TAMING LATENT ILL-CONDITIONING IN GAUSSIAN GRAPHICAL MODEL ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent-variable Gaussian graphical models separate a sparse conditional graph from the collective influence of unobserved variables. Although low-rank factorization makes this decomposition convenient to optimize, latent effects of unequal strength can slow gradient descent even when the observed precision matrix is well-conditioned. To address this difficulty, we combine scaled gradient descent with symmetric hard thresholding and a joint feasibility and descent line search. At the correctly specified rank, we establish a local linear convergence bound for the separate sparse and low-rank matrix errors, up to statistical accuracy. The contraction factor and statistical-error multiplier are independent of the latent eigenvalues, with precision conditioning and sparse–low-rank transversality held fixed. The analysis connects an exact cancellation in the scaled gradient to the geometry of fixed-rank matrices, allowing both components to be controlled along the same trajectory. We further provide sufficient conditions under which a thresholded spectral initializer identifies the support and enters the convergence neighborhood. In simulations, ScaledGD maintains similar late-stage convergence rates as latent conditioning worsens while gradient descent slows, and runs faster than AltGD in the sampled-data comparisons.

1 INTRODUCTION

Graph learning seeks to uncover conditional dependence among variables from their observations. Graphical models encode these relationships through a graph (Lauritzen, 2026; Koller and Friedman, 2009), connecting statistical representation with inference through their factorization properties (Wainwright and Jordan, 2008; Maathuis et al., 2019). When the graph is unknown and the number of variables is large, structural assumptions become essential to learning from limited data (Drton and Maathuis, 2017).

In Gaussian graphical models (GGMs), this task reduces to precision-matrix estimation: off-diagonal zeros encode conditional independence. Sparsity, originating in covariance selection (Dempster, 1972), thus gives both a statistical simplification and an interpretable graph. This viewpoint underlies neighborhood regression (Meinshausen and Bühlmann, 2006), penalized likelihood (Yuan and Lin, 2007; Rothman et al., 2008), graphical-lasso algorithms (Friedman et al., 2008; Banerjee et al., 2008), and guarantees for log-determinant regularization (Ravikumar et al., 2011) and constrained ℓ_1 estimation (Cai et al., 2011). However, some influences on the system may be unobserved. A common hidden influence can induce dependence across many variables, turning a sparse conditional graph into a dense marginal precision matrix.

The effect of these hidden variables is visible directly in the Gaussian model. Partition the joint precision of observed and hidden variables as

$$\begin{pmatrix} \mathbf{\Theta}_{OO} & \mathbf{\Theta}_{OH} \\ \mathbf{\Theta}_{HO} & \mathbf{\Theta}_{HH} \end{pmatrix}.$$

The marginal observed precision is the Schur complement

$$\mathbf{\Theta}^\star = \mathbf{S}^\star - \mathbf{L}^\star, \qquad \mathbf{S}^\star = \mathbf{\Theta}_{OO}, \quad \mathbf{L}^\star = \mathbf{\Theta}_{OH}\mathbf{\Theta}_{HH}^{-1}\mathbf{\Theta}_{HO} \succeq 0. \tag{1}$$

Here $\mathbf{S}^\star$ retains the conditional graph, while $\mathbf{L}^\star$ collects the effects of marginalization and has low rank when only a few latent directions are present. Recovering both components would distinguish sparse

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

conditional relationships from shared latent influences. This requires identifiability: a perturbation that is both sparse and low rank could otherwise be assigned to either component (Chandrasekaran et al., 2011; 2012a).

To estimate this decomposition, Chandrasekaran et al. (2010; 2012a) introduced the latent-variable GGM (LVGGM) framework with sparse and nuclear-norm regularization. This formulation prompted a broader examination of what can be recovered from the observed distribution. The resulting discussions address estimation rates and tuning (Giraud and Tsybakov, 2012), partial identifiability (Wainwright, 2012), and the signal strength needed to recover support and rank (Ren and Zhou, 2012). They also raise the question of whether an estimator can adapt to few or no latent variables (Candés and Soltanolkotabi, 2012), an issue revisited alongside identifiability and model selection in the authors' rejoinder (Chandrasekaran et al., 2012b). Together, these developments show why separating the components involves more than fitting the marginal precision accurately.

Building on this foundation, subsequent work has broadened both the statistical formulation and the structures that can be represented. Statistical error bounds quantify the accuracy of LVGGM estimation (Meng et al., 2014), while penalized D-trace loss provides an alternative to Gaussian likelihood (Wu et al., 2017). Additional structural information can enter through sparse connections involving hidden variables (Vinyes and Obozinski, 2018), Laplacian constraints (Li et al., 2023), or the adaptive penalties, sign constraints, and prescribed sparsity patterns supported by the Golazo formulation (Echave-Sustaeta Rodríguez and Röttger, 2024). Other extensions address covariance uncertainty in robust identification (Ciccone et al., 2018) and low-rank community blocks in heterogeneous graphical models (Shi et al., 2024).

Alongside these modeling developments, efficient computation remains central to making LVGGM estimation useful. Rank-constrained likelihood and EM-based fitting provide ways to work with a prescribed latent dimension (Yuan, 2012; Lauritzen and Meinshausen, 2012), and alternating-direction methods exploit the structure of the convex estimator (Ma et al., 2013). A particularly direct approach is to represent the latent matrix by a compact factor. Following this approach, Xu et al. (2017) combine alternating gradient descent with hard thresholding and establish local geometric convergence at the true rank. Factorization therefore provides a concrete route from an estimator to an iterative algorithm. Its convergence speed, however, must also account for the relative strengths of the latent directions.

A small number of hidden influences need not have comparable strengths. Writing $\mathbf{C} = \mathbf{\Theta}_{OH}\mathbf{\Theta}_{HH}^{-1/2}$ gives $\mathbf{L}^\star = \mathbf{C}\mathbf{C}^\top$, whose positive eigenvalues are the squared nonzero singular values of the effective observed–hidden coupling. Thus low rank controls the number of latent directions, but not their relative strengths; we measure this spectral spread by the ratio of the positive eigenvalues.

This source of ill-conditioning can persist even when estimating the marginal precision is locally well-conditioned. To illustrate, let $p > 2$, let $\mathbf{U} \in \mathbb{R}^{p\times 2}$ have orthonormal columns, and take

$$\mathbf{S}^\star = 2\mathbf{I}, \qquad \mathbf{L}^\star = \mathbf{U}\,\mathrm{diag}(1,\epsilon)\mathbf{U}^\top, \qquad 0 < \epsilon < 1. \tag{2}$$

Then $\kappa_L = 1/\epsilon$ can grow arbitrarily large, while $\mathbf{\Theta}^\star = \mathbf{S}^\star - \mathbf{L}^\star$ has eigenvalues in $[1, 2]$ and hence condition number 2. This is a valid LVGGM by taking $\mathbf{\Theta}_{HH} = \mathbf{I}_2$ and $\mathbf{\Theta}_{OH} = \mathbf{U}\,\mathrm{diag}(1, \sqrt{\epsilon})$. Thus latent ill-conditioning can be arbitrarily severe even when the marginal precision remains well-conditioned.

Under the factorization $\mathbf{L} = \mathbf{B}\mathbf{B}^\top$, a perturbation $\mathbf{H}$ changes the latent matrix to first order by $\mathbf{B}\mathbf{H}^\top + \mathbf{H}\mathbf{B}^\top$, so the local sensitivity depends on the factor singular values. For the example in (2), fix $\mathbf{S} = \mathbf{S}^\star$ and vary the two factor amplitudes along $\mathbf{U}$. The population likelihood has curvatures 4 and $4\epsilon/(2-\epsilon)^2$ at the truth. A common gradient step must accommodate the stronger direction, leaving slower progress along the weaker one as ϵ decreases, despite the fixed condition number of $\mathbf{\Theta}^\star$. This fixed-$\mathbf{S}$ calculation isolates the factor-conditioning mechanism; Theorem 1 addresses the coupled updates of $\mathbf{S}$ and $\mathbf{B}$.

Scaled gradient descent compensates for this imbalance through the factor Gram matrix. Tong et al. (2021) establish spectral independence of local rates for matrix sensing, robust PCA, matrix completion, and restricted strongly convex losses. Their PSD specialization suggests a natural LVGGM update. Here, however, the sparse and latent estimates evolve together: hard thresholding can change the support, and each joint step must keep $\mathbf{S} - \mathbf{B}\mathbf{B}^\top$ positive definite. A result for a fixed sparse component leaves these interactions unresolved.

In this paper, we study ScaledGD for joint sparse-low-rank LVGGM estimation at the correctly specified rank, locally after support identification. Our contributions are:

1. We establish local linear convergence of the separate sparse and latent matrix errors up to statistical accuracy (Theorem 1). The contraction factor is independent of $\kappa_L = \lambda_1(\mathbf{L}^\star)/\lambda_{r_\star}(\mathbf{L}^\star)$, with precision conditioning and transversality held fixed.
2. We analyze the joint updates of $\mathbf{S}$ and $\mathbf{B}$, including symmetric hard thresholding and the shared line search. Under the local conditions, the iterates retain the correct sparse support and factor rank, remain positive definite, and accept the full trial step.
3. We give sufficient conditions for the spectral initializer to identify the sparse support and enter the local convergence neighborhood. A Gaussian score bound then yields a finite-sample matrix-error guarantee.

Organization and notation. Section 2 presents the estimator and algorithm, Section 3 establishes the local convergence guarantee, and Section 4 gives initialization and statistical consequences. Section 5 reports the numerical study; complete proofs and supplementary experiments appear in the appendix. We use p for the number of observed variables, n for the sample size, and $r_\star$ for the rank of the true latent component. A star denotes a population quantity, a hat an estimate, and a subscript t an iterate. The true off-diagonal support of $\mathbf{S}^\star$ is Ω, and $s_\star = \|\mathbf{S}^\star\|_{0,\text{off}}$ counts its nonzero entries, including both symmetric entries of each edge. We write $\|\cdot\|_F$ and $\|\cdot\|_{\text{op}}$ for the Frobenius and spectral norms, respectively, and $\langle \mathbf{A}, \mathbf{C}\rangle = \operatorname{tr}(\mathbf{A}^\top \mathbf{C})$ for the matrix inner product. The relation $\mathbf{A} \succeq \mathbf{C}$ means that $\mathbf{A} - \mathbf{C}$ is positive semidefinite. Eigenvalues of symmetric matrices are ordered decreasingly.

2 The Proposed Estimator and Algorithm

2.1 Likelihood and Fixed-Rank Formulation

We begin with the Gaussian likelihood associated with the sparse–low-rank decomposition in (1). Given independent observations $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}^\star)$, $i = 1, \dots, n$, where $\mathbf{\Sigma}^\star = (\mathbf{\Theta}^\star)^{-1}$, let $\widehat{\mathbf{\Sigma}} = n^{-1}\sum_i \mathbf{x}_i\mathbf{x}_i^\top$ and define

$$\ell_n(\mathbf{\Theta}) = -\log\det\mathbf{\Theta} + \langle \widehat{\mathbf{\Sigma}}, \mathbf{\Theta}\rangle, \quad \mathbf{G}_n(\mathbf{\Theta}) = \widehat{\mathbf{\Sigma}} - \mathbf{\Theta}^{-1}, \quad f_n(\mathbf{S}, \mathbf{B}) = \ell_n(\mathbf{S} - \mathbf{B}\mathbf{B}^\top). \tag{3}$$

The estimator minimizes f_n over symmetric $\mathbf{S}$ with $\|\mathbf{S}\|_{0,\text{off}} \le s$ and factors $\mathbf{B} \in \mathbb{R}^{p\times r_\star}$ satisfying $\mathbf{S} - \mathbf{B}\mathbf{B}^\top \succ 0$. Here $\|\mathbf{S}\|_{0,\text{off}}$ counts symmetric off-diagonal entries, so s is even and the diagonal is unrestricted. Accordingly, $\mathcal{H}_s$ retains the diagonal and the $s/2$ off-diagonal pairs of largest magnitude.

