Submitted to the Annals of Statistics
arXiv: math.ST/0000000

LATENT VARIABLE GRAPHICAL MODEL SELECTION VIA CONVEX OPTIMIZATION – SUPPLEMENTARY MATERIAL

By Venkat Chandrasekaran *, Pablo A. Parrilo , and Alan S. Willsky

Massachusetts Institute of Technology

1. Matrix perturbation bounds. Given a low-rank matrix we consider what happens to the invariant subspaces when the matrix is perturbed by a small amount. We assume without loss of generality that the matrix under consideration is square and symmetric, and our methods can be extended to the general non-symmetric non-square case. We refer the interested reader to [1, 3] for more details, as the results presented here are only a brief summary of what is relevant for this paper. In particular the arguments presented here are along the lines of those presented in [1]. The appendices in [1] also provide a more refined analysis of second-order perturbation errors.

The resolvent of a matrix M is given by $(M - \zeta I)^{-1}$ [3], and it is well-defined for all $\zeta \in \mathbb{C}$ that do not coincide with an eigenvalue of M. If M has no eigenvalue with magnitude equal to η, then we have by the Cauchy residue formula that the projector onto the invariant subspace of a matrix M corresponding to all singular values smaller than η is given by

$$P_{M,\eta} = \frac{-1}{2\pi i} \oint_{\mathcal{C}_\eta} (M - \zeta I)^{-1} d\zeta, \tag{1.1}$$

where $\mathcal{C}_\eta$ denotes the positively-oriented circle of radius η centered at the origin. Similarly, we have that the weighted projection onto the invariant subspace corresponding to the smallest singular values is given by

$$P^w_{M,\eta} = M P_{M,\eta} = \frac{-1}{2\pi i} \oint_{\mathcal{C}_\eta} \zeta \ (M - \zeta I)^{-1} d\zeta, \tag{1.2}$$

Suppose that M is a low-rank matrix with smallest nonzero singular value σ, and let Δ be a perturbation of M such that $\|\Delta\|_2 \leq \kappa < \frac{\sigma}{2}$. We have the following identity for any $|\zeta| = \kappa$, which will be used repeatedly:

$$[(M + \Delta) - \zeta I]^{-1} - [M - \zeta I]^{-1} = -[M - \zeta I]^{-1} \Delta [(M + \Delta) - \zeta I]^{-1}. \tag{1.3}$$

*Corresponding author.

We then have that

$$\begin{aligned} P_{M+\Delta,\kappa} - P_{M,\kappa} &= \frac{-1}{2\pi i} \oint_{\mathcal{C}_\kappa} [(M+\Delta) - \zeta I]^{-1} - [M - \zeta I]^{-1} d\zeta \\ &= \frac{1}{2\pi i} \oint_{\mathcal{C}_\kappa} [M - \zeta I]^{-1} \Delta [(M+\Delta) - \zeta I]^{-1} d\zeta. \end{aligned} \tag{1.4}$$

Similarly, we have the following for $P^w_{M,\kappa}$:

$$\begin{aligned} P^w_{M+\Delta,\kappa} - P^w_{M,\kappa} &= \frac{-1}{2\pi i} \oint_{\mathcal{C}_\kappa} \zeta \ \left\{ [(M+\Delta) - \zeta I]^{-1} - [M - \zeta I]^{-1} \right\} d\zeta \\ &= \frac{1}{2\pi i} \oint_{\mathcal{C}_\kappa} \zeta \ \left\{ [M - \zeta I]^{-1} \Delta [(M+\Delta) - \zeta I]^{-1} \right\} d\zeta \\ &= \frac{1}{2\pi i} \oint_{\mathcal{C}_\kappa} \zeta \ [M - \zeta I]^{-1} \Delta [M - \zeta I]^{-1} d\zeta \\ &\quad - \frac{1}{2\pi i} \oint_{\mathcal{C}_\kappa} \zeta \ [M - \zeta I]^{-1} \Delta [M - \zeta I]^{-1} \Delta [(M+\Delta) - \zeta I]^{-1} d\zeta. \end{aligned} \tag{1.5}$$

Given these expressions, we have the following two results.

PROPOSITION 1.1. *Let $M \in \mathbb{R}^{p \times p}$ be a rank-r matrix with smallest nonzero singular value equal to σ, and let Δ be a perturbation to M such that $\|\Delta\|_2 \leq \frac{\kappa}{2}$ with $\kappa < \frac{\sigma}{2}$. Then we have that*

$$\|P_{M+\Delta,\kappa} - P_{M,\kappa}\|_2 \leq \frac{\kappa}{(\sigma - \kappa)(\sigma - \frac{3\kappa}{2})} \ \|\Delta\|_2.$$

Proof: This result follows directly from the expression (1.4), and the sub-multiplicative property of the spectral norm:

$$\begin{aligned} \|P_{M+\Delta,\kappa} - P_{M,\kappa}\|_2 &\leq \frac{1}{2\pi} \ 2\pi \ \kappa \ \frac{1}{\sigma - \kappa} \ \|\Delta\|_2 \ \frac{1}{\sigma - \frac{3\kappa}{2}} \\ &= \frac{\kappa}{(\sigma - \kappa)(\sigma - \frac{3\kappa}{2})} \ \|\Delta\|_2. \end{aligned}$$

Here, we used the fact that $\|[M - \zeta I]^{-1}\|_2 \leq \frac{1}{\sigma - \kappa}$ and $\|[(M+\Delta) - \zeta I]^{-1}\|_2 \leq \frac{1}{\sigma - \frac{3\kappa}{2}}$ for $|\zeta| = \kappa$. $\square$

Next, we develop a similar bound for $P^w_{M,\kappa}$. Let $U(M)$ denote the invariant subspace of M corresponding to the nonzero singular values, and let $P_{U(M)}$ denote the projector onto this subspace.

PROPOSITION 1.2. *Let $M \in \mathbb{R}^{p\times p}$ be a rank-r matrix with smallest nonzero singular value equal to σ, and let Δ be a perturbation to M such that $\|\Delta\|_2 \leq \frac{\kappa}{2}$ with $\kappa < \frac{\sigma}{2}$. Then we have that*

$$\|P^w_{M+\Delta,\kappa} - P^w_{M,\kappa} - (I - P_{U(M)})\Delta(I - P_{U(M)})\|_2 \leq \frac{\kappa^2}{(\sigma-\kappa)^2(\sigma - \frac{3\kappa}{2})}\ \|\Delta\|_2^2.$$

Proof: One can check that

$$\frac{1}{2\pi i}\oint_{\mathcal{C}_\kappa} \zeta\ [M - \zeta I]^{-1}\Delta[M - \zeta I]^{-1}d\zeta = (I - P_{U(M)})\Delta(I - P_{U(M)}).$$

Next we use the expression (1.5), and the sub-multiplicative property of the spectral norm:

$$\begin{aligned}\|P^w_{M+\Delta,\kappa} - P^w_{M,\kappa} &- (I - P_{U(M)})\Delta(I - P_{U(M)})\|_2 \\ &\leq \frac{1}{2\pi}\ 2\pi\ \kappa\ \kappa\ \frac{1}{\sigma-\kappa}\ \|\Delta\|_2\ \frac{1}{\sigma-\kappa}\ \|\Delta\|_2\ \frac{1}{\sigma - \frac{3\kappa}{2}} \\ &= \frac{\kappa^2}{(\sigma-\kappa)^2(\sigma - \frac{3\kappa}{2})}\ \|\Delta\|_2^2.\end{aligned}$$

As with the previous proof, we used the fact that $\|[M - \zeta I]^{-1}\|_2 \leq \frac{1}{\sigma-\kappa}$ and $\|[(M+\Delta) - \zeta I]^{-1}\|_2 \leq \frac{1}{\sigma - \frac{3\kappa}{2}}$ for $|\zeta| = \kappa$. □

We will use these expressions to derive bounds on the "twisting" between the tangent spaces at M and at $M + \Delta$ with respect to the rank variety.

