecified p_h^{fit} does cost test log-loss because higher-order CPT structure exists. *Latent-parent recovery*, defined in Section 4, counts how many ground-truth confounded observed pairs are recovered through a common fitted latent parent in the learned graph; it tracks the actual scientific question of whether the algorithm identifies the hidden structure and discriminates strongly among methods.

Results: warm-start saturates the oracle on every seed. Figure 3 reports latent-parent recovery at the true p_h for both networks alongside a representative p_h^{fit} sweep. The factor warm start recovers the maximum possible structure on every seed of every cell: 16/16 confounded pairs on confounder-8×3 and 6/6 on mediator-9×2, with zero seed-to-seed variation. Random initialization plateaus at roughly 78% recovery on the confounder network (12.1–12.9 pairs out of 16) and 33% on the mediator network (1.8–2.4 pairs out of 6), and this gap does *not* close as n_{train} grows from 500 to 4000. The mediator network



(a) Recovery at the true p_h, both networks.



(b) Recovery vs p_h^{fit} on confounder-8×3.

Figure 3: Friedman §5.3 replication: latent-parent recovery. **(a)** At the true p_h, factor warm start recovers all confounded pairs on every seed of every cell (zero error bars); random initialization plateaus at 33–78% and does not improve with more data. **(b)** True $p_h = 3$ marked dashed. Recovery is reported only for $p_h^{\text{fit}} \geq p_h^{\text{true}}$, since alignment is ill-defined when the fit has fewer latents than the truth. Mediator-network results are in Appendix K.

is the stronger test of the two. The truth violates the warm-start's implicit "latents-as-roots" bias, since the hidden mediators H_1, H_2 are children of X_1, X_2, X_3 rather than roots. Despite this mismatch, warm-start recovers the latent structure perfectly, indicating that

the eigenvector initialization works as long as the latents explain a substantial fraction of observed marginal variance, regardless of whether they are roots or mediators. Within the recoverable regime ($p_h^{\text{fit}} \geq p_h^{\text{true}}$), warm-start maintains exact recovery at $p_h^{\text{fit}} = 4$ on the confounder network (the extra latent produces no spurious structure) while random initialization improves only modestly with over-specification (14.2/16 at $p_h^{\text{fit}} = 4$ vs 12.1/16 at the true p_h, on the $n = 500$ cell). Mediator-network results are in Appendix K. One caveat: on the mediator network, warm-start exhibits a slightly *higher* test KL than random init despite the exact structural recovery (e.g. 0.045 vs 0.029 at $n = 4000$; Appendix K). The warm-start basin better matches the qualitative confounding structure but slightly less well matches the marginal density — evidence that in the linear–Gaussian setting the second-order density fit and the structural recovery objective are not perfectly aligned.

Comparison to Friedman. Friedman's reported implementation used greedy search over the discrete DAG space, supplemented by random restarts and edge perturbations to escape local optima. In our continuous linear–Gaussian implementation, the factor warm start substantially reduces the need for these implementation-level heuristics and, in the experiments reported here, eliminates the need for random restarts and perturbation-based escapes. The two approaches are therefore complementary: Friedman's experiments show that score-based discrete search can reach good models when combined with suitable search heuristics, whereas our results show that, in this linear–Gaussian regime, a principled spectral initialization can guide continuous optimization toward a favorable latent-structure basin without that restart machinery. We make no claim that this conclusion extends to non-Gaussian, nonlinear, or substantially larger-scale settings.

6. Limitations

Scalability: Each iteration scales as $O(p^3)$ through inversion of Σ_{XX} and matrix-exp / log-det evaluation in the M-step. At our scale ($p_x = 8$, $p_h \leq 3$) this is inexpensive, but the cost grows rapidly with the number of latent and observed nodes. Sparse W induces exploitable precision-matrix structure; very large graphs or many latents will likely require additional approximations (blockwise optimization, screening rules, approximate posteriors), which we do not explore. **Identifiability:** Observational linear–Gaussian DAGs are not identifiable without further assumptions, and explicit latents enlarge the equivalence class. The strong empirical recovery in Sections 4–5 relies on the equal-error-variance setting of (Peters and Bühlmann, 2014), which we present as an optional strengthening condition rather than an implicit default. Non-Gaussian identifiability tools (Shimizu et al., 2006; Wang and Drton, 2023) are an orthogonal route we do not pursue here. **Scope and empirical reach:** Our framework treats latent variables as explicit nodes in a fully specified acyclic SEM over (X, H) with p_h specified in advance — a narrower problem than discovering an unknown number of hidden confounders or learning ancestral graphs (Richardson and Spirtes, 2002). All experiments use synthetic data from the same family the method models, which is the right setting for validating Theorem 1 but does not establish performance under misspecification, non-Gaussian noise, or unknown p_h (cf. Reisach et al., 2021; Ng et al., 2024). Extending to real domains and to selecting p_h are natural next steps.

Acknowledgments

The authors thank the anonymous reviewers for their thoughtful comments and constructive feedback, which helped strengthen the current version of the paper. Generative AI tools were used to assist with language editing and to improve coding and implementation tasks. The authors critically reviewed, revised, and verified all AI-assisted material, including the resulting code and outputs, and take full responsibility for the content, correctness, originality, and scientific integrity of the paper.

References

Theodore W. Anderson. *An Introduction to Multivariate Statistical Analysis.* Wiley-Interscience, 3rd edition, 2003.

Matthew Ashman, Chao Ma, Agrin Hilmkil, Joel Jennings, and Cheng Zhang. Causal reasoning in the presence of latent confounders via neural ADMG learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Jushan Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71(1): 135–171, 2003.

Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45 (1):77–120, 2017.

David J Bartholomew, Martin Knott, and Irini Moustaki. *Latent variable models and factor analysis: A unified approach.* John Wiley & Sons, 2011.

Kevin Bello, Bryon Aragam, Pradeep Ravikumar, and Jinchi Yu. 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.

Alexis Bellot and Mihaela van der Schaar. Deconfounded score method: Scoring DAGs with dense unobserved confounding. *arXiv preprint arXiv:2103.15106*, 2021.

Rohit Bhattacharya, Tushar Nagarajan, Daniel Malinsky, and Ilya Shpitser. Differentiable causal discovery under unmeasured confounding. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2314–2322, 2021.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Venkat Chandrasekaran, Pablo A. Parrilo, and Alan S. Willsky. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967, 2012.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–38, 1977.

Mathias Drton, Rina Foygel, and Seth Sullivant. Global identifiability of linear structural equation models. *The Annals of Statistics*, 39(2):865 – 886, 2011. doi: 10.1214/10-AOS859. URL https://doi.org/10.1214/10-AOS859.

Nir Friedman. Learning belief networks in the presence of missing values and hidden variables. In *Proceedings of the Fourteenth International Conference on Machine Learning (ICML)*, pages 125–133. Morgan Kaufmann, 1997.

Nir Friedman. The Bayesian structural EM algorithm. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 129–138. Morgan Kaufmann, 1998.

Benjamin Frot, Preetam Nandy, and Marloes H. Maathuis. Robust causal structure learning with some hidden variables. *Journal of the Royal Statistical Society: Series B*, 81(3):459–487, 2019.

David Kaltenpoth and Jilles Vreeken. Nonlinear causal discovery with latent confounders. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural DAG learning. In *International Conference on Learning Representations (ICLR)*, 2020.

Thomas A. Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society: Series B*, 44(2):226–233, 1982.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B*, 72(4):417–473, 2010.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and DAG constraints for learning linear DAGs. In *Advances in Neural Information Processing Systems*, volume 33, pages 17943–17954, 2020.

Ignavier Ng, Biwei Huang, and Kun Zhang. Structure learning with continuous optimization: A sober look and beyond. In *Proceedings of the 3rd Conference on Causal Learning and Reasoning (CLeaR)*, 2024.

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

Alexander G. Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated DAG! Causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34:27772–27784, 2021.

Thomas Richardson and Peter Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

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

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

G. W. Stewart and Ji-guang Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Y. Samuel Wang and Mathias Drton. Causal discovery with unobserved confounding and non-Gaussian data. *Journal of Machine Learning Research*, 24(271):1–61, 2023.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

C. F. Jeff Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 7154–7163, 2019.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.

Xun Zheng, Bryon Aragam, Pradeep K. 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 (AISTATS)*, pages 3414–3425, 2020.

Appendix A. Variational decomposition: detailed derivation

For completeness we reproduce the standard derivation underlying the variational identity (10). Starting from the marginal likelihood,

$$\log p(X;\phi) = \log \int p_\phi(X,H)\, dH = \log \int q(H)\, \frac{p_\phi(X,H)}{q(H)}\, dH$$

for any density $q(H)$ whose support is contained in that of $p(H \mid X;\phi)$. Using $\log p_\phi(X,H) = \log p(H \mid X;\phi) + \log p(X;\phi)$ and expanding $\mathrm{KL}(q \,\|\, p(\cdot \mid X;\phi))$,

$$\mathrm{KL}\big(q \,\|\, p(\cdot \mid X;\phi)\big) = \mathbb{E}_q[\log q(H)] - \mathbb{E}_q[\log p_\phi(X,H)] + \log p(X;\phi).$$

Rearranging and using $\mathcal{H}(q) = -\mathbb{E}_q[\log q(H)]$ gives

$\log p(X;\phi) = \mathbb{E}_q[\log p_\phi(X,H)] + \mathcal{H}(q) + \mathrm{KL}\big(q \,\|\, p(\cdot \mid X;\phi)\big) = \mathcal{F}(q,\phi) + \mathrm{KL}\big(q \,\|\, p(\cdot \mid X;\phi)\big)$,

which is (10). Non-negativity of the KL gives $\log p(X;\phi) \geq \mathcal{F}(q,\phi)$ with equality iff $q = p(\cdot \mid X;\phi)$ almost everywhere.

Appendix B. Acyclicity implies invertibility

The following result establishes that acyclicity is sufficient for the probabilistic validity required by (3). The converse does not hold: $I - W^\top$ may be invertible even when the directed graph associated with W contains cycles.

Lemma 3 *If the directed graph associated with $W \in \mathbb{R}^{n\times n}$ is acyclic, then W is nilpotent. Consequently, both $I - W$ and $I - W^\top$ are invertible, with*

$$(I - W)^{-1} = \sum_{k=0}^{n-1} W^k$$

and

$$(I - W^\top)^{-1} = \sum_{k=0}^{n-1} (W^\top)^k.$$

Proof Because the directed graph associated with W is acyclic, its vertices admit a topological ordering. Let P be a permutation matrix implementing such an ordering and define

$$\widetilde{W} = PWP^\top.$$

Under the convention that $W_{ij} \neq 0$ represents the edge $i \to j$, every edge points from an earlier vertex to a later vertex in this ordering. Hence $\widetilde{W}$ is strictly upper triangular.