The factorization ensures that the latent estimate is positive semidefinite with rank at most $r_\star$. Since $\mathbf{B}$ and $\mathbf{B}\mathbf{Q}$ represent the same latent matrix for any orthogonal $\mathbf{Q}$, the identifiable object is $\mathbf{L} = \mathbf{B}\mathbf{B}^\top$. Our analysis therefore measures recovery directly in $(\mathbf{S}, \mathbf{L})$; the transversality condition in Section 3 ensures their local separation.

2.2 Spectral Initialization

We follow the spectral initialization of Xu et al. (2017), with the residual sign adapted to our model $\mathbf{\Theta} = \mathbf{S} - \mathbf{B}\mathbf{B}^\top$. The input $\mathbf{P}$ is a preliminary estimate of the marginal precision; $\mathbf{P} = \widehat{\mathbf{\Sigma}}^{-1}$ is one choice when the sample covariance is invertible.

Algorithm 1 Thresholded spectral initialization

Require: Symmetric preliminary precision $\mathbf{P}$; threshold $q > 0$; supplied rank $r_\star$; margin $\underline{\theta} > 0$.

1: Set $(\mathbf{S}_0)_{ij} \leftarrow P_{ij}\mathbf{1}\{i = j \text{ or } |P_{ij}| > q\}$ and $\mathbf{Y}_0 \leftarrow \mathbf{S}_0 - \mathbf{P}$.
2: Compute the leading $r_\star$ eigenpairs $(\nu_j, \mathbf{u}_j)$ of $\mathbf{Y}_0$, ordered by decreasing eigenvalue.
3: Set $\mathbf{B}_0 \leftarrow [\sqrt{(\nu_1)_+}\mathbf{u}_1, \dots, \sqrt{(\nu_{r_\star})_+}\mathbf{u}_{r_\star}]$, where $(x)_+ = \max\{x, 0\}$.
4: Set $s_0 \leftarrow \|\mathbf{S}_0\|_{0,\text{off}}$ and $\mathbf{\Theta}_0 \leftarrow \mathbf{S}_0 - \mathbf{B}_0\mathbf{B}_0^\top$.
5: If $\operatorname{rank}(\mathbf{B}_0) < r_\star$ or $\mathbf{\Theta}_0 \not\succeq 2\underline{\theta}\mathbf{I}$, report an inadmissible initializer.
6: **return** $(\mathbf{S}_0, \mathbf{B}_0, s_0)$.

The residual $\mathbf{S}_0 - \mathbf{P}$ has the sign of the positive semidefinite latent component under our convention. Its leading positive eigenpairs therefore supply the initial factor. The admissibility check prevents an indefinite precision or a rank-deficient factor from being passed to ScaledGD.

2.3 Scaled updates and shared line search

To optimize this likelihood, we take a gradient step in the sparse matrix and rescale the factor gradient by its Gram matrix. The resulting gradients and factor direction are

$$\nabla_{\mathbf{S}} f_n = \mathbf{G}_n, \qquad \nabla_{\mathbf{B}} f_n = -2\mathbf{G}_n\mathbf{B}, \qquad \mathbf{D}_B = 2\mathbf{G}_n\mathbf{B}(\mathbf{B}^\top\mathbf{B})^{-1}. \tag{4}$$

The scaling only requires an $r_\star \times r_\star$ positive definite solve, adding a small matrix computation to the factor update.

The role of this solve becomes clearer from a local optimization perspective. With $\mathbf{H} = \mathbf{B}^\top\mathbf{B} \succ 0$ and the sparse block fixed, the scaled direction solves

$$\mathbf{D}_B = \operatorname*{arg\,min}_{\mathbf{D}\in\mathbb{R}^{p\times r_\star}} \left\{ \langle \nabla_{\mathbf{B}} f_n, \mathbf{D}\rangle + \tfrac{1}{2}\left\|\mathbf{D}\mathbf{H}^{1/2}\right\|_F^2 \right\}. \tag{5}$$

Thus the update measures factor displacement in a metric determined by the current Gram matrix. This metric respects the factorization: replacing $\mathbf{B}$ by $\mathbf{B}\mathbf{Q}$ replaces $\mathbf{D}_B$ by $\mathbf{D}_B\mathbf{Q}$ and produces the same next latent matrix. The sparse and factor blocks may use different base steps; a shared backtracking multiplier coordinates their joint update. Since a change in either block affects the precision, the line search checks feasibility and likelihood decrease jointly. Algorithm 2 gives the complete procedure.

Algorithm 2 ScaledGD with symmetric thresholding for LVGGMs

Require: Correct rank $r_\star$; feasible $(\mathbf{S}_0, \mathbf{B}_0)$ with full-column-rank $\mathbf{B}_0$; even budget s; base steps $\eta_S, \eta_B > 0$; margin $\underline{\theta} > 0$; shrinkage $\beta \in (0,1)$.

1: Check $\|\mathbf{S}_0\|_{0,\mathrm{off}} \le s$ and $\mathbf{S}_0 - \mathbf{B}_0\mathbf{B}_0^\top \succeq 2\underline{\theta}\mathbf{I}$.
2: **for** $t = 0, 1, \ldots$ **do**
3: $\quad \mathbf{H}_t \leftarrow \mathbf{B}_t^\top\mathbf{B}_t$, $\mathbf{\Theta}_t \leftarrow \mathbf{S}_t - \mathbf{B}_t\mathbf{B}_t^\top$, $\mathbf{G}_t \leftarrow \widehat{\mathbf{\Sigma}} - \mathbf{\Theta}_t^{-1}$.
4: $\quad \mathbf{D}_{B,t} \leftarrow 2\mathbf{G}_t\mathbf{B}_t\mathbf{H}_t^{-1}$.
5: $\quad$ For shared multiplier $\gamma \in \{1, \beta, \beta^2, \ldots\}$, form

$$\mathbf{B}^+(\gamma) = \mathbf{B}_t + \gamma\eta_B\mathbf{D}_{B,t}, \qquad \mathbf{S}^+(\gamma) = \mathcal{H}_s(\mathbf{S}_t - \gamma\eta_S\mathbf{G}_t).$$

6: $\quad$ Accept the first trial with full-column-rank $\mathbf{B}^+$, $\mathbf{\Theta}^+ = \mathbf{S}^+ - \mathbf{B}^+\mathbf{B}^{+\top} \succeq \underline{\theta}\mathbf{I}$, and (6).
7: $\quad (\mathbf{S}_{t+1}, \mathbf{B}_{t+1}) \leftarrow (\mathbf{S}^+, \mathbf{B}^+)$.

For an accepted shared multiplier γ, let $\Delta\mathbf{S} = \mathbf{S}^+ - \mathbf{S}_t$ and $\Delta\mathbf{B} = \mathbf{B}^+ - \mathbf{B}_t$. The decrease test is

$$f_n(\mathbf{S}^+, \mathbf{B}^+) \le f_n(\mathbf{S}_t, \mathbf{B}_t) - \frac{1}{4\gamma}\left(\frac{\|\Delta\mathbf{S}\|_F^2}{\eta_S} + \frac{\left\|\Delta\mathbf{B}\mathbf{H}_t^{1/2}\right\|_F^2}{\eta_B} \right). \tag{6}$$

Backtracking terminates at any full-rank feasible point strictly above the precision margin; see Appendix D. In computational terms, scaling adds $O(pr_\star^2 + r_\star^3)$ work for Gram formation and the factor solve. Dense likelihood evaluation still generally requires an $O(p^3)$ precision factorization, so this cost remains part of the full iteration.

3 Local convergence without latent-condition-number dependence

3.1 Assumptions and error measures

Let $s_\star = \|\mathbf{S}^\star\|_{0,\mathrm{off}}$, $s_{\mathrm{tot}} = s_\star + p$, and $\Omega = \mathrm{supp}_{\mathrm{off}}(\mathbf{S}^\star)$. The linear space $\mathcal{M}_\Omega$ contains symmetric matrices supported on Ω and the diagonal; $\mathcal{P}_\Omega$ is its Frobenius orthogonal projection. Write $\mathbf{L}^\star =$

$\mathbf{B}^\star\mathbf{B}^{\star\top}$ with rank $r_\star \geq 1$, and set

$$\sigma_\star = \lambda_{r_\star}(\mathbf{L}^\star) > 0, \qquad \kappa_L = \lambda_1(\mathbf{L}^\star)/\sigma_\star, \qquad \beta_{\min} = \min_{(i,j)\in\Omega} |S^\star_{ij}|.$$

If Ω is empty, take $\beta_{\min} = +\infty$. Let $\mathcal{T}_\star = \{\mathbf{B}^\star\mathbf{Y}^\top + \mathbf{Y}\mathbf{B}^{\star\top} : \mathbf{Y} \in \mathbb{R}^{p\times r_\star}\}$.

Assumption 1 (Precision conditioning)**.** *For $0 < m \leq M$, $m\mathbf{I} \preceq \mathbf{\Theta}^\star \preceq M\mathbf{I}$.*

Assumption 2 (Sparse–low-rank transversality)**.** *For $\gamma_\star \in (0, 1]$, $\|\mathcal{P}_\Omega\mathcal{P}_{\mathcal{T}_\star}\|_{\mathrm{op}} \leq 1 - \gamma_\star$, where this operator norm is induced by the Frobenius norm.*

Transversality prevents sparse and latent perturbations from nearly canceling in the precision. Since the sparse space includes the diagonal, a diagonal latent tangent direction would violate this assumption.

To capture recovery of both components while respecting the rotational ambiguity of the factor, we use the matrix error and population likelihood gap

$$d(\mathbf{S}, \mathbf{B})^2 = \|\mathbf{S} - \mathbf{S}^\star\|_F^2 + \left\|\mathbf{B}\mathbf{B}^\top - \mathbf{L}^\star\right\|_F^2, \quad \Psi(\mathbf{S}, \mathbf{B}) = \ell(\mathbf{S} - \mathbf{B}\mathbf{B}^\top) - \ell(\mathbf{\Theta}^\star), \tag{7}$$

where $\ell(\mathbf{\Theta}) = -\log\det\mathbf{\Theta} + \langle\mathbf{\Sigma}^\star, \mathbf{\Theta}\rangle$ is the population loss. Write $d_t = d(\mathbf{S}_t, \mathbf{B}_t)$ and $\Psi_t = \Psi(\mathbf{S}_t, \mathbf{B}_t)$.

