Supplementary Material: Interpreting Latent Variables in Factor Models via Convex Optimization

Armeen Taeb † and Venkat Chandrasekaran †,‡ *

† Department of Electrical Engineering
‡ Department of Computing and Mathematical Sciences
California Institute of Technology
Pasadena, CA 91125

1 A Numerical Approach for Verifying Assumptions 1, 2, and 3 (main paper)

We begin by considering Assumption 1 in (15)(main paper). Let $f_1 \triangleq 2\max\{\sqrt{p}, \sqrt{2k_u}, \sqrt{2k_x}\gamma, \sqrt{q}\}$, $f_2 \triangleq \max\{\sqrt{2k_u}, \sqrt{2k_x}\gamma\}$ and $\omega \triangleq \max\{\omega_y, \omega_{yx}\}$. Let $Z = (Z_1, Z_2, Z_3, Z_4) \in \mathbb{H}'$ with $\Phi_\gamma(Z_1, Z_2, Z_3, Z_4) = 1$. It is straightforward to check that:

$$\begin{aligned}\Phi_\gamma[\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}\mathcal{P}_{\mathbb{H}'}(Z_1, Z_2, Z_3, Z_4)] \;\; &\geq \;\; f_1^{-1}\sigma_{\min}(\mathcal{P}_{\mathbb{H}^\star}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}\mathcal{P}_{\mathbb{H}^\star}) \\ &- \;\; \max\left\{1, \frac{1}{\gamma}\right\}(\sqrt{3}\omega + \omega + \sqrt{3}\omega^2)f_2\psi^2 \triangleq T_1\end{aligned}$$

Notice that the quantity $\sigma_{\min}(\mathcal{P}_{\mathbb{H}^\star}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}\mathcal{P}_{\mathbb{H}^\star})$ (and henceforth the quantity T_1) is computable given the population model. Thus a trivial lower bound for α is given by:

$$\inf_{\mathbb{H}' \in U(\omega_y, \omega_{yx})} \chi(\mathbb{H}', \Phi_\gamma) \geq \alpha \geq T_1$$

We now consider Assumption 2 in (16) (main paper). Let $Z = (Z_1, Z_2) \in \mathbb{H}[2,3]'$ with $\Gamma_\gamma(Z_1, Z_2) = 1$. Using triangle inequality, it is straightforward to check the following bound:

$$\begin{aligned}\Gamma_\gamma[\mathcal{P}_{\mathbb{H}[2,3]'}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}[2,3]'}(Z_1, Z_2)] \;\; &\geq \;\; \min\left\{1, \frac{1}{\gamma}\right\}(\sqrt{3}f_2)^{-1} \\ &\quad\; \sigma_{\min}(\mathcal{P}_{\mathbb{H}[2,3]^\star}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}[2,3]^\star}) \\ &- \;\; \max\left\{1, \frac{1}{\gamma}\right\}(\sqrt{3}\omega + \omega + \sqrt{3}\omega^2)f_2\psi^2 \triangleq T2\end{aligned}$$

Notice that the quantity T_2 is computable giving the population model. Then,

$$\inf_{\mathbb{H}' \in U(\omega_y, \omega_{yx})} \Xi(\mathbb{H}', \Gamma_\gamma) \geq T2$$

*Email: ataeb@caltech.edu, venkatc@caltech.edu

Now we consider Assumption 3 in (17) (main paper). Using triangle inequality, it is straightforward to check that:

$$\begin{aligned}\Gamma_\gamma[\mathcal{P}_{\mathbb{H}[2,3]'^\perp}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}[2,3]'}(Z_1,Z_2)] &\leq \sqrt{3}f_2\max\Big\{1,\frac{1}{\gamma}\Big\} \\ &\quad \sigma_{\max}(\mathcal{P}_{\mathbb{H}[2,3]^{\star\perp}}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}[2,3]^\star}) \\ &+ \max\Big\{1,\frac{1}{\gamma}\Big\}(\sqrt{3}\omega+\omega+\sqrt{3}\omega^2)f_2\psi^2 \triangleq T3\end{aligned}$$

Similarly, the quantity T_3 can be computed given the population model. Then, an upper bound for $\varphi(\mathbb{H}',\Gamma_\gamma)$ is given by:

$$\sup_{\mathbb{H}'\in U(\omega_{yx},\omega_{yx})}\varphi(\mathbb{H}',\Gamma_\gamma) \leq 1-\frac{2}{1+\beta}\leq\frac{T3}{T2}\Longrightarrow\beta\leq\frac{2}{1-\frac{T3}{T2}}-1$$

2 Proof of Proposition 1 (main paper)

Proof. We note that:

$$\|\Delta\|_2 \leq \|\Delta D_y\|_2+\|\Delta L_y\|_2+\|\Delta\Theta_{yx}\|_2+\|\Delta\Theta_x\|_2\leq(3+\gamma)\Phi_\gamma(\Delta)$$

Furthermore, recall that

$$R_{\Sigma^\star}(\mathcal{F}(\Delta))=\Sigma^{\star-1}\Big[\sum_{k=2}^{\infty}(-\mathcal{F}(\Delta)\Sigma^{\star-1})^k\Big].$$

Using this observation and some algebra, we have that:

$$\begin{aligned}\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\mathcal{F}(\Delta))]\leq m\psi\Big[\sum_{k=2}^{\infty}(\psi\|\Delta\|_2)^k\Big] &\leq m\psi^3\frac{(3+\gamma)^2\Phi_\gamma[\Delta]^2}{1-(3+\gamma)\Phi_\gamma[\Delta]\psi} \\ &\leq 2m\psi C'^2\Phi_\gamma[\Delta]^2\end{aligned}$$

□

3 Proof of Proposition 2 (main paper)

Proof. The proof of this result uses Brouwer's fixed-point theorem, and is inspired by the proof of a similar result in [5, 2]. The optimality conditions of (20) (main paper) suggest that there exist Lagrange multipliers $Q_{D_y}\in\mathcal{W}$, $Q_{T_y}\in {T'_y}^\perp$, and $Q_{T_{yx}}\in {T'_{yx}}^\perp$ such that

$$\begin{aligned}{}[\Sigma_n-\tilde{\Theta}^{-1}]_y+Q_{D_y}=0;\quad[\Sigma_n-\tilde{\Theta}^{-1}]_y+Q_{T_y} &\in \lambda_n\partial\|\tilde{L}_y\|_\star \\ [\Sigma_n-\tilde{\Theta}^{-1}]_{yx}+Q_{T_{yx}}\in-\lambda_n\gamma\partial\|\tilde{\Theta}_{yx}\|_\star;\quad[\Sigma_n-\tilde{\Theta}^{-1}]_x &= 0\end{aligned}$$

Letting the SVD of $\tilde{L}$ and $\tilde{\Theta}_{yx}$ be given by $\tilde{L}_y=\bar{U}\bar{D}\bar{V}'$ and $\tilde{\Theta}_{yx}=\breve{U}\breve{D}\breve{V}'$ respectively, and $Z\triangleq(0,\ \lambda_n\bar{U}\bar{V}',\ -\lambda_n\gamma\breve{U}\breve{V}',\ 0)$, we can restrict the optimality conditions of (15) (main paper) to the space $\mathbb{H}'$ to obtain, $\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger(\Sigma_n-\tilde{\Theta}^{-1})=Z$. Further, by appealing to the matrix inversion lemma, this condition can be restated as $\mathcal{P}_{\mathbb{H}_\mathcal{M}}\mathcal{F}^\dagger(E_n-R_{\Sigma^\star}(\mathcal{F}\Delta)+\mathbb{I}^\star\mathcal{F}(\Delta))=Z$. Based on the