2. Curvature of rank variety. For a symmetric rank-r matrix M, the projection onto the tangent space $T(M)$ (restricted to the variety of symmetric matrices with rank less than or equal to r) can be written in terms of the projection $P_{U(M)}$ onto the row space $U(M)$. For any matrix N

$$\mathcal{P}_{T(M)}(N) = P_{U(M)}N + NP_{U(M)} - P_{U(M)}NP_{U(M)}.$$

One can then check that the projection onto the normal space $T(M)^\perp$

$$\mathcal{P}_{T(M)^\perp}(N) = [I - \mathcal{P}_{T(M)}](N) = (I - P_{U(M)})\ N\ (I - P_{U(M)}).$$

PROPOSITION 2.1. *Let $M \in \mathbb{R}^{p\times p}$ be a rank-r matrix with smallest nonzero singular value equal to σ, and let Δ be a perturbation to M such that $\|\Delta\|_2 \leq \frac{\sigma}{8}$. Further, let $M + \Delta$ be a rank-r matrix. Then we have that*

$$\rho(T(M+\Delta), T(M)) \leq \frac{2}{\sigma}\ \|\Delta\|_2.$$

Proof: For any matrix N, we have that

$$[\mathcal{P}_{T(M+\Delta)} - \mathcal{P}_{T(M)}](N) = [P_{U(M+\Delta)} - P_{U(M)}]\ N\ [I - P_{U(M)}] + [I - P_{U(M+\Delta)}]\ N\ [P_{U(M+\Delta)} - P_{U(M)}].$$

Further, we note that for $\kappa < \frac{\sigma}{2}$

$$\begin{aligned} P_{U(M+\Delta)} - P_{U(M)} &= [I - P_{U(M)}] - [I - P_{U(M+\Delta)}] \\ &= P_{M,\kappa} - P_{M+\Delta,\kappa}, \end{aligned}$$

where $P_{M,\kappa}$ is defined in the previous section. Thus, we have the following sequence of inequalities for $\kappa = \frac{\sigma}{4}$:

$$\begin{aligned} \rho(T(M+\Delta), T(M)) &= \max_{\|N\|_2 \leq 1} \|[P_{U(M+\Delta)} - P_{U(M)}]\ N\ [I - P_{U(M)}] \\ &\qquad + [I - P_{U(M+\Delta)}]\ N\ [P_{U(M+\Delta)} - P_{U(M)}]\|_2 \\ &\leq \max_{\|N\|_2 \leq 1} \|[P_{U(M+\Delta)} - P_{U(M)}]\ N\ [I - P_{U(M)}]\|_2 \\ &\quad + \max_{\|N\|_2 \leq 1} \|[I - P_{U(M+\Delta)}]\ N\ [P_{U(M+\Delta)} - P_{U(M)}]\|_2 \\ &\leq 2\ \|P_{M+\Delta,\frac{\sigma}{4}} - P_{M,\frac{\sigma}{4}}\|_2 \\ &\leq \frac{2}{\sigma}\ \|\Delta\|_2, \end{aligned}$$

where we obtain the last inequality from Proposition 1.1. □

PROPOSITION 2.2. *Let $M \in \mathbb{R}^{p \times p}$ be a rank-r matrix with smallest nonzero singular value equal to σ, and let Δ be a perturbation to M such that $\|\Delta\| \leq \frac{\sigma}{8}$. Further, let $M + \Delta$ be a rank-r matrix. Then we have that*

$$\|\mathcal{P}_{T(M)^\perp}(\Delta)\|_2 \leq \frac{\|\Delta\|_2^2}{\sigma}.$$

Proof: Since both M and $M+\Delta$ are rank-r matrices, we have that $\mathcal{P}^w_{M+\Delta,\kappa} = \mathcal{P}^w_{M,\kappa} = 0$ for $\kappa = \frac{\sigma}{4}$. Consequently,

$$\begin{aligned} \|\mathcal{P}_{T(M)^\perp}(\Delta)\|_2 &= \|(I - P_{U(M)})\ \Delta\ (I - P_{U(M)})\|_2 \\ &\leq \frac{\|\Delta\|_2^2}{\sigma}, \end{aligned}$$

where we obtain the last inequality from Proposition 1.2 with $\kappa = \frac{\sigma}{4}$. □

3. Proof of supplementary results of main theorem. Throughout this section we denote $m = \max\{1, \frac{1}{\gamma}\}$. Further $\Omega = \Omega(K_O^*)$ and $T = T(K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*)$ denote the tangent spaces at the true sparse matrix $S^* = K_O^*$ and low-rank matrix $L^* = K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$. We assume that

$$\gamma \in \left[\frac{3\beta(2-\nu)\xi(T)}{\nu\alpha}, \frac{\nu\alpha}{2\beta(2-\nu)\mu(\Omega)}\right] \tag{3.1}$$

We also let $E_n = \Sigma_O^n - \Sigma_O^*$ denote the difference between the true marginal covariance and the sample covariance. Finally we let $D = \max\{1, \frac{\nu\alpha}{3\beta(2-\nu)}\}$ throughout this section. For γ in the above range we note that

$$m \leq \frac{D}{\xi(T)}. \tag{3.2}$$

Standard facts that we use throughout this section are that $\xi(T) \leq 1$ and that $\|M\|_\infty \leq \|M\|_2$ for any matrix M.

We study the following convex program:

$$\begin{aligned}(\bar{S}_n, \bar{L}_n) = \arg\min_{S,L} \ & \operatorname{tr}[(S-L)\,\Sigma_O^n] - \log\det(S-L) \ + \ \lambda_n[\gamma\|S\|_1 + \|L\|_*] \\ \text{s.t.} \quad & S - L \succ 0.\end{aligned} \tag{3.3}$$

Comparing (3.3) with the convex program (1.2) (main paper), the main difference is that we do not constrain the variable L to be positive semidefinite in (3.3) (recall that the nuclear norm of a positive semidefinite matrix is equal to its trace). However we show that the unique optimum $(\bar{S}_n, \bar{L}_n)$ of (3.3) under the hypotheses of Theorem 4.1 (main paper) is such that $\bar{L}_n \succeq 0$ (with high probability). Therefore we conclude that $(\bar{S}_n, \bar{L}_n)$ is also the unique optimum of (1.2) (main paper). The subdifferential with respect to the nuclear norm at a matrix M with (reduced) SVD given by $M = UDV^T$ is as follows:

$$N \in \partial\|M\|_* \quad \Leftrightarrow \quad \mathcal{P}_{T(M)}(N) = UV^T, \ \|\mathcal{P}_{T(M)^\perp}(N)\|_2 \leq 1.$$

The proof of this theorem consists of a number of steps, each of which is analyzed in separate sections below. We explicitly keep track of the constants α, β, ν, ψ. The key ideas are as follows:

1. We show that if we solve the convex program (3.3) subject to the additional constraints that $S \in \Omega$ and $L \in T'$ for some T' "close to" T (measured by $\rho(T', T)$), then the error between the optimal solution $(\bar{S}_n, \bar{L}_n)$ and the underlying matrices (S^*, L^*) is small. This result is discussed in Appendix 3.2.