For a strictly upper-triangular matrix, an elementary induction shows that

$$(\widetilde{W}^{\,k})_{ij} = 0 \qquad \text{whenever } j < i + k.$$

For $k = n$, no pair (i,j) can satisfy $j \geq i + n$, and therefore

$$\widetilde{W}^{\,n} = 0.$$

Since nilpotence is preserved under similarity,

$$W^n = (P^\top \widetilde{W} P)^n = P^\top \widetilde{W}^{\,n} P = 0.$$

The finite geometric-series identity now gives

$$(I - W) \sum_{k=0}^{n-1} W^k = I - W^n = I$$

and

$$\left(\sum_{k=0}^{n-1} W^k\right)(I - W) = I - W^n = I.$$

Thus,

$$(I - W)^{-1} = \sum_{k=0}^{n-1} W^k.$$

Finally,

$$I - W^\top = (I - W)^\top,$$

so $I - W^\top$ is invertible and

$$(I - W^\top)^{-1} = \big((I - W)^{-1}\big)^\top = \sum_{k=0}^{n-1} (W^\top)^k.$$

■

Appendix C. Closed-form Gaussian marginal density for test-set evaluation

This appendix derives the closed-form expression for the marginal density $p(X \mid W, \sigma^2)$ used to compute test-set log-likelihood in Sections 4–5 and in Appendix K. Recall the SEM $Z = W^\top Z + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_p)$ and W acyclic. Rearranging gives $(I - W^\top)Z = \varepsilon$, so $Z = (I - W^\top)^{-1}\varepsilon$, and

$$\Sigma_Z(W, \sigma^2) = \mathrm{Cov}(Z) = \sigma^2 (I - W)^{-\top} (I - W)^{-1}.$$

Partitioning into observed and latent blocks, the marginal of X obtained by integrating out H is

$$X \sim \mathcal{N}\big(0,\, \Sigma_X(W, \sigma^2)\big), \qquad \Sigma_X(W, \sigma^2) = [\Sigma_Z(W, \sigma^2)]_{1:p_x,\, 1:p_x}.$$

For a held-out test point $x \in \mathbb{R}^{p_x}$, the log-density under the fitted SEM $(\widehat{W}, \widehat{\sigma}^2)$ is the standard zero-mean multivariate Gaussian

$$\log p(x \mid \widehat{W}, \widehat{\sigma}^2) = -\tfrac{1}{2}\Big[p_x \log(2\pi) + \log\det \Sigma_X(\widehat{W}, \widehat{\sigma}^2) + x^\top \Sigma_X(\widehat{W}, \widehat{\sigma}^2)^{-1} x\Big].$$

Following Friedman (1998) we report the test-set log-loss *difference* between the fitted model and the ground truth, averaged over an independent test set $\{x^{(j)}\}_{j=1}^{N_{\text{test}}}$:

$$\mathrm{KL}_{\text{test}}(\widehat{W}, \widehat{\sigma}^2 \,\|\, W^\star, \sigma^{2\star}) = \frac{1}{N_{\text{test}}} \sum_{j=1}^{N_{\text{test}}} \Big[\log p(x^{(j)} \mid W^\star, \sigma^{2\star}) - \log p(x^{(j)} \mid \widehat{W}, \widehat{\sigma}^2)\Big].$$

This is the empirical KL divergence $\widehat{\mathrm{KL}}(p_{\text{true}} \,\|\, p_{\text{fit}})$ up to the constant test-set entropy of p_{true}, which cancels when comparing methods on the same test set.

Numerical considerations. $(I - W)^{-1}$ is well-defined whenever W is acyclic but can become ill-conditioned near the boundary of the DAG constraint where the spectral radius of W approaches 1. We add a ridge $10^{-8} I$ to Σ_X before inversion; this affects the log det and quadratic form by amounts negligible relative to the seed-to-seed standard error. When $\Sigma_X(\widehat{W}, \widehat{\sigma}^2)$ is not numerically positive definite (an occasional issue with under-regularized fits at small n), we fall back to a pseudo-inverse with eigenvalue floor $\max(\lambda_i, 10^{-8})$; this affected fewer than 0.5% of cells. For the no-latent baseline ($p_h^{\text{fit}} = 0$), $\widehat{\sigma}^2$ must be the value estimated by the fitting procedure itself, not a default value such as $\widehat{\sigma}^2 = 1$, since a misspecified $\widehat{\sigma}^2$ inflates the log-loss by an amount that scales with p_x and dominates the comparison across methods.

Appendix D. Hyperparameters and implementation details

All experiments use the following settings unless stated otherwise. The proposed method runs for 15 outer EM iterations with 300 inner M-step gradient steps per outer iteration, optimized by Adam (Kingma and Ba, 2015) with learning rate 0.01 in float64 precision on CPU. The ℓ_1 regularization weight is $\lambda = 0.02$ on observed rows of W. Latent rows are unregularized by default ($\lambda_{\text{lat}} = 0$); the symmetric-ℓ_1 ablation (Appendix H) sets $\lambda_{\text{lat}} = \lambda$. The acyclicity penalty coefficient is $\rho = 0.5$ inside the M-step; the observed-only baselines use a short rho schedule $\rho \in \{1, 10, 100\}$ with 800 Adam iterations per stage, matched between NOTEARS and DAGMA. Stability selection uses $n_{\text{inits}} = 16$ random initializations with per-fit binarization threshold 0.15 and consensus frequency threshold 0.75. The factor warm start uses the top-p_h eigenvectors of $\frac{1}{n}X^\top X$ scaled by the square roots of the corresponding eigenvalues, with random sign per latent. The OLS warm start uses ridge regression with ridge coefficient 10^{-3}, applied per variable. For the monotonicity check we count an EM step as non-monotone if the observed-data log-likelihood decreases by more than 10^{-3}. Edges are binarized at threshold 0.30 for observed-subgraph metrics and 0.15 for latent-parent identification, reflecting the empirically smaller magnitudes of fitted latent–observed weights. Latent labels are aligned to the ground truth via a greedy bipartite match on observed-children Jaccard score before any full-graph metric is computed.

Appendix E. Experiment 1: complete results table

Table 1 reports all metrics for all twelve methods at all three sample sizes. Bold entries mark the best non-oracle method on each (metric, n) cell. The table is the source of every numerical claim in Section 4.

Two patterns are worth highlighting against the table. First, per-method behaviour is essentially flat in n across the range studied; this is consistent with the interpretation in Section 4 that observed-only error is structural (model-class limitation, not statistical), and the latent factor-warmstart variant has already converged to a near-oracle local optimum. Second, recall is roughly comparable across all non-oracle methods (0.76–0.85); the dramatic differences are in precision and in the spurious-edge / latent-recovery diagnostics. This is what one would expect if the difference between methods is which sparse structure they propose to fit, rather than how much of the true structure they cover.

Appendix F. Full-graph recovery metrics

For latent-modelling methods we additionally report SHD, precision, and recall on the full $(X, H) \times (X, H)$ graph. Latent labels are aligned to the ground truth via greedy bipartite match on observed-children Jaccard score before computing these metrics (Figure 4).

The full-graph metrics are weaker than the latent-parent recovery diagnostic in the main text and we treat them as secondary. Two reasons: full-graph SHD penalizes latent-block edge weight errors that do not materially affect the qualitative recovery of which pairs share a hidden common cause; and the alignment step is a greedy match on a small $(p_h \times p_h)$ bipartite problem, robust at $p_h = 2$ but not unconditionally extrapolable to larger p_h without studying the alignment quality itself.

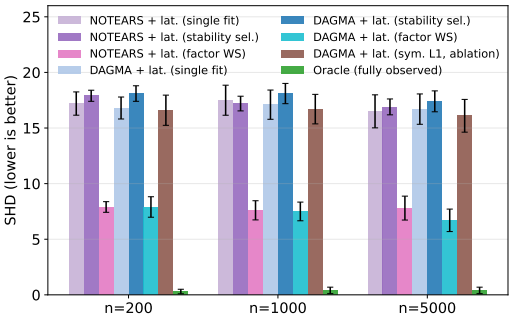
Table 1: Experiment 1 results: mean ± standard error over 10 seeds. Best non-oracle method per (metric, n) cell in bold. Spurious = spurious confounder-induced edges; SHD_{obs} = SHD on the observed subgraph; P, R = directed-edge precision and recall on the observed subgraph; #recov = latent-parent recoveries (n/a for observed-only methods).

Method	n	Spurious	SHD_{obs}	P	R	#recov
NOTEARS obs-only	200	6.4 ± 0.45	8.1 ± 0.52	0.21 ± 0.02	0.76 ± 0.08	—
	1000	5.9 ± 0.43	7.5 ± 0.53	0.23 ± 0.03	0.78 ± 0.08	—
	5000	6.4 ± 0.49	8.0 ± 0.60	0.21 ± 0.03	0.76 ± 0.08	—
NOTEARS obs-only + OLS warm	200	6.4 ± 0.45	8.1 ± 0.52	0.21 ± 0.02	0.76 ± 0.08	—
	1000	5.9 ± 0.44	7.5 ± 0.53	0.23 ± 0.03	0.78 ± 0.08	—
	5000	6.4 ± 0.49	8.0 ± 0.60	0.21 ± 0.03	0.76 ± 0.08	—
DAGMA obs-only	200	6.8 ± 0.42	8.3 ± 0.45	0.20 ± 0.03	0.76 ± 0.08	—
	1000	6.5 ± 0.49	8.1 ± 0.65	0.22 ± 0.03	0.76 ± 0.08	—
	5000	6.6 ± 0.55	8.0 ± 0.68	0.22 ± 0.03	0.76 ± 0.08	—
DAGMA obs-only + OLS warm	200	6.8 ± 0.42	8.3 ± 0.45	0.20 ± 0.03	0.76 ± 0.08	—
	1000	6.5 ± 0.49	8.1 ± 0.65	0.22 ± 0.03	0.76 ± 0.08	—
	5000	6.6 ± 0.55	8.0 ± 0.68	0.22 ± 0.03	0.76 ± 0.08	—
NOTEARS + lat (single fit)	200	5.4 ± 0.40	8.3 ± 0.72	0.23 ± 0.03	0.85 ± 0.08	3.0 ± 0.94
	1000	5.7 ± 0.53	8.7 ± 0.91	0.23 ± 0.04	0.84 ± 0.08	3.3 ± 1.03
	5000	5.3 ± 0.49	8.0 ± 1.06	0.26 ± 0.03	0.88 ± 0.08	3.5 ± 1.10
NOTEARS + lat (stability sel.)	200	6.6 ± 0.49	9.7 ± 0.51	0.19 ± 0.03	0.79 ± 0.09	0.0 ± 0.00
	1000	6.6 ± 0.62	9.2 ± 0.63	0.20 ± 0.02	0.82 ± 0.08	0.1 ± 0.09
	5000	6.5 ± 0.71	8.8 ± 0.72	0.21 ± 0.02	0.82 ± 0.08	0.1 ± 0.09
DAGMA + lat (single fit)	200	5.3 ± 0.40	8.1 ± 0.75	0.23 ± 0.03	0.82 ± 0.08	3.1 ± 0.91
	1000	5.6 ± 0.51	8.5 ± 0.94	0.22 ± 0.04	0.79 ± 0.09	3.0 ± 0.87
	5000	5.2 ± 0.39	8.1 ± 0.98	0.23 ± 0.03	0.78 ± 0.08	3.4 ± 0.90
DAGMA + lat (stability sel.)	200	6.6 ± 0.64	9.7 ± 0.79	0.19 ± 0.03	0.79 ± 0.09	0.2 ± 0.13
	1000	6.8 ± 0.70	10.0 ± 0.87	0.19 ± 0.03	0.84 ± 0.07	0.4 ± 0.29
	5000	6.7 ± 0.72	9.5 ± 0.86	0.20 ± 0.03	0.82 ± 0.08	0.6 ± 0.38
DAGMA + lat (sym. ℓ_1)	200	5.0 ± 0.40	7.7 ± 0.99	0.26 ± 0.05	0.84 ± 0.08	1.4 ± 0.57
	1000	4.8 ± 0.39	7.5 ± 1.09	0.26 ± 0.05	0.79 ± 0.09	1.2 ± 0.51
	5000	4.7 ± 0.49	7.3 ± 1.10	0.28 ± 0.06	0.82 ± 0.08	1.5 ± 0.71
NOTEARS + lat (factor warm)	200	$\mathbf{1.9 \pm 0.38}$	$\mathbf{3.2 \pm 0.47}$	0.50 ± 0.06	0.82 ± 0.07	11.2 ± 0.73
	1000	$\mathbf{1.3 \pm 0.37}$	$\mathbf{2.7 \pm 0.58}$	0.55 ± 0.07	0.83 ± 0.07	11.3 ± 0.67
	5000	1.0 ± 0.33	2.2 ± 0.57	$\mathbf{0.67 \pm 0.06}$	0.85 ± 0.07	11.5 ± 0.70
DAGMA + lat (factor warm)	200	$\mathbf{1.6 \pm 0.38}$	$\mathbf{2.7 \pm 0.45}$	$\mathbf{0.54 \pm 0.08}$	0.84 ± 0.07	$\mathbf{11.6 \pm 0.64}$
	1000	$\mathbf{1.4 \pm 0.35}$	$\mathbf{2.5 \pm 0.49}$	$\mathbf{0.54 \pm 0.06}$	0.82 ± 0.07	$\mathbf{11.6 \pm 0.68}$
	5000	$\mathbf{1.1 \pm 0.39}$	$\mathbf{2.1 \pm 0.48}$	$\mathbf{0.64 \pm 0.06}$	0.82 ± 0.08	$\mathbf{11.6 \pm 0.68}$
Oracle (fully observed)	200	0.0 ± 0.00	0.1 ± 0.09	1.00 ± 0.00	0.97 ± 0.02	11.9 ± 0.64
	1000	0.0 ± 0.00	0.3 ± 0.20	0.96 ± 0.04	0.95 ± 0.03	11.8 ± 0.63
	5000	0.0 ± 0.00	0.3 ± 0.20	0.96 ± 0.04	0.95 ± 0.03	11.8 ± 0.63