Fisher information assumption 1 in (15) (main paper), the optimum of (20) (main paper) is unique (this is because the Hessian of the negative log-likelihood term is positive definite restricted to the tangent space constraints). Moreover, using standard Lagrangian duality, one can show that the set of variables $(\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y)$ that satisfy the restricted optimality conditions are unique. Consider the following function $S(\underline{\delta})$ restricted to $\underline{\delta} \in \mathcal{W} \times T'_y \times T'_{yx} \times \mathbb{S}^q$ with $\rho(T(L^\star_y), T'_y) \leq \omega_y$ and $\rho(T(\Theta^\star_{yx}), T'_{yx}) \leq \omega_{yx}$:

$$S(\underline{\delta}) = \underline{\delta} - (\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}\mathcal{P}_{\mathbb{H}'})^{-1}\Big(\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger[E_n \quad - \quad R_{\Sigma^\star}\mathcal{F}(\underline{\delta} + C_T) \\ +\mathbb{I}^\star\mathcal{F}(\underline{\delta} + C_T)] - Z\Big)$$

The function $S(\underline{\delta})$ is well-defined since the operator $\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}\mathcal{P}_{\mathbb{H}'}$ is bijective due to Fisher information assumption 1 in (15) (main paper). As a result, $\underline{\delta}$ is a fixed point of $S(\underline{\delta})$ if and only if $\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger[E_n - R_{\Sigma^\star}(\mathcal{F}(\underline{\delta}+C_T)) + \mathbb{I}^\star\mathcal{F}(\underline{\delta}+C_T)] = Z$. Since the pair $(\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y)$ are the unique solution to (20) (main paper), the only fixed point of S is $\mathcal{P}_{\mathbb{H}'}[\Delta]$. Next we show that this unique optimum lives inside the ball $\mathbb{B}_{r_1^u, r_2^u} = \{\underline{\delta} \;\mid\; \max\{\|\delta_2\|_2, \frac{1}{\gamma}\|\delta_3\|_2\} \leq r_1^u, \max\{\|\delta_1\|_2, \|\delta_4\|_2\} \leq r_2^u \; \underline{\delta} \in \mathbb{H}'\}$. In particular, we show that under the map S, the image of $\mathbb{B}_{r_1^u, r_2^u}$ lies in $\mathbb{B}_{r_1^u, r_2^u}$ and appeal to Brouwer's fixed point theorem to conclude that $\mathcal{P}_{\mathbb{H}'}[\Delta] \in \mathbb{B}_{r_1^u, r_2^u}$. For $\underline{\delta} \in \mathbb{B}_{r_1^u, r_2^u}$, the first component of $S(\underline{\delta})$, denoted by $S(\underline{\delta})_1$, can be bounded as follows:

$$\begin{aligned} \|S(\underline{\delta})_1\|_2 &= \Big\|\Big[(\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}\mathcal{P}_{\mathbb{H}'})^{-1}\Big(\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger[E_n - R_{\Sigma^\star}(\mathcal{F}(\underline{\delta} + C_T)) \\ &+ \mathbb{I}^\star\mathcal{F}C_T] + Z\Big)\Big]_1\Big\|_2 \leq \frac{2}{\alpha}\Big[\Phi_\gamma[\mathcal{F}^\dagger(E_n + \mathbb{I}^\star\mathcal{F}(C_T))]\Big] \\ &+ \frac{2}{\alpha}\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\underline{\delta} + C_T)] \leq \frac{r_2^u}{2} + \frac{2}{\alpha}\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\underline{\delta} + C_T)] \end{aligned}$$

The first inequality holds because of Fisher Information Assumption 1 in (15) (main paper), and the properties that $\Phi_\gamma[\mathcal{P}_{\mathbb{H}_\mathcal{M}}(.)] \leq 2\Phi_\gamma(.)$ (since projecting into the tangent space of a low-rank matrix variety increases the spectral norm by a factor of at most two) and $\Phi_\gamma(Z) = \lambda_n$. Moreover, since $r_1^u \leq \frac{1}{4C'}$, we have $\Phi_\gamma(\underline{\delta} + C_T) \leq \Phi_\gamma(\underline{\delta}) + \Phi_\gamma(C_T) \leq 2r_1^u \leq \frac{1}{2C'}$. Moreover, $r_1^u \leq r_2^u \max\{1 + \frac{\kappa}{2}, \frac{\alpha}{8}\}$. We can now appeal to Proposition 1 (main paper) to obtain:

$$\begin{aligned} \frac{2}{\alpha}\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\underline{\delta} + C_T)] &\leq \frac{4}{\alpha}m\psi C'^2[\Phi_\gamma(\underline{\delta} + C_T)]^2 \\ &\leq \frac{16}{\alpha}m\psi C'^2(r_2^u)^2\max\{1 + \frac{\kappa}{2}, \frac{\alpha}{8}\}^2 \\ &\leq \frac{r_2^u}{2} \end{aligned}$$

Thus, we conclude that $\|S(\delta)_1\|_2 \leq r_2^u$. Similarly, we check that:

$$\begin{aligned} \|[S(\underline{\delta})_2]\|_2 &= \Big\|\Big[(\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}\mathcal{P}_{\mathbb{H}'})^{-1}\Big(\mathcal{P}_{\mathbb{H}'}\mathcal{F}^\dagger[E_n - R_{\Sigma^\star}(\mathcal{F}(\underline{\delta} + C_T)) \\ &+ \mathbb{I}^\star\mathcal{F}C_T] + Z\Big)\Big]_2\Big\|_2 \leq \frac{2}{\alpha}\Big[\Phi_\gamma[\mathcal{F}^\dagger(E_n + \mathbb{I}^\star\mathcal{F}(C_T)] + \lambda_n\Big] \\ &+ \frac{2}{\alpha}\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\underline{\delta} + C_T)] \leq \frac{r_1^u}{2} + \frac{2}{\alpha}\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\underline{\delta} + C_T)] \leq r_1^u \end{aligned}$$

Using a similar approach, we can conclude that $\frac{1}{\gamma}\|S(\delta)_3\|_2 \leq r_1^u$ and $\|S(\delta)_3\|_2 \leq r_2^u$. Therefore, Brouwer's fixed point theorem suggests that $\mathcal{P}_{\mathbb{H}'}(\Delta) \in \mathcal{B}_{r_1^u, r_2^u}$. Hence, $\|\Delta_1\|_2 \leq r_2^u$, $\|\Delta_4\|_2 \leq r_2^u$, $\|\Delta_2\|_2 \leq \|\mathcal{P}_{\mathbb{H}'[2]}(\Delta_2)\|_2 + \|\mathcal{P}_{\mathbb{H}'[2]^\perp}(\Delta_2)\|_2 \leq 2r_1^u$, and
$\frac{1}{\gamma}\|\Delta_3\|_2 \leq \frac{1}{\gamma}\|\mathcal{P}_{\mathbb{H}'[3]}(\Delta_3)\|_2 + \frac{1}{\gamma}\|\mathcal{P}_{\mathbb{H}'[3]^\perp}(\Delta_2)\|_2 \leq 2r_1^u$. □