Sampling enters through $\mathbf{E} = \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}^\star$. With $u = \max_{i,j}|E_{ij}|$ and $v = \|\mathbf{E}\|_{\mathrm{op}}$, its effect is summarized by

$$\zeta = \sqrt{s_{\mathrm{tot}}}\, u + 2\sqrt{r_\star}\, v. \tag{8}$$

3.2 Convergence guarantee

Theorem 1 (Local ScaledGD optimization–statistical bound)**.** *Suppose Assumptions 1–2 hold. Let $\mathsf{L}, a_0, \omega, R$, and ζ_0 be as defined in Appendix A.1. Use Algorithm 2 with $r = r_\star$, $s = s_\star$, $\underline{\theta} = m/4$, and equal base steps $\eta_S = \eta_B = \eta$ satisfying $0 < \eta \leq 1/(72\mathsf{L})$. Suppose further that $\zeta \leq \zeta_0$, $\mathbf{S}_0 \in \mathcal{M}_\Omega$, $\mathrm{rank}(\mathbf{B}_0) = r_\star$, and*

$$d_0 \leq R/2, \qquad \Psi_0 \leq a_0 R^2/4. \tag{9}$$

Then every first trial is accepted. For all $t \geq 0$, $\mathbf{S}_t$ has the correct signed off-diagonal support, $\mathrm{rank}(\mathbf{B}_t) = r_\star$, $d_t \leq R/2$, and $\mathbf{\Theta}_t \succeq 3m\mathbf{I}/4$. With $\rho = 1 - \eta\omega \in (0, 1)$,

$$\Psi_t \leq \rho^t\Psi_0 + \frac{2}{\omega}(1 - \rho^t)\zeta^2, \tag{10}$$

$$d_t^2 \leq \frac{\mathsf{L}}{a_0}\rho^t d_0^2 + \frac{2}{\omega a_0}\zeta^2. \tag{11}$$

The contraction factor and both coefficients in (11) depend only on $(m, M, \gamma_\star, \eta)$, not on the latent eigenvalues.

The Gram-matrix scaling compensates for unequal factor amplitudes, so the local descent bound controls perturbations of $\mathbf{B}\mathbf{B}^\top$ without a latent-condition-number factor. With $\mathbf{E} = 0$, both components converge to the truth; under sampling noise, the bound gives a geometric transient followed by a statistical neighborhood, not convergence to an empirical minimizer. The radius R and admissible noise ζ_0 still depend on signal strength. Although Algorithm 2 permits unequal base steps, the theorem covers only equal steps. Theorem 1 assumes that the initial pair already has the correct support and lies in this neighborhood. The next section gives sufficient conditions under which Algorithm 1 supplies such a pair, then translates the score term into a sample-size-dependent error bound.

4 Initialization and statistical consequences

The initializer's accuracy depends on the preliminary precision estimate, the entrywise size of the latent component, and the strength of the nonzero sparse edges.

A support-identifying spectral pilot. Let the preliminary precision estimate $\mathbf{P}$ satisfy the deterministic bounds

$$\|\mathbf{P} - \mathbf{\Theta}^\star\|_{\max} \le \delta_\infty, \qquad \|\mathbf{P} - \mathbf{\Theta}^\star\|_{\mathrm{op}} \le \delta_{\mathrm{op}}, \qquad \|\mathbf{L}^\star\|_{\max} \le \ell_\infty. \tag{12}$$

Here $\|\mathbf{A}\|_{\max} = \max_{i,j} |A_{ij}|$. For the guarantee below, run Algorithm 1 with $q = \ell_\infty + \delta_\infty > 0$. The threshold in this sufficient condition depends on unknown population bounds.

Theorem 2 (Sufficient pilot conditions)**.** *Under Assumptions 1–2 and (12), suppose* $\beta_{\min} > 2q$. *Define* $h = \sqrt{s_{\mathrm{tot}}}\, q, \Delta = h + \delta_{\mathrm{op}}, d_{\mathrm{sp}}^2 = h^2 + 8r_\star\Delta^2$. *If*

$$\Delta \le \sigma_\star/2, \qquad d_{\mathrm{sp}} \le \frac{R}{2}\sqrt{a_0/\mathsf{L}}, \tag{13}$$

then Algorithm 1 returns a pair with the correct signed support, $s_0 = s_\star$, *full-column-rank* $\mathbf{B}_0$, *and (9). Consequently, under the step and score conditions of Theorem 1, its conclusions hold for the spectral run with* $s = s_0$ *and* $d_0 \le d_{\mathrm{sp}}$.

The threshold accounts for both estimation error and latent entry size. More samples reduce the former but need not reduce the latter, so even an accurate precision pilot can require strong edges for support identification. Appendix G reports how support recovery varies with sample size and edge strength.

Once a suitable start is available, the effect of sampling enters through the fixed score matrix $\mathbf{E}$. The following concentration event controls its entrywise and spectral norms simultaneously.

Lemma 3 (One Gaussian score event)**.** *For* $\delta \in (0,1)$, *let* $h_\delta = \log(4p^2/\delta)$ *and* $q_\delta = p\log 9 + \log(4/\delta)$. *With probability at least* $1-\delta$,

$$u \le u_\delta = \frac{2}{m}\left(\sqrt{h_\delta/n} + h_\delta/n\right), \qquad v \le v_\delta = \frac{4}{m}\left(\sqrt{q_\delta/n} + q_\delta/n\right). \tag{14}$$

Thus $\zeta \le \zeta_\delta = \sqrt{s_{\mathrm{tot}}}u_\delta + 2\sqrt{r_\star}v_\delta$.

This event controls projections onto every rank-$r_\star$ subspace, including data-dependent iterates. It therefore combines with Theorem 1 to yield a statistical bound while reusing the same data throughout.

Corollary 4 (Statistical matrix error)**.** *Suppose the initialization conditions of Theorem 1 hold on the score event and* $\zeta_\delta \le \zeta_0$. *For* $n \ge \max\{h_\delta, q_\delta\}$, *with probability at least* $1-\delta$, *simultaneously for every* $t \ge 0$,

$$d_t \le C\left[\rho^{t/2}d_0 + \sqrt{\frac{(s_\star + p)h_\delta}{n}} + \sqrt{\frac{r_\star q_\delta}{n}}\right], \tag{15}$$

where C *depends only on* $(m, M, \gamma_\star)$. *For fixed* $\delta = p^{-c}$ *and* $p \ge 2$, *the statistical terms have order* $\varepsilon_n = \sqrt{(s_\star + p)\log p/n} + \sqrt{pr_\star/n}$.

The sparse and low-rank statistical terms match the dimension, sparsity, and rank dependence of established LVGGM bounds (Meng et al., 2014; Xu et al., 2017). Here the guarantee concerns scaled iterates under stronger local assumptions, rather than a new minimax theorem. Initialization and noise restrictions may still introduce sample-size dependence on $\sigma_\star$, while C and the contraction factor remain independent of the latent spectrum.

A concrete pilot and stopping interpretation. If $v_\delta \le 1/(2M)$, the sample-inverse pilot $\mathbf{P} = \widehat{\mathbf{\Sigma}}^{-1}$ satisfies (12) with $\delta_\infty = \delta_{\mathrm{op}} = 2M^2 v_\delta$. Theorem 2 applies when its remaining conditions hold. Singular sample covariance, including $n < p$, requires another pilot with its own guarantee. After initialization, sampling error dominates the bound once $\rho^{t/2}d_0 \lesssim \varepsilon_n$. This gives a theoretical optimization benchmark; a data-driven stopping rule remains a separate practical choice.

Corollary 5 (Precision, covariance, and graph consequences)**.** *Along the trajectory of Theorem 1,*

$$\|\mathbf{\Theta}_t - \mathbf{\Theta}^\star\|_F \le \sqrt{2}d_t, \qquad \left\|\mathbf{\Theta}_t^{-1} - \mathbf{\Sigma}^\star\right\|_F \le \frac{2\sqrt{2}}{m^2}d_t.$$

The signed off-diagonal support of $\mathbf{S}_t$ *equals that of* $\mathbf{S}^\star$ *for all* t. *This is preservation of support identified at initialization, not a theorem about correcting an initially wrong support.*

These consequences return to the observed distribution and conditional graph. Spectral independence applies to matrix recovery; unweighted factor error has a different sensitivity to weak latent eigenvalues.

5 Simulation Study

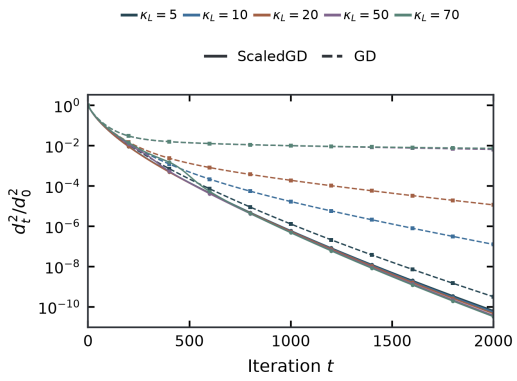
Data generation. We generate a diagonally dominant sparse matrix $\mathbf{S}^\star$ from a random graph. Let $\mathbf{U} \in \mathbb{R}^{p \times r_\star}$ have orthonormal columns and let $\mathbf{D}_\kappa$ have geometrically spaced diagonal entries from 1 to $1/\kappa_L$. With $\widetilde{\mathbf{L}} = \mathbf{U}\mathbf{D}_\kappa\mathbf{U}^\top$, we set

$$\mathbf{L}^\star = \frac{\chi\widetilde{\mathbf{L}}}{\left\|(\mathbf{S}^\star)^{-1/2}\widetilde{\mathbf{L}}(\mathbf{S}^\star)^{-1/2}\right\|_{\mathrm{op}}}, \qquad \mathbf{\Theta}^\star = \mathbf{S}^\star - \mathbf{L}^\star, \qquad \mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, (\mathbf{\Theta}^\star)^{-1}). \tag{16}$$

For $0 < \chi < 1$, this construction gives $\mathbf{\Theta}^\star \succeq (1-\chi)\mathbf{S}^\star \succ 0$. Each experiment specifies its dimensions, signal strength, and sample size.

5.1 Effect of latent conditioning

We set $p = 200$, $r_\star = 5$, mean graph degree four, edge magnitude $\omega_e = 0.2$, and $\chi = 0.5$. We vary $\kappa_L \in \{5, 10, 20, 50, 70\}$ over 20 paired Monte Carlo replications, keeping the graph and latent subspace fixed across conditions within each replication. To isolate optimization from covariance estimation error, we use the true covariance $\mathbf{\Sigma}^\star = (\mathbf{\Theta}^\star)^{-1}$ in the likelihood and set the preliminary precision to $\mathbf{P} = \mathbf{\Theta}^\star$. The spectral threshold is $q = \|\mathbf{P}\|_{\mathrm{op}}\sqrt{\frac{\log p}{6400}}$, with denominator 6400 fixed across conditions; the rank is set to $r_\star = 5$. Thus the initialization follows the same thresholded spectral construction at each condition, while its input reflects the corresponding ground-truth precision. Every resulting initial precision is positive definite and every initial factor has full column rank. ScaledGD and GD share the initial pair, sparsity budget, and backtracking rule; GD omits the Gram preconditioner from the factor update. Both use $\eta_S = \eta_B = 0.1$ and run for 2000 iterations without early stopping.



κ_L	ScaledGD	GD
5	-0.00914	-0.00832
10	-0.00927	-0.00490
20	-0.00939	-0.00281
50	-0.00950	-0.000424
70	-0.00959	-0.000313

Figure 1: Convergence under the true covariance at $\chi = 0.5$ (20 paired replications). Left: mean normalized joint matrix error d_t^2/d_0^2; solid lines denote ScaledGD and dashed lines GD. Right: secant slope of the log mean curve between iterations 1000 and 2000.