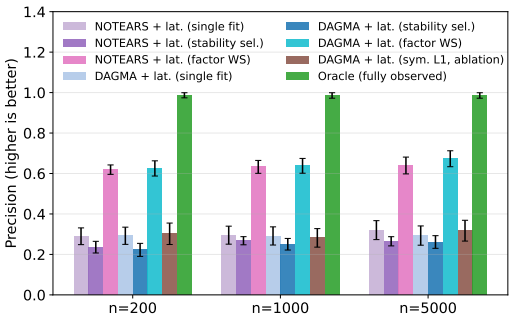
Appendix G. Stability selection: a negative result

Figure 5 compares the stability-selected estimator with the random-init single-fit estimator for both NOTEARS and DAGMA. Stability selection performs worse on every metric: spurious confounder-induced edges increase from roughly 5.3–5.7 for the single-fit variant to roughly 6.5–6.8 for the stability-selected variant, observed-subgraph SHD increases correspondingly, and latent-parent recovery essentially collapses (from $\sim$ 3–3.5 pairs to fewer than 1).

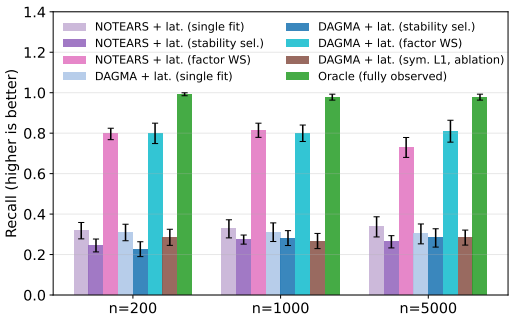
The mechanism is straightforward. Stability selection (Meinshausen and Bühlmann, 2010) assumes that the true graph is the one most consistently recovered across resamples or restarts. Here, the variance across random initializations is precisely the basin variance



(*a*) SHD.



(*b*) Precision.

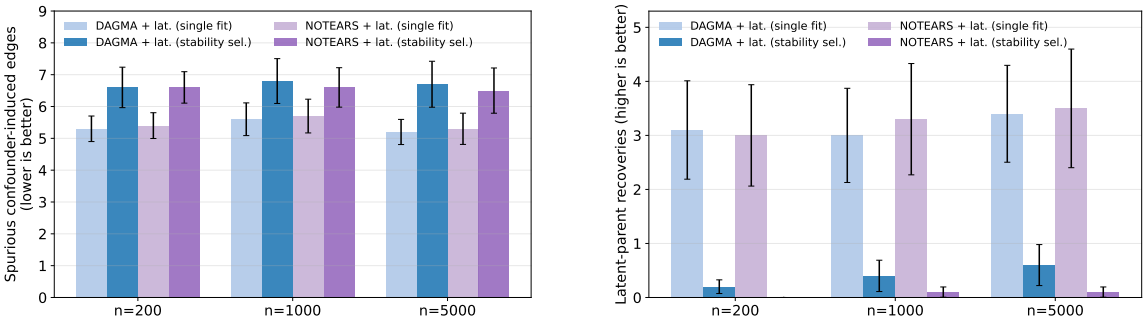


(*c*) Recall.

Figure 4: Full-graph metrics on $(X, H) \times (X, H)$ after latent-label alignment. Factor-warmstart variants reduce SHD and improve precision substantially over random init but leave a residual gap to the oracle on these stricter metrics, indicating recovery of the qualitative confounding structure but not exact latent weights or finer latent topology. Recall is high across all latent-modelling methods including random init, consistent with random-init producing dense graphs that cover the true edges but also fabricate many spurious ones.

that the latent factor warm start is designed to eliminate: most random-init runs land in basins where latent rows of W are weakly identified and where the specific latent edges that emerge are different from run to run. Aggregating these by edge frequency averages away the rare initializations that did discover the right latent structure. This is not an indictment of stability selection in general; it states that, in the explicit-latent SEM setting and at this scale, basin variance dominates estimation variance, and warm-starting the optimization is a more direct fix than averaging over basins.

Cross-fit alignment uses no ground truth. A subtlety worth disclosing explicitly: because the fitted latent indices are arbitrary, comparing the latent block of W across random restarts requires first aligning each fit's latent labels to a common reference. We use the *first* fit's latent labels as the reference and align fits $2, \ldots, n_{\text{inits}}$ to it via the same greedy bipartite match described in Section 2 (overlap minus symmetric difference of thresholded observed-children sets). This is alignment of fits to each other; ground truth is not used at any point during fitting or aggregation. The observed block requires no permutation.



(*a*) Spurious confounder-induced edges. (*b*) Latent-parent recovery.

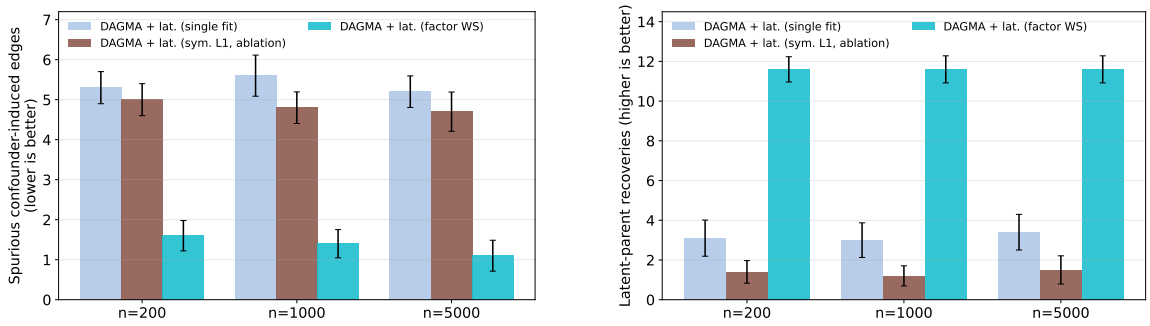
Figure 5: Stability selection vs single-fit estimation, both acyclicity penalties. Stability selection is uniformly worse than the single fit it is trying to stabilize: spurious-edge counts increase and latent-parent recovery essentially collapses, because consensus across $n_{\text{inits}} = 16$ random initializations averages away the rare runs that did discover the right latent structure.

Appendix H. Symmetric-ℓ_1 ablation

The asymmetric latent ℓ_1 choice in our default configuration ($\lambda_{\text{lat}} = 0$) is motivated by the linear–Gaussian latent identifiability issue: with $\lambda_{\text{lat}} = \lambda$, ℓ_1 shrinkage tends to drive latent–observed weights to zero and the model collapses to an observed-only explanation. The symmetric-ℓ_1 variant (Figure 6) is slightly better than the random-init single fit on the spurious-edge metric (4.7–5.0 vs 5.2–5.6 across sample sizes) but performs worse on latent-parent recovery (1.2–1.5 vs 3.0–3.5). Increasing latent ℓ_1 in the random-init regime suppresses both spurious observed edges and the latents themselves, at roughly comparable rates—consistent with the original motivation for $\lambda_{\text{lat}} = 0$. Either way, the gap to the factor-warmstart variant remains an order of magnitude larger than the gap between these two regularization choices, supporting the conclusion that initialization, not the latent ℓ_1 choice, is the load-bearing factor.

Appendix I. 2×2 acyclicity-orthogonality grid (single fit, no warm start)

Before introducing the warm-start methods, the cleanest test of the framework's claimed orthogonality to the choice of acyclicity penalty is the 2×2 grid crossing acyclicity (NOTEARS, DAGMA) with observability (obs-only, +latents), all with random initialization and single-fit estimation (Figure 7). Within each acyclicity family, adding explicit latents (without warm start) reduces spurious confounder-induced edges modestly; the magnitude of the reduction is roughly the same under NOTEARS and under DAGMA, supporting the framework's claimed orthogonality to the acyclicity choice. The reduction is much smaller than the additional gain achieved by adding the factor warm start (see main-text Figure 2).



(*a*) Spurious confounder-induced edges. (*b*) Latent-parent recovery.

Figure 6: Effect of latent ℓ_1 on DAGMA latent variants. Symmetric-ℓ_1 recovers about 1.5 pairs out of $\sim$12, intermediate between random-init single-fit ($\sim$3) and the factor warm start ($\sim$11.6). Even when latents are not aggressively regularized toward zero, random initialization rarely places them in informative directions.

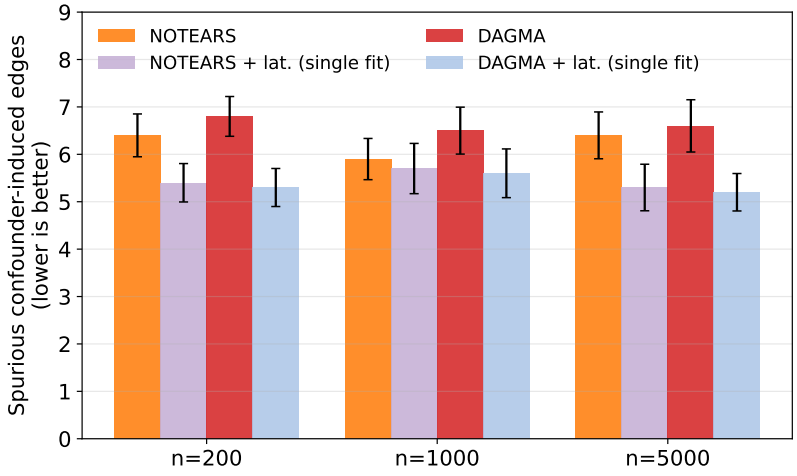
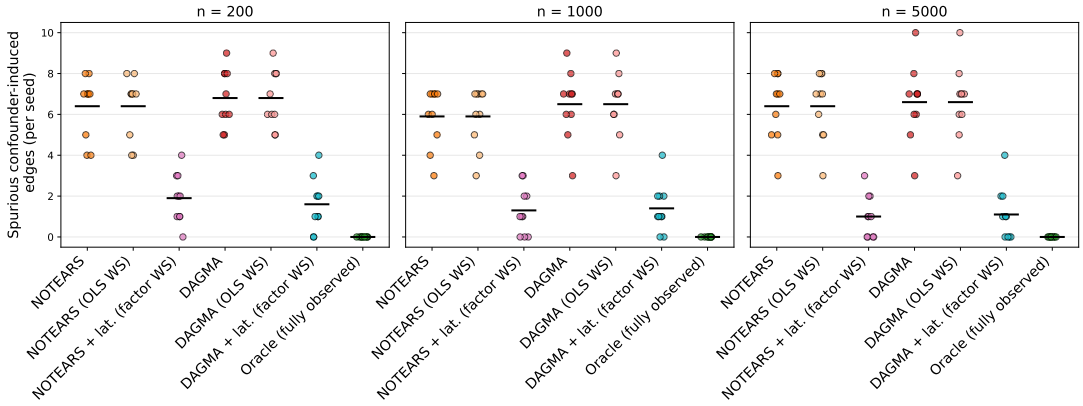


Figure 7: Acyclicity-orthogonality grid on the spurious-edge metric, single fit with random initialization.

