Targeted Principal Components Regression

Karl Oskar Ekvall[1, 2]

[1]Applied Statistics Research Unit, Institute of Statistics and Mathematical Methods in Economics, Faculty of Mathematics and Geoinformation, TU Wien

[2]Division of Biostatistics, Institute of Environmental Medicine Karolinska Institute
Nobels väg 13, 171 77 Stockholm, Sweden

karl.oskar.ekvall@ki.se

1 Preliminaries

Recall the functions

$$g(\theta) = g(\theta; y, x) = -\log|\Omega| + (y - \beta^\mathsf{T} x)^\mathsf{T}\Omega(y - \beta^\mathsf{T} x) + \log|\tau I_p + \Psi| + x^\mathsf{T}(\tau I_p + \Psi)^{-1}x,$$

$$\begin{aligned} G(\theta) = G(\theta; \theta_*) = -\log|\Omega| + \operatorname{tr}\{(\beta - \beta_*)^\mathsf{T}(\tau_* I_p + \Psi_*)(\beta - \beta_*)\Omega\} + \operatorname{tr}(\Omega\Omega_*^{-1}) \\ + \log|\tau I_p + \Psi| + \operatorname{tr}\{(\tau I_p + \Psi)^{-1}(\tau_* I_p + \Psi_*)\}, \end{aligned}$$

and $G_n(\theta) = G_n(\theta; Y_1, X_1, \dots, Y_n, X_n) = n^{-1}\sum_{i=1}^n g(\theta; Y_i, X_i)$. Let $\Sigma_X = \tau I_p + \Psi$; then

$$G_n(\theta) = -\log|\Omega| + n^{-1}\operatorname{tr}\{(Y - X\beta)^\mathsf{T}(Y - X\beta)\Omega\} + \log|\Sigma_X| + n^{-1}\operatorname{tr}(X^\mathsf{T} X \Sigma_X^{-1}).$$

The gradient $\nabla g = \nabla_\theta g$ is characterized by

$$\begin{aligned} \nabla_\beta g(\theta) &= -2x(y - \beta^\mathsf{T} x)^\mathsf{T}\Omega; \\ \nabla_\Omega g(\theta) &= -\Omega^{-1} + (y - \beta^\mathsf{T} x)(y - \beta^\mathsf{T} x)^\mathsf{T}; \\ \nabla_\tau g(\theta) &= \operatorname{tr}(\Sigma_X^{-1} - \Sigma_X^{-1} x x^\mathsf{T} \Sigma_X^{-1}); \\ \nabla_\Psi g(\theta) &= \Sigma_X^{-1} - \Sigma_X^{-1} x x^\mathsf{T} \Sigma_X^{-1}. \end{aligned}$$

By differentiating these expressions, one gets that the non-zero blocks of the Hessian $\nabla^2 g$

are given by

$$
\begin{aligned}
\nabla^2_{\beta} g(\theta) &= 2\Omega \otimes xx^{\mathsf{T}};\\
\nabla^2_{\beta\Omega} g(\theta) &= -2I_r \otimes x(y-\beta^{\mathsf{T}}x)^{\mathsf{T}};\\
\nabla^2_{\Omega} g(\theta) &= \Omega^{-1} \otimes \Omega^{-1};\\
\nabla^2_{\tau} g(\theta) &= \mathrm{tr}\{-\Sigma_X^{-2} + \Sigma_X^{-2}xx^{\mathsf{T}}\Sigma_X^{-1} + \Sigma_X^{-1}xx^{\mathsf{T}}\Sigma_X^{-2}\};\\
\nabla^2_{\tau\Psi} g(\theta) &= -\Sigma_X^{-2} + \Sigma_X^{-2}xx^{\mathsf{T}}\Sigma_X^{-1} + \Sigma_X^{-1}xx^{\mathsf{T}}\Sigma_X^{-2};\\
\nabla^2_{\Psi} g(\theta) &= -\Sigma_X^{-1} \otimes \Sigma_X^{-1} + \Sigma_X^{-1} \otimes \Sigma_X^{-1}xx^{\mathsf{T}}\Sigma_X^{-1} + \Sigma_X^{-1}xx^{\mathsf{T}}\Sigma_X^{-1} \otimes \Sigma_X^{-1}.
\end{aligned}
$$

The gradient ∇G is characterized by

$$
\begin{aligned}
\nabla_{\beta} G(\theta) &= 2\Sigma_{X*}(\beta-\beta_*)\Omega;\\
\nabla_{\Omega} G(\theta) &= -\Omega^{-1} + (\beta-\beta_*)^{\mathsf{T}}\Sigma_{X*}(\beta-\beta_*) + \Omega_*^{-1};\\
\nabla_{\tau} G(\theta) &= \mathrm{tr}\{\Sigma_X^{-1} - \Sigma_X^{-1}\Sigma_{X*}\Sigma_X^{-1}\};\\
\nabla_{\Psi} G(\theta) &= \Sigma_X^{-1} - \Sigma_X^{-1}\Sigma_{X*}\Sigma_X^{-1}.
\end{aligned}
$$

The non-zero blocks of the Hessian $\nabla^2 G$ are given by

$$
\begin{aligned}
\nabla^2_{\beta} G(\theta) &= 2\Omega \otimes \Sigma_{X*};\\
\nabla^2_{\beta\Omega} G(\theta) &= 2I_r \otimes \Sigma_{X*}(\beta-\beta_*);\\
\nabla^2_{\Omega} G(\theta) &= \Omega^{-1} \otimes \Omega^{-1};\\
\nabla^2_{\tau} G(\theta) &= \mathrm{tr}\{-\Sigma_X^{-2} + \Sigma_X^{-2}\Sigma_{X*}\Sigma_X^{-1} + \Sigma_X^{-1}\Sigma_{X*}\Sigma_X^{-2}\};\\
\nabla^2_{\tau\Psi} G(\theta) &= -\Sigma_X^{-2} + \Sigma_X^{-2}\Sigma_{X*}\Sigma_X^{-1} + \Sigma_X^{-1}\Sigma_{X*}\Sigma_X^{-2};\\
\nabla^2_{\Psi} G(\theta) &= -\Sigma_X^{-1} \otimes \Sigma_X^{-1} + \Sigma_X^{-1} \otimes \Sigma_X^{-1}\Sigma_{X*}\Sigma_X^{-1} + \Sigma_X^{-1}\Sigma_{X*}\Sigma_X^{-1} \otimes \Sigma_X^{-1}.
\end{aligned}
$$