Figure 1 shows that the ScaledGD curves nearly coincide after the initial transient: over iterations 1000–2000, their log-error slopes range from -0.00914 to -0.00959 as κ_L varies. By contrast, GD's slope changes from -0.00832 at $\kappa_L = 5$ to -0.000313 at $\kappa_L = 70$. The nearly unchanged ScaledGD slopes empirically support the κ_L-independent local contraction factor in Theorem 1.

5.2 Recovery and convergence from sampled data

For sampled-data experiments, we set $\chi = 0.9$, mean graph degree four, edge magnitude $\omega_e = 0.2$, and $\kappa_L = 5$. ScaledGD and AltGD (Xu et al., 2017) use the same data and initialization at the supplied rank. We also compare with PPA for the convex LVGGM estimator (Chandrasekaran et al., 2010; Wang et al., 2010) and graphical lasso (Friedman et al., 2008) as a sparse-only reference for marginal precision recovery. We form the preliminary precision as $\mathbf{P} = \widehat{\mathbf{\Sigma}}^{-1}$ from the training sample. In these synthetic comparisons, the initial sparse matrix retains the known number $s_\star/2$ of off-diagonal pairs with largest magnitude; this fixes the budget but does not reveal their locations. The initial factor is obtained from the residual in Algorithm 1, and its scale is selected by validation likelihood. ScaledGD and AltGD receive exactly the same initial pair. Their step sizes are selected on validation data; ScaledGD tunes η_S and η_B separately and uses shared backtracking.

For Table 1, ScaledGD and AltGD run with no damping, a 10^{-4} relative-iterate stopping tolerance, and a cap of 1000 iterations. We report the iterate with the lowest validation likelihood along each trajectory, while the runtime covers the full optimization run. PPA retains its native stopping rule.

For the recovery comparison, we set $p = 200$, $r = r_\star = 5$, and vary $n \in \{800, 1600, 3200, 6400\}$. Table 1 reports estimation errors and runtime. At $n = 1600$ and 3200, ScaledGD has the lowest sparse-component and precision errors among the four methods and takes less time than AltGD and PPA. At $n = 6400$, its sparse-component and precision errors are nearly identical to AltGD's, with a shorter runtime. PPA has the lowest latent-component error throughout. Graphical lasso estimates only Θ; for its sparse-component error we use $\widehat{\mathbf{S}} = \widehat{\mathbf{\Theta}}$, and we do not report a latent-component error.

Table 1: Relative Frobenius errors and optimization time in seconds (mean $\pm$ standard error over 20 replications). A dash indicates that the method has no latent estimate.

n	Method	$\|\widehat{\mathbf{S}} - \mathbf{S}^\star\|_F / \|\mathbf{S}^\star\|_F$	$\|\widehat{\mathbf{L}} - \mathbf{L}^\star\|_F / \|\mathbf{L}^\star\|_F$	$\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^\star\|_F / \|\mathbf{\Theta}^\star\|_F$	Time (s)
800	ScaledGD	0.203 ± 0.001	1.365 ± 0.016	0.209 ± 0.001	3.82 ± 0.17
	AltGD	0.310 ± 0.007	1.674 ± 0.039	0.310 ± 0.007	3.97 ± 0.23
	PPA	0.204 ± 0.001	$\mathbf{0.719 \pm 0.004}$	0.205 ± 0.001	4.00 ± 0.04
	Glasso	$\mathbf{0.200 \pm 0.001}$	—	$\mathbf{0.195 \pm 0.001}$	$\mathbf{0.30 \pm 0.04}$
1600	ScaledGD	$\mathbf{0.111 \pm 0.001}$	1.055 ± 0.013	$\mathbf{0.124 \pm 0.001}$	2.01 ± 0.11
	AltGD	0.140 ± 0.002	1.086 ± 0.014	0.149 ± 0.002	3.10 ± 0.22
	PPA	0.161 ± 0.001	$\mathbf{0.591 \pm 0.004}$	0.163 ± 0.001	3.71 ± 0.04
	Glasso	0.156 ± 0.0005	—	0.148 ± 0.0005	$\mathbf{0.56 \pm 0.05}$
3200	ScaledGD	$\mathbf{0.060 \pm 0.001}$	0.804 ± 0.009	$\mathbf{0.076 \pm 0.001}$	$\mathbf{0.63 \pm 0.05}$
	AltGD	0.069 ± 0.002	0.775 ± 0.011	0.082 ± 0.001	1.97 ± 0.24
	PPA	0.122 ± 0.001	$\mathbf{0.485 \pm 0.003}$	0.123 ± 0.001	3.46 ± 0.04
	Glasso	0.125 ± 0.0004	—	0.115 ± 0.0004	2.17 ± 0.46
6400	ScaledGD	$\mathbf{0.040 \pm 0.001}$	0.571 ± 0.008	0.053 ± 0.001	$\mathbf{0.24 \pm 0.02}$
	AltGD	$\mathbf{0.040 \pm 0.001}$	0.549 ± 0.013	$\mathbf{0.052 \pm 0.001}$	1.31 ± 0.32
	PPA	0.089 ± 0.001	$\mathbf{0.401 \pm 0.003}$	0.090 ± 0.001	3.45 ± 0.04
	Glasso	0.103 ± 0.001	—	0.089 ± 0.0004	6.07 ± 0.80

Figure 2 reports a separate convergence experiment, in which p, $r_\star$, and n increase together rather than varying n at fixed p and $r_\star$ as in Table 1. It tracks the Frobenius errors of $\mathbf{S}_t$, $\mathbf{B}_t\mathbf{B}_t^\top$, and $\mathbf{\Theta}_t$ over 1000 iterations, with the same data and initialization for ScaledGD and AltGD in each setting. The three panels distinguish recovery of the sparse and latent components from fitting their difference: a small precision error alone need not imply accurate recovery of both components. ScaledGD reduces the sparse-component and precision errors faster than AltGD; the difference in latent-component error is smaller.

6 Conclusion and discussion

We studied joint sparse and low-rank estimation in latent-variable Gaussian graphical models and proposed ScaledGD, which combines a Gram-scaled factor update with sparse hard thresholding and a joint feasibility line search. At the correctly specified rank, our main theorem gives local linear

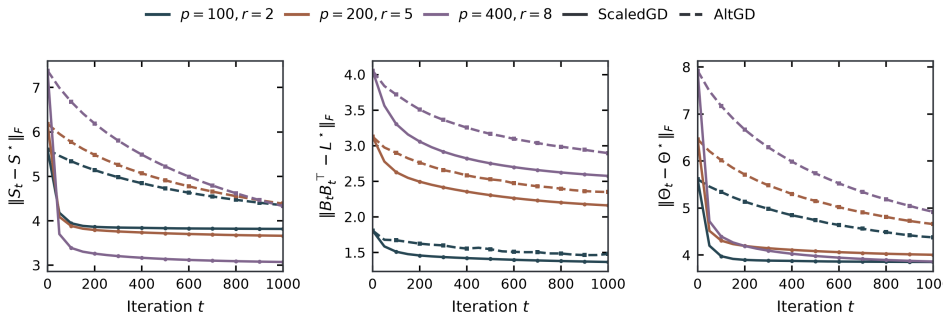


Figure 2: Sparse, latent, and precision errors along ScaledGD and AltGD trajectories for $(p, r_\star, n) = (100, 2, 800), (200, 5, 1600), (400, 8, 3200)$. Solid and dashed lines denote ScaledGD and AltGD; methods share data and initialization within each setting.

convergence for both matrix components with a contraction factor independent of the latent condition number; we also give sufficient conditions for a thresholded spectral initializer and derive a statistical error bound. The experiments illustrate the predicted conditioning behavior and compare recovery and convergence with existing methods.

The analysis isolates two distinct roles of latent conditioning. The factor-Gram scaling removes the latent eigenvalues from the local contraction bound, while the smallest nonzero eigenvalue still enters the sufficient initialization radius and noise level. Thus the result concerns local optimization after the iterates enter the prescribed neighborhood; it does not remove the signal conditions needed to identify the sparse and latent components.

The guarantee assumes an identified sparse support, the correct latent rank, a sufficiently accurate initialization, and controlled sampling noise. The analysis does not address how to learn the support during optimization or how to select the rank from data. Rank misspecification is especially relevant for ScaledGD because redundant factor columns can make the Gram matrix singular. Extending the method and its analysis to jointly handle support and rank selection is a natural next step.

AI USE STATEMENT

We used generative AI tools to discuss experimental design, implement and debug simulation code. We also used them to examine mathematical arguments, improve the presentation of proofs. The authors take responsibility for the mathematical claims, cited sources, experimental results, and final text, including AI-assisted material.

REPRODUCIBILITY STATEMENT

The supplement provides complete proofs, explicit constants, and the full simulation settings and summary results. The experiments use the data-generating process, initialization, stopping rules, and parameter choices specified in Section 5 and Appendix G.

REFERENCES

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

Tony Cai, Weidong Liu, and Xi Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011. doi: 10.1198/jasa.2011.tm10155.

Emmanuel J. Candés and Mahdi Soltanolkotabi. Discussion: Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1997–2004, 2012. doi: 10.1214/12-AOS1001.

Venkat Chandrasekaran, Pablo A Parrilo, and Alan S Willsky. Latent variable graphical model selection via convex optimization. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1610–1613. IEEE, 2010.

Venkat Chandrasekaran, Sujay Sanghavi, Pablo A. Parrilo, and Alan S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

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, 2012a.

Venkat Chandrasekaran, Pablo A. Parrilo, and Alan S. Willsky. Rejoinder: Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):2005–2013, 2012b. doi: 10.1214/12-AOS1020.

Valentina Ciccone, Augusto Ferrante, and Mattia Zorzi. Robust identification of "sparse plus low-rank" graphical models: An optimization approach. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 2241–2246. IEEE, 2018. doi: 10.1109/CDC.2018.8619796.

Arthur P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972. doi: 10.2307/2528966.

Mathias Drton and Marloes H. Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393, 2017. doi: 10.1146/annurev-statistics-060116-053803.

Ignacio Echave-Sustaeta Rodríguez and Frank Röttger. Latent Gaussian graphical models with Golazo penalty. In *Proceedings of the 12th International Conference on Probabilistic Graphical Models*, volume 246 of *Proceedings of Machine Learning Research*, pages 199–212. PMLR, 2024.

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

Christophe Giraud and Alexandre Tsybakov. Discussion: Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1984–1988, 2012.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

Steffen Lauritzen and Nicolai Meinshausen. Discussion: Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1973–1977, 2012. doi: 10.1214/12-AOS980.

Steffen L. Lauritzen. *Graphical Models*. Oxford University Press, second edition, 2026. doi: 10.1093/9780191775284.001.0001.

Ran Li, Jiming Lin, Hongbing Qiu, Wenhui Zhang, and Junyi Wang. Graph learning for latent-variable Gaussian graphical models under Laplacian constraints. *Neurocomputing*, 532:67–76, 2023.

Shiqian Ma, Lingzhou Xue, and Hui Zou. Alternating direction methods for latent variable Gaussian graphical model selection. *Neural Computation*, 25(8):2172–2198, 2013.

Marloes Maathuis, Mathias Drton, Steffen L. Lauritzen, and Martin J. Wainwright, editors. *Handbook of Graphical Models*. CRC Press, 2019.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006. doi: 10.1214/009053606000000281.