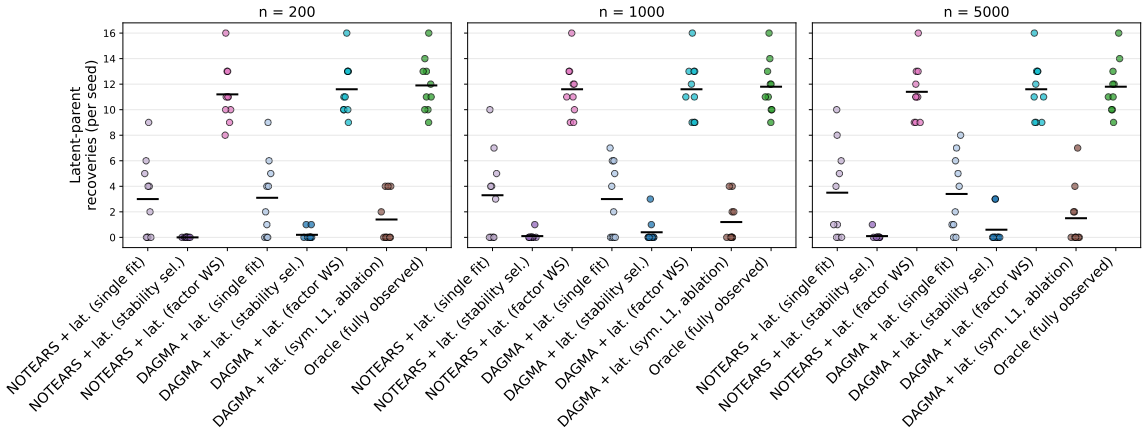
Appendix J. Per-seed spread for Experiment 1

The aggregate bars in the main text figures hide the per-seed distribution of the metrics. Figure 8 shows the underlying per-seed values for the headline metric and for latent-parent recovery, restricted to the methods that appear in the main-text headline figures.

The per-seed view supports two interpretations the aggregate bars obscure. First, the factor-warmstart improvement is consistent across seeds, not driven by a few favourable seeds. Second, near-oracle latent-parent recovery holds on essentially every seed—the DAGMA factor-warmstart variant recovers the maximum possible number of pairs on most seeds and is at most one or two pairs short on the rest.



(*a*) Spurious confounder-induced edges.



(*b*) Latent-parent recoveries.

Figure 8: Per-seed spread for the two headline diagnostics of Experiment 1. **(a)** Factor-warmstart variants concentrate near the bottom with low spread; observed-only and random-init latent variants concentrate near the top with moderate spread; oracle (rightmost) is at zero on every seed. **(b)** Factor-warmstart variants sit immediately below the oracle on essentially every seed; the DAGMA variant tracks the oracle particularly tightly.

Appendix K. Friedman §5.3 replication: additional results

Figure 10 reports the test-set KL divergence to ground truth as p_h^{fit} varies from 0 to 4, on both networks and all four sample sizes. The two panels back the main-text claim that test KL does not discriminate among different p_h^{fit} values in the linear–Gaussian setting: curves are essentially flat in p_h^{fit}, with absolute values clustered between 0.01 and 0.10 across all cells. Any latent model with sufficient rank matches the second-order moments of X to within sampling noise, because Gaussian distributions are fully characterized by their first two moments.

The mediator network warrants attention. Warm-start exhibits test KL between 0.05 and 0.10, while random initialization stays between 0.03 and 0.07. The gap is small (~ 0.02

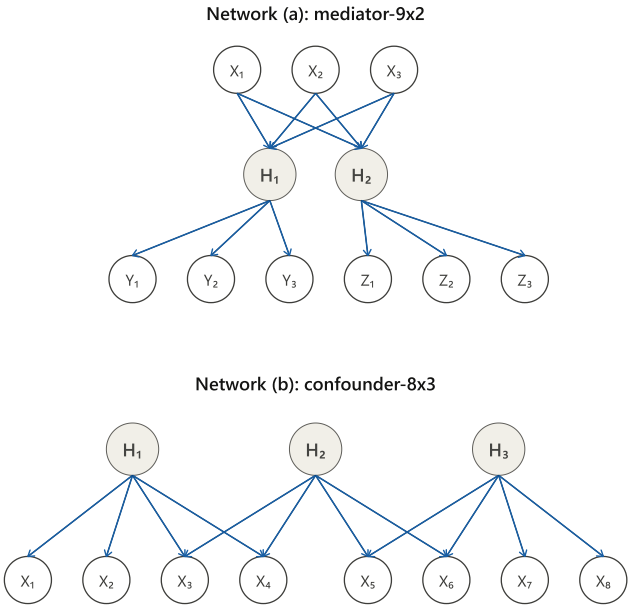


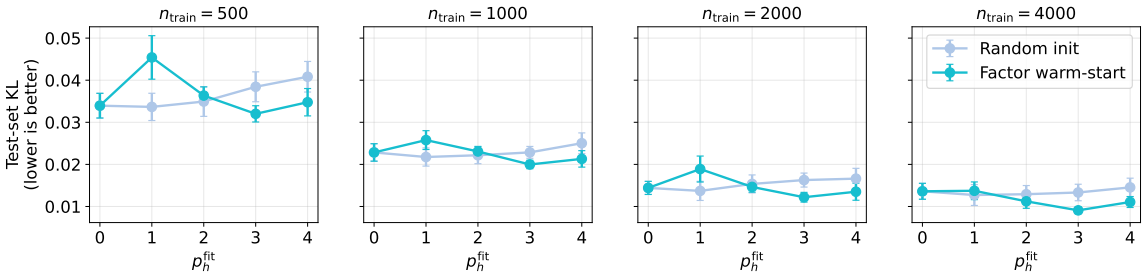
Figure 9: The two networks adapted from Friedman (1998, §5.3). Top: mediator-9×2 (Network A); bottom: confounder-8×3 (Network B). Shaded nodes are latent.

on average) but consistent and goes in the opposite direction to the structural recovery results, which is direct empirical evidence that, in the linear–Gaussian setting, the second-order density fit and the structural recovery objective are not perfectly aligned.

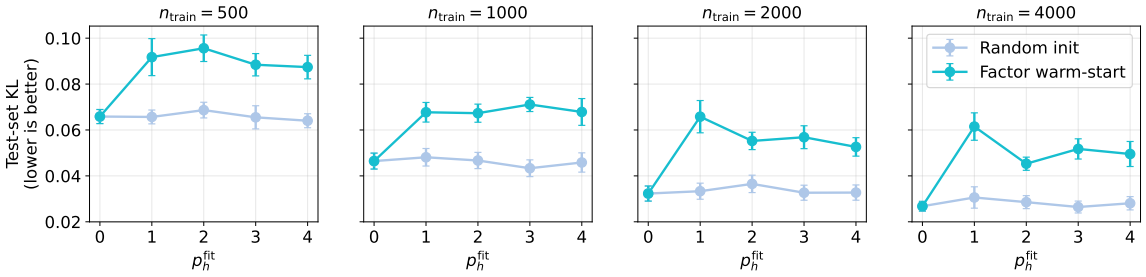
Figure 11 shows the p_h^{fit} sweep of latent-parent recovery on the mediator-9×2 network. The pattern mirrors the confounder-8×3 result in the main text: warm-start saturates the oracle ceiling (6/6) at $p_h^{\text{fit}} \in \{2, 3, 4\}$ on every seed, while random initialization recovers between 1.4 and 3.2 pairs depending on p_h^{fit} and n_{train}, without monotone improvement in either direction. Notably, random initialization does *not* converge toward warm-start's recovery level as n_{train} grows from 500 to 4000, consistent with the main-text interpretation that the bottleneck is optimization-basin choice, not statistical estimation: more data does not help the optimizer escape a wrong basin.

Numerical summary. Table 2 reports the headline numbers for both networks at the true p_h. Warm-start has zero standard error on recovery in every cell because the maximum was reached on every seed.

The table makes the structural-vs-density tension explicit. On the confounder network, warm-start dominates random initialization on both metrics. On the mediator network, warm-start dominates on recovery (perfect 6/6 on every seed) but is dominated on test KL (by roughly 0.02 on average). The bolding in each cell marks the better method on that metric; it is not a global statement about which method is preferred, which depends on whether the application values structural recovery or marginal density fit.



(*a*) Confounder-8×3 (true $p_h = 3$, dashed).



(*b*) Mediator-9×2 (true $p_h = 2$, dashed).

Figure 10: Test-set KL vs p_h^{fit}. **(a)** Both methods produce KL within $[0.01, 0.05]$ regardless of p_h^{fit}, with no systematic preference for the true p_h; warm-start has slightly lower KL than random initialization in most cells. **(b)** On this harder network warm-start produces *higher* test KL than random init across all non-zero p_h^{fit}—the structure-vs-density tradeoff discussed in Section 5.

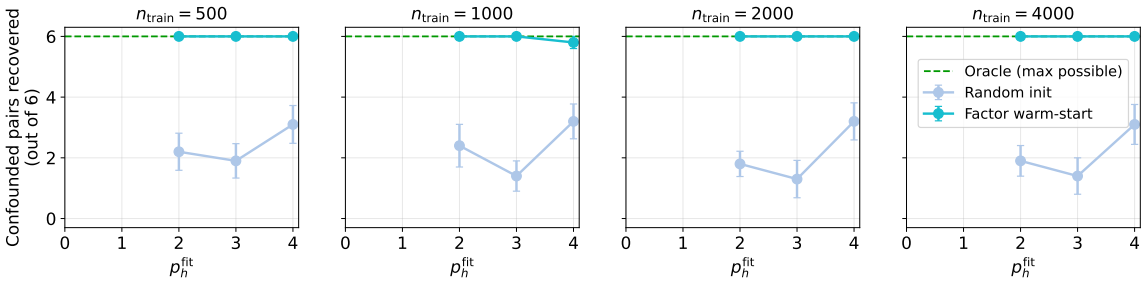


Figure 11: Latent-parent recovery vs p_h^{fit} on mediator-9×2 (true $p_h = 2$, dashed). Warm-start saturates the oracle ceiling (6/6) at the true and over-specified p_h on every seed (zero error bars). Random init plateaus at roughly 1.4–3.2 pairs with no improvement from larger n_{train}.

Appendix L. Acyclicity of fitted outputs

The optimization treats acyclicity as a penalty, so we audit the claim that fitted models are DAGs directly on the fitted matrices. For every fit we report the raw-matrix penalty under both smooth characterizations, $h_{\text{NT}}(\widehat{W})$ (matrix exponential) and $h_{\text{DAGMA}}(\widehat{W})$ (log-det), regardless of which was optimized; the probabilistic-validity margin $\sigma_{\min}(I - \widehat{W}^{\top})$ required

Table 2: Friedman §5.3 replication: results at the true p_h. Recovery is the number of ground-truth confounded observed pairs recovered through a common fitted latent parent (out of 16 for confounder-8×3, 6 for mediator-9×2). Bolding marks the better method per cell; on the mediator network this is split between KL and recovery, illustrating the structure-vs-density tension.