2 Proofs

2.1 Proof of Proposition 2.1

Consider an enlarged parameter set Θ_1 where Ψ is of rank at most k and has a spectral decomposition UDU^{T}, $D \in \mathbb{R}^{k\times k}$, such that $\beta = U\gamma$ for some $\gamma \in \mathbb{R}^{k\times r}$; that is, β is in the column space of U, but not necessarily in that of Ψ if some diagonal elements of D are equal to zero. Clearly, G_n can be defined the same way on Θ_1 as on Θ. The enlarged parameter set is useful because it is closed. To see this, pick a convergent sequence $\{\theta_m\} \in \Theta_1$; we need to show the limit point $\theta \in \Theta_1$. It straightforward to show Ω, τ, and Ψ must be symmetric and positive semi-definite, so we omit the details. To see the rank of Ψ is at most k, suppose for contradiction it has $k+1$ strictly positive eigenvalues. Then by Weyl's perturbation theorem (Bhatia, 2012, Corollary III.2.6), so does Ψ_m for all large enough m, which is a contradiction to $\Psi_m \in \mathcal{S}$. Because $\Psi \in \mathcal{S}$, we can write $\Psi = UDU^{\mathsf{T}}$ and it remains to show $P_U\beta = \beta$. To that end, write $\Psi_m = U_m D_m U_m^{\mathsf{T}}$ and $\beta_m = U_m\gamma_m$ for every m. Since $\|\beta_m\| = \|\gamma_m\|$ and

β_m converges, $\{\gamma_m\}$ is bounded. Similarly, $\{D_m\}$ is bounded and $\{U_m\}$ is a sequence in the compact set of $p \times k$ semi-orthogonal matrices. Thus, we can pick out a subsequence along which γ_m, D_m, and U_m converge to some limits γ, D, and U. The diagonal elements of D are non-negative since those of D_m are and U is a semi-orthogonal matrix by closedness. Thus, along this subsequence we get by taking limits on both sides of the identities $\beta_m = U_m\gamma_m$ and $\Psi_m = U_m D_m U_m^\mathsf{T}$ that $\beta = U\gamma$ and $\Psi = UDU^\mathsf{T}$, which proves Θ_1 is closed.

Next, we show there exists a compact $A \subseteq \Theta_1$ such that $G_n(\theta) > G_n(\theta_0)$ for all $\theta \in \Theta_1 \setminus A$ and some arbitrary but fixed $\theta_0 \in \Theta$. Unconstrained partial minimization in β and Σ_X shows $G_n(\theta) \geq -\log|\Omega| + \operatorname{tr}(Y^\mathsf{T} Q_X Y \Omega) + \log|X^\mathsf{T} X/n| + \operatorname{tr}(I_p)$, which tends to ∞ if $\lambda_{\max}(\Omega) \to \infty$ or $\lambda_{\min}(\Omega) \to 0$; this is so because $Y^\mathsf{T} Q_X Y$ is positive definite when $[Y, X]$ has full column rank. Similarly, unconstrained partial minimization in β and Ω shows $G_n(\theta) \geq \log|Y^\mathsf{T} Q_X Y/n| + \operatorname{tr}(I_r) + \log|\tau I_p + \Psi| + n^{-1}\operatorname{tr}\{X^\mathsf{T} X(\tau I_p + \Psi)^{-1}\}$, which since $X^\mathsf{T} X$ is positive definite when $[Y, X]$ and hence X has full column rank, tends to ∞ if $\tau \to \infty$, $\tau \to 0$, or $\lambda_{\max}(\Psi) \to \infty$. Unconstrained partial minimization in Ω and Σ_X gives $G_n(\theta) \geq \log|(Y - X\beta)^\mathsf{T}(Y - X\beta)/n| + \operatorname{tr}(I_r) + \log|X^\mathsf{T} X/n| + \operatorname{tr}(I_p)$, which tends to ∞ if $\|\beta\| \to \infty$ since $X^\mathsf{T} X$ is positive definite. Thus, we can take $A = \{\theta \in \Theta_1 : \|\beta\| \leq c, c^{-1} \leq \lambda_{\min}(\Omega) \leq \lambda_{\max}(\Omega) \leq c, c^{-1} \leq \tau \leq c, \lambda_{\max}(\Psi) \leq c\}$ for some large enough $c < \infty$. Using that Θ_1 is closed and A is bounded by construction, it is routine to show A is a closed and bounded subset of $\mathbb{R}^d$ and hence compact.

Now, G_n is continuous at points where $\tau > 0$ and $\lambda_{\min}(\Omega) > 0$ and hence attains its minimum over A, which must be a minimum over Θ_1 by the above. To show G_n attains its minimum over Θ it thus suffices to show that for any minimizer $\hat\theta$ over Θ_1, it must hold that $\hat\Psi$ has rank k so that $\hat\theta \in \Theta$. Consider a point θ such that Ψ has rank $s < k$. Then for any $v \neq 0$ such that $\Psi v = 0$, $\Psi + vv^\mathsf{T}$ has rank $s + 1 \leq k$ and $\Psi + vv^\mathsf{T} \in \mathcal{S}$. Moreover, β is in the column space of $[U, v]$ since it is in that of U. Thus, if we can show setting $v \neq 0$ decreases the objective function, no point with the rank of Ψ equal to $s < k$ can be a minimizer, and hence we are done. Consider the function of v defined for a fixed θ by $\log|\Sigma_X + vv^\mathsf{T}| + n^{-1}\operatorname{tr}\{(\Sigma_X + vv^\mathsf{T})^{-1}X^\mathsf{T} X\} = \log|\Sigma_X| + \log(1 + v^\mathsf{T}\Sigma_X^{-1}v) + n^{-1}\operatorname{tr}(X^\mathsf{T} X\Sigma_X^{-1}) - n^{-1}v^\mathsf{T}\Sigma_X^{-1}X^\mathsf{T} X\Sigma_X^{-1}v(1 + v^\mathsf{T}\Sigma_X^{-1}v)^{-1}$, where we used the matrix determinant lemma and the Sherman-Morrison formula. Now observe that if v is orthogonal to the s leading eigenvectors of Σ_X, then $v^\mathsf{T}\Sigma_X v = \tau\|v\|^2$ and $v^\mathsf{T}\Sigma_X^{-1}v = \tau^{-1}\|v\|^2$. Let $c = \|v\|^2$. Then, with the spectral decomposition $\Sigma_X^{-1} = \sum_{j=1}^p \lambda_j^{-1}(\Sigma_X)u_j u_j^\mathsf{T}$, the terms of the objective that depend on v are

$$\log(1 + \tau^{-1}c) - \frac{c\tau^{-1}\left(\sum_{j=s+1}^p v^\mathsf{T} u_j u_j^\mathsf{T}\right)(S_X/\tau)\left(\sum_{j=s+1}^p u_j u_j^\mathsf{T} v\right)}{1 + \tau^{-1}c}$$