Zhaoshi Meng, Brian Eriksson, and Alfred O. Hero. Learning latent variable gaussian graphical models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 1269–1277, 2014.

Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011. doi: 10.1214/11-EJS631.

Zhao Ren and Harrison H. Zhou. Discussion: Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1989–1996, 2012. doi: 10.1214/12-AOS985.

Adam J. Rothman, Peter J. Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008. doi: 10.1214/08-EJS176.

Dapeng Shi, Tiandong Wang, and Zhiliang Ying. Simultaneous identification of sparse structures and communities in heterogeneous graphical models. *arXiv preprint arXiv:2405.09841*, 2024.

Tian Tong, Cong Ma, and Yuejie Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *Journal of Machine Learning Research*, 22(150):1–63, 2021.

Marina Vinyes and Guillaume Obozinski. Learning the effect of latent variables in Gaussian graphical models with unobserved variables. *arXiv preprint arXiv:1807.07754*, 2018.

Martin J. Wainwright. Discussion: Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1978–1983, 2012. doi: 10.1214/12-AOS981.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. doi: 10.1561/2200000001.

Chengjing Wang, Defeng Sun, and Kim-Chuan Toh. Solving log-determinant optimization problems by a Newton-CG primal proximal point algorithm. *SIAM Journal on Optimization*, 20(6):2994–3013, 2010. doi: 10.1137/090772514.

Changjing Wu, Hongyu Zhao, Huaying Fang, and Minghua Deng. Graphical model selection with latent variables. *Electronic Journal of Statistics*, 11(2):3485–3521, 2017. doi: 10.1214/17-EJS1331.

Pan Xu, Jian Ma, and Quanquan Gu. Speeding up latent variable gaussian graphical model estimation via nonconvex optimization. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ming Yuan. Discussion: Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1968–1972, 2012. doi: 10.1214/12-AOS979.

Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007. doi: 10.1093/biomet/asm018.

Appendix contents

Appendix A Notation and constants **12**

Appendix B Proofs of supporting lemmas **13**

Appendix C Proof of Theorem 1 **15**

Appendix D Proof of backtracking termination **17**

Appendix E Proof of Theorem 2 **17**

Appendix F Proofs of the statistical results **18**

Appendix G Supplementary simulation results **19**

Guide. Appendix A gives notation and constants. Appendices B–C prove Theorem 1, beginning with the supporting lemmas. Appendices D–F cover backtracking, Theorem 2, and the statistical results, respectively. Appendix G reports supplementary simulations.

A Notation and constants

Throughout this appendix, all unspecified matrix norms are Frobenius norms, while $\|\cdot\|_{\mathrm{op}}$ is the spectral norm. Let $\mathbf{L} = \mathbf{B}\mathbf{B}^\top$, $\mathbf{H} = \mathbf{B}^\top\mathbf{B}$, $\mathbf{P} = \mathbf{P}_{\mathbf{B}}$, $\mathbf{D} = \mathbf{L} - \mathbf{L}^\star$, $\mathbf{A} = \mathbf{S} - \mathbf{S}^\star$, and $\mathbf{Z} = \mathbf{\Theta} - \mathbf{\Theta}^\star = \mathbf{A} - \mathbf{D}$. The symbol $\mathbf{P}$ in the geometric proofs denotes the column-space projector, not the precision pilot used in Appendix E. The sparse projection $\mathcal{P}_\Omega$ always includes the diagonal.

A.1 Constants and local neighborhood

For the local result, define

$$\mu = \frac{1}{4M^2}, \qquad \mathsf{L} = \frac{4}{m^2}, \qquad a_0 = \frac{\mu\gamma_\star}{8}, \qquad \omega = \frac{a_0^2}{2\mathsf{L}}, \qquad \kappa_\ell = \mathsf{L}/\mu = 16(M/m)^2. \tag{17}$$

Here κ_ℓ is a ratio of likelihood curvature bounds; χ in the simulation design denotes latent signal strength and is unrelated to this ratio. Choose any positive radius satisfying

$$R \le \min\left\{\frac{\sigma_\star}{8}, \frac{m}{4\sqrt{2}}, \frac{\beta_{\min}}{4}, \frac{\sqrt{\gamma_\star}\sigma_\star}{8}, \frac{\mu\gamma_\star\sigma_\star}{16\sqrt{2}\mathsf{L}}\right\}. \tag{18}$$

For the score ζ defined in (8), the admissible noise level is

$$\zeta_0 = R\min\left\{\mathsf{L}, \sqrt{\frac{\omega a_0}{8}}\right\}. \tag{19}$$

At the largest permitted step $\eta = 1/(72\mathsf{L})$, the contraction factor is

$$\rho = 1 - \frac{\gamma_\star^2}{9216\kappa_\ell^2}. \tag{20}$$

Thus the displayed contraction and error multipliers depend only on precision bounds and transversality, whereas the neighborhood and admissible noise also depend on the weakest latent eigenvalue and sparse edge strength.

B Proofs of supporting lemmas

B.1 Likelihood identities

For a symmetric direction $\mathbf{V}$ and positive definite $\mathbf{\Theta}$,

$$\nabla \ell(\mathbf{\Theta}) = \mathbf{\Sigma}^\star - \mathbf{\Theta}^{-1}, \qquad \nabla^2 \ell(\mathbf{\Theta})[\mathbf{V}] = \mathbf{\Theta}^{-1}\mathbf{V}\mathbf{\Theta}^{-1}.$$

The empirical and population Hessians coincide, since their difference is linear in $\mathbf{\Theta}$. On $m\mathbf{I}/2 \preceq \mathbf{\Theta} \preceq 2M\mathbf{I}$,

$$\mu \|\mathbf{V}\|_F^2 \le \langle \mathbf{V}, \nabla^2 \ell(\mathbf{\Theta})[\mathbf{V}] \rangle = \left\| \mathbf{\Theta}^{-1/2}\mathbf{V}\mathbf{\Theta}^{-1/2} \right\|_F^2 \le \mathsf{L} \|\mathbf{V}\|_F^2 . \tag{21}$$

If $d \le R$, then $\|\mathbf{Z}\|_{\mathrm{op}} \le \sqrt{2}d \le m/4$, so

$$3m\mathbf{I}/4 \preceq \mathbf{\Theta} \preceq (M + m/4)\mathbf{I} \preceq 2M\mathbf{I}. \tag{22}$$

The line segment from $\mathbf{\Theta}^\star$ to $\mathbf{\Theta}$ lies in the same spectral band. Taylor's integral formula, with $\nabla \ell(\mathbf{\Theta}^\star) = 0$, gives

$$\frac{\mu}{2} \|\mathbf{Z}\|_F^2 \le \Psi \le \frac{\mathsf{L}}{2} \|\mathbf{Z}\|_F^2 , \quad \langle \mathbf{G}, \mathbf{Z} \rangle \ge \mu \|\mathbf{Z}\|_F^2 , \quad \|\mathbf{G}\|_F \le \mathsf{L} \|\mathbf{Z}\|_F . \tag{23}$$

The chain rule applied to $\mathbf{\Theta} = \mathbf{S} - \mathbf{B}\mathbf{B}^\top$ proves (4). These identities also explain why convexity in the precision alone does not imply strong convexity in the separate components: cancellation in $\mathbf{A} - \mathbf{D}$ must be excluded geometrically.

B.2 Fixed-rank geometry

For a rank-$r_\star$ PSD matrix $\mathbf{L}$ with range projector $\mathbf{P}$, its tangent space in the symmetric matrices is

$$\mathcal{T}(\mathbf{L}) = \{\mathbf{B}\mathbf{Y}^\top + \mathbf{Y}\mathbf{B}^\top : \mathbf{Y} \in \mathbb{R}^{p \times r_\star}\}, \qquad \mathcal{P}_{\mathcal{T}(\mathbf{L})}\mathbf{V} = \mathbf{P}\mathbf{V} + \mathbf{V}\mathbf{P} - \mathbf{P}\mathbf{V}\mathbf{P}.$$

In an orthonormal basis adapted to $\mathbf{P}$, these tangent matrices have arbitrary upper-left and off-diagonal blocks and a zero lower-right block. Consequently the orthogonal normal projection is $(\mathbf{I} - \mathbf{P})\mathbf{V}(\mathbf{I} - \mathbf{P})$.

Lemma 6 (Fixed-rank chord remainder)**.** *Suppose $\mathbf{L}, \mathbf{L}^\star \succeq 0$ have rank $r_\star$, and $e = \|\mathbf{L} - \mathbf{L}^\star\|_F \le \sigma_\star/4$. Then*

$$\|(\mathbf{I} - \mathbf{P})(\mathbf{L} - \mathbf{L}^\star)(\mathbf{I} - \mathbf{P})\|_F \le 2e^2/\sigma_\star, \tag{24}$$

$$\|(\mathbf{I} - \mathbf{P}_\star)(\mathbf{L} - \mathbf{L}^\star)(\mathbf{I} - \mathbf{P}_\star)\|_F \le 2e^2/\sigma_\star, \tag{25}$$

where $\mathbf{P}_\star$ projects onto the range of $\mathbf{L}^\star$.

Proof. Write $\mathbf{L}$ in an orthonormal basis for the range of $\mathbf{L}^\star$ and its complement as

$$\mathbf{L} = \begin{pmatrix} \mathbf{A}_1 & \mathbf{C}^\top \\ \mathbf{C} & \mathbf{D}_1 \end{pmatrix}, \qquad \mathbf{L}^\star = \begin{pmatrix} \mathbf{\Lambda}_\star & 0 \\ 0 & 0 \end{pmatrix}.$$

Here $\mathbf{A}_1 \succeq (\sigma_\star - e)\mathbf{I} \succeq \sigma_\star \mathbf{I}/2$. Block Gaussian elimination is a congruence that changes $\mathbf{L}$ into $\mathrm{diag}(\mathbf{A}_1, \mathbf{D}_1 - \mathbf{C}\mathbf{A}_1^{-1}\mathbf{C}^\top)$. Since $\mathbf{L}$ and $\mathbf{A}_1$ both have rank $r_\star$, the Schur complement is zero. Thus

$$\|\mathbf{D}_1\|_F \le \|\mathbf{C}\|_{\mathrm{op}} \left\|\mathbf{A}_1^{-1}\right\|_{\mathrm{op}} \|\mathbf{C}\|_F \le 2e^2/\sigma_\star,$$

which proves (25). To prove (24), interchange the two matrices. Weyl's inequality gives $\lambda_{r_\star}(\mathbf{L}) \ge \sigma_\star - e$. The upper-left block of $\mathbf{L}^\star$ in the range basis of $\mathbf{L}$ is therefore bounded below by $(\sigma_\star - 2e)\mathbf{I} \succeq \sigma_\star \mathbf{I}/2$. The same Schur-complement argument applies, with a possible minus sign on the normal block, which does not affect its norm. □

Lemma 7 (Component geometry)**.** *Under the two model assumptions, if $\mathbf{S} \in \mathcal{M}_\Omega$, $\mathrm{rank}(\mathbf{B}) = r_\star$, and $d \le R$, then*

$$\frac{\sqrt{\gamma_\star}}{2} d \le \|\mathbf{Z}\|_F \le \sqrt{2}d, \quad a_0 d^2 \le \Psi \le \mathsf{L} d^2, \quad \lambda_{\min}(\mathbf{H}) \ge \sigma_\star/2. \tag{26}$$