Network	n	Test KL (random)	Test KL (warm-start)	Recovery (random / warm-start)
confounder-8×3	500	0.038 ± 0.004	$\mathbf{0.032 \pm 0.002}$	12.1 ± 0.59 / $\mathbf{16.0 \pm 0.00}$
	1000	0.023 ± 0.001	$\mathbf{0.020 \pm 0.001}$	12.4 ± 0.56 / $\mathbf{16.0 \pm 0.00}$
	2000	0.016 ± 0.002	$\mathbf{0.012 \pm 0.001}$	12.9 ± 0.67 / $\mathbf{16.0 \pm 0.00}$
	4000	0.013 ± 0.002	$\mathbf{0.009 \pm 0.001}$	12.7 ± 0.56 / $\mathbf{16.0 \pm 0.00}$
mediator-9×2	500	$\mathbf{0.069 \pm 0.003}$	0.096 ± 0.006	2.2 ± 0.61 / $\mathbf{6.0 \pm 0.00}$
	1000	$\mathbf{0.047 \pm 0.004}$	0.067 ± 0.004	2.4 ± 0.70 / $\mathbf{6.0 \pm 0.00}$
	2000	$\mathbf{0.037 \pm 0.004}$	0.055 ± 0.004	1.8 ± 0.42 / $\mathbf{6.0 \pm 0.00}$
	4000	$\mathbf{0.029 \pm 0.003}$	0.045 ± 0.003	1.9 ± 0.50 / $\mathbf{6.0 \pm 0.00}$

by the exact E-step; and an exact cycle check (Kahn's algorithm) on the graph thresholded at the evaluation thresholds (0.30 observed rows, 0.15 latent rows).

A fixed penalty leaves thresholded cycles. Under a fixed weight $\rho = 0.5$, $h(\widehat{W})$ plateaus at the penalty–likelihood equilibrium ($\sim 10^{-2}$; Fig. 13, left) rather than converging to zero, and the thresholded estimate contains a directed cycle in 104 of 390 fits (27%), roughly constant across sample sizes and worst for stability-selection consensus graphs (20–21 of 30), which vote edge-wise and do not preserve acyclicity (Fig. 12, gray bars). Probabilistic validity was never at risk: $\sigma_{\min}(I - W_t^\top) \in [0.10, 0.45]$ at every outer iterate of every fit (Fig. 15), so every E-step was well-posed — acyclicity is sufficient but not necessary for the model to define a valid Gaussian law (Appendix B).

Decoding-time DAG projection. As a decoding step (the fitted $\widehat{W}$ is untouched), we project the thresholded graph onto a DAG by iteratively deleting the smallest-magnitude edge on a detected cycle. The projection removes 1.45 edges per affected fit; over all 390 fits and 1,050 metric entries it never worsens SHD or spurious-confounder counts, improves 85 entries, and costs a single latent-parent recovery in 2 entries.

Eliminating the cause: a continuation schedule inside EM. Fixing the plateau at its source requires the penalty weight to grow. A schedule interacts with the generalized-EM acceptance rule: if M-step candidates are accepted only when the *unpenalized* Q improves, every feasibility-restoring step (which trades Q for h) is vetoed and any schedule degenerates to fixed-ρ behavior, which we verified empirically. Judging acceptance by the scheduled objective $-Q + \rho_t h$ resolves the conflict. Under this rule, the generalized-EM guarantee of Theorem 1 applies to the scheduled objective $-Q + \rho_t h$ rather than to the observed-data log-likelihood; the fixed-ρ formulation of Section 2 retains the original guarantee, and empirically the schedule's final observed-data log-likelihoods remain within 0.4% of the fixed-ρ values. Table 3 compares regimes at $n = 1000$ (DAGMA, 5 seeds); the continuation schedule $\rho_t = \min(0.5 \cdot 1.5^t, 100)$ drives $h(\widehat{W})$ to $\sim 10^{-5}$ (Fig. 14) with exactly acyclic thresholded graphs and equal or better structure recovery.

Full-grid results under the continuation schedule. Rerunning the complete grid with the schedule (all non-EM methods bit-identical between runs, an exact control): 0 of 210 EM fits have a cyclic thresholded graph (Fig. 12, blue bars); $\max h_{\mathrm{NT}}(\widehat{W}) \leq 5 \times 10^{-4}$; margins in $[0.18, 0.32]$. The warm-start configuration attains near-oracle accuracy (SHD ≤ 1, zero spurious confounder edges for $n \geq 1000$, latent recovery preserved at $\approx 11/11.5$), and stability selection — which under the fixed penalty erased all latent-parent recoveries — recovers ≈ 8 of 11 confounded pairs, indicating the schedule also collapses cross-restart multimodality. Residual cycles occur only outside the proposed method: the one-shot NOTEARS observed-only baseline (3/30, removed by the same projection) and PC (1/30, a known finite-sample Meek-rule orientation artifact).

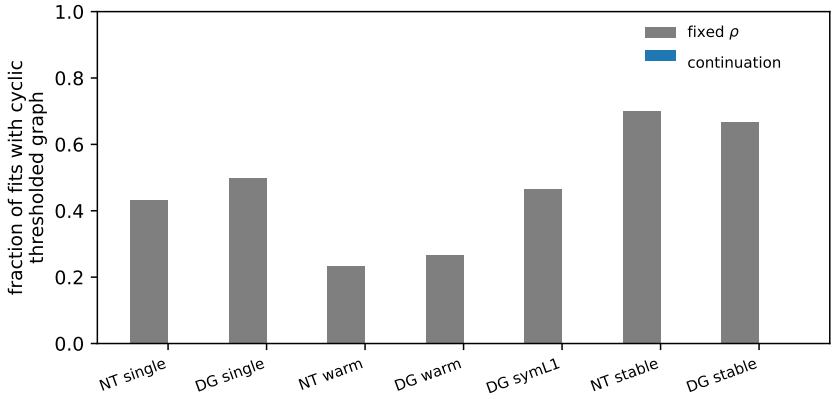


Figure 12: Fraction of fits whose thresholded graph contains a directed cycle, per EM configuration (30 fits each): fixed ρ (gray) vs. the continuation schedule (blue; all bars at zero).

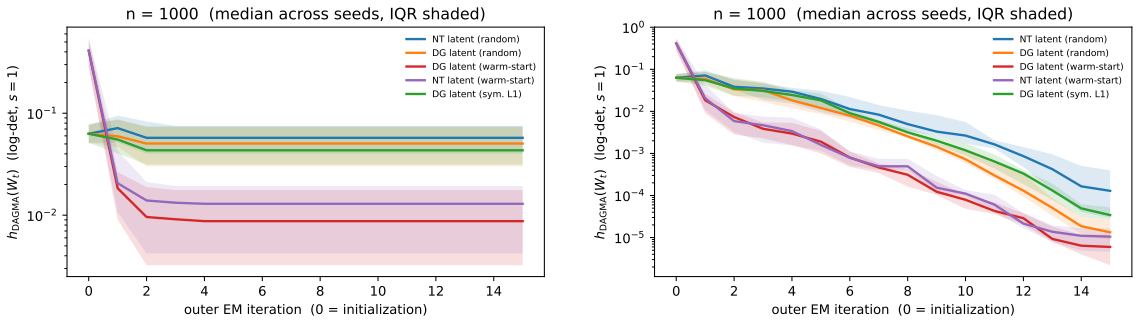


Figure 13: $h_{\mathrm{DAGMA}}(W_t)$ per outer EM iteration ($n = 1000$; median across 10 seeds, IQR shaded). Left: fixed $\rho = 0.5$ equilibrates at $h \sim 10^{-2}$. Right: the continuation schedule drives h to $\sim 10^{-5}$.

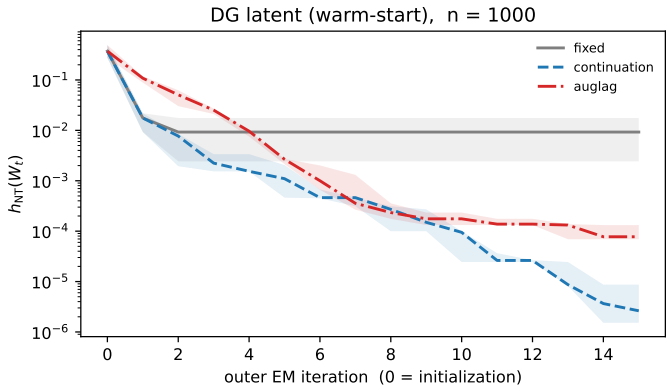


Figure 14: Penalty-regime ablation (warm start, $n = 1000$): fixed ρ plateaus; the continuation schedule and augmented Lagrangian drive $h(W_t)$ down four or more decades.

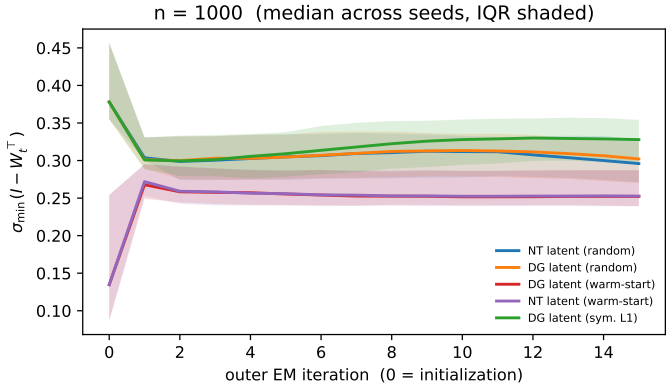


Figure 15: Probabilistic-validity margin $\sigma_{\min}(I - W_t^\top)$ per outer iteration ($n = 1000$): bounded well away from zero for every method at every iterate, so every E-step was well-posed.

Table 3: Penalty-regime ablation (n=1000, DAGMA, 5 seeds; medians).

Init	Regime	final h_{NT}	cyclic	SHD	spur.	log-lik
random	fixed	4.9×10^{-2}	2/5	8.0	7.0	−13707
random	continuation	1.3×10^{-5}	0/5	7.0	6.0	−13763
random	aug. Lagr.	5.6×10^{-5}	0/5	9.0	7.0	−13762
warm	fixed	9.3×10^{-3}	1/5	1.0	1.0	−13715
warm	continuation	2.6×10^{-6}	0/5	0.0	0.0	−13710
warm	aug. Lagr.	7.8×10^{-5}	0/5	0.0	0.0	−13709

Appendix M. Null-latent integrity check

On data whose true graph has *no* latent variables, the explicit-latent model with unregularized latent rows fabricates hidden confounding: 4–17 falsely confounded pairs on average,

increasing with the number of fitted latents and not vanishing with n (Fig. 16), with every fitted latent becoming active (Fig. 17). Three mitigations are effective. (i) Symmetric ℓ_1 on latent rows reduces fabrication 3–4× (near zero for $p_h^{\text{fit}} = 1$) and restores the active-latent count as a model-selection signal. (ii) Held-out fit exposes the fabrication: the latent fits never improve test KL over the observed-only model (Fig. 18; 0.0026 vs. 0.004–0.03 at $n = 5000$), so we recommend selecting p_h by held-out likelihood and interpreting latent structure only when it improves held-out fit. (iii) The factor warm start — the best configuration when latents exist — is the *worst* on null data. The mechanism is systematic rather than stochastic: on null data the leading principal directions of $\widehat{\Sigma}_X$ reflect covariance generated by genuine observed–observed edges, so the warm start initializes each latent as a proxy for real observed structure. The fit then routes that dependence through the latent, simultaneously fabricating hidden confounding and absorbing true direct edges into latent pathways — the only null-data configuration that degrades the observed subgraph — and, consistent with a systematic cause, the effect does not diminish as n grows (Fig. 16). Random initialization with symmetric ℓ_1 should therefore be preferred when the existence of latent confounders is in question.