2. We analyze the optimization problem (3.3) with the additional constraint that the variables S and L belong to the algebraic varieties of sparse and low-rank matrices respectively, and that the corresponding tangent spaces are close to the tangent spaces at (S^*, L^*). We show that under suitable conditions on the minimum nonzero singular value of the true low-rank matrix L^* and on the minimum magnitude nonzero entry of the true sparse matrix S^*, the optimum of this modified program is achieved at a *smooth* point of the underlying varieties. In particular the bound on the minimum nonzero singular value of L^* helps bound the curvature of the low-rank matrix variety locally around L^* (we use the results described in Appendix 2). These results are described in Appendix 3.3.
3. The next step is to show that the variety constraint can be linearized and changed to a tangent-space constraint (see Appendix 3.4), thus giving us a *convex program*. Under suitable conditions this tangent-space constrained program also has an optimum that has the same support/rank as the true (S^*, L^*). Based on the previous step these tangent spaces in the constraints are close to the tangent spaces at the true (S^*, L^*). Therefore we use the first step to conclude that the resulting error in the estimate is small.
4. Finally we show that under suitable identifiability conditions these tangent-space constraints are inactive at the optimum. Therefore we conclude with the statement that the optimum of the convex program (3.3) without any variety constraints is achieved at a pair of matrices that have the same support/rank as the true (S^*, L^*) (with high probability). Further the low-rank component of the solution is positive semidefinite, thus allowing us to conclude that the original convex program (1.2) (main paper) also provides estimates that are algebraically correct.

3.1. *Proof of main paper Proposition 5.1 – Bounded curvature of matrix inverse.* Consider the Taylor series of the inverse of a matrix:

$$(M + \Delta)^{-1} = M^{-1} - M^{-1}\Delta M^{-1} + R_{M^{-1}}(\Delta),$$

where

$$R_{M^{-1}}(\Delta) = M^{-1}\left[\sum_{k=2}^{\infty}(-\Delta M^{-1})^k\right].$$

This infinite sum converges for Δ sufficiently small. The following proposition provides a bound on the second-order term specialized to our setting:

PROPOSITION 3.1. *Suppose that γ is in the range given by (3.1). Let $g_\gamma(\Delta_S, \Delta_L) \leq \frac{1}{2C_1}$ for $C_1 = \psi(1 + \frac{\alpha}{6\beta})$, and for any (Δ_S, Δ_L) with $\Delta_S \in \Omega$. Then we have that*

$$g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L))) \leq \frac{2D\psi C_1^2 g_\gamma(\Delta_S, \Delta_L)^2}{\xi(T)}.$$

Proof: We have that

$$\begin{aligned}
\|\mathcal{A}(\Delta_S, \Delta_L)\|_2 &\leq \|\Delta_S\|_2 + \|\Delta_L\|_2 \\
&\leq \gamma\mu(\Omega)\frac{\|\Delta_S\|_\infty}{\gamma} + \|\Delta_L\|_2 \\
&\leq (1 + \gamma\mu(\Omega))g_\gamma(\Delta_S, \Delta_L) \\
&\leq (1 + \frac{\alpha}{6\beta})g_\gamma(\Delta_S, \Delta_L) \\
&\leq \frac{1}{2\psi},
\end{aligned}$$

where the second-to-last inequality follows from the range for γ (3.1) and that $\nu \in (0, \frac{1}{2}]$, and the final inequality follows from the bound on $g_\gamma(\Delta_S, \Delta_L)$. Therefore,

$$\begin{aligned}
\|R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L))\|_2 &\leq \psi\sum_{k=2}^{\infty}(\|\Delta_S + \Delta_L\|_2\psi)^k \\
&\leq \psi^3\|\Delta_S + \Delta_L\|_2^2\frac{1}{1 - \|\Delta_S + \Delta_L\|_2\psi} \\
&\leq 2\psi^3(1 + \frac{\alpha}{6\beta})^2 g_\gamma(\Delta_S, \Delta_L)^2 \\
&= 2\psi C_1^2 g_\gamma(\Delta_S, \Delta_L)^2.
\end{aligned}$$

Here we apply the last two inequalities from above. Since the $\|\cdot\|_\infty$-norm is bounded above by the spectral norm $\|\cdot\|_2$, we have the desired result. □

3.2. *Proof of main paper Proposition 5.2 – Bounded errors.* Next we analyze the following convex program subject to certain additional tangent-space constraints:

(3.4)

$$\begin{aligned}
(\hat{S}_\Omega, \hat{L}_{T'}) = \arg\min_{S,L} \operatorname{tr}[(S - L)\ \Sigma_O^n] - \log\det(S - L)\ +\ \lambda_n[\gamma\|S\|_1 + \|L\|_*] \\
\text{s.t.} \quad S - L \succ 0, \quad S \in \Omega, \quad L \in T',
\end{aligned}$$

for some subspace T'. We show that if T' is any tangent space to the low-rank matrix variety such that $\rho(T, T') \leq \frac{\xi(T)}{2}$, then we can bound the error

$(\Delta_S, \Delta_L) = (\hat{S}_\Omega - S^*, L^* - \hat{L}_{T'})$. Let $\mathcal{C}_{T'} = \mathcal{P}_{T'^\perp}(L^*)$ denote the normal component of the true low-rank matrix at T', and recall that $E_n = \Sigma_O^n - \Sigma_O^*$ denotes the difference between the true marginal covariance and the sample covariance. The proof of the following result uses Brouwer's fixed-point theorem [4], and is inspired by the proof of a similar result in [5] for standard sparse graphical model recovery without latent variables.

PROPOSITION 3.2. *Let the error (Δ_S, Δ_L) in the solution of the convex program (3.4) (with T' such that $\rho(T', T) \leq \frac{\xi(T)}{2}$) be as defined above. Further let $C_1 = \psi(1 + \frac{\alpha}{6\beta})$, and define*

$$r = \max\left\{\frac{8}{\alpha}\left[g_\gamma(\mathcal{A}^\dagger E_n) + g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{C}_{T'}) + \lambda_n\right], \|\mathcal{C}_{T'}\|_2\right\}.$$

If we have that

$$r \leq \min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\right\},$$

for γ in the range given by (3.1), then

$$g_\gamma(\Delta_S, \Delta_L) \leq 2r.$$

Proof: Based on Proposition 3.3 (main paper) we note that the convex program (3.4) is strictly convex (because the negative log-likelihood term has a strictly positive-definite Hessian due to the constraints involving transverse tangent spaces), and therefore the optimum is unique. Applying the optimality conditions of the convex program (3.4) at the optimum $(\hat{S}_\Omega, \hat{L}_{T'})$, we have that there exist Lagrange multipliers $Q_{\Omega^\perp} \in \Omega^\perp$, $Q_{T'^\perp} \in T'^\perp$ such that

$$\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1} + Q_{\Omega^\perp} \in -\lambda_n\gamma\partial\|\hat{S}_\Omega\|_1, \;\; \Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1} + Q_{T'^\perp} \in \lambda_n\partial\|\hat{L}_{T'}\|_*.$$

Restricting these conditions to the space $\mathcal{Y} = \Omega \times T'$, one can check that

$$\mathcal{P}_\Omega[\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1}] = Z_\Omega, \;\; \mathcal{P}_{T'}[\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1}] = Z_{T'},$$

where $Z_\Omega \in \Omega, Z_{T'} \in T'$ and $\|Z_\Omega\|_\infty = \lambda_n\gamma, \|Z_{T'}\|_2 \leq 2\lambda_n$ (we use here the fact that projecting onto a tangent space T' increases the spectral norm by at most a factor of two). Denoting $Z = [Z_\Omega, Z_{T'}]$, we conclude that

$$\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger[\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1}] = Z, \tag{3.5}$$

with $g_\gamma(Z) \leq 2\lambda_n$. Since the optimum $(\hat{S}_\Omega, \hat{L}_{T'})$ is unique, one can check using Lagrangian duality theory [6] that $(\hat{S}_\Omega, \hat{L}_{T'})$ is the unique solution

of the equation (3.5). Rewriting $\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1}$ in terms of the errors (Δ_S, Δ_L), we have using the Taylor series of the matrix inverse that

$$\begin{aligned}\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T'})^{-1} &= \Sigma_O^n - [\mathcal{A}(\Delta_S, \Delta_L) + (\Sigma_O^*)^{-1}]^{-1} \\ &= E_n - R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)) + \mathcal{I}^*\mathcal{A}(\Delta_S, \Delta_L) \\ &= E_n - R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)) + \mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L) + \mathcal{I}^*\mathcal{C}_{T'}. \quad (3.6)\end{aligned}$$