4 Proof of Proposition 3 (main paper)

Below, we outline our proof strategy:

1. We proceed by analyzing (19) (main paper) with additional constraints that the variables L_y, and Θ_{yx} belong to the algebraic varieties low-rank matrices (specified by rank of $L_y^\star$, and $\Theta_{yx}^\star$) , and that the tangent spaces $T(L_y)$, $T(\Theta_{yx})$ are close to the nominal tangent spaces $T(L_y^\star)$, and $T(\Theta_{yx}^\star)$ respectively. We prove that under suitable conditions on the minimum nonzero singular value of $L_y^\star$, and minimum nonzero singular value of $\Theta_{yx}^\star$, any optimum pair of variables (Θ, D_y, L_y) of this non-convex program are smooth points of the underlying varieties; that is $\mathrm{rank}(\mathrm{L_y}) = \mathrm{rank}(\mathrm{L_y^\star})$ and $\mathrm{rank}(\Theta_{\mathrm{yx}}) = \mathrm{rank}(\Theta_{\mathrm{yx}}^\star)$. Further, we show that L_y has the same inertia as $L_y^\star$ so that $L_y \succeq 0$.

2. Conclusions of the previous step imply the the variety constraints can be "linearized" at the optimum of the non-convex program to obtain tangent-space constraints. Under the specified conditions on the regularization parameter λ_n, we prove that with high probability, the unique optimum of this "linearized" program coincides with the global optimum of the non-convex program.

3. Finally, we show that the tangent-space constraints of the linearized program are inactive at the optimum. Therefore the optimal solution of (19) (main paper) has the property that with high probability: $\mathrm{rank}(\bar{L}_y) = \mathrm{rank}(L_y^\star)$ and $\mathrm{rank}(\bar{\Theta}_{yx}) = \mathrm{rank}(\Theta_{yx}^\star)$. Since $\bar{L}_y \succeq 0$, we conclude that the variables $(\bar{\Theta}, \bar{D}_y, \bar{L}_y)$ are the unique optimum of (4) (main paper).

4.1 Variety Constrained Program

We begin by considering a variety-constrained optimization program. Letting $(M, N, P, Q) \subset \mathbb{S}^p \times \mathbb{S}^p \times \mathbb{R}^{p\times q} \times \mathbb{S}^q$, we denote $\mathcal{P}_{[2,3]}(M, N, P, Q) = (N, P) \subset \mathbb{S}^p \times \mathbb{R}^{p\times q}$. The variety-constrained optimization program is given by:

$$(\Theta^{\mathcal{M}}, D_y^{\mathcal{M}}, L_y^{\mathcal{M}}) = \underset{\substack{\Theta\in\mathbb{S}^{q+p},\ \Theta\succ 0\\ D_y, L_y\in\mathbb{S}^p}}{\mathrm{argmin}} \quad -\ell(\Theta; \mathcal{D}_n^+) + \lambda_n[\|L_y\|_\star + \gamma\|\Theta_{yx}\|_\star]$$
$$\text{s.t.} \quad \Theta_y = D_y - L_y, (\Theta, D_y, L_y) \in \mathcal{M}. \tag{1}$$

Here, the set $\mathcal{M} = \mathcal{M}_1 \cap \mathcal{M}_2$, where the sets $\mathcal{M}_1$ and $\mathcal{M}_2$ are given by:

$$\begin{aligned}
\mathcal{M}_1 \quad &\triangleq \quad \Big\{(\Theta, D_y, L_y) \in \mathbb{S}^{(p+q)} \times \mathbb{S}^p \times \mathbb{S}^p \Big| D_y \text{ is diagonal}, \quad \mathrm{rank}(L_y) \le \mathrm{rank}(L_y^\star) \\
&\mathrm{rank} \quad (\Theta_{yx}) \le \mathrm{rank}(\Theta_{yx}^\star); \|\mathcal{P}_{T(L_y^\star)^\perp}(L_y - L_y^\star)\|_2 \le \frac{\lambda_n}{2\psi^2} \\
&\qquad \|\mathcal{P}_{T(\Theta_{yx}^\star)^\perp}(\Theta_{yx} - \Theta_{yx}^\star)\|_2 \le \frac{\lambda_n}{2\psi^2}\Big\} \\
\mathcal{M}_2 \quad &\triangleq \quad \Big\{(\Theta, D_y, L_y) \in \mathbb{S}^{(p+q)} \times \mathbb{S}^p \times \mathbb{S}^p \Big| \\
&\qquad \|\mathbb{I}^\star\mathcal{F}(\Delta)\|_2 \le 6\bar{m}\psi^2\lambda_n\Big(\frac{8}{\alpha\kappa} + \frac{4}{\alpha} + \frac{1}{\kappa}\Big)\Big\},
\end{aligned}$$

The optimization program (1) is non-convex due to the rank constraints $\mathrm{rank}(L_y) \le \mathrm{rank}(L_y^\star)$ and $\mathrm{rank}(\Theta_{yx}) \le \mathrm{rank}(\Theta_{yx}^\star)$ in the set $\mathcal{M}$. These constraints ensure that the matrices L_y, and Θ_{yx} belong to appropriate varieties. The constraints in $\mathcal{M}$ along $T(L_y^\star)^\perp$ and $T(\Theta_{yx}^\star)^\perp$ ensure that the tangent spaces $T(L_y)$ and $T(\Theta_{yx})$ are "close" to $T(L_y^\star)$ and $T(\Theta_{yx}^\star)$ respectively. Finally, the last conditions roughly controls the error. We begin by proving the following useful proposition:

Proposition 4.1. *Let (Θ, D_y, L_y) be a set of feasible variables of (1). Let $\Delta = (D_y - D_y^\star, L_y - L_y^\star, \Theta_{yx} - \Theta_{yx}^\star, \Theta_x - \Theta_x^\star)$ and recall that $C_1' = \frac{2\bar{m}m}{\kappa\alpha}\Big(6\psi^2 + \frac{5}{\alpha}\psi^2 + \frac{46\psi^2\kappa}{\alpha} + \kappa\Big) + \frac{1}{\psi^2}$. Then, $\Phi_\gamma[\Delta] \leq C_1'\lambda_n$*

Proof. Let $\mathbb{H}^\star = \mathcal{W} \times T(L_y^\star) \times T(\Theta_{yx}^\star) \times \mathbb{S}^q$. Then,

$$\begin{aligned}
\Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^\star \mathcal{F}\mathcal{P}_{\mathbb{H}^\star}(\Delta)] &\leq \Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^\star \mathcal{F}(\Delta)] + \Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^\star \mathcal{F}\mathcal{P}_{\mathbb{H}^{\star\perp}}(\Delta)] \\
&\leq 6\bar{m}m\psi^2\lambda_n\Big(\frac{8}{\alpha\kappa} + \frac{4}{\alpha} + \frac{1}{\kappa}\Big) \\
&+ m\psi^2\Big(\frac{\omega_y\lambda_n}{2\psi^2} + \frac{\omega_{yx}\lambda_n}{2\psi^2}\Big) \\
&\leq \frac{\bar{m}m\lambda_n}{\kappa}\Big(6\psi^2 + \frac{24}{\alpha}\psi^2 + \frac{48\psi^2\kappa}{\alpha} + \kappa\Big)
\end{aligned}$$