Consider a $v \propto u_j$ for some $j \in \{s+1, \ldots, p\}$. The last display then becomes $\log(1 + \tau^{-1}c) - (1 + \tau^{-1}c)^{-1}c\tau^{-1}u_j^\mathsf{T}(S_X/\tau)u_j$. By making the change of variables $t = c\tau^{-1}$ and letting $a = u_j^\mathsf{T} S_X u_j/\tau$ one gets $\log(1 + t) - (1 + t)^{-1}ta$ which is minimized by $t = a - 1$, which is feasible and non-zero if $a > 1$. To see $\hat\tau > \lambda_{\min}(S_X)$ and hence $a > 1$, let VHV^T be a spectral decomposition of Σ_X, where the diagonal elements of H are the eigenvalues $h_j = \lambda_j(\Sigma_X)$. It follows that $h_{k+1} = \cdots = h_p = \tau$ and the first k columns of V are those of U. The part of G_n depending on H is $\log|\Sigma_X| + \operatorname{tr}(S_X\Sigma_X^{-1}) = \sum_{j=1}^p\{\log h_j + h_j^{-1}a_j\}$, where $a_j = V_j^\mathsf{T} X^\mathsf{T} X V_j/n$.

The derivative of this with respect to h_j is $h_j^{-1} - h_j^{-2} a_j$ if $j \geq k$ and $(p-k)h_j^{-1} - h_j^{-2} \sum_{j=k+1}^{p} a_j$ if $j = k+1$. Suppose that $h_p < \lambda_{\min}(S_X)$; then since $\sum_{j=k+1}^{p} a_j \geq (p-k)\lambda_{\min}(S_X)$, the derivative for h_p is negative. That is, moving $h_p = \tau$ towards h_k does not affect the ordering of the eigenvalues but decreases the objective function. The same holds for every h_j, so for a minimizer it must be that $\hat{h}_1 \geq \cdots \geq \hat{h}_p = \hat{\tau} \geq \lambda_{\min}(S_X)$. A similar argument shows $\hat{h}_1 \leq \lambda_{\max}(S_X)$. It remains to show $\hat{\tau} \neq \lambda_{\min}(S_X)$. Suppose $\hat{h}_1 = \lambda_{\min}(S_X)$, then in fact $h_j = \lambda_{\min}(S_X)$ for all j giving $\lambda_{\min}(S_X)$ multiplicity p, contradicting $\lambda_{k+1}(S_X) > \lambda_p(S_X)$. But then, by the same argument as when showing $\hat{\tau} \geq \lambda_{\min}(S_X)$, it must be that $\hat{h}_2 > \lambda_{\min}(S_X)$ since otherwise the objective could be decreased by moving $\hat{h}_2$ towards $\hat{h}_1$. Continuing this process shows $\hat{h}_{k+1} > \lambda_{\min}(S_X)$ as desired.

2.2 Proof of Theorem 2.2

We start with a lemma.

Lemma A. *Condition (i) of Theorem 2.2 implies, for a generic $0 < c < \infty$ that may change between claims but does not depend on r or p: (a) $c^{-1} \leq \lambda_{\min}(\Omega_*) \leq \lambda_{\max}(\Omega_*) \leq c$, (b) $c^{-1} \leq \lambda_{\min}(\Omega_*^{-1} + \beta_*^{\mathsf{T}}\Sigma_{X*}\beta) \leq \lambda_{\max}(\Omega_*^{-1} + \beta_*^{\mathsf{T}}\Sigma_{X*}\beta) \leq c$, (c) $c^{-1} \leq \lambda_{\min}(\Sigma_{X*}) \leq \lambda_{\min}(\Sigma_{X*}) \leq c$, and (d) $\|\beta_*\| \leq c$.*

Proof. Claim (a) follows from observing that $\Omega_*^{-1} = \Sigma_*/\Sigma_{X*}$ is the Schur-complement of Σ_{X*} in Σ_* (Smith, 1992, Theorem 5), while (b) and (c) are by the Cauchy interlacing theorem. Now (d) follows since $c \geq v^{\mathsf{T}}(\Omega_*^{-1} + \beta_*^{\mathsf{T}}\Sigma_{X*}\beta_*)v \geq \lambda_{\min}(\Omega_*) + \|\beta_*\|\lambda_{\min}(\Sigma_{X*}) \geq c^{-1} + c^{-1}\|\beta_*\|$ so that $\|\beta_*\| \leq c^2 - 1$. □

We next show G_n concentrates around G on suitably chosen subsets of Θ. For any $c > 1$ define the set $A = A(c)$ by

$$A = \{\theta \in \Theta : \|\beta\| \leq c, c^{-1} \leq \lambda_{\min}(\Omega), \leq \lambda_{\max}(\Omega) \leq c, c^{-1} \leq \tau \leq c, \lambda_{\max}(\Psi) \leq c\}. \quad (1)$$

Lemma B. *Under the conditions of Theorem 2.2, for all $c < \infty$ large enough and any $\epsilon > 0$,*

$$\mathrm{pr}\left(\sup_{\theta \in A(c)} |G_n(\theta) - G(\theta)| \geq \epsilon\right) \to 0.$$

Proof. We have

$$\begin{aligned} G_n(\theta) - G(\theta) = {} & n^{-1}\operatorname{tr}\{(Y - X\beta)^{\mathsf{T}}(Y - X\beta)\Omega\} - \operatorname{tr}\{(\beta - \beta_*)^{\mathsf{T}}\Sigma_{X*}(\beta - \beta_*)\Omega\} - \operatorname{tr}(\Omega\Omega_*^{-1}) \\ & + n^{-1}\operatorname{tr}(X^{\mathsf{T}}X\Sigma_X^{-1}) - \operatorname{tr}(\Sigma_{X*}\Sigma_X^{-1}). \end{aligned}$$