Since T' is a tangent space such that $\rho(T', T) \leq \frac{\xi(T)}{2}$, we have from Proposition 3.3 (main paper) that the operator $\mathcal{B} = \left(\mathcal{P}_{\mathcal{Y}}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}\right)^{-1}$ from $\mathcal{Y}$ to $\mathcal{Y}$ is bijective and is well-defined. Now consider the following matrix-valued function from $(\delta_S, \delta_L) \in \mathcal{Y}$ to $\mathcal{Y}$:

$$F(\delta_S, \delta_L) = (\delta_S, \delta_L) - \mathcal{B}\left\{\mathcal{P}_{\mathcal{Y}}\mathcal{A}^\dagger[E_n - R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'})) + \mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\delta_S, \delta_L) + \mathcal{I}^*\mathcal{C}_{T'}] - Z\right\}.$$

A point $(\delta_S, \delta_L) \in \mathcal{Y}$ is a fixed-point of F if and only if $\mathcal{P}_{\mathcal{Y}}\mathcal{A}^\dagger[E_n - R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'})) + \mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\delta_S, \delta_L) + \mathcal{I}^*\mathcal{C}_{T'}] = Z$. Applying equations (3.5) and (3.6) above, we then see that the only fixed-point of F by construction is the "true" error $\mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)$ restricted to $\mathcal{Y}$. The reason for this is that, as discussed above, $(\hat{S}_\Omega, \hat{L}_{T'})$ is the unique optimum of (3.4) and therefore is the *unique solution* of (3.5). Next we show that this unique fixed-point of F lies in the ball $\mathbb{B}_r = \{(\delta_S, \delta_L) \mid g_\gamma(\delta_S, \delta_L) \leq r, \quad (\delta_S, \delta_L) \in \mathcal{Y}\}$.

In order to prove this step, we resort to Brouwer's fixed point theorem [4]. In particular we show that the function F maps the ball $\mathbb{B}_r$ onto itself. Since F is a continuous function and $\mathbb{B}_r$ is a compact set, we can conclude the proof of this proposition. Simplifying the function F, we have that

$$F(\delta_S, \delta_L) = \mathcal{B}\left\{\mathcal{P}_{\mathcal{Y}}\mathcal{A}^\dagger[-E_n + R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'})) - \mathcal{I}^*\mathcal{C}_{T'}] + Z\right\}.$$

Consequently, we have from Proposition 3.3 (main paper) that

$$\begin{aligned}g_\gamma(F(\delta_S, \delta_L)) &\leq \frac{2}{\alpha}\, g_\gamma\left(\mathcal{P}_{\mathcal{Y}}\mathcal{A}^\dagger[E_n - R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'})) + \mathcal{I}^*\mathcal{C}_{T'}] - Z\right) \\ &\leq \frac{4}{\alpha}\, \left\{g_\gamma(\mathcal{A}^\dagger[E_n - R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'})) + \mathcal{I}^*\mathcal{C}_{T'}]) + \lambda_n\right\} \\ &\leq \frac{r}{2} \;+\; \frac{4}{\alpha}\, g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'}))),\end{aligned}$$

where in the second inequality we use the fact that $g_\gamma(\mathcal{P}_{\mathcal{Y}}(\cdot, \cdot)) \leq 2g_\gamma(\cdot, \cdot)$ and that $g_\gamma(Z) \leq 2\lambda_n$, and in the final inequality we use the assumption on r.

We now bound the term $g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L)))$ using Proposition 3.1 as $g_\gamma(\Delta_S, \Delta_L) \leq \frac{1}{2C_1}$:

$$\begin{aligned}
\frac{4}{\alpha}\, g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\delta_S, \delta_L + \mathcal{C}_{T'}))) &\leq \frac{8D\psi C_1^2 (g_\gamma(\delta_S, \delta_L) + \|\mathcal{C}_{T'}\|_2)^2}{\xi(T)\alpha} \\
&\leq \frac{32 D\psi C_1^2 r^2}{\xi(T)\alpha} \\
&\leq \frac{32 D\psi C_1^2 r}{\xi(T)\alpha}\, \frac{\alpha\xi(T)}{64 D\psi C_1^2} \\
&\leq \frac{r}{2},
\end{aligned}$$

where we have used the fact that $r \leq \frac{\alpha\xi(T)}{64D\psi C_1^2}$. Hence $g_\gamma(\mathcal{P}_\mathcal{Y}(\Delta_S, \Delta_L)) \leq r$ by Brouwer's fixed-point theorem. Finally we observe that

$$\begin{aligned}
g_\gamma(\Delta_S, \Delta_L) &\leq g_\gamma(\mathcal{P}_\mathcal{Y}(\Delta_S, \Delta_L)) + \|\mathcal{C}_{T'}\|_2 \\
&\leq 2r.
\end{aligned}$$

□

3.3. *Solving a variety-constrained problem.* In order to prove that the solution $(\bar{S}_n, \bar{L}_n)$ of (3.3) has the same sparsity pattern/rank as (S^*, L^*), we will study an optimization problem that explicitly enforces these constraints. Specifically, we consider the following *non-convex* constraint set:

$$\begin{aligned}
\mathcal{M} \;=\; \{(S, L) \mid\; & S \in \Omega(S^*),\ \operatorname{rank}(L) \leq \operatorname{rank}(L^*), \\
& \|\mathcal{P}_{T^\perp}(L - L^*)\|_2 \leq \frac{\xi(T)\lambda_n}{D\psi^2},\ g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{A}(S - S^*, L^* - L)) \leq 11\lambda_n\}
\end{aligned}$$

Recall that $S^* = K_O^*$ and $L^* = K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$. The first constraint ensures that the tangent space at S is the same as the tangent space at S^*; therefore the support of S is contained in the support of S^*. The second and third constraints ensure that L lives in the appropriate low-rank variety, but has a tangent space "close" to the tangent space T. The final constraint roughly bounds the sum of the errors $(S - S^*) + (L^* - L)$; note that this does not necessarily bound the individual errors. Notice that the only non-convex constraint is that $\operatorname{rank}(L) \leq \operatorname{rank}(L^*)$. We then have the following nonlinear program:

(3.7)
$$\begin{aligned}
(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M}) = \arg\min_{S,L}\ & \operatorname{tr}[(S - L)\ \Sigma_O^n] - \log\det(S - L)\ +\ \lambda_n[\gamma\|S\|_1 + \|L\|_*] \\
\text{s.t.}\quad & S - L \succ 0,\quad (S, L) \in \mathcal{M}.
\end{aligned}$$

Under suitable conditions this nonlinear program is shown to have a unique solution. Each of the constraints in $\mathcal{M}$ is useful for proving the consistency of the solution of the convex program (3.3). We show that under suitable conditions the constraints in $\mathcal{M}$ are actually inactive at the optimal $(\hat{S}_{\mathcal{M}}, \hat{L}_{\mathcal{M}})$, thus allowing us to conclude that the solution of (3.3) is also equal to $(\hat{S}_{\mathcal{M}}, \hat{L}_{\mathcal{M}})$; hence the solution of (3.3) shares the consistency properties of $(\hat{S}_{\mathcal{M}}, \hat{L}_{\mathcal{M}})$. A number of interesting properties can be derived simply by studying the constraint set $\mathcal{M}$.