Since $\Phi_\gamma[\mathcal{P}_{\mathbb{H}^\star}(\cdot)] \leq 2\Phi_\gamma(\cdot)$, we have that $\Phi_\gamma[\mathcal{P}_{\mathbb{H}^\star}\mathcal{F}^\dagger \mathbb{I}^\star \mathcal{F}\mathcal{P}_{\mathbb{H}^\star}(\Delta)] \leq \frac{2\bar{m}m\lambda_n}{\kappa\alpha}\Big(6\psi^2 + \frac{24}{\alpha}\psi^2 + \frac{48\psi^2\kappa}{\alpha} + \kappa\Big)$. Consequently, we apply Fisher Information Assumption 1 in (15) (main paper) to conclude that $\Phi_\gamma[\mathcal{P}_{\mathbb{H}^\star}(\Delta)] \leq \frac{2\bar{m}m\lambda_n}{\kappa\alpha}\Big(6\psi^2 + \frac{24}{\alpha}\psi^2 + \frac{48\psi^2\kappa}{\alpha} + \kappa\Big)$. Moreover:

$$\begin{aligned}
\Phi_\gamma[\Delta] \leq \Phi_\gamma[\mathcal{P}_{\mathbb{H}^\star}(\Delta)] + \Phi_\gamma[\mathcal{P}_{\mathbb{H}^{\star\perp}}(\Delta)] &\leq \frac{2\bar{m}m\lambda_n}{\kappa\alpha}\Big(6\psi^2 + \frac{24}{\alpha}\psi^2 + \frac{48\psi^2\kappa}{\alpha} + \kappa\Big) \\
&+ \frac{\lambda_n}{\psi^2} = C_1'\lambda_n
\end{aligned}$$

□

Proposition 4.1 leads to powerful implications. In particular, under additional conditions on the minimum nonzero singular values of $L_y^\star$ and $\Theta_{yx}^\star$, any feasible set of variables (Θ, D_y, L_y) of (1) has two key properties: (*a*) The variables (Θ_{yx}, L_y) are smooth points of the underlying varieties, (*b*) The constraints in $\mathcal{M}$ along $T(L_y^\star)^\perp$ and $T(\Theta_{yx}^\star)^\perp$ are locally inactive at Θ_{yx} and L_y. These properties, among others, are proved in the following corollary.

Corollary 4.2. *Consider any feasible variables (Θ, D_y, L_y) of (1). Let σ_y be the smallest nonzero singular value of $L_y^\star$ and σ_{yx} be the smallest nonzero singular value of $\Theta_{yx}^\star$. Let $\mathbb{H}' = \mathcal{W} \times T(L_y) \times T(\Theta_{yx}) \times \mathbb{S}^q$ and $C_{T'} = \mathcal{P}_{\mathbb{H}'^\perp}(0, L_y^\star, \Theta_{yx}^\star, 0)$. Furthermore, recall that $C_1' = \frac{2\bar{m}m}{\kappa\alpha}\Big(6\psi^2 + \frac{24}{\alpha}\psi^2 + \frac{48\psi^2\kappa}{\alpha} + \kappa\Big) + \frac{1}{\psi^2}$, $C_2' = \frac{4}{\alpha}(1 + \frac{2}{\kappa})$, $C_{\sigma_y}' = C_1'^2\psi^2 \max\{2\kappa + 1, \frac{2}{C_2'\psi^2} + 1\}$ and $C_{\sigma_{yx}}' = C_1'^2\psi^2 \max\{2\kappa + \frac{\kappa}{\gamma}, \frac{2}{C_2'\psi^2} + \frac{\kappa}{\gamma}\}$. Suppose that the following inequalities are met: $\sigma_y \geq \frac{m}{\omega_y}C_{\sigma_y}\lambda_n$,*
$\sigma_{yx} \geq \frac{m\gamma^2}{\omega_{yx}}C_{\sigma_{yx}}'\lambda_n$. Then,

1. *L_y and Θ_{yx} are smooth points of their underlying varieties, i.e.* $\mathrm{rank(L_y)} = \mathrm{rank(L_y^\star)}$, $\mathrm{rank(\Theta_{yx})} = \mathrm{rank(\Theta_{yx}^\star)}$*; Moreover L_y has the same inertia as $L_y^\star$.*
2. *$\|\mathcal{P}_{T(L_y^\star)^\perp}(L_y - L_y^\star)\|_2 \leq \frac{\lambda_n}{48m\psi^2}$ and $\|\mathcal{P}_{T(\Theta_{yx}^\star)^\perp}(\Theta_{yx} - \Theta_{yx}^\star)\|_2 \leq \frac{\lambda_n}{48m\psi^2}$*
3. *$\rho(T(L_y), T(L_y^\star)) \leq \omega_y$; $\rho(T(\Theta_{yx}), T(\Theta_{yx}^\star)) \leq \omega_{yx}$; that is, the tangent spaces at L_y and Θ_{yx} is "close" to the tangent space $L_y^\star$ and $\Theta_{yx}^\star$.*
4. *$\Phi_\gamma[C_{T'}] \leq \min\{\frac{\lambda_n}{\kappa\psi^2}, C_2'\lambda_n\}$*

Proof. We note the following relations before proving each step: $C'_1 \geq \frac{1}{\psi^2} \geq \frac{1}{m\psi^2}$, $\omega_y, \omega_{yx} \in (0,1)$, and $\kappa \geq 6$. We also appeal to the results of regarding perturbation analysis of the low-rank matrix variety [1].

1. Based on the assumptions regarding the minimum nonzero singular values of $L_y^\star$ and $\Theta_{yx}^\star$, one can check that:

$$\begin{aligned} \sigma_y &\geq \frac{C_1'^2\lambda_n}{\omega_y} m\psi^2(\kappa+1) \geq \frac{C_1'\lambda_n}{\omega_y}(2\kappa+1) \geq 8\|L - L_y^\star\|_2 \\ \sigma_{yx} &\geq \frac{C_1'^2\lambda_n}{\omega_{yx}}\gamma^2 m\psi^2\Big(\frac{6\beta}{\gamma}+2\kappa\Big) \geq 8\|\Theta_{yx} - \Theta_{yx}^\star\|_2 \end{aligned}$$

Combining these results and Proposition 4.1, we conclude that L_y and Θ_{yx} are smooth points of their respective varieties, i.e. $\text{rank}(L_y) = \text{rank}(L_y^\star)$, and $\text{rank}(\Theta_{yx}) = \text{rank}(\Theta_{yx}^\star)$. Furthermore, L_y has the same inertia as $L_y^\star$.