We show that the suprema of lines one and two over A are both $o_p(1)$ and start with the first. Let $\varepsilon = Y - X\beta_*$, $\tilde{\beta} = \beta_* - \beta$, and $S_Y = Y^{\mathsf{T}}Y/n$. Then the first line is

$$n^{-1}\operatorname{tr}\{(\varepsilon^{\mathsf{T}}\varepsilon + 2\varepsilon^{\mathsf{T}}X\tilde{\beta} + \tilde{\beta}^{\mathsf{T}}X^{\mathsf{T}}X\tilde{\beta})\Omega\} - \operatorname{tr}(\tilde{\beta}^{\mathsf{T}}\Sigma_{X*}\tilde{\beta}\Omega) - \operatorname{tr}(\Omega\Omega_*^{-1})$$

or

$$\operatorname{tr}\{(S_\varepsilon - \Omega_*^{-1})\Omega\} + 2\operatorname{tr}\{(S_{\varepsilon X}\tilde{\beta})\Omega\} + \operatorname{tr}\{\tilde{\beta}(S_X - \Sigma_{X*})\tilde{\beta}\Omega\}$$

Thus, repeatedly using that $\operatorname{tr}(A) \leq r\|A\|$ for any $A \in \mathbb{R}^{r\times r}$ and that operator norms are sub-multiplicative, the absolute value of the first line is upper bounded on A by

$$rc\|S_\varepsilon - \Omega_*^{-1}\| + 4rc\|S_{\varepsilon X}\| + 4rc^3\|S_X - \Sigma_{X*}\|,$$

where we implicitly assumed that the c in the definition of A and condition (i) of Theorem 2.2 are the same, which can always be arranged by picking the larger of the two, so that $\|\beta - \beta_*\| \leq 2c$ on A. But the last display is $o_p(1)$ by condition (ii) of Theorem 2.2.

Now, the second line whose supremum we need to show is $o_p(1)$ is, by the Woodbury identity $\Sigma_X = \tau^{-1}I_p - \tau^{-2}U(D^{-1} + \tau^{-1}I_k)^{-1}U^\mathsf{T}$ with spectral decomposition $\Psi = UDU^\mathsf{T}$,

$$\operatorname{tr}\{(S_X - \Sigma_{X*})\Sigma_X^{-1}\} = \tau^{-1}\operatorname{tr}(S_X - \Sigma_{X*}) - \tau^{-2}\operatorname{tr}\{U^\mathsf{T}(S_X - \Sigma_{X*})U(D^{-1} + \tau^{-1}I_k)^{-1}\}.$$

The absolute value of first term is, on A, less than $c\operatorname{tr}(S_X - \Sigma_{X*})$ which is $o_p(1)$ by condition (iii) of Theorem 2.2. Also on A, the absolute value of the second term is, since $U \in \mathbb{R}^{p\times k}$, less than $c^2k\|(S_X - \Sigma_{X*})(D^{-1} + \tau^{-1}I_k)^{-1}\| \leq c^3k\|S_X - \Sigma_{X*}\|$, which is $o_p(1)$ by condition (ii) of Theorem 2.2. □

Lemma C. *Under the conditions of Theorem 2.2, there exists a $0 < c < \infty$ such that*

$$\operatorname{pr}\left(\arg\min_{\theta\in\Theta} G_n(\theta) \subseteq A(c)\right) \to 1.$$

Proof. Because $\tau_* \geq c^{-1} > 0$, condition (ii) of Theorem 2.2 implies $S_X = X^\mathsf{T}X/n$ is invertible with probability tending to one, so it suffices to consider outcomes with invertible S_X. Pick $\hat{\theta} \in \arg\min_{\theta\in\Theta} G_n(\theta)$; if none exists we are done trivially.

Let $\hat{\Psi} = \hat{U}\hat{D}\hat{U}^\mathsf{T}$ by spectral decomposition and pick a $\hat{\gamma}$ such that $\hat{\beta} = \hat{U}\hat{\gamma}$. Since $\hat{\theta}$ is a minimizer $\hat{\gamma}$ minimizes $\gamma \mapsto \operatorname{tr}\{(Y - X\hat{U}\gamma)^\mathsf{T}(Y - X\hat{U}\gamma)\hat{\Omega}\}$; that is, $\hat{\gamma} = (\hat{U}^\mathsf{T}S_X\hat{U})^{-1}\hat{U}^\mathsf{T}S_{XY}$. Thus, using that the spectral norm is submultiplicative,

$$\|\hat{\beta}\| = \|\hat{\gamma}\| \leq \|(\hat{U}^\mathsf{T}S_X\hat{U})^{-1}\|\|\hat{U}^\mathsf{T}S_{XY}\| \leq \lambda_{\min}(S_X)^{-1}\|S_X\|^{1/2}\|S_Y\|^{1/2},$$

which by condition (ii) of Theorem 2.2 tends in probability to $\tau_*^{-1}\{\tau_* + \lambda_{\max}(\Psi_*)\}^{1/2}\|\Omega_*^{-1} + \beta_*\Sigma_{X*}\beta_*\|^{1/2} \leq 2^{1/2}c^2$, where the inequality is by condition (i). Thus, with probability tending to one, every minimizer satisfies $\|\hat{\beta}\| \leq c$ for some large enough c. Next, since $\hat{\Omega} = (Y^\mathsf{T}Q_{X\hat{U}}Y/n)^{-1}$ and the column space of $X\hat{U}$ is a subset of that of X, it follows that

$$\lambda_{\min}(Y^\mathsf{T}Q_XY/n) \leq \lambda_{\min}(\hat{\Omega}^{-1}) \leq \lambda_{\max}(\hat{\Omega}^{-1}) \leq \lambda_{\max}(Y^\mathsf{T}Y/n).$$

By condition (ii), the left-most and right-most expressions tend to, respectively, $0 < \lambda_{\min}(\Omega_*^{-1})$ and $\lambda_{\max}(\Omega_*^{-1} + \beta_*^\mathsf{T}\Sigma_{X*}\beta_*) < \infty$, from which it follows, by condition (i), that $c^{-1} \leq \lambda_{\min}(\hat{\Omega}) \leq \lambda_{\max}(\hat{\Omega}) \leq c$ with probability tending to one for some large enough c. That $\hat{\tau} \leq c$ and $\lambda_{\max}(\hat{\Psi}) \leq c$ follows similarly from Proposition 2.1 and conditions (i) and (ii). □