PROPOSITION 3.3. *Consider* any $(S, L) \in \mathcal{M}$, *and let* $\Delta_S = S - S^*, \Delta_L = L^* - L$. *For* γ *in the range specified by* (3.1) *and letting* $C_2 = \frac{48}{\alpha} + \frac{1}{\psi^2}$, *we have that* $g_\gamma(\Delta_S, \Delta_L) \leq C_2 \lambda_n$.

Proof: We have by the triangle inequality that

$$\begin{aligned} g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{A}(\mathcal{P}_\Omega(\Delta_S), \mathcal{P}_T(\Delta_L))) &\leq 11\lambda_n + g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{A}(\mathcal{P}_{\Omega^\perp}(\Delta_S), \mathcal{P}_{T^\perp}(\Delta_L))) \\ &\leq 11\lambda_n + m\psi^2 \|\mathcal{P}_{T^\perp}(\Delta_L)\|_2 \\ &\leq 12\lambda_n, \end{aligned}$$

as $m \leq \frac{D}{\xi(T)}$. Therefore, we have that $g_\gamma(\mathcal{P}_{\mathcal{Y}} \mathcal{A}^\dagger \mathcal{I}^* \mathcal{A} \mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)) \leq 24\lambda_n$, where $\mathcal{Y} = \Omega \times T$. Consequently, we can apply Proposition 3.3 (main paper) to conclude that

$$g_\gamma(\mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)) \leq \frac{48\lambda_n}{\alpha}.$$

Finally, we use the triangle inequality again to conclude that

$$\begin{aligned} g_\gamma(\Delta_S, \Delta_L) &\leq g_\gamma(\mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)) + g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}(\Delta_S, \Delta_L)) \\ &\leq \frac{48\lambda_n}{\alpha} + m\|\mathcal{P}_{T^\perp}(\Delta_L)\|_2 \\ &\leq C_2 \lambda_n. \end{aligned}$$

□

This simple result immediately leads to a number of useful corollaries. For example we have that under a suitable bound on the minimum nonzero singular value of $L^* = K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$, the constraint in $\mathcal{M}$ along the normal direction $T^\perp$ is locally inactive. Next we list several useful consequences of Proposition 3.3.

COROLLARY 3.4. *Consider* any $(S, L) \in \mathcal{M}$, *and let* $\Delta_S = S - S^*, \Delta_L = L^* - L$. *Suppose* γ *is in the range specified by* (3.1), *and let* $C_3 = \left(\frac{6(2-\nu)}{\nu} + 1\right) C_2^2 \psi^2 D$ *and* $C_4 = C_2 + \frac{3\alpha C_2^2 (2-\nu)}{16(3-\nu)}$ *(where* C_2 *is as defined in Proposition 3.3). Let the*

minimum nonzero singular value σ *of* $L^* = K^*_{O,H}(K^*_H)^{-1}K^*_{H,O}$ *be such that* $\sigma \geq \frac{C_5\lambda_n}{\xi(T)^2}$ *for* $C_5 = \max\{C_3, C_4\}$*, and suppose that the smallest magnitude nonzero entry of* S^* *is greater than* $\frac{C_6\lambda_n}{\mu(\Omega)}$ *for* $C_6 = \frac{C_2\nu\alpha}{\beta(2-\nu)}$*. Setting* $T' = T(L)$ *and* $\mathcal{C}_{T'} = \mathcal{P}_{T'^{\perp}}(L^*)$*, we then have that:*

1. *L has rank equal to* $\mathrm{rank}(L^*)$*, i.e., L is a smooth point of the variety of matrices with rank less than or equal to* $\mathrm{rank}(L^*)$*. In particular L has the same inertia as* L^**.*
2. $\|\mathcal{P}_{T^{\perp}}(\Delta_L)\|_2 \leq \frac{\xi(T)\lambda_n}{19D\psi^2}$.
3. $\rho(T, T') \leq \frac{\xi(T)}{4}$.
4. $g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T'}) \leq \frac{\lambda_n\nu}{6(2-\nu)}$.
5. $\|\mathcal{C}_{T'}\|_2 \leq \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)}$.
6. $\mathrm{sign}(S) = \mathrm{sign}(S^*)$.

Proof: We note the following facts before proving each step. First $C_2 \geq \frac{1}{\psi^2} \geq \frac{1}{m\psi^2} \geq \frac{\xi(T)}{D\psi^2}$. Second $\xi(T) \leq 1$. Third we have from Proposition 3.3 that $\|\Delta_L\|_2 \leq C_2\lambda_n$. Finally $\frac{6(2-\nu)}{\nu} \geq 18$ for $\nu \in (0, \frac{1}{2}]$. We prove each step separately.

For the first step, we note that

$$\sigma \geq \frac{C_3\lambda_n}{\xi(T)^2} \geq \frac{19C_2^2\psi^2 D\lambda_n}{\xi(T)^2} \geq \frac{19C_2\lambda_n}{\xi(T)} \geq 8C_2\lambda_n \geq 8\|\Delta_L\|_2.$$

Hence L is a smooth point with rank equal to $\mathrm{rank}(L^*)$, and specifically has the same inertia as L^*.

For the second step, we use the fact that $\sigma \geq 8\|\Delta_L\|_2$ to apply Proposition 2.2:

$$\|\mathcal{P}_{T^{\perp}}(\Delta_L)\| \leq \frac{\|\Delta_L\|_2^2}{\sigma} \leq \frac{C_2^2\xi(T)^2\lambda_n^2}{C_3\lambda_n} \leq \frac{\xi(T)\lambda_n}{19D\psi^2}.$$

For the third step we apply Proposition 2.1 (by using the conclusion from above that $\sigma \geq 8\|\Delta_L\|_2$) so that

$$\rho(T, T') \leq \frac{2\|\Delta_L\|_2}{\sigma} \leq \frac{2C_2\xi(T)^2}{C_3} \leq \frac{2\xi(T)^2}{19C_2D\psi^2} \leq \frac{\xi(T)}{4}.$$

For the fourth step let σ' denote the minimum singular value of L. Consequently,

$$\sigma' \geq \frac{C_3\lambda_n}{\xi(T)^2} - C_2\lambda_n \geq C_2\lambda_n\left[\frac{19C_2D\psi^2}{\xi(T)^2} - 1\right] \geq 8\|\Delta_L\|_2.$$

Using the same reasoning as in the proof of the second step, we have that

$$\begin{aligned}\|\mathcal{C}_{T'}\|_2 &\leq \frac{\|\Delta_L\|_2^2}{\sigma'} \leq \frac{C_2^2\lambda_n^2}{(\frac{C_3}{\xi(T)^2} - C_2)\lambda_n} = \frac{C_2^2\xi(T)^2\lambda_n}{C_2^2 D\psi^2(\frac{6(2-\nu)}{\nu}) + C_2^2 D\psi^2 - C_2\xi(T)^2} \\ &\leq \frac{C_2^2\xi(T)^2\lambda_n}{C_2^2 D\psi^2(\frac{6(2-\nu)}{\nu})} \leq \frac{\nu\xi(T)\lambda_n}{6(2-\nu)D\psi^2}.\end{aligned}$$

Hence

$$g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T'}) \leq m\psi^2\|\mathcal{C}_{T'}\|_2 \leq \frac{\lambda_n\nu}{6(2-\nu)}.$$

For the fifth step the bound on σ' implies that

$$\sigma' \geq \frac{C_4\lambda_n}{\xi(T)^2} - C_2\lambda_n \geq \frac{3C_2^2\alpha(2-\nu)}{16(3-\nu)}\lambda_n$$

Since $\sigma' \geq 8\|\Delta_L\|_2$, we have from Proposition 2.2 and some algebra that

$$\|\mathcal{C}_{T'}\|_2 \leq \frac{C_2^2\lambda_n^2}{\sigma'} \leq \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)}.$$

For the final step since $\|\Delta_S\|_\infty \leq \gamma C_2\lambda_n$, the assumed lower bound on the minimum magnitude nonzero entry of S^* guarantees that $\operatorname{sign}(S) = \operatorname{sign}(S^*)$. □

Notice that this corollary applies to *any* $(S, L) \in \mathcal{M}$, and is hence applicable to *any solution* $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$ of the $\mathcal{M}$-constrained program (3.7). For now we choose an arbitrary solution $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$ and proceed. In the next steps we show that $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$ is *the unique* solution to the convex program (3.3), thus showing that $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$ is also the unique solution to (3.7).