2. Since $\sigma_y \geq 8\|L_y - L_y^\star\|_2$, and $\sigma_{yx} \geq 8\|\Theta_{yx} - \Theta_{yx}^\star\|_2$, we can appeal to Proposition 2.2 of [2] to conclude that the constraints in $\mathcal{M}$ along $\mathcal{P}_{T(L_y^\star)^\perp}$ and $\mathcal{P}_{T(\Theta_{yx}^\star)^\perp}$ are strictly feasible:

$$\begin{aligned} \|\mathcal{P}_{T(L_y^\star)^\perp}(L_y - L_y^\star)\|_2 &\leq \frac{\|L_y - L_y^\star\|_2^2}{\sigma_y} \leq \frac{\lambda_n}{48m\psi^2} \\ \|\mathcal{P}_{T(\Theta_{yx}^\star)^\perp}(\Theta_{yx} - \Theta_{yx}^\star)\|_2 &\leq \frac{\|\Theta_{yx} - \Theta_{yx}^\star\|_2^2}{\sigma_{yx}} \leq \frac{\lambda_n}{48m\psi^2} \end{aligned}$$

3. Appealing to Proposition 2.1 of [2], we prove that the tangent spaces $T(L_y)$ and $T(\Theta_{yx})$ are close to $T(L_y^\star)$ and $T(\Theta_{yx}^\star)$ respectively:

$$\begin{aligned} \rho(T(L_y), T(L_y^\star)) &\leq \frac{2\|L_y - L_y^\star\|_2}{\sigma_y} \leq \frac{2C_1'\lambda_n\omega_y}{C_1'^2\lambda_n m\psi^2(2\kappa+1)} \leq \omega_y \\ \rho(T(\Theta_{yx}), T(\Theta_{yx}^\star)) &\leq \frac{2\|\Theta_{yx} - \Theta_{yx}^\star\|_2}{\sigma_{yx}} \leq \frac{2C_1'\lambda_n\gamma\omega_{yx}}{\frac{C_1'^2\lambda_n}{\omega_{yx}}\gamma^2 m\psi^2\Big(\frac{\kappa}{\gamma}+2\kappa\Big)} \leq \omega_{yx} \end{aligned}$$

4. Letting σ'_y and σ'_{yx} be the minimum nonzero singular value of L_y and Θ_{yx} respectively, one can check that:

$$\sigma'_y \geq \sigma_y - \|L_y - L_y^\star\|_2 \geq 8C_1'\lambda_n \geq 8\|L_y - L_y^\star\|_2$$

$$\sigma'_{yx} \geq \sigma_{yx} - \|\Theta_{yx} - \Theta_{yx}^\star\|_2 \geq 8C_1'\lambda_n\gamma \geq 8\|\Theta_{yx} - \Theta_{yx}^\star\|_2$$

Once again appealing to Proposition 2.2 of [2] and simple algebra, we have:

$$\begin{aligned} \Phi_\gamma(C_{T'}) &\leq m\|\mathcal{P}_{T(L_y)^\perp}(L_y - L_y^\star)\|_2 + m\|\mathcal{P}_{T(\Theta_{yx})^\perp}(\Theta_{yx} - \Theta_{yx}^\star)\|_2 \\ &\leq m\frac{\|L_y - L_y^\star\|_2^2}{\sigma'_y} + m\frac{\|\Theta_{yx} - \Theta_{yx}^\star\|_2^2}{\sigma'_{yx}} \leq \min\Big\{\frac{\lambda_n}{\kappa\psi^2}, C_2'\lambda_n\Big\} \end{aligned}$$

□

4.2 Variety Constrained Program to Tangent Space Constrained Program

Consider any optimal solution $(\Theta^{\mathcal{M}}, D_y^{\mathcal{M}}, L_y^{\mathcal{M}})$ of (1). In Corollary 4.2, we concluded that the variables $(\Theta_{yx}^{\mathcal{M}}, L_y^{\mathcal{M}})$ are smooth points of their respective varieties. As a result, the rank constraints $\text{rank}(\text{L}_\text{y}) \leq \text{rank}(\text{L}_\text{y}^\star)$ and $\text{rank}(\Theta_\text{yx}) \leq \text{rank}(\Theta_\text{yx}^\star)$ can be "linearized" to $L_y \in T(L_y^{\mathcal{M}})$ and $\Theta_{yx} \in T(\Theta_{yx}^{\mathcal{M}})$ respectively. Since all the remaining constraints are convex, the optimum of this linearized program is also the optimum of (1). Moreover, we once more appeal to Corollary 4.2 to conclude that the constraints in $\mathcal{M}$ along $\mathcal{P}_{T(L_y^\star)^\perp}$ and $\mathcal{P}_{T(\Theta_{yx}^\star)^\perp}$ are strictly feasible at $(\Theta^{\mathcal{M}}, D_y^{\mathcal{M}}, L_y^{\mathcal{M}})$. As a result, these constraints are locally inactive and can be removed without changing the optimum. Therefore the constraint $(\Theta^{\mathcal{M}}, D_y^{\mathcal{M}}, L_y^{\mathcal{M}}) \in \mathcal{M}_1$ is inactive and can be removed. We now argue that the constraint $(\Theta^{\mathcal{M}}, D_y^{\mathcal{M}}, L_y^{\mathcal{M}}) \in \mathcal{M}_2$ in (1) can also removed in this "linearized" convex program. In particular, letting $\mathbb{H}_{\mathcal{M}} \triangleq \mathcal{W} \times T(L_y^{\mathcal{M}}) \times T(\Theta_{yx}^{\mathcal{M}}) \times \mathbb{S}^q$, consider the following convex optimization program:

$$(\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y) = \underset{\substack{\Theta \in \mathbb{S}^{q+p},\ \Theta \succ 0 \\ D_y, L_y \in \mathbb{S}^p}}{\text{argmin}} \quad -\ell(\Theta; \mathcal{D}_n^+) + \lambda_n [\|L_y\|_\star + \gamma \|\Theta_{yx}\|_\star]$$
$$\text{s.t.} \quad \Theta_y = D_y - L_y,\ (D_y, L_y, \Theta_{yx}, \Theta_x) \in \mathbb{H}_{\mathcal{M}} \tag{2}$$

We prove that under conditions imposed on the regularization parameter λ_n, the pair of variables $(\Theta^{\mathcal{M}}, D_y^{\mathcal{M}}, L_y^{\mathcal{M}})$ is the unique optimum of (2). That is, we show that

1. $\|\mathbb{I}^\star \mathcal{F}(\Delta)\|_2 < 6\bar{m}\psi^2 \lambda_n \left(\frac{8}{\alpha\kappa} + \frac{4}{\alpha} + \frac{1}{\kappa}\right)$