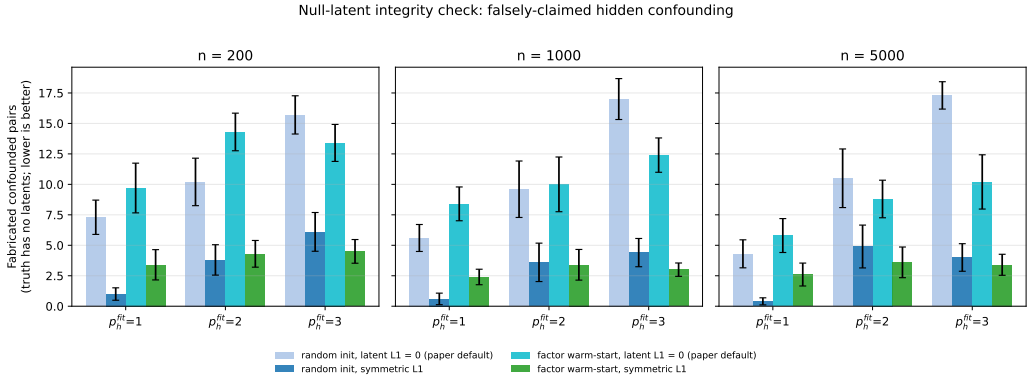


Figure 16: Null-latent integrity check: falsely-claimed hidden confounding per configuration (10 seeds; truth has no latents; lower is better).

Appendix N. Comparison with alternative latent-DAG methods

We compare against three alternative approaches to structure learning under latent confounding, run on the same data as the main experiment (identical seeds, sample sizes, and confounded pairs). *PC* (Spirtes et al., 2000) (Fisher-z tests, $\alpha = 0.05$) assumes causal sufficiency and therefore serves as the constraint-based analogue of our observed-only baselines: any adjacency it places between a truly confounded pair is a spurious claim. *FCI* (Spirtes et al., 2000) drops causal sufficiency and is the natural constraint-based competitor; a latent common cause of X_i and X_j should surface as a bidirected edge $X_i \leftrightarrow X_j$. *Friedman's Structural EM* (Friedman, 1998) is defined for discrete multinomial networks and cannot be run verbatim on linear-Gaussian data, so we implement a faithful adaptation that shares our exact Gaussian E-step and differs only in the M-step: greedy add/delete/reverse hill climbing on the decomposable Gaussian BIC score computed from the expected sufficient

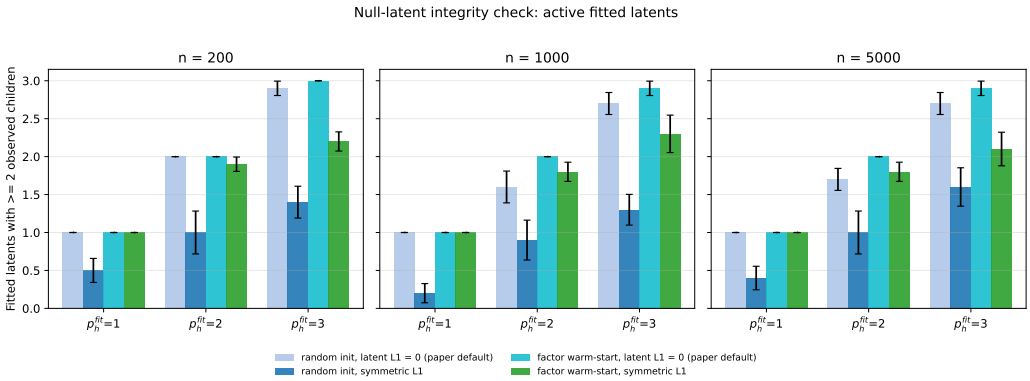


Figure 17: Active fitted latents (≥ 2 observed children). Unregularized latent rows saturate ($\approx p_h^{\text{fit}}$); symmetric ℓ_1 keeps the count informative.

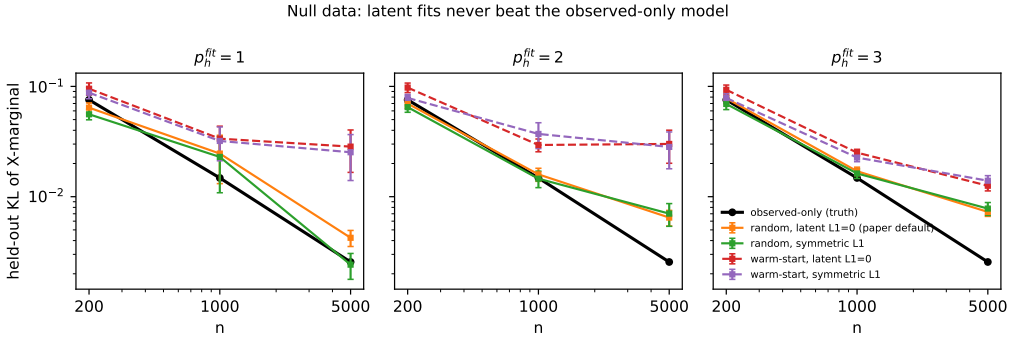


Figure 18: Held-out KL of the X-marginal on null data. No latent configuration improves on the observed-only model at any n or p_h^{fit}, so held-out model selection rejects the fabricated latents.

statistics, with random restarts and a generalized-EM acceptance check. The comparison therefore isolates exactly the discrete-search-vs-differentiable M-step distinction. Because PC and FCI output equivalence classes (CPDAGs/PAGs) rather than weighted DAGs, metrics are adapted accordingly: SHD is computed on the skeleton, spurious confounder edges count adjacencies for PC but only definitely-directed edges for FCI (bidirected and circle marks correctly allow confounding and are not penalized), and bidirected detections of truly confounded pairs serve as the PAG analogue of latent-parent recovery; neither method estimates parameters or a likelihood, so density-fit comparisons do not apply to them.

Figure 19 summarizes the results. FCI detects latent confounding in principle but weakly in practice, recovering 1.5–4.5 of the ≈11 truly confounded pairs at $\alpha = 0.05$, while PC, lacking causal sufficiency, incurs spurious confounder adjacencies comparable to the observed-only baselines. The Friedman-style Structural EM behaves as a credible likelihood-based competitor — it uses its latent variables and attains SHD 7.5–8.5 with 1.5–2 pairs recovered — but lands close to our *random-initialization* single fit, consistent with our claim that the advantage of the differentiable M-step lies less in per-step optimization quality

than in basin selection: the factor warm start, which has no analogue in discrete search over graph space, moves the shared E-step into the correct basin, and the resulting configuration attains SHD 0–0.5 with ≈11/11.5 pairs recovered, matching the fully supervised oracle. We note finally that these baselines also contextualize the acyclicity audit of Appendix L: the Friedman variant maintains a DAG by construction and passes the audit trivially (0/30 cyclic, $h = 0$), whereas PC exhibits one CPDAG orientation cycle in 30 runs, a known finite-sample artifact of the Meek orientation rules under faithfulness violations.

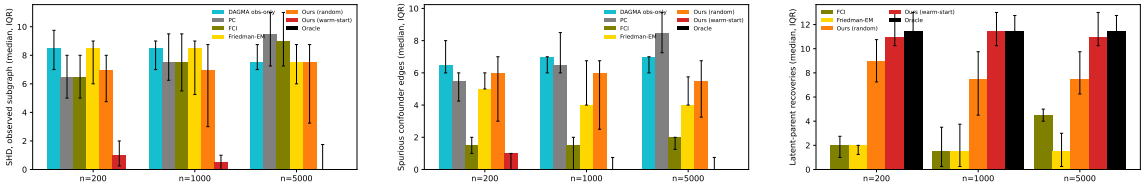


Figure 19: Comparison with alternative latent-DAG methods on identical data (10 seeds; median, IQR). Left: SHD on the observed subgraph (skeleton SHD for PC/FCI). Center: spurious confounder edges. Right: latent-parent recoveries (bidirected detections for FCI); the proposed warm-start configuration matches the oracle.

Appendix O. Spectral Warm Starts and Local EM Basins

This appendix gives a conditional theoretical justification for the factor warm start in Eq. (8). The result should be read as a local basin statement, not as a global convergence theorem. It has two parts. First, if the observed marginal covariance contains a dominant rank-p_h component attributable to latent structure, then the empirical principal directions initialize the explicit latent rows near the corresponding covariance-scale loading subspace. Second, once the iterate is inside a local EM/GEM basin, exact posterior moments together with accepted GEM-type updates refine the parameters geometrically, up to statistical error.

The formulation below is stated directly in terms of the observed marginal covariance $\Sigma_X^\star$, rather than in terms of exogenous hidden causes. This distinction matters. In hidden-confounder examples, the low-rank component arises from root latent common causes. In mediator examples, the hidden variables may be descendants of observed variables, yet their presence can still explain a substantial fraction of the observed marginal covariance. The spectral warm start only sees $\widehat{\Sigma}_X$; therefore the relevant primitive condition is a dominant low-rank component of $\Sigma_X^\star$, regardless of whether the corresponding hidden variables are roots, mediators, or mixtures of both.

O.1. Marginal low-rank condition

Let $X \in \mathbb{R}^{p_x}$ denote the observed variables and let

$$\widehat{\Sigma}_X := \frac{1}{n} X^\top X$$

be the empirical observed covariance. We assume that the population observed covariance admits the decomposition

$$\Sigma_X^\star = L^\star L^{\star\top} + \Sigma_{\mathrm{bg}}, \tag{11}$$

where $L^\star \in \mathbb{R}^{p_x \times p_h}$ is a covariance-scale latent loading matrix and Σ_{bg} contains idiosyncratic noise, weak observed–observed effects, and any covariance not explained by the dominant rank-p_h latent component. The matrix $L^\star$ is not necessarily the raw SEM loading matrix. In a pure hidden-confounder factor model, it coincides with the latent loading matrix up to scale. In a general linear–Gaussian latent DAG, it is the covariance-scale object induced by the marginalization over hidden variables. This is the object identified by PCA. Let

$$A^\star := L^\star L^{\star\top}, \qquad \alpha := \lambda_{p_h}(A^\star) > 0,$$

where α is the weakest nonzero latent-factor covariance eigenvalue. Define the marginal separation gap

$$\Delta := \lambda_{p_h}(A^\star) - \|\Sigma_{\mathrm{bg}}\|_{\mathrm{op}}. \tag{12}$$

The quantity α measures the scale of the weakest latent factor, while Δ measures its separation from the background covariance. The separation condition ensures that the top p_h eigenspace of $\Sigma_X^\star$ is identifiable as a marginal latent subspace. The loading perturbation itself scales with $\sqrt{\alpha}$, the natural scale of the factor loadings.

Assumption 1 (Dominant marginal latent component) *The observed covariance satisfies*

$$\Sigma_X^\star = L^\star L^{\star\top} + \Sigma_{\mathrm{bg}}, \qquad \mathrm{rank}(L^\star) = p_h,$$

and

$$\Delta := \lambda_{p_h}(L^\star L^{\star\top}) - \|\Sigma_{\mathrm{bg}}\|_{\mathrm{op}} > 0.$$

Assumption 2 (Empirical covariance concentration) *With probability at least* $1-\delta$,

$$\|\widehat{\Sigma}_X - \Sigma_X^\star\|_{\mathrm{op}} \le \varepsilon_n(\delta). \tag{13}$$

For centered Gaussian or sub-Gaussian samples one may take

$$\varepsilon_n(\delta) \lesssim \|\Sigma_X^\star\|_{\mathrm{op}} \left[\sqrt{\frac{p_x + \log(1/\delta)}{n}} + \frac{p_x + \log(1/\delta)}{n}\right],$$

by standard covariance concentration bounds (Vershynin, 2018; Wainwright, 2019).