3.4. *From variety constraint to tangent-space constraint.* Given the solution $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$, we show that the solution to the convex program (3.4) with the tangent space constraint $L \in T_\mathcal{M} \triangleq T(\hat{L}_\mathcal{M})$ is the same as $(\hat{S}_\mathcal{M}, \hat{L}_\mathcal{M})$ under suitable conditions:

(3.8)
$$\begin{aligned}(\hat{S}_\Omega, \hat{L}_{T_\mathcal{M}}) = \arg\min_{S,L} \operatorname{tr}[(S-L)\,\Sigma_O^n] - \log\det(S-L) \;+\; \lambda_n[\gamma\|S\|_1 + \|L\|_*] \\ \text{s.t.} \quad S - L \succ 0, \quad S \in \Omega, \quad L \in T_\mathcal{M}.\end{aligned}$$

Assuming the bound of Corollary 3.4 on the minimum singular value of L^* the uniqueness of the solution $(\hat{S}_\Omega, \hat{L}_{T_\mathcal{M}})$ is assured. This is because we have from Proposition 3.3 (main paper) and from Corollary 3.4 that $\mathcal{I}^*$ is

injective on $\Omega \oplus T_{\mathcal{M}}$. Therefore the Hessian of the convex objective function of (3.8) is strictly positive-definite at $(\hat{S}_\Omega, \hat{L}_{T_{\mathcal{M}}})$.

We let $\mathcal{C}_{\mathcal{M}} = \mathcal{P}_{T_{\mathcal{M}}^\perp}(L^*)$. Recall that $E_n = \Sigma_O^n - \Sigma_O^*$ denotes the difference between the sample covariance matrix and the marginal covariance matrix of the observed variables.

PROPOSITION 3.5. *Let γ be in the range specified by (3.1). Suppose that the minimum nonzero singular value σ of $L^* = K_{O,H}^*(K_H^*)^{-1}K_{H,O}^*$ is such that $\sigma \geq \frac{C_5\lambda_n}{\xi(T)^2}$ (C_5 is defined in Corollary 3.4). Suppose also that the minimum magnitude nonzero entry of S^* is greater than or equal to $\frac{C_6\lambda_n}{\mu(\Omega)}$ (C_6 is defined in Corollary 3.4). Let $g_\gamma(\mathcal{A}^\dagger E_n) \leq \frac{\lambda_n\nu}{6(2-\nu)}$. Further suppose that*

$$\lambda_n \leq \frac{3\alpha(2-\nu)}{16(3-\nu)} \min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\right\}.$$

Then we have that

$$(\hat{S}_\Omega, \hat{L}_{T_{\mathcal{M}}}) = (\hat{S}_{\mathcal{M}}, \hat{L}_{\mathcal{M}}).$$

Proof: Note first that the condition on the minimum singular value of L^* in Corollary 3.4 is satisfied. Therefore we proceed with the following two steps:

1. First we can change the non-convex constraint $\text{rank}(L) \leq \text{rank}(L^*)$ to the linear constraint $L \in T(\hat{L}_{\mathcal{M}})$. This is because the lower bound assumed for σ implies that $\hat{L}_{\mathcal{M}}$ is a smooth point of the algebraic variety of matrices with rank less than or equal to $\text{rank}(L^*)$ (from Corollary 3.4). Due to the convexity of all the other constraints and the objective, the optimum of this "linearized" convex program will still be $(\hat{S}_{\mathcal{M}}, \hat{L}_{\mathcal{M}})$.
2. Next we can again apply Corollary 3.4 (based on the bound on σ) to conclude that the constraint $\|\mathcal{P}_{T^\perp}(L - L^*)\|_2 \leq \frac{\xi(T)\lambda_n}{D\psi^2}$ is *locally inactive* at the point $(\hat{S}_{\mathcal{M}}, \hat{L}_{\mathcal{M}})$.

Consequently, we have that $(\hat{S}_{\mathcal{M}}, \hat{L}_{\mathcal{M}})$ can be written as the solution of a *convex program*:

(3.9)

$$\begin{aligned}(\hat{S}_{\mathcal{M}}, \hat{L}_{\mathcal{M}}) = \arg\min_{S,L} \ & \text{tr}[(S-L)\ \Sigma_O^n] - \log\det(S-L) \ + \ \lambda_n[\gamma\|S\|_1 + \|L\|_*] \\ \text{s.t.} \quad & S - L \succ 0, \quad S \in \Omega, \quad L \in T_{\mathcal{M}}, \\ & g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}(S - S^*, L^* - L)) \leq 11\lambda_n.\end{aligned}$$

We now need to argue that the constraint $g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}(S-S^*, L^*-L)) \leq 11\lambda_n$ is also inactive in the convex program (3.9). We proceed by showing that the solution $(\hat{S}_\Omega, \hat{L}_{T_\mathcal{M}})$ of the convex program (3.8) has the property that $g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}(\hat{S}_\Omega - S^*, L^* - \hat{L}_{T_\mathcal{M}})) < 11\lambda_n$, which concludes the proof of this proposition. We have from Corollary 3.4 that $g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_\mathcal{M}}) \leq \frac{\lambda_n\nu}{6(2-\nu)}$. Since $g_\gamma(\mathcal{A}^\dagger E_n) \leq \frac{\lambda_n\nu}{6(2-\nu)}$ by assumption, one can verify that

$$\begin{aligned}\frac{8}{\alpha}\left[\lambda_n + g_\gamma(\mathcal{A}^\dagger E_n) + g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_\mathcal{M}})\right] &\leq \frac{8\lambda_n}{\alpha}\left[1 + \frac{\nu}{3(2-\nu)}\right] \\ &= \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)} \\ &\leq \min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\right\}.\end{aligned}$$

The last line follows from the assumption on λ_n. We also note that $\|C_{T_\mathcal{M}}\|_2 \leq \frac{16(3-\nu)\lambda_n}{3\alpha(2-\nu)}$ from Corollary 3.4, which implies that $\|C_{T_\mathcal{M}}\|_2 \leq \min\left\{\frac{1}{4C_1}, \frac{\alpha\xi(T)}{64D\psi C_1^2}\right\}$. Letting $(\Delta_S, \Delta_L) = (S_\Omega - S^*, L^* - \hat{L}_{T_\mathcal{M}})$, we can conclude from Proposition 3.2 that $g_\gamma(\Delta_L, \Delta_S) \leq \frac{32(3-\nu)\lambda_n}{3\alpha(2-\nu)}$. Next we apply Proposition 3.1 (as $g_\gamma(\Delta_L, \Delta_S) \leq \frac{1}{2C_1}$) to conclude that

$$\begin{aligned}g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\Delta_S + \Delta_L)) &\leq \frac{2D\psi C_1^2 g_\gamma(\Delta_S, \Delta_L)^2}{\xi(T)} \\ &\leq \frac{2D\psi C_1^2}{\xi(T)}\frac{32(3-\nu)\lambda_n}{3\alpha(2-\nu)}\frac{\alpha\xi(T)}{32D\psi C_1^2} \\ &\leq \frac{2(3-\nu)\lambda_n}{3(2-\nu)}. \qquad (3.10)\end{aligned}$$