Proof. Let $\mathbf{T}_\star = \mathcal{P}_{\mathcal{T}_\star}\mathbf{D}$ and $\mathbf{N}_\star = \mathbf{D} - \mathbf{T}_\star$. Transversality implies, for $\mathbf{A} \in \mathcal{M}_\Omega$ and $\mathbf{T}_\star \in \mathcal{T}_\star$,

$$|\langle \mathbf{A}, \mathbf{T}_\star \rangle| \le (1-\gamma_\star)\,\|\mathbf{A}\|_F\,\|\mathbf{T}_\star\|_F\,, \quad \|\mathbf{A}-\mathbf{T}_\star\|_F^2 \ge \gamma_\star(\|\mathbf{A}\|_F^2 + \|\mathbf{T}_\star\|_F^2).$$

The triangle inequality in the product space gives $\sqrt{\|\mathbf{A}\|_F^2 + \|\mathbf{T}_\star\|_F^2} \ge d - \|\mathbf{N}_\star\|_F$. By Lemma 6,

$$\|\mathbf{Z}\|_F \ge \sqrt{\gamma_\star}d - (1+\sqrt{\gamma_\star})\,\|\mathbf{N}_\star\|_F \ge \sqrt{\gamma_\star}d - 4d^2/\sigma_\star \ge \frac{\sqrt{\gamma_\star}}{2}d.$$

The last inequality uses $R \le \sqrt{\gamma_\star}\sigma_\star/8$. The upper bound is Cauchy–Schwarz. Substitution in (23) proves the potential bounds, with $a_0 = \mu\gamma_\star/8$. Finally, the positive eigenvalues of $\mathbf{L}$ are those of $\mathbf{H}$; Weyl's inequality and $R \le \sigma_\star/8$ imply the stated Gram lower bound. □

B.3 Scaled tangent lift and gradient bound

Lemma 8 (A lift without latent spectral factors)**.** *Let* $\mathbf{B}$ *have full column rank and* $\mathbf{T} \in \mathcal{T}(\mathbf{L})$*. There exists a matrix* $\mathbf{Y}$ *such that*

$$\mathbf{B}\mathbf{Y}^\top + \mathbf{Y}\mathbf{B}^\top = \mathbf{T}, \qquad \left\|\mathbf{Y}\mathbf{H}^{1/2}\right\|_F^2 \le \frac{1}{2}\,\|\mathbf{T}\|_F^2\,. \tag{27}$$

Moreover, for any symmetric $\mathbf{V}$*,*

$$\left\|\mathbf{V}\mathbf{B}\mathbf{H}^{-1/2}\right\|_F^2 = \|\mathbf{V}\mathbf{P}\|_F^2\,. \tag{28}$$

Proof. Write $\mathbf{B} = \mathbf{U}\mathbf{R}$ with $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$ and invertible $\mathbf{R}$. Choose an orthonormal complement $\mathbf{U}_\perp$ and write

$$\mathbf{T} = \mathbf{U}\mathbf{A}_1\mathbf{U}^\top + \mathbf{U}_\perp\mathbf{C}\mathbf{U}^\top + \mathbf{U}\mathbf{C}^\top\mathbf{U}_\perp^\top, \qquad \mathbf{A}_1 = \mathbf{A}_1^\top.$$

Take $\mathbf{Y} = (\mathbf{U}\mathbf{A}_1/2 + \mathbf{U}_\perp\mathbf{C})\mathbf{R}^{-\top}$. Multiplication gives the first identity in (27). Since $\mathbf{H} = \mathbf{R}^\top\mathbf{R}$,

$$\left\|\mathbf{Y}\mathbf{H}^{1/2}\right\|_F^2 = \tfrac{1}{4}\,\|\mathbf{A}_1\|_F^2 + \|\mathbf{C}\|_F^2 \le \tfrac{1}{2}(\|\mathbf{A}_1\|_F^2 + 2\,\|\mathbf{C}\|_F^2) = \tfrac{1}{2}\,\|\mathbf{T}\|_F^2\,.$$

For the second identity, $\mathbf{B}\mathbf{H}^{-1/2}$ has orthonormal columns and its range projector is $\mathbf{P}$. Taking traces gives $\left\|\mathbf{V}\mathbf{B}\mathbf{H}^{-1/2}\right\|_F^2 = \operatorname{tr}(\mathbf{V}^\top\mathbf{V}\mathbf{P}) = \|\mathbf{V}\mathbf{P}\|_F^2$. □

For $\mathbf{L} = \mathbf{B}\mathbf{B}^\top$, write $\mathbf{H} = \mathbf{B}^\top\mathbf{B}$, $\mathbf{P}_\mathbf{B} = \mathbf{B}\mathbf{H}^{-1}\mathbf{B}^\top$, and $\mathbf{G} = \mathbf{\Sigma}^\star - \mathbf{\Theta}^{-1}$. Define the scaled descent energy

$$Q^2 = \|\mathcal{P}_\Omega\mathbf{G}\|_F^2 + \left\|(-2\mathbf{G}\mathbf{B})\mathbf{H}^{-1/2}\right\|_F^2 = \|\mathcal{P}_\Omega\mathbf{G}\|_F^2 + 4\,\|\mathbf{G}\mathbf{P}_\mathbf{B}\|_F^2\,. \tag{29}$$

Lemma 9 (Joint scaled gradient dominance)**.** *In the neighborhood of Lemma 7, the quantity* Q *in* (29) *satisfies*

$$a_0 d \le Q \le 4\mathsf{L}d, \qquad Q^2 \ge \frac{a_0^2}{\mathsf{L}}\Psi = 2\omega\Psi. \tag{30}$$

Proof. Set $\mathbf{T} = \mathcal{P}_{\mathcal{T}(\mathbf{L})}\mathbf{D}$ and $\mathbf{N} = \mathbf{D} - \mathbf{T}$. Lemma 6 bounds $\|\mathbf{N}\|_F \le 2d^2/\sigma_\star$. Choose the lift $\mathbf{Y}$ of $\mathbf{T}$ in Lemma 8. Because $\mathbf{Z} = \mathbf{A} - \mathbf{T} - \mathbf{N}$,

$$\begin{aligned}
\langle \mathbf{G}, \mathbf{Z}\rangle &= \langle \mathcal{P}_\Omega\mathbf{G}, \mathbf{A}\rangle + \langle -2\mathbf{G}\mathbf{B}, \mathbf{Y}\rangle - \langle \mathbf{G}, \mathbf{N}\rangle \\
&\le Q\left(\|\mathbf{A}\|_F^2 + \left\|\mathbf{Y}\mathbf{H}^{1/2}\right\|_F^2\right)^{1/2} + \|\mathbf{G}\|_F\,\|\mathbf{N}\|_F \\
&\le Qd + \frac{2\sqrt{2}\mathsf{L}}{\sigma_\star}d^3.
\end{aligned}$$

The second line is Cauchy–Schwarz with the dual pair $\mathbf{H}^{-1/2}, \mathbf{H}^{1/2}$. The last line uses $\|\mathbf{T}\|_F \le \|\mathbf{D}\|_F$ and (23). On the other hand, (23) and (26) give $\langle \mathbf{G}, \mathbf{Z}\rangle \ge \mu\gamma_\star d^2/4$. Our radius guarantees

$2\sqrt{2}\mathsf{L}R/\sigma_\star \le \mu\gamma_\star/8$. Subtracting the remainder and dividing by $d > 0$ yields $Q \ge a_0 d$; at $d = 0$ the conclusion is immediate. Also

$$Q \le \sqrt{5}\,\|\mathbf{G}\|_F \le \sqrt{10}\mathsf{L}d \le 4\mathsf{L}d.$$

Finally $\Psi \le \mathsf{L}d^2$ proves the potential form. No bound on an unweighted factor distance, or on $\lambda_1(\mathbf{L}^\star)$, has entered this argument. □

C Proof of Theorem 1

C.1 Score bound

At any full-column-rank $\mathbf{B}$, define

$$g = (\mathcal{P}_\Omega \mathbf{G}, -2\mathbf{G}\mathbf{B}), \qquad e_g = (\mathcal{P}_\Omega \mathbf{E}, -2\mathbf{E}\mathbf{B}).$$

For a pair $(\mathbf{U}, \mathbf{V})$, let $\|(\mathbf{U}, \mathbf{V})\|_{*,\mathbf{B}}^2 = \|\mathbf{U}\|_F^2 + \left\|\mathbf{V}\mathbf{H}^{-1/2}\right\|_F^2$. By Lemma 8,

$$\|e_g\|_{*,\mathbf{B}}^2 \le s_{\mathrm{tot}} u^2 + 4 r_\star v^2 \le \zeta^2. \tag{31}$$

Indeed $\|\mathbf{E}\mathbf{P}\|_F \le \sqrt{r_\star}\,\|\mathbf{E}\|_{\mathrm{op}}$. This bound holds for every current subspace, including one selected using the data. Let

$$Q_n^2 = \|\mathcal{P}_\Omega \mathbf{G}_n\|_F^2 + 4\,\|\mathbf{G}_n \mathbf{P}\|_F^2\,.$$

Then $Q_n \le Q + \zeta$, and the scaled inner product between g and e_g has magnitude at most $Q\zeta$.

C.2 Support preservation

Assume at the current iterate that $\mathbf{S} \in \mathcal{M}_\Omega$, $\mathrm{rank}(\mathbf{B}) = r_\star$, $d \le R/2$, and $\Psi \le a_0 R^2/4$. Since $\zeta \le \mathsf{L}R$, $v \le \zeta/(2\sqrt{r_\star}) \le \mathsf{L}R/2$. Thus

$$\|\mathbf{G}_n\|_{\mathrm{op}} \le \|\mathbf{G}\|_F + v \le 2\mathsf{L}R, \qquad Q_n \le 4\mathsf{L}d + \zeta \le 3\mathsf{L}R. \tag{32}$$

For the full sparse trial $\mathbf{S} - \eta\mathbf{G}_n$, an off-support entry has magnitude at most $2\eta\mathsf{L}R \le R/36$. Every true off-diagonal entry has magnitude at least

$$\beta_{\min} - R/2 - 2\eta\mathsf{L}R \ge \beta_{\min} - 19R/36 > R/36,$$

and has the correct sign. There are exactly $s_\star$ such entries. Hence hard thresholding retains precisely these pairs, and

$$\mathbf{S}^+ = \mathbf{S} - \eta\mathcal{P}_\Omega \mathbf{G}_n. \tag{33}$$

The same conclusion holds for every smaller trial step. If $s_\star = 0$, thresholding simply keeps the diagonal and the conclusion holds without a gap argument.

C.3 Rank and feasibility

Set $\mathbf{D}_S = -\mathcal{P}_\Omega \mathbf{G}_n$, $\mathbf{D}_B = 2\mathbf{G}_n\mathbf{B}\mathbf{H}^{-1}$, and $\mathbf{W} = \mathbf{D}_B\mathbf{D}_B^\top \succeq 0$. Then

$$\|\mathbf{D}_S\|_F^2 + \left\|\mathbf{D}_B\mathbf{H}^{1/2}\right\|_F^2 = Q_n^2, \tag{34}$$

$$\left\|\mathbf{D}_B\mathbf{B}^\top\right\|_F = \left\|\mathbf{D}_B\mathbf{H}^{1/2}\right\|_F,$$

$$\mathrm{tr}(\mathbf{W}) = \|\mathbf{D}_B\|_F^2 \le 2Q_n^2/\sigma_\star. \tag{35}$$

For $0 \le \alpha \le \eta$, define $\mathbf{S}(\alpha) = \mathbf{S} + \alpha\mathbf{D}_S$ and $\mathbf{B}(\alpha) = \mathbf{B} + \alpha\mathbf{D}_B$. First, $\mathbf{B}(\alpha)$ has full column rank: by (32),

$$\alpha Q_n \le R/24, \qquad \frac{\alpha\,\|\mathbf{D}_B\|_{\mathrm{op}}}{\sigma_{\min}(\mathbf{B})} \le \frac{2\alpha Q_n}{\sigma_\star} \le \frac{R}{12\sigma_\star} \le \frac{1}{96} < 1.$$

Here $\sigma_{\min}(\mathbf{B})^2 = \lambda_{\min}(\mathbf{H}) \ge \sigma_\star/2$.