O.2. The factor warm start

Let

$$\widehat{\Sigma}_X = \sum_{k=1}^{p_x} \mu_k u_k u_k^\top, \qquad \mu_1 \ge \mu_2 \ge \cdots.$$

Define the empirical rank-p_h spectral factor

$$\widehat{L} := \left[\sqrt{\mu_1}u_1, \ldots, \sqrt{\mu_{p_h}}u_{p_h}\right] \in \mathbb{R}^{p_x \times p_h}. \tag{14}$$

The factor warm start writes the columns of $\widehat{L}$ into the latent rows of W:

$$W^{(0)}_{p_x+k,\,1:p_x} = \epsilon_k \sqrt{\mu_k}\, u_k^\top, \qquad k = 1, \ldots, p_h, \tag{15}$$

where $\epsilon_k \in \{-1, +1\}$ accounts for the sign indeterminacy of Gaussian latent factors. This is the usual spectral initialization for a Gaussian latent-factor component (Anderson, 2003; Bartholomew et al., 2011; Bai, 2003), adapted to our SEM row convention.

Because $L^\star L^{\star\top}$ is invariant under $L^\star \mapsto L^\star R$ for $R \in \mathbb{O}(p_h)$, the natural subspace-level distance is

$$d_{\mathrm{load}}(\widehat{L}, L^\star) := \min_{R \in \mathbb{O}(p_h)} \|\widehat{L} - L^\star R\|_F. \tag{16}$$

If the model or regularizer fixes a coordinate system for the latent variables, this orthogonal ambiguity reduces to sign and permutation ambiguity.

Lemma 4 (Spectral basin-entry bound) *Suppose Assumptions 1 and 2 hold. Let*

$$\tau_n(\delta) := \varepsilon_n(\delta) + \|\Sigma_{\mathrm{bg}}\|_{\mathrm{op}}.$$

If $\tau_n(\delta) \le c\Delta$ for a sufficiently small numerical constant $c > 0$, then, with probability at least $1 - \delta$,

$$d_{\mathrm{load}}(\widehat{L}, L^\star) \le C\sqrt{p_h}\,\frac{\tau_n(\delta)}{\sqrt{\alpha}}, \qquad \alpha = \lambda_{p_h}(L^\star L^{\star\top}), \tag{17}$$

for a universal constant $C > 0$.

Proof Write

$$\widehat{\Sigma}_X = L^\star L^{\star\top} + E, \qquad E := (\widehat{\Sigma}_X - \Sigma_X^\star) + \Sigma_{\mathrm{bg}}.$$

On the event (13),

$$\|E\|_{\mathrm{op}} \le \varepsilon_n(\delta) + \|\Sigma_{\mathrm{bg}}\|_{\mathrm{op}} = \tau_n(\delta).$$

The condition $\tau_n(\delta) \le c\Delta$ guarantees that the top p_h eigenspace of $\widehat{\Sigma}_X$ remains separated from the background eigenspace. Let

$$\widehat{A} = \widehat{L}\widehat{L}^\top = \sum_{k=1}^{p_h} \mu_k u_k u_k^\top$$

be the best rank-p_h positive semidefinite spectral approximation to $\widehat{\Sigma}_X$. Since $A^\star = L^\star L^{\star\top}$ has rank p_h, Weyl's inequality and the Eckart–Young theorem imply

$$\|\widehat{A} - A^\star\|_{\mathrm{op}} \le 2\|E\|_{\mathrm{op}} \le 2\tau_n(\delta).$$

Moreover, $\widehat{A} - A^\star$ has rank at most $2p_h$, so

$$\|\widehat{A} - A^\star\|_F \le 2\sqrt{2p_h}\,\tau_n(\delta).$$

A standard Davis–Kahan/Wedin/Procrustes perturbation bound for positive semidefinite factors then gives

$$\min_{R \in \mathbb{O}(p_h)} \|\widehat{L} - L^\star R\|_F \le C\frac{\|\widehat{A} - A^\star\|_F}{\sqrt{\alpha}}.$$

Combining the last two displays proves (17). The perturbation tools used here are standard; see Davis and Kahan (1970), Stewart and Sun (1990), and Yu et al. (2015). ■

Remark 5 (Scale versus separation) *The separation condition is governed by the eigengap Δ, whereas the factor perturbation scale is governed by $\sqrt{\alpha}$, since the loadings are square roots of covariance eigenvalues. Replacing $\sqrt{\alpha}$ by $\sqrt{\Delta}$ gives a valid but looser bound when $\Delta \le \alpha$.*

O.3. From loading proximity to an EM basin

Let $\Phi^\star$ denote the local equivalence class of full population parameters $\phi = (W, \Sigma_\varepsilon)$ that induce the same relevant latent covariance structure, modulo latent sign, permutation, and, where appropriate, rotation. Each representative $\phi^\star \in \Phi^\star$ is assumed to satisfy the EM self-consistency condition

$$\nabla_1 Q(\phi^\star \mid \phi^\star) = 0,$$

in local coordinates orthogonal to the latent equivalence directions. The spectral bound concerns the covariance-scale factor $L^\star$, whereas the algorithm initializes an SEM parameter W. We therefore separate spectral error from SEM-parameterization error.

Assumption 3 (Covariance-to-SEM mismatch) *There exists a representative $\phi^\star \in \Phi^\star$ whose latent rows are compatible with the covariance-scale factor $L^\star$ up to mismatch*

$$\zeta_{\mathrm{SEM}} := \min_{R \in \mathbb{O}(p_h)} \|L^\star R - W_{HX}^{\star\top}\|_F.$$

In the pure factor/confounder case with standardized latents and no observed propagation, $\zeta_{\mathrm{SEM}} = 0$. In general, ζ_{SEM} measures the price of initializing SEM loadings from marginal covariance directions.

Assumption 4 (Remaining coordinates) *The initialization of observed rows of W and of Σ_ε satisfies*

$$d_{\mathrm{rest}}(\phi^{(0)}, \Phi^\star) \le r_{\mathrm{rest}}$$

with probability at least $1 - \delta$.

Lemma 6 (Entry into a local basin) *Let r_{EM} be the radius of a local EM/GEM contraction basin. If*

$$C\sqrt{p_h}\frac{\tau_n(\delta)}{\sqrt{\alpha}} + \zeta_{\mathrm{SEM}} + r_{\mathrm{rest}} \le r_{\mathrm{EM}}, \tag{18}$$

then, with probability at least $1 - 2\delta$,

$$d(\phi^{(0)}, \Phi^\star) \le r_{\mathrm{EM}}.$$

Proof The full initialization error is bounded by the latent-loading error, the covariance-to-SEM mismatch, and the remaining-coordinate error:

$$d(\phi^{(0)}, \Phi^\star) \le d_{\mathrm{load}}(\widehat{L}, L^\star) + \zeta_{\mathrm{SEM}} + d_{\mathrm{rest}}(\phi^{(0)}, \Phi^\star).$$

Apply Lemma 4, Assumption 3, and Assumption 4, then use a union bound. ■

O.4. A non-vacuous local-radius condition

The basin radius r_{EM} should not be viewed as arbitrary. The following standard regularity condition gives a concrete sufficient radius. Let

$$Q(\phi \mid \phi') := \mathbb{E}_{\phi^\star}\left[\mathbb{E}_{\phi'}\{\log p_\phi(X, H) \mid X\}\right]$$

be the population EM surrogate, formed using the exact Gaussian posterior of $H \mid X$ under ϕ'.

Assumption 5 (Primitive local regularity) *On a neighborhood of $\Phi^\star$, choose local coordinates orthogonal to the latent equivalence directions. There exist constants $\lambda_0, \mu_0, \gamma_0, \beta > 0$ such that:*

1. *The complete-data curvature at the population optimum satisfies*
$$-\mu_0 I \preceq \nabla_1^2 Q(\phi^\star \mid \phi^\star) \preceq -\lambda_0 I.$$
2. *The cross-gradient stability at the optimum satisfies*
$$\|\nabla_1 Q(\phi^\star \mid \phi) - \nabla_1 Q(\phi^\star \mid \phi^\star)\|_2 \le \gamma_0 d(\phi, \Phi^\star).$$
3. *The Hessian and cross-gradient maps are locally β-Lipschitz.*
4. $\gamma_0 < \lambda_0$.

Lemma 7 (A sufficient EM basin radius) *Under Assumption 5, any radius satisfying*

$$0 < r_{\mathrm{EM}} \le \frac{\lambda_0 - \gamma_0}{4\beta}$$

is a valid local basin radius. On this ball, the EM surrogate is λ-strongly concave and μ-smooth with

$$\lambda \ge \lambda_0 - \beta r_{\mathrm{EM}}, \qquad \mu \le \mu_0 + \beta r_{\mathrm{EM}},$$

and the gradient-stability constant satisfies

$$\gamma \le \gamma_0 + \beta r_{\mathrm{EM}} < \lambda.$$

Proof The Lipschitz condition implies that, throughout the ball of radius r_{EM},

$$\nabla_1^2 Q(\phi \mid \phi')$$

differs from $\nabla_1^2 Q(\phi^\star \mid \phi^\star)$ by at most βr_{EM} in operator norm. Hence the local curvature bounds become

$$-(\mu_0 + \beta r_{\mathrm{EM}})I \preceq \nabla_1^2 Q(\phi \mid \phi') \preceq -(\lambda_0 - \beta r_{\mathrm{EM}})I.$$

Similarly, the gradient-stability constant increases from γ_0 to at most $\gamma_0 + \beta r_{\mathrm{EM}}$. If

$$r_{\mathrm{EM}} \le \frac{\lambda_0 - \gamma_0}{4\beta},$$

then

$$\gamma_0 + \beta r_{\mathrm{EM}} < \lambda_0 - \beta r_{\mathrm{EM}},$$

so $\gamma < \lambda$ on the ball. ■

O.5. SNR and missing information

The condition $\gamma < \lambda$ is the usual low-missing-information condition for EM. In Gaussian latent-variable models, stronger latent signal typically reduces the posterior uncertainty of the hidden variables, thereby reducing missing information. To make this connection explicit, consider the local Gaussian factor block

$$X = LH + \varepsilon, \qquad H \sim N(0, I), \qquad \varepsilon \sim N(0, \Psi), \qquad \Psi \succ 0.$$

The posterior covariance of $H \mid X$ is

$$V(L, \Psi) = \left(I + L^\top \Psi^{-1} L\right)^{-1}. \tag{19}$$

Thus

$$\|V(L, \Psi)\|_{\mathrm{op}} \le \frac{1}{1 + \lambda_{\min}(L^\top \Psi^{-1} L)}. \tag{20}$$

The same latent signal strength that produces a large eigengap in $\Sigma_X^\star$ therefore also makes the posterior variance of the latent variables small.

Lemma 8 (SNR-controlled missing information; interpretive factor block) *Consider the local Gaussian factor block above. On a compact, well-conditioned neighborhood of the population loading block, Louis' missing-information identity implies the qualitative bound*

$$\frac{\gamma}{\lambda} \le C_{\mathrm{mis}} \, \|V(L^\star, \Psi^\star)\|_{\mathrm{op}},$$

where C_{mis} absorbs local conditioning, observed-covariance, posterior-mean, and fourth-moment constants. In particular, if $\Psi^\star \preceq \psi_{\max} I$, then

$$\|V(L^\star, \Psi^\star)\|_{\mathrm{op}} \le \frac{1}{1 + \alpha/\psi_{\max}} \le \frac{1}{1 + \Delta/\psi_{\max}}, \qquad \alpha = \lambda_{p_h}(L^\star L^{\star\top}).$$

Thus the same latent signal strength that separates the leading marginal eigenspace also reduces posterior uncertainty about the latent variables, making the BWY-style condition (Balakrishnan et al., 2017), i.e., $\gamma < \lambda$ more plausible.