From the optimality conditions of (3.8) one can also check that for $\mathcal{Y} = \Omega \times T_\mathcal{M}$,

$$\begin{aligned}g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_\mathcal{Y}(\Delta_S, \Delta_L)) &\leq 2\lambda_n + g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger R_{\Sigma_O^*}(\Delta_S + \Delta_L)) \\ &\quad + g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_\mathcal{M}}) + g_\gamma(\mathcal{P}_\mathcal{Y}\mathcal{A}^\dagger E_n) \\ &\leq 2[\lambda_n + g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\Delta_S + \Delta_L)) \\ &\quad + g_\gamma(\mathcal{A}^\dagger E_n) + g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_\mathcal{M}})] \\ &\leq 4\left[\frac{2(3-\nu)\lambda_n}{3(2-\nu)}\right].\end{aligned}$$

Here we used (3.10) in the last inequality, and also that $g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_\mathcal{M}}) \leq \frac{\lambda_n\nu}{6(2-\nu)}$ (as noted above from Corollary 3.4) and that $g_\gamma(\mathcal{A}^\dagger E_n) \leq \frac{\lambda_n\nu}{6(2-\nu)}$.

Therefore,

$$g_\gamma(\mathcal{P}_{\mathcal{Y}}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\Delta_S,\Delta_L)) \quad \leq \quad \frac{16\lambda_n}{3}, \tag{3.11}$$

because $\nu \in (0, \frac{1}{2}]$. Based on Proposition 3.3 in the main paper (the second part), we also have that

$$g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\Delta_S,\Delta_L)) \leq (1-\nu)\frac{16\lambda_n}{3} \leq \frac{16\lambda_n}{3}. \tag{3.12}$$

Summarizing steps (3.11) and (3.12),

$$\begin{aligned} g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}(\Delta_S,\Delta_L)) &\leq g_\gamma(\mathcal{P}_{\mathcal{Y}}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\Delta_S,\Delta_L)) \\ &\quad + g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\Delta_S,\Delta_L)) + g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_{\mathcal{M}}}) \\ &\leq \frac{16\lambda_n}{3} + \frac{16\lambda_n}{3} + \frac{\lambda_n\nu}{6(2-\nu)} \\ &\leq \frac{32\lambda_n}{3} + \frac{\lambda_n}{18} \\ &< 11\lambda_n. \end{aligned}$$

This concludes the proof of the proposition. □

This proposition has the following important consequence.

COROLLARY 3.6. *Under the assumptions of Proposition 3.5 we have that* $\mathrm{rank}(\hat{L}_{T_{\mathcal{M}}}) = \mathrm{rank}(L^*)$ *and that* $T(\hat{L}_{T_{\mathcal{M}}}) = T_{\mathcal{M}}$. *Moreover,* $\hat{L}_{T_{\mathcal{M}}}$ *actually has the same inertia as* L^*. *We also have that* $\mathrm{sign}(\hat{S}_\Omega) = \mathrm{sign}(S^*)$.

3.5. *Proof of main paper Proposition 5.3 – Removing the tangent-space constraints.* The following lemma provides a simple set of sufficient conditions under which the optimal solution $(\hat{S}_\Omega, \hat{L}_{T_{\mathcal{M}}})$ of (3.8) satisfies the optimality conditions of the convex program (3.3) (without the tangent space constraints).

LEMMA 3.7. *Let* $(\hat{S}_\Omega, \hat{L}_{T_{\mathcal{M}}})$ *be the solution to the tangent-space constrained convex program* (3.8). *Suppose that the assumptions of Proposition 3.5 hold. If in addition we have that*

$$g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\Delta_S,\Delta_L))) \leq \frac{\lambda_n\nu}{6(2-\nu)},$$

then $(\hat{S}_\Omega, \hat{L}_{T_{\mathcal{M}}})$ *is also the unique optimum of the convex program* (3.3).

Proof: Recall from Corollary 3.6 that the tangent space at $\hat{L}_{T_{\mathcal{M}}}$ is equal to $T_{\mathcal{M}}$. Applying the optimality conditions of the convex program (3.8) at the optimum $(\hat{S}_\Omega, \hat{L}_{T_{\mathcal{M}}})$, we have that there exist Lagrange multipliers $Q_{\Omega^\perp} \in \Omega^\perp$, $Q_{T_{\mathcal{M}}^\perp} \in T_{\mathcal{M}}^\perp$ such that

$$\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1} + Q_{\Omega^\perp} \in -\lambda_n \gamma \partial \|\hat{S}_\Omega\|_1, \;\; \Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1} + Q_{T_{\mathcal{M}}^\perp} \in \lambda_n \partial \|\hat{L}_{T_{\mathcal{M}}}\|_*.$$

Restricting these conditions to the space $\mathcal{Y} = \Omega \times T_{\mathcal{M}}$, one can check that

$$\mathcal{P}_\Omega[\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1}] = -\lambda_n \gamma \operatorname{sign}(S^*), \;\; \mathcal{P}_{T_{\mathcal{M}}}[\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1}] = \lambda_n U V^T,$$

where $\hat{L}_{T_{\mathcal{M}}} = UDV^T$ is a reduced SVD of $\hat{L}_{T_{\mathcal{M}}}$. Denoting $Z = [-\lambda_n \gamma \operatorname{sign}(S^*), \lambda_n U V^T]$, we conclude that

$$\mathcal{P}_{\mathcal{Y}} \mathcal{A}^\dagger [\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1}] = Z, \tag{3.13}$$

with $g_\gamma(Z) = \lambda_n$. It is clear that the optimality condition of the convex program (3.3) (without the tangent-space constraints) on $\mathcal{Y}$ is satisfied. All we need to show is that

$$g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp} \mathcal{A}^\dagger [\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1}]) < \lambda_n. \tag{3.14}$$

Rewriting $\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1}$ in terms of the error $(\Delta_S, \Delta_L) = (\hat{S}_\Omega - S^*, L^* - \hat{L}_{T_{\mathcal{M}}})$, we have that

$$\Sigma_O^n - (\hat{S}_\Omega - \hat{L}_{T_{\mathcal{M}}})^{-1} = E_n - R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)) + \mathcal{I}^* \mathcal{A}(\Delta_S, \Delta_L).$$

Restating the condition (3.13) on $\mathcal{Y}$, we have that

$$\mathcal{P}_{\mathcal{Y}} \mathcal{A}^\dagger \mathcal{I}^* \mathcal{A} \mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L) = Z + \mathcal{P}_{\mathcal{Y}} \mathcal{A}^\dagger [-E_n + R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)) - \mathcal{I}^* \mathcal{C}_{T_{\mathcal{M}}}]. \tag{3.15}$$

(Recall that $\mathcal{C}_{T_{\mathcal{M}}} = \mathcal{P}_{T_{\mathcal{M}}^\perp}(L^*)$.) A sufficient condition to show (3.14) and complete the proof of this lemma is that

$$g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp} \mathcal{A}^\dagger \mathcal{I}^* \mathcal{A} \mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)) < \lambda_n - g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp} \mathcal{A}^\dagger [-E_n + R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)) - \mathcal{I}^* \mathcal{C}_{T_{\mathcal{M}}}]).$$

We prove this inequality next. Recall from Corollary 3.4 that $g_\gamma(\mathcal{A}^\dagger \mathcal{I}^* \mathcal{C}_{T_{\mathcal{M}}}) \leq \frac{\lambda_n \nu}{6(2-\nu)}$. Therefore, from equation (3.15) we can conclude that

$$\begin{aligned} g_\gamma(\mathcal{P}_{\mathcal{Y}} \mathcal{A}^\dagger \mathcal{I}^* \mathcal{A} \mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)) &\leq \lambda_n + 2(g_\gamma(\mathcal{A}^\dagger [-E_n + R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)) - \mathcal{I}^* \mathcal{C}_{T_{\mathcal{M}}}])) \\ &\leq \lambda_n + 2 \left[\frac{3\lambda_n \nu}{6(2-\nu)} \right] \\ &= \frac{2\lambda_n}{2-\nu}. \end{aligned}$$

Here we used the bounds assumed on $g_\gamma(\mathcal{A}^\dagger E_n)$ and on $g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)))$.