Appealing to Corollary 4.2 and Proposition 4 (main paper), we have that $\Phi_\gamma[\mathcal{F}^\dagger \mathbb{I}^\star \mathcal{F} C_{T_{\mathcal{M}}}] \leq \frac{\lambda_n}{\kappa}$, $\Phi_\gamma[C_{T_{\mathcal{M}}}] \leq C_2' \lambda_n$ and (with high probability) $\Phi_\gamma[\mathcal{F}^\dagger E_n] \leq \frac{\lambda_n}{\kappa}$. Consequently, based on the bound on λ_n in assumption of Theorem 1 (main paper), it is straightforward to show that $r_1^u \leq \min\left\{\frac{1}{4C'}, \frac{\alpha}{32 \max\{1+\frac{\kappa}{2}, \frac{\alpha}{8}\}^2 m\psi C'^2}\right\}$ so that $\Phi_\gamma[\Delta] \leq \frac{1}{2C'}$. Hence by Proposition 2 (main paper), we have that $\|\Delta_1\|_2, \|\Delta_4\|_2 \leq r_2^u < r_1^u$, $\|\Delta_2\|_2 \leq 2r_1^u$ and $\|\Delta\|_3 \leq 2\gamma r_1^u$. Therefore:

$$\begin{aligned} \|\mathbb{I}^\star \mathcal{F}(\Delta)\|_2 &\leq \psi^2(\|\Delta_1\|_2 + \|\Delta_2\|_2 + \|\Delta_3\|_2 + \|\Delta_4\|_2) \\ &< 6\bar{m}\psi^2 r_1^u \leq 6\bar{m}\psi^2 \lambda_n \Big(\frac{8}{\alpha\kappa} + \frac{4}{\alpha} + \frac{1}{\kappa}\Big) \end{aligned}$$

4.3 From Tangent Space Constraints to the Original Problem

The optimality conditions of (2) suggest that there exist Lagrange multipliers $Q_{D_y} \in \mathcal{W}$, $Q_{T_y} \in T(L_y^{\mathcal{M}})^\perp$, and $Q_{T_{yx}} \in T(\Theta_{yx}^{\mathcal{M}})^\perp$ such that

$$\begin{aligned} [\Sigma_n - \tilde{\Theta}^{-1}]_y + Q_{D_y} = 0; \quad [\Sigma_n - \tilde{\Theta}^{-1}]_y + Q_{T_y} &\in \lambda_n \partial \|\tilde{L}_y\|_\star \\ [\Sigma_n - \tilde{\Theta}^{-1}]_{yx} + Q_{T_{yx}} \in -\lambda_n \gamma \partial \|\tilde{\Theta}_{yx}\|_\star; \quad [\Sigma_n - \tilde{\Theta}^{-1}]_x &= 0 \end{aligned}$$

Letting the SVD of $\tilde{L}_y$ and $\tilde{\Theta}_{yx}$ be given by $\tilde{L}_y = \bar{U}\bar{O}\bar{V}'$ and $\tilde{\Theta}_{yx} = \breve{U}\breve{O}\breve{V}'$ respectively, and $Z \triangleq (0,\ \lambda_n \bar{U}\bar{V}',\ -\lambda_n \gamma \breve{U}\breve{V}',\ 0)$, we can restrict the optimality conditions to the space $\mathbb{H}_{\mathcal{M}}$ to obtain, $\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{F}^\dagger(\Sigma_n - \tilde{\Theta}^{-1}) = Z$. We proceed by proving that the variables $(\tilde{\Theta}, \tilde{D}_y, \tilde{L}_y)$ satisfy the optimality conditions of the original convex program (4) (main paper). That is:

1. $\mathcal{P}_{\mathbb{H}_{\mathcal{M}}} \mathcal{F}^\dagger(\Sigma_n - \tilde{\Theta}^{-1}) = Z$
2. $\max\left\{\|\mathcal{P}_{T_y'^\perp}(\Sigma_n - \tilde{\Theta}^{-1})_y\|_2, \frac{1}{\gamma}\|\mathcal{P}_{T_{yx}'^\perp}(\Sigma_n - \tilde{\Theta}^{-1})_{yx}\|_2\right\} < \lambda_n$

It is clear that the first condition is satisfied since the pair $(\tilde{\Theta}, \tilde{S}_y, \tilde{L}_y)$ is optimum for (2). To prove that the second condition, we must prove that $\Gamma_\gamma[\mathcal{P}_{\mathbb{H}^\perp_{\mathcal{M}}[2,3]}\mathcal{G}^\dagger(\Sigma_n - \tilde{\Theta}^{-1})] < \lambda_n$. In particular, denoting $\Delta = (\tilde{D}_y - D^\star_y, \tilde{L}_y - L^\star_y, \tilde{\Theta}_{yx} - \Theta^\star_{yx}, \tilde{\Theta}_x - \Theta^\star_x)$ we show that:

$$\begin{aligned} \Gamma_\gamma[\mathcal{P}_{\mathbb{H}^\perp_{\mathcal{M}}[2,3]}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}_{\mathcal{M}}[2,3]}(\Delta)] &< \lambda_n - \Phi_\gamma[\mathcal{P}_{\mathbb{H}^\perp_{\mathcal{M}}}\mathcal{F}^\dagger E_n] \\ &- \Phi_\gamma[\mathcal{P}_{\mathbb{H}^\perp_{\mathcal{M}}}\mathcal{F}^\dagger R_{\Sigma^\star}(\mathcal{F}(\Delta))] \\ &- \Phi_\gamma[\mathcal{P}_{\mathbb{H}^\perp_{\mathcal{M}}}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}C_{T_{\mathcal{M}}}] \\ &- \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}[2,3]^\perp}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{F}(\Delta_1, 0, 0, \Delta_4)] \end{aligned} \tag{3}$$

Using the fact that $\Gamma_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}[2,3]^\perp}\mathcal{G}^\dagger(N)] \le \Phi_\gamma[\mathcal{P}_{\mathbb{H}^\perp_{\mathcal{M}}}\mathcal{F}^\dagger(N)]$ for any matrix $N \in \mathbb{S}^{p+q}$, this would in turn imply that:

$$\begin{aligned} \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}[2,3]^\perp}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}_{\mathcal{M}}[2,3]}(\Delta)] &< \lambda_n - \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}[2,3]^\perp}\mathcal{G}^\dagger E_n] \\ &- \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}[2,3]^\perp}\mathcal{G}^\dagger R_{\Sigma^\star}(\mathcal{F}(\Delta))] \\ &- \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}[2,3]^\perp}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{F}C_{T_{\mathcal{M}}}] \\ &- \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}[2,3]^\perp}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{F}(\Delta_1, 0, 0, \Delta_4)] \end{aligned} \tag{4}$$

Indeed (4) implies that the second optimality condition is satisfied. So we focus on showing that (4) is satisfied. Since $\Phi_\gamma[\Delta] \le \frac{1}{2C'}$, we can appeal to Proposition 1 (main paper) and the bound on λ_n to conclude $\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\mathcal{F}(\Delta))] \le 2m\psi C'^2\Phi_\gamma[\Delta]^2 \le 2m\psi C'^2 C_1'^2\lambda_n^2 \le \frac{\lambda_n}{\kappa}$. Using the first optimality condition, the fact that projecting into tangent spaces with respect to rank variety increase the spectral norm by at most a factor of two (i.e. $\Phi_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}}(\cdot)] \le 2\Phi_\gamma(\cdot)$), the fact that $\Gamma_\gamma[\mathcal{G}^\dagger(\cdot)] \le \Phi_\gamma[\mathcal{F}^\dagger(\cdot)]$, and that $\kappa = \beta(6 + \frac{16\psi^2 m}{\alpha})$, we have that:

$$\begin{aligned} \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}[2,3]}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}_{\mathcal{M}}[2,3]}(\Delta)] &\le \lambda_n + 2\Gamma_\gamma[\mathcal{G}^\dagger R_{\Sigma^\star}(\Delta)] + 2\Gamma_\gamma[\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{F}C_{T_M}] \\ &+ 2\Gamma_\gamma[\mathcal{G}^\dagger E_n] + \Gamma_\gamma[\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{F}(\Delta_1, 0, 0, \Delta_4)] \\ &\le \lambda_n + 2\Phi_\gamma[\mathcal{F}^\dagger R_{\Sigma^\star}(\Delta)] + 2\Phi_\gamma[\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}C_{T_M}] \\ &+ 2\Phi_\gamma[\mathcal{F}^\dagger E_n] + \Phi_\gamma[\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}(\Delta_1, 0, 0, \Delta_4)] \\ &\le \lambda_n + \frac{\lambda_n}{\beta} \end{aligned}$$