Writing $\mathbf{L}(\alpha) = \mathbf{B}(\alpha)\mathbf{B}(\alpha)^\top$, the factor expansion is exact:

$$\mathbf{L}(\alpha) - \mathbf{L} = \alpha(\mathbf{D}_B\mathbf{B}^\top + \mathbf{B}\mathbf{D}_B^\top) + \alpha^2\mathbf{W}. \tag{36}$$

The product-space change in $(\mathbf{S}, \mathbf{L})$, and also the precision change, are each bounded by

$$3\alpha Q_n + \frac{2\alpha^2 Q_n^2}{\sigma_\star} \le 4\alpha Q_n. \tag{37}$$

To justify the factor 3, if $x = \|\mathbf{D}_S\|_F$ and $y = \left\|\mathbf{D}_B\mathbf{H}^{1/2}\right\|_F$, then the precision linear increment is at most $x + 2y \le \sqrt{5}\sqrt{x^2+y^2} \le 3Q_n$; the component-pair bound is no larger. The quadratic term uses (35); $2\alpha Q_n/\sigma_\star \le 1/96$ absorbs it into the extra αQ_n. It follows, before applying any descent inequality, that

$$d(\mathbf{S}(\alpha), \mathbf{B}(\alpha)) \le R/2 + 4\eta Q_n \le 2R/3 < R.$$

Consequently the whole trial curve lies in the matrix neighborhood, has precision at least $3m\mathbf{I}/4$, and has a full-rank factor. The straight precision segment between its endpoints also lies in the spectral band by convexity. Thus the likelihood Taylor bound is applicable without a circular basin assumption.

C.4 Full-step acceptance

Define $\mathbf{V} = \mathbf{D}_S - \mathbf{D}_B\mathbf{B}^\top - \mathbf{B}\mathbf{D}_B^\top$, so $\mathbf{\Theta}^+ - \mathbf{\Theta} = \eta\mathbf{V} - \eta^2\mathbf{W}$. Direct substitution gives

$$\langle \mathbf{G}_n, \mathbf{V}\rangle = -\|\mathcal{P}_\Omega \mathbf{G}_n\|_F^2 - 4\,\mathrm{tr}(\mathbf{G}_n^2\mathbf{P}) = -Q_n^2.$$

The linearity of $\ell_n - \ell$ and (21) yield

$$\begin{aligned} f_n(\mathbf{S}^+, \mathbf{B}^+) - f_n(\mathbf{S}, \mathbf{B}) &\le -\eta Q_n^2 + \eta^2 \|\mathbf{G}_n\|_{\mathrm{op}}\,\mathrm{tr}(\mathbf{W}) + \frac{\mathsf{L}}{2}\left\|\mathbf{\Theta}^+ - \mathbf{\Theta}\right\|_F^2 \\ &\le -\eta Q_n^2 + \eta^2\left(\frac{2\|\mathbf{G}_n\|_{\mathrm{op}}}{\sigma_\star} + 8\mathsf{L}\right)Q_n^2 \\ &\le -\eta Q_n^2 + 9\mathsf{L}\eta^2 Q_n^2 \le -\frac{7}{8}\eta Q_n^2. \end{aligned}$$

The penultimate bound uses $\|\mathbf{G}_n\|_{\mathrm{op}} \le 2\mathsf{L}R$ and $R \le \sigma_\star/8$. The final inequality uses $\eta \le 1/(72\mathsf{L})$. Since the decrement requested in (6) is $\eta Q_n^2/4$, the full trial passes the descent test. Rank and feasibility were already verified. Therefore this is the accepted algorithmic step, not an auxiliary projected iteration.

C.5 Contraction and induction

The same Taylor expansion for ℓ, now retaining the empirical perturbation in the direction, gives

$$\langle \mathbf{G}, \mathbf{V}\rangle \le -Q^2 + Q\zeta.$$

Indeed, the exact linear term is the negative inner product of g with $g + e_g$ in the scaled dual metric. The other terms are bounded by $9\mathsf{L}\eta^2 Q_n^2$ as before, since $\|\mathbf{G}\|_{\mathrm{op}} \le \mathsf{L}R \le 2\mathsf{L}R$. Therefore

$$\begin{aligned} \Psi^+ &\le \Psi - \eta Q^2 + \eta Q\zeta + 9\mathsf{L}\eta^2 Q_n^2 \\ &\le \Psi - \tfrac{3}{4}\eta Q^2 + \eta Q\zeta + \tfrac{1}{4}\eta\zeta^2 \\ &\le \Psi - \tfrac{1}{2}\eta Q^2 + \tfrac{5}{4}\eta\zeta^2 \\ &\le (1 - \eta\omega)\Psi + 2\eta\zeta^2. \end{aligned}$$

For the second line use $Q_n^2 \le 2Q^2 + 2\zeta^2$ and $9\mathsf{L}\eta \le 1/8$; for the third use $Q\zeta \le Q^2/4 + \zeta^2$; for the last use Lemma 9. The noise restriction ensures

$$\frac{2\zeta^2}{\omega} \le a_0 R^2/4.$$

It follows that $\Psi^+ \le a_0 R^2/4$. The endpoint was already shown to be in the radius-R neighborhood. Its lower potential bound therefore implies $d^+ \le R/2$. This closes the induction on d, Ψ, support, rank, and feasibility, starting from (9). Signed support follows either from the threshold gap or from $\|\mathbf{S}_t - \mathbf{S}^\star\|_F \le R/2 < \beta_{\min}$.

Since $\mu \le \mathsf{L}$ and $\gamma_\star \le 1$, $0 < \eta\omega \le 1/9216 < 1$. Unrolling the scalar recursion gives

$$\Psi_t \le \rho^t \Psi_0 + 2\eta\zeta^2 \sum_{j=0}^{t-1} \rho^j = \rho^t \Psi_0 + \frac{2}{\omega}(1-\rho^t)\zeta^2.$$

Using $a_0 d_t^2 \le \Psi_t$ and $\Psi_0 \le \mathsf{L} d_0^2$ proves Theorem 1. Finally, substituting $a_0 = \mu\gamma_\star/8$ and $\eta = 1/(72\mathsf{L})$ gives (20) exactly. This also audits every possible source of latent spectral dependence: $\sigma_\star^{-1}$ occurs only in the chord remainder, the factor-step remainder, and the rank check, where it is absorbed by $R/\sigma_\star$. No such term remains in η, ω, or the matrix-error multipliers.

D Proof of backtracking termination

We prove only termination of a trial search here, not a global convergence guarantee. Suppose the current $\mathbf{S}$ is symmetric with at most s off-diagonal nonzeros, $\mathbf{B}$ has full column rank, and $\mathbf{\Theta} \succ \underline{\theta}\mathbf{I}$. For trial multiplier γ, set $a_S = \gamma\eta_S$, $a_B = \gamma\eta_B$, $\Delta\mathbf{S} = \mathcal{H}_s(\mathbf{S} - a_S\mathbf{G}_n) - \mathbf{S}$, and $\Delta\mathbf{B} = -a_B \nabla_{\mathbf{B}} f_n \mathbf{H}^{-1}$. Hard thresholding by symmetric pairs is a Frobenius nearest projection onto the sparsity set. Comparing with the feasible current $\mathbf{S}$ yields

$$\|\Delta\mathbf{S} + a_S\mathbf{G}_n\|_F^2 \le a_S^2 \|\mathbf{G}_n\|_F^2, \quad \langle \mathbf{G}_n, \Delta\mathbf{S}\rangle \le -\frac{\|\Delta\mathbf{S}\|_F^2}{2a_S}, \quad \|\Delta\mathbf{S}\|_F \le 2a_S \|\mathbf{G}_n\|_F.$$

Also

$$\langle \nabla_{\mathbf{B}} f_n, \Delta\mathbf{B}\rangle = -\frac{\left\|\Delta\mathbf{B}\mathbf{H}^{1/2}\right\|_F^2}{a_B}.$$

Thus both increments tend to zero with α. Full column rank and strict feasibility are open conditions, so all sufficiently short trial segments retain both. On such a fixed neighborhood the factor-coordinate Hessian of f_n is bounded by some finite L_{loc}. With $W = \|\Delta\mathbf{S}\|_F^2 + \left\|\Delta\mathbf{B}\mathbf{H}^{1/2}\right\|_F^2$ and $c_H = \min\{1, \lambda_{\min}(\mathbf{H})\} > 0$,

$$f_n(\mathbf{S}+\Delta\mathbf{S}, \mathbf{B}+\Delta\mathbf{B}) - f_n(\mathbf{S},\mathbf{B}) \le -\frac{1}{2}W_\gamma + CW_\gamma^2, \qquad W_\gamma = \frac{\|\Delta\mathbf{S}\|_F^2}{a_S} + \frac{\left\|\Delta\mathbf{B}\mathbf{H}^{1/2}\right\|_F^2}{a_B},$$

for some finite local constant C (with fixed positive η_S, η_B and fixed current $\mathbf{H} \succ 0$). The increments are $O(\gamma)$, while $W_\gamma = O(\gamma)$; hence, for sufficiently small positive γ, the Taylor remainder is at most $W_\gamma/4$. This gives the required decrease $-W_\gamma/4$, exactly the test in (6). Full column rank and strict feasibility are open conditions, so sufficiently short trials retain both. Since the backtracking multipliers tend to zero, a trial is accepted after finitely many reductions. A current point exactly at the numerical margin is not covered by this auxiliary statement; Theorem 1 keeps all iterates strictly above that margin.