Applying the second part of Proposition 3.3 (main paper), we have that

$$\begin{aligned} g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger\mathcal{I}^*\mathcal{A}\mathcal{P}_{\mathcal{Y}}(\Delta_S, \Delta_L)) &\leq \frac{2\lambda_n(1-\nu)}{2-\nu} \\ &= \lambda_n - \frac{\nu\lambda_n}{2-\nu} \\ &< \lambda_n - \frac{\nu\lambda_n}{2(2-\nu)} \\ &\leq \lambda_n - g_\gamma(\mathcal{A}^\dagger[-E_n + R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)) - \mathcal{I}^*\mathcal{C}_{T_\mathcal{M}}]) \\ &\leq \lambda_n - g_\gamma(\mathcal{P}_{\mathcal{Y}^\perp}\mathcal{A}^\dagger[-E_n + R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)) - \mathcal{I}^*\mathcal{C}_{T_\mathcal{M}}]). \end{aligned}$$

Here the second-to-last inequality follows from the bounds on $g_\gamma(\mathcal{A}^\dagger E_n)$, $g_\gamma(\mathcal{A}^\dagger R_{\Sigma_O^*}(\mathcal{A}(\Delta_S, \Delta_L)))$, and $g_\gamma(\mathcal{A}^\dagger\mathcal{I}^*\mathcal{C}_{T_\mathcal{M}})$. This concludes the proof of the lemma. □

3.6. *Proof of main paper Lemma 5.4 – Probabilistic analysis.* All the analysis described so far in this section has been completely deterministic in nature. Here we present the probabilistic component of our proof. Specifically, we study the rate at which the sample covariance matrix converges to the true covariance matrix. The following result from [2] plays a key role in our analysis:

THEOREM 3.8. *Given natural numbers n, p with $p \leq n$, let Γ be a $p \times n$ matrix with i.i.d. Gaussian entries that have zero-mean and variance $\frac{1}{n}$. Then the largest and smallest singular values $s_1(\Gamma)$ and $s_p(\Gamma)$ of Γ are such that*

$$\max\left\{\Pr\left[s_1(\Gamma) \geq 1 + \sqrt{\tfrac{p}{n}} + t\right], \Pr\left[s_p(\Gamma) \leq 1 - \sqrt{\tfrac{p}{n}} - t\right]\right\} \leq \exp\left\{-\tfrac{nt^2}{2}\right\},$$

for any $t > 0$.

Using this result the next lemma provides a probabilistic bound between the sample covariance Σ_O^n formed using n samples and the true covariance Σ_O^* in spectral norm. This result is well-known, and we mainly discuss it here for completeness and also to show explicitly the dependence on $\psi = \|\Sigma_O^*\|_2$.

LEMMA 3.9. *Let $\psi = \|\Sigma_O^*\|_2$. Given any $\delta > 0$ with $\delta \leq 8\psi$, let the number of samples n be such that $n \geq \frac{64p\psi^2}{\delta^2}$. Then we have that*

$$\Pr\left[\|\Sigma_O^n - \Sigma_O^*\|_2 \geq \delta\right] \leq 2\exp\left\{-\tfrac{n\delta^2}{128\psi^2}\right\}.$$

Proof: Since the spectral norm is unitarily invariant, we can assume that Σ_O^* is diagonal without loss of generality. Let $\bar{\Sigma}^n = (\Sigma_O^*)^{-\frac{1}{2}}\Sigma_O^n(\Sigma_O^*)^{-\frac{1}{2}}$, and let $s_1(\bar{\Sigma}^n), s_p(\bar{\Sigma}^n)$ denote the largest/smallest singular values of $\bar{\Sigma}^n$. Note that $\bar{\Sigma}^n$ can be viewed as the sample covariance matrix formed from n independent samples drawn from a model with identity covariance, i.e., $\bar{\Sigma}^n = \Gamma\Gamma^T$ where Γ denotes a $p \times n$ matrix with i.i.d. Gaussian entries that have zero-mean and variance $\frac{1}{n}$. We then have that

$$\begin{aligned}
\Pr\left[\|\Sigma_O^n - \Sigma_O^*\|_2 \geq \delta\right] &\leq \Pr\left[\|\bar{\Sigma}^n - I\|_2 \geq \tfrac{\delta}{\psi}\right] \\
&\leq \Pr\left[s_1(\bar{\Sigma}^n) \geq 1 + \tfrac{\delta}{\psi}\right] + \Pr\left[s_p(\bar{\Sigma}^n) \leq 1 - \tfrac{\delta}{\psi}\right] \\
&= \Pr\left[s_1(\Gamma)^2 \geq 1 + \tfrac{\delta}{\psi}\right] + \Pr\left[s_p(\Gamma)^2 \leq 1 - \tfrac{\delta}{\psi}\right] \\
&\leq \Pr\left[s_1(\Gamma) \geq 1 + \tfrac{\delta}{4\psi}\right] + \Pr\left[s_p(\Gamma) \leq 1 - \tfrac{\delta}{4\psi}\right] \\
&\leq \Pr\left[s_1(\Gamma) \geq 1 + \sqrt{\tfrac{p}{n}} + \tfrac{\delta}{8\psi}\right] + \Pr\left[s_p(\Gamma) \leq 1 - \sqrt{\tfrac{p}{n}} - \tfrac{\delta}{8\psi}\right] \\
&\leq 2\exp\left\{-\tfrac{n\delta^2}{128\psi^2}\right\}.
\end{aligned}$$

Here we used the fact that $n \geq \frac{64p\psi^2}{\delta^2}$ in the fourth inequality, and we applied Theorem 3.8 to obtain the final inequality by setting $t = \frac{\delta}{8\psi}$. □

The following corollary describes relates the number of samples required for an error bound to hold with probability $1 - 2\exp\{-p\}$.

References.

[1] Bach, F. (2008). Consistency of trace norm minimization. *J. Mach. Lear. Res.*. **9** 1019–1048.

[2] Davidson, K. R. and Szarek, S.J. (2001). Local operator theory, random matrices and Banach spaces. *Handbook of the Geometry of Banach Spaces*. **I** 317–366.

[3] Kato, T. (1995). *Perturbation theory for linear operators*. Springer.

[4] Ortega, J. M. and Rheinboldt, W. G. (1970). *Iterative solution of nonlinear equations in several variables*. Academic Press.

[5] Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Elec. Jour. of Stat.* **4** 935–980.

[6] Rockafellar, R. T. (1996). *Convex Analysis*. Princeton University Press.

Laboratory for Information and Decision Systems
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139
USA
E-mail: venkatc@mit.edu
parrilo@mit.edu
willsky@mit.edu