Lemma D. *Under condition (i) of Theorem 2.2, there exists a $\delta > 0$, which can depend on c but not p, such that, for every $\theta \in A(c)$,*

$$G(\theta) - G(\theta_*) \geq \delta \|\theta - \theta_*\|_M^2.$$

Proof. The inequality is an equality if $\theta = \theta_*$, so pick a $\theta \neq \theta_*$ and let $\epsilon = \|\theta - \theta_*\|_M > 0$. By definition of $\|\cdot\|_M$, it must hold that (a) $\|\beta - \beta_*\| = \epsilon$, (b) $\|\Omega - \Omega_*\| = \epsilon$, (c) $|\tau - \tau_*| = \epsilon$, or (d) $\|\Psi - \Psi_*\| = \epsilon$. Let $G_1(\theta) = -\log|\Omega| + \operatorname{tr}\{(\beta - \beta_*)^\mathsf{T}(\tau_* I_p + \Psi_*)(\beta - \beta_*)\Omega\} + \operatorname{tr}(\Omega\Omega_*^{-1})$ and $G_2(\theta) = \log|\Sigma_X| + \operatorname{tr}(\Sigma_X^{-1}\Sigma_{X*})$. Since both G_1 and G_2 are minimized by θ_*, we have

$$G(\theta) - G(\theta_*) \geq \max\{G_1(\theta) - G_1(\theta_*), G_2(\theta) - G_2(\theta_*)\}.$$

Thus, it suffices to show that if either of (a) – (d) holds, then at least one of the terms in the maximum on the right-hand side are greater than $\epsilon^2\delta$ for some $\delta > 0$ not depending on p.

Consider first G_2 and let $\Omega_X = \Sigma_X^{-1} = (\tau I_p + \Psi)^{-1}$. The map $\operatorname{vec}(\Omega_X) \mapsto G_2(\theta)$ is convex with gradient vanishing at $\operatorname{vec}(\Omega_{X*})$ and Hessian $\Sigma_X \otimes \Sigma_X$. Thus, $G_2(\theta) - G(\theta_*) \geq 2^{-1}\lambda_{\min}(\Sigma_X \otimes \Sigma_X)\|\operatorname{vec}(\Omega_X) - \operatorname{vec}(\Omega_X)\|^2 \geq 2^{-1}\tau^2\|\Omega_X - \Omega_{X*}\|^2$. Now $\|\Sigma_X - \Sigma_{X*}\| = \|\Sigma_X(\Omega_X - \Omega_{X*})\Sigma_{X*}\| \leq \|\Sigma_X\|\|\Sigma_{X*}\|\|\Omega_X - \Omega_{X*}\|$, so $G_2(\theta) - G_2(\theta_*) \geq 2^{-1}\tau^2\|\Sigma_X - \Sigma_{X*}\|^2/(\|\Sigma_X\| + \|\Sigma_{X*}\|)^2 \geq 2^{-3}c^{-4}\|\Sigma_X - \Sigma_{X*}\|$. Now suppose (c) holds, then $\|\Sigma_X - \Sigma_{X*}\| \geq \epsilon$ by Weyl's inequalities, so we can take $\delta = 2^{-3}c^{-4}$. Next suppose (d) holds. If $|\tau - \tau_*| \geq \epsilon/2$, then we can take $\delta = 2^{-4}c^{-4}$ by the same argument as before, so suppose $|\tau - \tau_*| < \epsilon/2$. Write $\Sigma_X - \Sigma_{X*} = \Psi - \Psi_* + (\tau - \tau_*)I_p$. Then for any unit-length v, $v^\mathsf{T}(\Sigma_X - \Sigma_{X*})v = v^\mathsf{T}(\Psi - \Psi_*)v + (\tau - \tau_*)$. It follows, since the spectral norm of a symmetric matrix is its largest absolute eigenvalue, that $\|\Sigma_X - \Sigma_{X*}\| \geq \epsilon/2$, and hence we can take $\delta = 2^{-4}c^{-4}$.

Consider now G_1 and suppose (a) holds. Minimize partially in Ω, which amounts to setting $\Omega^{-1} = (\beta - \beta_*)^\mathsf{T}\Sigma_{X*}(\beta - \beta_*) + \Omega_*^{-1}$. One obtains that $G_1(\theta) - G_1(\theta_*)$ is lower bounded by $\log|(\beta - \beta_*)^\mathsf{T}\Sigma_{X*}(\beta - \beta_*) + \Omega_*^{-1}| - \log|\Omega_*^{-1}|$. By the mean value theorem and using that the gradient of $\Omega^{-1} \mapsto \log|\Omega^{-1}|$ is $\operatorname{tr}(\Omega)$, the last display is equal to $\operatorname{tr}\{\tilde{\Omega}(\beta - \beta_*)^\mathsf{T}\Sigma_{X*}(\beta - \beta_*)\}$, where $\tilde{\Omega}^{-1} = \Omega_*^{-1} + s(\beta - \beta_*)^\mathsf{T}\Sigma_{X*}(\beta - \beta_*)$ for some $s \in [0, 1]$. But the last trace is a quadratic in $\operatorname{vec}(\beta)$ with Hessian $\tilde{\Omega} \otimes \Sigma_X$. The eigenvalues of $\tilde{\Omega}^{-1}$ are less than $c + c^3$, so the eigenvalues of the Hessian, which are the products of the eigenvalues of the terms, are greater than $c^{-1}(c + c^3)^{-1}$. Thus, we can take $\delta = 2^{-1}(c^{-2} + c^{-4})$.

Finally, suppose (b) holds and minimize partially in β; that is, set $\beta = \beta_*$. One gets $G_1(\theta) - G(\theta_*) \geq -\log|\Omega| + \operatorname{tr}(\Omega\Omega_*^{-1}) + \log|\Omega_*| - \operatorname{tr}(I_r)$. We already know this is a convex function of $\operatorname{vec}(\Omega)$ which is minimized at $\operatorname{vec}(\Omega_*)$ and with Hessian $\Omega^{-1} \otimes \Omega^{-1}$. Thus, $G_1(\theta) - G_1(\theta_*) \geq \|\Omega - \Omega_*\|_F^2 2^{-1}c^{-2} \geq \|\Omega - \Omega_*\|^2 2^{-1}c^{-2}$, so we can take $\delta = 2^{-1}c^{-2}$. To conclude, we have shown the claim holds with $\delta = \min\{2^{-4}c^{-4}, 2^{-1}(c^{-2} + c^{-4}), 2^{-1}c^{-2}\}$ □