E Proof of Theorem 2

Proof of Theorem 2. For $(i,j) \notin \Omega$, $i \ne j$, $\mathbf{\Theta}^\star_{ij} = -L^\star_{ij}$, so $|P_{ij}| \le \ell_\infty + \delta_\infty = q$. Such entries are removed by the strict threshold. For $(i,j) \in \Omega$,

$$|P_{ij} - S^\star_{ij}| \le q, \qquad |P_{ij}| \ge \beta_{\min} - q > q,$$

so every true edge is retained with the correct sign. The diagonal is retained by definition. Thus $s_0 = s_\star$, $\mathbf{S}_0 \in \mathcal{M}_\Omega$, and

$$\|\mathbf{S}_0 - \mathbf{S}^\star\|_F \le \sqrt{s_{\mathrm{tot}}}\, q = h.$$

The raw low-rank pilot satisfies

$$\mathbf{Y}_0 - \mathbf{L}^\star = (\mathbf{S}_0 - \mathbf{S}^\star) - (\mathbf{P} - \mathbf{\Theta}^\star), \qquad \|\mathbf{Y}_0 - \mathbf{L}^\star\|_{\mathrm{op}} \le \Delta.$$

Weyl's inequality gives $\nu_{r_\star} \geq \sigma_\star - \Delta \geq \sigma_\star/2 > 0$ and $|\nu_j| \leq \Delta$ for $j > r_\star$. Therefore the top-$r_\star$ PSD truncation $\mathbf{L}_0 = \mathbf{B}_0\mathbf{B}_0^\top$ has rank $r_\star$, and

$$\|\mathbf{L}_0 - \mathbf{Y}_0\|_{\mathrm{op}} \leq \Delta, \qquad \|\mathbf{L}_0 - \mathbf{L}^\star\|_{\mathrm{op}} \leq 2\Delta.$$

The difference has rank at most $2r_\star$, hence $\|\mathbf{L}_0 - \mathbf{L}^\star\|_F \leq 2\sqrt{2r_\star}\Delta$ and $d_0 \leq d_{\mathrm{sp}}$. Since $a_0/\mathsf{L} \leq 1$, condition (13) gives $d_0 \leq R/2$. Lemma 7 then yields $\Psi_0 \leq \mathsf{L}d_0^2 \leq a_0R^2/4$ and $\mathbf{\Theta}_0 \succeq 3m\mathbf{I}/4 \succ 2\underline{\theta}\mathbf{I}$. All initialization checks and hypotheses of Theorem 1 have now been verified. □

Sample-inverse pilot. On the score event, $v_\delta \leq 1/(2M)$ implies $\widehat{\mathbf{\Sigma}} \succeq (1/M - v_\delta)\mathbf{I} \succeq \mathbf{I}/(2M)$. The resolvent identity gives

$$\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Theta}^\star = -\widehat{\mathbf{\Sigma}}^{-1}\mathbf{E}\mathbf{\Theta}^\star, \qquad \left\|\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Theta}^\star\right\|_{\mathrm{op}} \leq 2M^2 v_\delta.$$

Every entry is bounded by the operator norm. Thus the deterministic bounds used in the main text hold on the same event, without independence or sample splitting. Substituting these bounds and an available latent spikiness bound into (13) is still required: increasing n alone need not make ℓ_∞ small enough. The pilot is an explicit sufficient construction, not a general high-dimensional initialization theorem.

F Proofs of the statistical results

Proof of Lemma 3. If $Z \sim \mathcal{N}(0,1)$, direct integration and a power-series expansion give

$$\log \mathbb{E}e^{t(Z^2-1)} = -t - \tfrac{1}{2}\log(1-2t) \leq \frac{t^2}{1-2t}, \quad 0 < t < 1/2,$$

and $\log \mathbb{E}e^{-t(Z^2-1)} \leq t^2$ for $t > 0$. Applying Chernoff's inequality to independent copies, with $t = \sqrt{x}/(\sqrt{n} + 2\sqrt{x})$ for the upper tail and $t = \sqrt{x/n}$ for the lower tail, yields

$$\Pr\left\{\left|n^{-1}\sum_{i=1}^n Z_i^2 - 1\right| > 2\sqrt{x/n} + 2x/n\right\} \leq 2e^{-x}. \tag{38}$$

For a fixed vector $\mathbf{v}$, divide $\mathbf{v}^\top\mathbf{x}_i$ by its population standard deviation. Thus (38) bounds $|\mathbf{v}^\top\mathbf{E}\mathbf{v}|$ by $2\|\mathbf{v}\|_2^2 m^{-1}(\sqrt{x/n} + x/n)$ except with probability $2e^{-x}$.

Apply this bound with $x = h_\delta$ to the p^2 vectors consisting of $\mathbf{e}_i$ and $\mathbf{e}_i \pm \mathbf{e}_j$ for $i < j$. The union failure probability is $2p^2e^{-h_\delta} = \delta/2$. Polarization,

$$E_{ij} = \tfrac{1}{4}\left[(\mathbf{e}_i + \mathbf{e}_j)^\top\mathbf{E}(\mathbf{e}_i + \mathbf{e}_j) - (\mathbf{e}_i - \mathbf{e}_j)^\top\mathbf{E}(\mathbf{e}_i - \mathbf{e}_j)\right],$$

proves $u \leq (2/m)(\sqrt{h_\delta/n} + h_\delta/n)$, including the diagonal case.

For the spectral norm, take a maximal $1/4$-separated subset $\mathcal{N}$ of the unit sphere. Balls of radius $1/8$ around its points are disjoint and lie in a ball of radius $9/8$, so $|\mathcal{N}| \leq 9^p$. Maximality gives a $1/4$-net. If $\mathbf{v}$ is unit length and $\mathbf{w} \in \mathcal{N}$ is within $1/4$, then

$$|\mathbf{v}^\top\mathbf{E}\mathbf{v} - \mathbf{w}^\top\mathbf{E}\mathbf{w}| \leq \tfrac{1}{2}\|\mathbf{E}\|_{\mathrm{op}}.$$

Taking the largest absolute Rayleigh quotient shows $\|\mathbf{E}\|_{\mathrm{op}} \leq 2\max_{\mathbf{w}\in\mathcal{N}}|\mathbf{w}^\top\mathbf{E}\mathbf{w}|$. Use (38) with $x = q_\delta$ at every net point. The union failure probability is at most $2 \cdot 9^p e^{-q_\delta} = \delta/2$, and the resulting bound is exactly $v \leq v_\delta$. Combining the entrywise and spectral events proves the lemma. □

Proof of Corollary 4. The entire trajectory theorem is deterministic on the score event. Taking square roots in (11) gives

$$d_t \leq \sqrt{\mathsf{L}/a_0}\,\rho^{t/2}d_0 + \sqrt{2/(\omega a_0)}\,\zeta_\delta.$$

When $n \geq \max\{h_\delta, q_\delta\}$, $x/n \leq \sqrt{x/n}$ for either logarithmic factor. Hence

$$\zeta_\delta \leq \frac{4}{m}\sqrt{\frac{s_{\mathrm{tot}} h_\delta}{n}} + \frac{16}{m}\sqrt{\frac{r_\star q_\delta}{n}}.$$

This proves (15), with all coefficients depending only on $(m, M, \gamma_\star)$. For fixed $c > 0$, $\delta = p^{-c}$ gives $h_\delta = O_c(\log p)$ and $q_\delta = O_c(p)$ for $p \geq 2$. The constant in this simplified order may additionally depend on c. No union bound over iterations is needed. The required sample size must make the score bound small relative to R and, if used, meet the pilot conditions; these are distinct from the latent-independent error multiplier. □

Proof of Corollary 5. The precision bound follows from $\|\mathbf{S}_t - \mathbf{S}^\star - (\mathbf{L}_t - \mathbf{L}^\star)\|_F \leq \sqrt{2} d_t$. The inverse identity

$$\mathbf{\Theta}_t^{-1} - (\mathbf{\Theta}^\star)^{-1} = -\mathbf{\Theta}_t^{-1}(\mathbf{\Theta}_t - \mathbf{\Theta}^\star)(\mathbf{\Theta}^\star)^{-1}$$

and $\mathbf{\Theta}_t \succeq 3m\mathbf{I}/4$, $\mathbf{\Theta}^\star \succeq m\mathbf{I}$ imply a multiplier $4/(3m^2) \leq 2/m^2$. Signed support was established in the induction. This conclusion relies on support identification at initialization and does not imply graph recovery from an arbitrary sparse starting point. □

G SUPPLEMENTARY SIMULATION RESULTS

Conditioning under the true covariance. At iteration 2000, the mean normalized joint errors range from 3.3×10^{-11} to 6.4×10^{-11} for ScaledGD and from 3.2×10^{-10} to 7.3×10^{-3} for GD across the five conditions. To maintain numerical accuracy near the optimum, the backtracking test evaluates the likelihood difference as $\langle \mathbf{G}_t, \Delta\rangle + \sum_j [\mu_j - \log(1+\mu_j)]$, where Δ is the candidate precision change and μ_j are the eigenvalues of $\mathbf{C}_t^{-1}\Delta\mathbf{C}_t^{-\top}$ for $\mathbf{C}_t\mathbf{C}_t^\top = \mathbf{\Theta}_t$. This is algebraically equivalent to subtracting the two likelihood values, but avoids cancellation when they are nearly equal; the update and sufficient-decrease condition are unchanged.

Comparator implementation. PPA solves the convex LVGGM estimator (Chandrasekaran et al., 2010; 2012a)

$$\min_{\mathbf{\Theta}=\mathbf{S}-\mathbf{L}\succ 0,\ \mathbf{L}\succeq 0} -\log\det\mathbf{\Theta} + \langle \widehat{\mathbf{\Sigma}}, \mathbf{\Theta}\rangle + \lambda_S \|\mathbf{S}\|_{1,\mathrm{off}} + \lambda_L \operatorname{tr}(\mathbf{L}), \tag{39}$$

where $\|\mathbf{S}\|_{1,\mathrm{off}} = \sum_{i\neq j}|S_{ij}|$ leaves the diagonal unpenalized, and the regularization parameters are selected by validation likelihood. We use a proximal-point implementation following LogdetPPA (Wang et al., 2010); reported times are from this implementation. Graphical lasso estimates only $\mathbf{\Theta}$; its sparse-component error in Table 1 uses $\widehat{\mathbf{S}} = \widehat{\mathbf{\Theta}}$. The latent-error entry is omitted because the method does not fit a latent component.

Support recovery of the spectral initializer. We evaluated Algorithm 1 at $n \in \{800, 1600, 3200, 6400\}$ and edge strengths $\omega_e \in \{0.02, 0.05, 0.1, 0.2, 0.4\}$, using $\mathbf{P} = \widehat{\mathbf{\Sigma}}^{-1}$ and selecting the threshold q by validation likelihood. Figure 3 reports the resulting sparse-support F1 over 20 replications per setting. At $n = 6400$, mean F1 increases from 0.012 for $\omega_e = 0.02$ to 0.999 for $\omega_e = 0.4$; at $\omega_e = 0.05$, it increases from 0 at $n = 800$ to 0.805 at $n = 6400$. In these settings, support recovery improves with both edge strength and sample size.

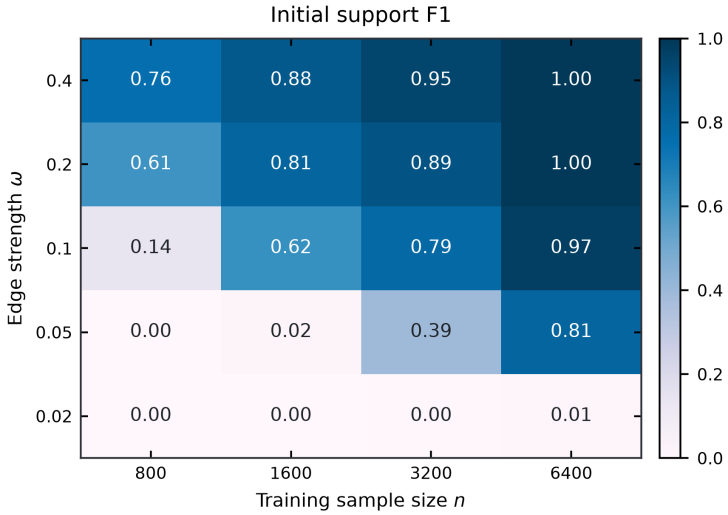


Figure 3: Sparse-support F1 of the spectral initializer versus sample size and edge strength. The threshold is selected by validation likelihood; entries are means over 20 replications.