Applying Fisher Information Assumption 2 in (16) (main paper), we obtain:

$$\begin{aligned} \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}[2,3]^\perp}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{G}\mathcal{P}_{\mathbb{H}_{\mathcal{M}}[2,3]}(\Delta)] &\le \frac{(\beta+1)\lambda_n}{\beta}\Big(1 - \frac{2}{\beta+1}\Big) = \lambda_n - \frac{\lambda_n}{\beta} \\ &< \lambda_n - \frac{\lambda_n}{2\beta} \\ &\le \lambda_n - \Phi_\gamma[\mathcal{F}^\dagger R_\Sigma(\mathcal{F}(\Delta))] - \Phi_\gamma[\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}C_{T_{\mathcal{M}}}] \\ &- \Phi_\gamma[\mathcal{F}^\dagger E_n] - \Gamma_\gamma[\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{F}(\Delta_1, 0, 0, \Delta_4)] \\ &\le \lambda_n - \Phi_\gamma[\mathcal{P}_{\mathbb{H}^\perp_{\mathcal{M}}}\mathcal{F}^\dagger R_{\Sigma^\star}(\mathcal{F}(\Delta))] \\ &- \Phi_\gamma[\mathcal{P}_{\mathbb{H}^\perp_{\mathcal{M}}}\mathcal{F}^\dagger\mathbb{I}^\star\mathcal{F}C_{T_{\mathcal{M}}}] \\ &- \Phi_\gamma[\mathcal{P}_{\mathbb{H}^\perp_{\mathcal{M}}}\mathcal{F}^\dagger E_n] \\ &- \Gamma_\gamma[\mathcal{P}_{\mathbb{H}_{\mathcal{M}}[2,3]^\perp}\mathcal{G}^\dagger\mathbb{I}^\star\mathcal{F}(\Delta_1, 0, 0, \Delta_4)] \end{aligned}$$

Here, we used the fact that $\|\mathcal{P}_{T^\perp}(.)\|_2 \le \|.\|_2$ for a tangent space T of the low-rank matrix variety.

5 Proof of Proposition 4 (main paper)

We must study the rate of convergence of the sample covariance matrix to the population covariance matrix. The following result from [3] plays a key role in obtaining this result.

Proposition 5.1. *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 $\sigma_1(\Gamma)$ and $\sigma_p(\Gamma)$ of Γ are such that:*

$$\max\left\{Prob[\sigma_1(\Gamma) \leq 1 + \sqrt{\frac{p}{n}} + t], Prob[\sigma_p(\Gamma) \leq 1 - \sqrt{\frac{p}{n}} - t]\right\}.$$

We now proceed with proving Proposition 4 (main paper). First, note that $\Phi_\gamma[\mathcal{F}^\dagger E_n] \leq m\|\Sigma_n - \Sigma^\star\|_2$. Using Proposition 5.1 and the fact that $\frac{\lambda_n}{m\kappa} \leq 8\psi$ and $n \geq \frac{64\kappa^2(p+q)m^2\psi^2}{\lambda_n^2}$, the following bound holds: $\Pr[m\|\Sigma_n - \Sigma^\star\|_2 \geq \frac{\lambda_n}{\kappa}] \leq 2\exp\left\{-\frac{n\lambda_n^2}{128\kappa^2 m^2\psi^2}\right\}$. Thus, $\Phi_\gamma[\mathcal{F}^\dagger E_n] \leq \frac{\lambda_n}{\kappa}$ with probability greater than $1 - 2\exp\left\{-\frac{n\lambda_n^2}{128\kappa^2 m^2\psi^2}\right\}$.

6 Consistency of the Convex Program (18) (main paper)

In this section, we prove the consistency of convex program (18) (main paper) for estimating a factor model. We first introduce some notation. We define the linear operator: $\tilde{\mathcal{F}} : \mathbb{S}^p \times \mathbb{S}^p \to \mathbb{S}^p$ and its adjoint $\tilde{\mathcal{F}}^\dagger : \mathbb{S}^p \to \mathbb{S}^p \times \mathbb{S}^p$ as follows:

$$\tilde{\mathcal{F}}(M, K) \triangleq M - K, \qquad \tilde{\mathcal{F}}^\dagger(Q) \triangleq (Q, Q). \tag{5}$$

We consider a population composite factor model (3) (main paper) $y = \mathcal{A}^\star x + \mathcal{B}_u^\star \zeta_u + \epsilon$ underlying a pair of random vectors $(y, x) \in \mathbb{R}^p \times \mathbb{R}^q$, with $\text{rank}(\mathcal{A}^\star) = k_x$, $\mathcal{B}_u^\star \in \mathbb{R}^{p \times k_u}$, and column-space$(\mathcal{A}^\star) \cap$ column-space$(\mathcal{B}_u^\star) = \{0\}$. As the convex relaxation (18) (main paper) is solved in the precision matrix parametrization, the conditions for our theorems are more naturally stated in terms of the joint precision matrix $\Theta^\star \in \mathbb{S}^{p+q}$, $\Theta^\star \succ 0$ of (y, x). The algebraic aspects of the parameters underlying the factor model translate to algebraic properties of submatrices of $\Theta^\star$. In particular, the submatrix $\Theta^\star_{yx}$ has rank equal to k_x, and the submatrix $\Theta^\star_y$ is decomposable as $D^\star_y - L^\star_y$ with $D^\star_y$ being diagonal and $L^\star_y \succeq 0$ having rank equal to k_u. Finally, the transversality of column-space$(\mathcal{A}^\star)$ and column-space$(\mathcal{B}_u^\star)$ translates to the fact that column-space$(\Theta^\star_{yx}) \cap$ column-space$(L^\star_y) = \{0\}$ have a transverse intersection. We consider the factor model underlying the random vector $y \in \mathbb{R}^p$ that is induced upon marginalization of x. In particular, the precision matrix of y is given by $\tilde{\Theta}^\star_y = D^\star_y - [L^\star_y + \Theta^\star_{yx}(\Theta^\star_x)^{-1}\Theta^\star_{xy}]$. To learn an accurate factor model, we seek an estimate $(\hat{\tilde{D}}_y, \hat{\tilde{L}}_y)$ from the convex program (18) (main paper) such that $\text{rank}(\hat{\tilde{L}}_y) = \text{rank}(L^\star_y + \Theta^\star_{yx}{\Theta^\star_x}^{-1}\Theta^\star_{xy})$, and the errors $\|\hat{\tilde{D}}_y - D^\star_y\|_2, \|\hat{\tilde{L}}_y - [L^\star_y + \Theta^\star_{yx}(\Theta^\star_x)^{-1}\Theta^\star_{xy}]\|_2$ are small.