Proof of Theorem 2.2. The existence part follows from Proposition 2.1 and conditions (i) and (ii). Pick an $\epsilon > 0$ and a c large enough that Lemma B and C hold on $A = A(c)$. By increasing c and decreasing ϵ if necessary, we may assume $B = \{\theta : \|\theta - \theta_*\|_M < \epsilon\} \subset A(c)$. On $A \setminus B$, $G(\theta) \geq G(\theta_*) + \delta\epsilon^2$ by Lemma D. Thus, by Lemma B, $G_n(\theta) > G(\theta_*) + \delta\epsilon^2/2$ with probability tending to one. Also with probability tending to one, $G_n(\theta_*) < G(\theta_*) + \delta\epsilon^2$. Thus, with probability tending to one, using Lemma C for the first equality, $\arg\min_{\theta \in \Theta} G_n(\theta) = \arg\min_{\theta \in A} G_n(\theta) = \arg\min_{\theta \in B} G_n(\theta)$, which completes the proof. □

2.3 Proof of Theorem 2.4

Lemma E. *Under the conditions of Theorem 2.4, $\sqrt{n}\nabla G_n(\theta_*)$ tends in distribution to a multivariate normal vector with mean zero and positive definite covariance matrix with finite entries.*

Proof. The gradient of $g(\cdot; z)$ at θ_*, with $\varepsilon = y - \beta_*^\mathsf{T} x$, has subvectors given by the vectorizations of $\nabla_\beta g(\theta_*; z) = -2x\varepsilon^\mathsf{T}\Omega_*$, $\nabla_\Omega g(\theta_*; z) = -\Omega_*^{-1} + \varepsilon\varepsilon^\mathsf{T}$, $\nabla_\Psi g(\theta_*; z) = \Sigma_{X*}^{-1} - \Sigma_{X*}^{-1}xx^\mathsf{T}\Sigma_{X*}^{-1}$, and $\nabla_\tau g(\theta_*; z) = \mathrm{tr}\{\nabla_\Psi g(\theta_*; z)\}$. Consider, for example, the subvector $\mathrm{vec}\{\nabla_\beta g(\theta_*; z)\} = -2(\Omega_* \otimes I_p)\,\mathrm{vec}(x\varepsilon^\mathsf{T})$. Observe $\mathbb{E}(X_i\varepsilon_i^\mathsf{T}) = 0$ and hence $n^{-1/2}\sum_{i=1}^n \mathrm{vec}\{\nabla_\beta g(\theta_*; Z_i)\} = -2(\Omega_* \otimes I_p)n^{-1/2}\sum_{i=1}^n \mathrm{vec}(X_i\varepsilon_i^\mathsf{T})$ tends to a multivariate normal vector by assumption (iii). All the other subvectors, and the full vector, can be treated similarly, using for $\nabla_\tau g(\theta_*; z)$ that $\mathrm{tr}(\Sigma_{X*}^{-1}xx^\mathsf{T}\Sigma_{X*}^{-1}) = \mathrm{tr}(\Sigma_{X*}^{-2}xx^\mathsf{T}) = \mathrm{vec}(\Sigma_{X*}^{-2})^\mathsf{T}\,\mathrm{vec}(xx^\mathsf{T})$. □

Lemma F. *Under the conditions of Theorem 2.4, $g(\theta; z) = g(\theta_*; z) + \nabla g(\theta_*; z)^\mathsf{T}(\theta - \theta_*) + \|\theta - \theta_*\| r(\theta; z)$ with a $r(\theta; z)$ that is stochastically equicontinuous in the sense that for every $\epsilon > 0$ and $\delta > 0$, there exists a $\rho > 0$ such that*

$$\limsup_{n\to\infty} \mathrm{pr}^*\left(\sup_{\|\theta-\theta_*\|<\rho}\left|n^{-1/2}\sum_{i=1}^n [r(\theta; Z_i) - \mathbb{E}\{r(\theta, Z_i)\}]\right| > \delta\right) < \epsilon,$$

where the superscript $*$ denotes outer probability.

Proof. Consider the function $h(s) = g(\theta_* + s(\theta - \theta_*); z)$, so that $g(\theta; z) - g(\theta_*; z) = h(1) - h(0)$. Taylor expansion with integral-form remainder gives $h(s) = h(0) + h'(0)s + \int_0^s h''(t)(s-t)\mathrm{d}t$, where $h'(s) = \nabla g(\theta_* + s(\theta - \theta_*); z)^\mathsf{T}(\theta - \theta_*)$ and $h''(s) = (\theta - \theta_*)^\mathsf{T}\nabla^2 g(\theta_* + s(\theta - \theta_*); z)(\theta - \theta_*)$. Thus, $r(\theta_*; z) = 0$ and for $\theta \neq \theta_*$

$$r(\theta; z) = \frac{(\theta - \theta_*)^\mathsf{T}}{\|\theta - \theta_*\|}\int_0^1 \nabla^2 g(\theta_* + s(\theta - \theta_*); z)(1 - s)\mathrm{d}s(\theta - \theta_*).$$

Denote the middle term (matrix) by $K(\theta; z)$ so that $r(\theta; z) = \|\theta - \theta_*\|^{-1}(\theta - \theta_*)^\mathsf{T}K(\theta; z)(\theta - \theta_*)$. Blocks of the matrix K correspond to blocks of $\nabla^2 g$. For example, the leading $pr \times pr$ block of K is $K_1(\theta; z) = \int\{\Omega_* + s(\Omega - \Omega_*)\}(1 - s)\mathrm{d}s \otimes xx^\mathsf{T}$, and hence $n^{-1}\sum_{i=1}^n [K_1(\theta; Z_i) - \mathbb{E}\{K_1(\theta; Z_i)\}] = \int\{\Omega_* + s(\Omega - \Omega_*)\}(1 - s)\mathrm{d}s \otimes n^{-1/2}\sum_{i=1}^n \{X_iX_i^\mathsf{T} - \Sigma_{X*}\}$. The elements of the right-hand matrix are $O_p(1)$ since the vectorization satisfies a central limit theorem by condition (iii). The elements of the left-hand matrix are bounded on a neighborhood of θ_*. Thus, the elements of $n^{-1}\sum_{i=1}^n [K_1(\theta; Z_i) - \mathbb{E}\{K_1(\theta; Z_i)\}]$ are $O_p(1)$ uniformly in θ on a neighborhood of θ_*. Similar arguments for the other blocks, using that the inverse covariance matrices in the Hessian have bounded eigenvalues on small enough neighborhoods of θ_* since $\lambda_{\min}(\Omega_*) > 0$ and $\tau_* > 0$, show the elements of $n^{-1/2}\sum_{i=1}^n [K(\theta; Z_i) - \mathbb{E}\{K(\theta; Z_i)\}]$ are $O_p(1)$ uniformly in θ on a neighborhood of θ_*; thus, the spectral norm is $O_p(1)$ uniformly in θ on a neighborhood of θ_*. The result follows since $|n^{-1/2}\sum_{i=1}^n [r(\theta; Z_i) - \mathbb{E}\{r(\theta, Z_i)\}]| = |\|\theta - \theta_*\|^{-1}(\theta - \theta_*)^\mathsf{T}n^{-1/2}\sum_{i=1}^n [K(\theta; Z_i) - \mathbb{E}\{K(\theta; Z_i)\}](\theta - \theta_*) \leq \|\theta - \theta_*\|\|n^{-1/2}\sum_{i=1}^n [K(\theta; Z_i) - \mathbb{E}\{K(\theta; Z_i)\}]\|$. □