Proof By Louis' missing-information identity (Louis, 1982),

$$I_{\text{obs}} = I_{\text{comp}} - I_{\text{mis}},$$

where I_{comp} is the complete-data information and I_{mis} is the conditional variance of the complete-data score given the observed data. For the loading block of a Gaussian factor model, the complete-data score depends on latent sufficient statistics such as $XH^\top$ and $HH^\top$. Their conditional fluctuations are controlled by $\mathrm{Var}(H \mid X) = V(L, \Psi)$, together with local bounds on Σ_X, posterior means, and fourth moments. On a compact, well-conditioned neighborhood, these additional terms are uniformly bounded and can be absorbed into a constant C_{mis}. Therefore

$$\|I_{\text{mis}}\|_{\text{op}} \le C \, \|V(L^\star, \Psi^\star)\|_{\text{op}}.$$

Since the complete-data information is uniformly positive definite on the identifiable tangent space, the local missing-information ratio satisfies

$$\gamma/\lambda \le C_{\text{mis}} \|V(L^\star, \Psi^\star)\|_{\text{op}}.$$

It remains to connect $\|V\|_{\text{op}}$ to the latent signal scale. From (19),

$$\|V(L^\star, \Psi^\star)\|_{\text{op}} = \frac{1}{1 + \lambda_{\min}(L^{\star\top} \Psi^{\star -1} L^\star)}.$$

If $\Psi^\star \preceq \psi_{\max} I$, then $\Psi^{\star -1} \succeq \psi_{\max}^{-1} I$, and hence, by the Rayleigh quotient,

$$\lambda_{\min}(L^{\star\top} \Psi^{\star -1} L^\star) \ge \frac{1}{\psi_{\max}} \lambda_{\min}(L^{\star\top} L^\star) = \frac{\alpha}{\psi_{\max}}.$$

Therefore

$$\|V(L^\star, \Psi^\star)\|_{\text{op}} \le \frac{1}{1 + \alpha/\psi_{\max}}.$$

Since $\Delta \le \alpha$, the weaker bound

$$\|V(L^\star, \Psi^\star)\|_{\text{op}} \le \frac{1}{1 + \Delta/\psi_{\max}}$$

also holds. ■

Lemma 8 is not used to prove monotonicity of Algorithm 1. Its role is explanatory: in the strong latent signal regime, the same marginal signal strength that helps the spectral initializer also reduces posterior latent uncertainty, making local EM contraction more plausible. Modulo the primitive local regularity in Assumption 5 and the accepted-step condition below, the warm start's behavior is therefore anchored to a single observable primitive: a separated low-rank component of $\Sigma_X^\star$.

O.6. Local EM and gradient-EM contraction

The following contraction arguments are standard in modern local analyses of EM and gradient EM (Wu, 1983; Balakrishnan et al., 2017). They are stated here to make explicit what the warm start buys: it places the iterate in a region where these local arguments can apply. Let

$$M(\phi') := \arg\max_{\phi} Q(\phi \mid \phi')$$

be the population EM operator. Also define the population gradient-EM map

$$G_\eta(\phi) := \phi + \eta \nabla_1 Q(\phi \mid \phi), \tag{21}$$

where the gradient is taken in local identifiable coordinates. For constrained parameterizations, G_η may be followed by projection onto the feasible set.

Lemma 9 (Population EM contraction) *On the local basin described above, the exact population EM operator satisfies*

$$d(M(\phi), \Phi^\star) \le \frac{\gamma}{\lambda}\, d(\phi, \Phi^\star). \tag{22}$$

Proof Choose local representatives so that $d(\phi, \Phi^\star) = \|\phi - \phi^\star\|_2$. Since $M(\phi)$ maximizes $Q(\cdot \mid \phi)$,

$$\nabla_1 Q(M(\phi) \mid \phi) = 0.$$

By self-consistency,

$$\nabla_1 Q(\phi^\star \mid \phi^\star) = 0.$$

Using λ-strong concavity of $Q(\cdot \mid \phi)$,

$$\lambda \|M(\phi) - \phi^\star\|_2 \le \|\nabla_1 Q(\phi^\star \mid \phi) - \nabla_1 Q(\phi^\star \mid \phi^\star)\|_2 .$$

The gradient-stability condition bounds the right-hand side by $\gamma\|\phi - \phi^\star\|_2$. Therefore

$$\|M(\phi) - \phi^\star\|_2 \le (\gamma/\lambda)\|\phi - \phi^\star\|_2.$$

Minimizing over representatives of $\Phi^\star$ gives the claim. ■

Lemma 10 (Population gradient-EM contraction) *Let G_η be defined by (21). If $0 < \eta \le 2/(\lambda + \mu)$, then*

$$d(G_\eta(\phi), \Phi^\star) \le \kappa_\eta\, d(\phi, \Phi^\star), \tag{23}$$

where

$$\kappa_\eta := \rho_\eta + \eta\gamma, \qquad \rho_\eta := \max\{|1 - \eta\lambda|,\ |1 - \eta\mu|\}. \tag{24}$$

In particular, with $\eta = 2/(\lambda + \mu)$,

$$\kappa_\eta = \frac{\mu - \lambda + 2\gamma}{\mu + \lambda} < 1$$

whenever $\gamma < \lambda$.

Proof Choose local representatives so that $d(\phi, \Phi^\star) = \|\phi - \phi^\star\|_2$. Decompose

$$G_\eta(\phi) - \phi^\star = \underbrace{\phi - \phi^\star + \eta\nabla_1 Q(\phi \mid \phi^\star)}_{\text{complete-data gradient step}} + \eta \underbrace{[\nabla_1 Q(\phi \mid \phi) - \nabla_1 Q(\phi \mid \phi^\star)]}_{\text{missing-information perturbation}}.$$

Because $Q(\cdot \mid \phi^\star)$ is λ-strongly concave and μ-smooth, a gradient-ascent step on this function contracts toward its maximizer $\phi^\star$ by factor

$$\rho_\eta = \max\{|1 - \eta\lambda|,\ |1 - \eta\mu|\}.$$

The gradient-stability condition bounds the perturbation term by $\eta\gamma\|\phi - \phi^\star\|_2$. Therefore

$$\|G_\eta(\phi) - \phi^\star\|_2 \le (\rho_\eta + \eta\gamma)\|\phi - \phi^\star\|_2.$$

The expression for $\eta = 2/(\lambda + \mu)$ follows by substitution. ■

Assumption 6 below should therefore be read as a tracking condition: Lemmas 9 and 10 certify that contracting EM and gradient-EM maps exist on the local basin, while Assumption 6 requires the accepted penalized-Adam proposal to follow one of these maps closely enough to inherit a distance contraction.

O.7. Rejection-aware local refinement

The main algorithm does not run exact EM. It runs an inner Adam optimizer on a penalized objective and then accepts the candidate only if the unpenalized Q improves. The penalties shape the proposal, while the acceptance check gives GEM monotonicity. This safeguard alone does not imply geometric contraction: a candidate may improve Q without decreasing distance to the local parameter class $\Phi^\star$, and a rejected candidate leaves the iterate unchanged.

Accepted proposals versus monotone safeguards. The unpenalized-Q safeguard is an EM monotonicity device, not by itself a contraction device. Therefore geometric refinement requires an additional local condition: accepted proposals must track a contracting EM or gradient-EM map closely enough. This is the content of Assumption 6 below. The empirical role of the safeguard is nevertheless important: in our monotonicity ablation, the full Gaussian E-step together with the unpenalized-Q acceptance check produced 0/150 likelihood-decreasing outer steps and 0/10 seeds with any decrease, whereas disabling the check or replacing posterior second moments by mean imputation caused frequent failures.

Assumption 6 (Accepted proposal tracks a contracting map) *Inside the local basin, suppose that whenever a proposal is accepted, it satisfies*

$$d(\phi^{(t+1)}, \Phi^\star) \le \kappa\, d(\phi^{(t)}, \Phi^\star) + \epsilon_n, \qquad 0 < \kappa < 1.$$

If the proposal is rejected, then

$$\phi^{(t+1)} = \phi^{(t)}.$$

Theorem 11 (Warm start plus accepted-step geometric refinement) *Suppose Lemma 6 places the factor warm start inside the local basin, and suppose Assumption 6 holds. Let N_t be the number of accepted contraction-like updates among the first t outer iterations. Then*

$$d(\phi^{(t)}, \Phi^\star) \le \kappa^{N_t} d(\phi^{(0)}, \Phi^\star) + \frac{\epsilon_n}{1-\kappa}.$$

Proof On rejected iterations, the distance is unchanged. On accepted iterations, the distance obeys

$$d_{s+1} \le \kappa d_s + \epsilon_n.$$

Applying this recursion only over the accepted updates gives

$$d(\phi^{(t)}, \Phi^\star) \le \kappa^{N_t} d(\phi^{(0)}, \Phi^\star) + \epsilon_n \sum_{j=0}^{N_t-1} \kappa^j \le \kappa^{N_t} d(\phi^{(0)}, \Phi^\star) + \frac{\epsilon_n}{1-\kappa}.$$

■

The theorem cleanly separates two facts. The unpenalized-Q safeguard guarantees monotone EM-style improvement of the observed-data likelihood. A geometric rate additionally requires that accepted proposals track a locally contracting EM or gradient-EM step. If many proposals are rejected, the algorithm may be monotone but stalled.

O.8. Mediator networks and the role of ζ_{SEM}

The marginal-spike assumption covers both root-confounder and mediator settings because it is stated directly on $\Sigma_X^\star$. However, the stronger SEM-parameter basin-entry inequality

may be easier to satisfy for root hidden confounders than for hidden mediators. In mediator networks, covariance-scale directions can differ substantially from the literal SEM latent rows because effects propagate through $X \to H \to Y$ pathways; this discrepancy is precisely what ζ_{SEM} measures. Thus mediator success in Section 5 should not be read as evidence that ζ_{SEM} is small. Rather, the warm start appears to place the model in the correct marginal covariance basin, after which exact posterior moments, the GEM safeguard, and the alignment-free latent-parent metric tolerate the remaining covariance-to-SEM mismatch.

O.9. Connection to structural recovery metrics

The spectral and EM arguments above are statements about covariance-scale latent structure and local likelihood refinement. They do not, by themselves, prove exact recovery of the true SEM support. This limitation is intrinsic: observational linear–Gaussian latent-variable models have sign, permutation, rotation, and Markov-equivalence ambiguities.

The alignment-free latent-parent metric is deliberately matched to this theory: it depends on whether observed variables share some fitted latent parent, not on the identity, sign, or ordering of that latent. Therefore it is less sensitive to within-subspace rotations than full latent-node support recovery. Exact directed-support recovery would require additional coordinate-identification and loading-margin assumptions.

For example, if all true shared-parent pairs have loading magnitudes above a threshold margin and all nonshared cross-loadings are below that margin, then sufficiently small $d_{\mathrm{load}}(\widehat{L}, L^\star)$ implies that thresholded fitted loadings preserve the common-parent relation. Without such a margin or coordinate-identification condition, the theorem should be interpreted as explaining why the warm start finds the correct latent covariance basin, not as a proof of exact directed support recovery.

O.10. Connection to negative results

The same theorem also explains why random initialization can fail even as the sample size increases. The statistical term $\varepsilon_n(\delta)$ shrinks with n, but the basin-entry condition also contains the initialization distance. A random initialization has no reason to satisfy

$$d(\phi^{(0)}, \Phi^\star) \le r_{\mathrm{EM}}.$$

If it starts outside the basin, increasing n reduces sampling error but does not necessarily move the iterate into the correct basin. Similarly, stability selection can average over runs that land in different local basins, which may improve edge stability in some settings but can dilute the explicit-latent structure when basin selection itself is the dominant source of variability.

Thus the factor warm start is not merely an optimization convenience. In the regime where latent variables explain a separated low-rank component of the observed covariance, it supplies the basin-selection mechanism that random initialization lacks. At the same time, the guarantee remains conditional: modulo standard local EM regularity and the accepted-step tracking condition, the warm-start behavior is anchored to a separated marginal low-rank component of $\Sigma_X^\star$; translating that covariance-level proximity into SEM-level structure further requires either bounded covariance-to-SEM mismatch or a recovery metric, such as latent-parent recovery, that is tolerant to this mismatch.