Following the same reasoning as the Fisher information conditions for consistency of the convex program (4) (main paper), A natural set of conditions on the population Fisher information at $\tilde{\Theta}^\star_y$

defined as $\mathbb{I}_y^\star = (\tilde{\Theta}_y^\star)^{-1} \otimes (\tilde{\Theta}_y^\star)^{-1}$ are given by:

$$\text{Assumption 4} \quad : \quad \inf_{\mathbb{H}' \in \tilde{U}(\tilde{\omega}_y)} \tilde{\chi}(\mathbb{H}', \tilde{\Phi}) \geq \tilde{\alpha}, \quad \text{for some } \tilde{\alpha} > 0 \tag{6}$$

$$\text{Assumption 5} \quad : \quad \inf_{\mathbb{H}' \in \tilde{U}(\tilde{\omega}_y)} \tilde{\Xi}(\mathbb{H}') > 0 \tag{7}$$

$$\text{Assumption 6} \quad : \quad \sup_{\mathbb{H}' \in \tilde{U}(\tilde{\omega}_y)} \tilde{\varphi}(\mathbb{H}') \leq 1 - \frac{2}{\tilde{\beta}+1} \quad \text{for some } \tilde{\beta} \geq 2, \tag{8}$$

where,

$$\begin{aligned}
\tilde{\chi}(\mathbb{H}, \|.\|_\Upsilon) &\triangleq \min_{\substack{Z \in \mathbb{H} \\ \|Z\|_\Upsilon = 1}} \|\mathcal{P}_{\mathbb{H}} \tilde{\mathcal{I}}^\dagger \mathbb{I}_y^\star \tilde{\mathcal{I}} \mathcal{P}_{\mathbb{H}}(Z)\|_\Upsilon \\
\tilde{\Xi}(\mathbb{H}) &\triangleq \min_{\substack{Z \in \mathbb{H}[2] \\ \|Z\|_2 = 1}} \|\mathcal{P}_{\mathbb{H}[2]} \mathbb{I}_y^\star \mathcal{P}_{\mathbb{H}[2]}(Z)\|_2 \\
\tilde{\varphi}(\mathbb{H}) &\triangleq \max_{\substack{Z \in \mathbb{H}[2] \\ \|Z\|_2 = 1}} \|\mathcal{P}_{\mathbb{H}^\perp[2]} \mathbb{I}_y^\star \mathcal{P}_{\mathbb{H}[2]} (\mathcal{P}_{\mathbb{H}[2]} \mathbb{I}_y^\star \mathcal{P}_{\mathbb{H}[2]})^{-1}(Z)\|_2 \\
\tilde{U}(\tilde{\omega}_y) &\triangleq \Big\{ \mathcal{W} \times T' \mid \rho(T', T(L_y^\star + \Theta_{yx}^\star (\Theta_x^\star)^{-1} \Theta_{xy}^\star)) \leq \tilde{\omega}_y \Big\} \\
\tilde{\Phi}(D, L) &\triangleq \max \{\|D\|_2, \|L\|_2\}.
\end{aligned}$$

Assumption 4 controls the gain of the Fisher information $\mathbb{I}_y^\star$ restricted to appropriate subspaces and Assumption 5 and 6 are in the spirit of irrepresentability conditions. As the variety of low-rank matrices is locally curved around $T(L_y^\star + \Theta_{yx}^\star(\Theta_x^\star)^{-1}\Theta_{xy}^\star)$, we control the Fisher information $\mathbb{I}_y^\star$ at nearby tangent spaces T' where $\rho(T', T(L_y^\star + \Theta_{yx}^\star(\Theta_x^\star)^{-1}\Theta_{xy}^\star)) \leq \tilde{\omega}_y$. We also note that measuring the gains of Fisher information $\mathbb{I}_y^\star$ with the norm $\tilde{\Phi}$ and $\|\cdot\|_2$ is natural as these are closely tied with dual norm of the regularizer trace($\tilde{L}_y$) in (18) (main paper).

We present a theorem of consistency of the convex relaxation (18) (main paper) under Assumptions 4, 5 and 6. We let σ denote the minimum nonzero singular value of $L_y^\star + \Theta_{yx}^\star(\Theta_x^\star)^{-1}\Theta_{xy}^\star$. The proof strategy is similar in spirit to the strategy for proving the consistency of the convex relaxation (4) (main paper).

Theorem 6.1. *Suppose that there exists $\tilde{\alpha} > 0$, $\tilde{\beta} \geq 2$, $\tilde{\omega}_y \in (0,1)$ so that the population Fisher information $\mathbb{I}_y^\star$ satisfies Assumptions 4, 5 and 6 in* (6), (7), *and* (8). *Suppose that the following conditions hold:*

1. $n \gtrsim \left[\frac{\tilde{\beta}^2}{\tilde{\alpha}^2}\right](p)$
2. $\tilde{\lambda}_n \sim \frac{\tilde{\beta}}{\tilde{\alpha}}\sqrt{\frac{p}{n}}$
3. $\sigma \gtrsim \frac{\tilde{\beta}}{\tilde{\alpha}^5 \tilde{\omega}_y}\tilde{\lambda}_n$

Then with probability greater than $1 - 2\exp\Big\{-C\frac{\tilde{\alpha}}{\tilde{\beta}} n \tilde{\lambda}_n^2\Big\}$, the optimal solution $(\hat{\Theta}, \hat{\tilde{D}}_y, \hat{\tilde{L}}_y)$ of (18) (main paper) with i.i.d. observations $\mathcal{D}_n = \{y^{(i)}\}_{i=1}^n$ satisfies the following properties:

1. *rank($\hat{\tilde{L}}_y$) = rank($L_y^\star + \Theta_{yx}^\star(\Theta_x^\star)^{-1}\Theta_{xy}^\star$)*
2. $\|\hat{\tilde{D}}_y - D_y^\star\|_2 \lesssim \frac{\tilde{\lambda}_n}{\tilde{\alpha}^2}$, $\|\hat{\tilde{L}}_y - L_y^\star - \Theta_{yx}^\star(\Theta_x^\star)^{-1}\Theta_{xy}^\star\|_2 \lesssim \frac{\tilde{\lambda}_n}{\tilde{\alpha}^2}$

References

[1] Bach, F. Consistency of trace norm minimization. *Journal of Machine Learning Research*, 9:1019–1048, 2008.

[2] Chandrasekaran, V., Parrilo, P. A. & Willsky, A. S.. Latent Variable Graphical Model Selection via Convex Optimization. *Annals of Statistics*, 40:1935–1967, 2012.

[3] Davidson, K.R. & Szarek, S.J. Local operator theory, random matrices and Banach spaces. *Handbook of the Geometry of Banach Spaces*, 1:317–366, 2001.

[4] Kato, T. Perturbation theory for linear operators. *Springer*, 1995.

[5] Ravikumar, P., Wainwright, M. J., Raskutti, G. & Yu, B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 4:935–980, 2011.