Proof of Theorem 2.4. We verify the conditions of Theorem 4.4 from Geyer (1994). Chernoff-regularity is from Theorem 2.6. Assumption A is verified by noting G is minimized at $\theta_* \in \Theta$, $\nabla G(\theta_*) = 0$, and that G has a local quadratic approximation with $o(\|\theta - \theta_*\|^2)$ remainder around θ_* by Taylor's theorem since the third order derivatives are bounded for θ close to θ_*. The last statement follows from differentiating the expressions for $\nabla^2 G(\theta)$ and observing that powers of Ω and Σ_X are bounded around θ_* since $\lambda_{\min}(\Omega_*) > 0$ and $\tau_* > 0$. Assumptions B and C are verified in Lemmas F and E, respectively. Assumption D holds since $\hat{\theta}$ is a minimizer by assumption. □

2.4 Proof of Theorem 2.6

The following lemma is from Li et al. (2019).

Lemma G. *Let $\mathcal{R} \subseteq \mathbb{R}^{p\times p}$ be the set of $p \times p$ matrices of rank k and A an arbitrary point in $\mathcal{R}$ with singular value decomposition $A = UDV^\top$, $D \in \mathbb{R}^{k\times k}$; then $T_{\mathcal{R}}(A) = \{B \in \mathbb{R}^{p\times p} : Q_U B Q_V = 0\}$.*

Proof of Lemma 2.5. Using the definition, if $C_m \to C$, $a_m \downarrow 0$, and $\Psi + a_m C_m \in \mathcal{S}$ for all n, then C_m and hence C must be symmetric. Next, let $\mathcal{R}$ be the set of $p \times p$ matrices of rank k. Since $\Psi \in \mathcal{S} \subseteq \mathcal{R}$, it is immediate from the definition that $\tilde{T}_{\mathcal{S}}(\Psi) \subseteq T_{\mathcal{S}}(\Psi) \subseteq T_{\mathcal{R}}(\Psi)$. But Lemma G says $T_{\mathcal{R}}(\Psi) = \{C \in \mathbb{R}^{p\times p} : Q_U C Q_U\}$, so we have proved $\tilde{T}_{\mathcal{S}}(\Psi) \subseteq T_{\mathcal{S}}(\Psi) \subseteq \{C \in \mathbb{R}^{p\times p} : C = C^\top, Q_U C Q_U = 0\}$. To prove the reverse inclusions, pick arbitrary symmetric C such that $Q_U C Q_U = 0$ and $a_m \downarrow 0$. We must find $C_m \to C$ satisfying $\Psi + a_m C_m \in \mathcal{S}$ for all m. To that end, consider $C_m = C$ and $\Psi_m = \Psi + a_m C$. For any v such that $P_U v = 0$, $v^\top \Psi_m v = v^\top Q_U(\Psi + a_m C) Q_U v = 0$, while for any v such that $P_U v \neq 0$, $v^\top \Psi_m v \geq \|P_U v\| \lambda_k(\Psi) - a_m \|C\|$, which is strictly positive for small enough a_m. It follows that the null spaces of Ψ_n and Ψ agree and that $v^\top \Psi_m v > 0$ for any v not in that null space. Thus, as desired, Ψ_m is positive semi-definite with rank k. For the at most finitely many m where a_m is not small enough, we may take $C_m = 0$ without affecting the conclusion, and this completes the proof. □

Proof of Theorem 2.6. The claims that $O = O^\top$ and $t \in \mathbb{R}$ are straightforward to verify using the definition so we omit the details. That $C = C^\top$ and $Q_\Psi C Q_\Psi = 0$ is by Lemma 2.5. To see that $P_\Psi B = B$, consider arbitrary sequences $a_m \downarrow 0$ and $(B_m, O_m, t_m, C_m) \to (B, O, t, C)$ satisfying Definition 2.1; that is, $\theta_m = \theta + a_m(B_m, O_m, t_m, C_m)$ is in the parameter set for all (large enough) m. Since Ψ has k strictly positive eigenvalues, so does $\Psi_m = \Psi + a_m C_m$ for all large enough m, and hence Ψ_m has rank at least k; thus, it in fact has rank k since θ_m is in the parameter set. Let $\Psi = UDU^\top$ and consider $Q_U \beta_m = Q_U(\beta + a_m B_m) = a_m Q_U B_m$. Since θ_m is in the parameter set, we can also write $\beta_m = (\Psi + a_m C_m)\gamma_m$ for some γ_m and get $Q_U \beta_m = a_m Q_U C_m \gamma_m$. Thus, dividing by a_m we get $Q_U B_m = Q_U C_m \gamma_m$. If we can ensure $\{\gamma_m\}$ is bounded so that it has a convergent subsequence, then we are done upon taking limits along that subsequence to get $Q_U B = Q_U C \gamma = 0$ since $Q_U C = 0$. To see that $\{\gamma_m\}$ can be selected to be bounded, note that we may restrict attention to γ_m in the row space of Ψ_m, which is also its column space. Thus, $\gamma_m = U_m \alpha_m = \sum_{j=1}^k \alpha_{mj} u_{mj}$ and $\beta_m = \Psi_m \gamma_m = \sum_{j=1}^k \lambda_j(\Psi_m) \alpha_{mj} u_j$. The norm $\|\beta_m\|$ is bounded since β_m converges and its square is equal to $\|\sum_{j=1}^{k}\lambda_j(\Psi_m)\alpha_{mj}u_j\|^2 = \sum_{j=1}^{k}\lambda_j(\Psi_m)^2\alpha_{mj}^2 \geq \lambda_k(\Psi_m)^2\sum_{j=1}^{k}\alpha_{mj}^2 = \lambda_k(\Psi_m)^2\|\gamma_m\|$, so $\|\gamma_m\| \leq \lambda_k(\Psi_m)^{-1}\|\beta_m\|$, which for all large enough m is bounded by, say, $2\|\beta\|\lambda_k(\Psi)^{-1} < \infty$. □

3 Additional details

The objective function in Section 3.1 is

$$-\log|\Omega| + n^{-1}\operatorname{tr}\{(Y - XL\gamma)^{\mathsf{T}}(Y - XL\gamma)\Omega\} + \log|\tau(I_p + LL^{\mathsf{T}})| + \tau^{-1}\operatorname{tr}\{S_X(I_p + LL^{\mathsf{T}})^{-1}\}.$$

Differentiating with respect to γ, Ω, and τ and setting those derivatives to zero gives

$$\log|Y^{\mathsf{T}}Q_{XL}Y/n| + \operatorname{tr}(I_r) + \log[\operatorname{tr}\{S_X(I_p + LL^{\mathsf{T}})^{-1}\}] + \log|I_p + LL^{\mathsf{T}}| + 1.$$

We derive the gradient of H_n assuming its argument L is an unconstrained matrix; the gradient under the restriction that $L_{i,j} = 0$ for $j > i$ is obtained by setting the corresponding elements of the unconstrained gradient to zero. The differential of $Q_{XL} = I_n - XL(L^{\mathsf{T}}X^{\mathsf{T}}XL)^{-1}L^{\mathsf{T}}X^{\mathsf{T}}$ is

$$\begin{aligned}
dQ_{XL} = &-X(dL)(L^{\mathsf{T}}X^{\mathsf{T}}XL)^{-1}L^{\mathsf{T}}X^{\mathsf{T}} \\
&+ XL(L^{\mathsf{T}}X^{\mathsf{T}}XL)^{-1}[(dL)^{\mathsf{T}}X^{\mathsf{T}}XL + L^{\mathsf{T}}X^{\mathsf{T}}XdL](L^{\mathsf{T}}X^{\mathsf{T}}XL)^{-1}L^{\mathsf{T}}X^{\mathsf{T}} \\
&- XL(L^{\mathsf{T}}X^{\mathsf{T}}XL)^{-1}(dL)^{\mathsf{T}}X^{\mathsf{T}}.
\end{aligned}$$

Thus, the differential of $\log|Y^{\mathsf{T}}Q_{XL}Y|$ is, with $S = Y^{\mathsf{T}}Q_{XL}Y$,

$$\begin{aligned}
d\log|Y^{\mathsf{T}}Q_{XL}Y| = &-\operatorname{tr}\left[S^{-1}Y^{\mathsf{T}}X(dL)(L^{\mathsf{T}}X^{\mathsf{T}}XL)^{-1}L^{\mathsf{T}}X^{\mathsf{T}}Y\right] \\
&+ \operatorname{tr}\left[S^{-1}Y^{\mathsf{T}}XL(L^{\mathsf{T}}X^{\mathsf{T}}XL)^{-1}[(dL)^{\mathsf{T}}X^{\mathsf{T}}XL + L^{\mathsf{T}}X^{\mathsf{T}}XdL](L^{\mathsf{T}}X^{\mathsf{T}}XL)^{-1}L^{\mathsf{T}}X^{\mathsf{T}}Y\right] \\
&- \operatorname{tr}\left[S^{-1}Y^{\mathsf{T}}XL(L^{\mathsf{T}}X^{\mathsf{T}}XL)^{-1}(dL)^{\mathsf{T}}X^{\mathsf{T}}Y\right] \\
= &-\operatorname{tr}\left[(L^{\mathsf{T}}X^{\mathsf{T}}XL)^{-1}L^{\mathsf{T}}X^{\mathsf{T}}YS^{-1}Y^{\mathsf{T}}XdL\right] \\
&+ \operatorname{tr}\left[(dL)^{\mathsf{T}}X^{\mathsf{T}}XL(L^{\mathsf{T}}X^{\mathsf{T}}XL)^{-1}L^{\mathsf{T}}X^{\mathsf{T}}YS^{-1}Y^{\mathsf{T}}XL(L^{\mathsf{T}}X^{\mathsf{T}}XL)^{-1}\right] \\
&+ \operatorname{tr}\left[(L^{\mathsf{T}}X^{\mathsf{T}}XL)^{-1}L^{\mathsf{T}}X^{\mathsf{T}}YS^{-1}Y^{\mathsf{T}}XL(L^{\mathsf{T}}X^{\mathsf{T}}XL)^{-1}L^{\mathsf{T}}X^{\mathsf{T}}XdL\right] \\
&- \operatorname{tr}\left[(dL)^{\mathsf{T}}X^{\mathsf{T}}YS^{-1}Y^{\mathsf{T}}XL(L^{\mathsf{T}}X^{\mathsf{T}}XL)^{-1}\right].
\end{aligned}$$

Thus, $\nabla\log|Y^{\mathsf{T}}Q_{XL}Y|$ is

$$\begin{aligned}
\nabla\log|Y^{\mathsf{T}}Q_{XL}Y| = &-2X^{\mathsf{T}}YS^{-1}Y^{\mathsf{T}}XL(L^{\mathsf{T}}X^{\mathsf{T}}XL)^{-1} \\
&+ 2X^{\mathsf{T}}XL(L^{\mathsf{T}}X^{\mathsf{T}}XL)^{-1}L^{\mathsf{T}}X^{\mathsf{T}}YS^{-1}Y^{\mathsf{T}}XL(L^{\mathsf{T}}X^{\mathsf{T}}XL)^{-1}
\end{aligned}$$

Finally,

$$\nabla\log|I_p + LL^{\mathsf{T}}| = 2(I_p + LL^{\mathsf{T}})^{-1}L,$$

and

$$\nabla\log\operatorname{tr}\left[X^{\mathsf{T}}X(I_p + LL^{\mathsf{T}})^{-1}\right] = -\frac{2}{\operatorname{tr}\left[X^{\mathsf{T}}X(I_p + LL^{\mathsf{T}})^{-1}\right]}(I_p + LL^{\mathsf{T}})^{-1}X^{\mathsf{T}}X(I_p + LL^{\mathsf{T}})^{-1}L.$$

References

Bhatia, R. (2012). *Matrix Analysis*. Springer New York.

Geyer, C. J. (1994). On the asymptotics of constrained M-estimation. *Annals of Statistics*, 22(4):1993–2010.

Li, X., Xiu, N., and Luo, Z. (2019). Low-rank matrix optimization over affine set. *arXiv preprint arXiv:1912.03029*.

Smith, R. L. (1992). Some interlacing properties of the Schur complement of a Hermitian matrix. *Linear Algebra and its Applications*, 177:137–144.