Appendix for High-dimensional union support recovery in multivariate regression

A Proof of Lemma 1

We begin by noting that the block-regularized problem (2) is convex, and not differentiable for all B. In particular, denoting by β_i the i^{th} row of B, the subdifferential of the norm ℓ_1/ℓ_2-block norm over row i takes the following form, which introduces the function ζ in the problem

$$[\partial \|B\|_{\ell_1/\ell_2}]_i \quad = \quad \begin{cases} \zeta(\beta_i) = \frac{\beta_i}{\|\beta_i\|_2} & \text{if } \beta_i \neq \vec{0} \\ Z_i \quad \text{such that } \|Z_i\|_2 \leq 1 & \text{otherwise.} \end{cases}$$

Using the notation β_i to denote a row of B and denoting by

$$\mathcal{K} := \{(w,v) \in \mathbb{R}^K \times \mathbb{R} \mid \|w\|_2 \leq v\}$$

the usual second-order cone (SOC), we can rewrite the original convex program (2) as the second order cone program (SOCP):

$$\min_{\substack{B \in \mathbb{R}^{p\times K} \\ b \in \mathbb{R}^p}} \ \frac{1}{2n} \|| Y - XB\||_F^2 + \lambda_n \sum_{i=1}^p b_i \quad \text{s.t.} \quad (\beta_i, b_i) \in \mathcal{K}, \quad 1 \leq i \leq p \qquad (1^\sharp)$$

We now dualize the conic constraints [BV04], using conic Lagrange multipliers belonging to the dual cone $\mathcal{K}^* = \{(z,t) \in \mathbb{R}^{K+1} \mid z^T\mathbf{w} + vt \geq 0, (\mathbf{w},v) \in \mathcal{K}\}$. The second-order cone $\mathcal{K}$ is self-dual [BV04], so that the convex program $(1^\sharp)$ is equivalent to

$$\min_{\substack{B \in \mathbb{R}^{p\times K} \\ b \in \mathbb{R}^p}} \quad \max_{\substack{Z \in \mathbb{R}^{p\times K} \\ t \in \mathbb{R}^p}} \quad \frac{1}{2n} \|| Y - XB\||_F^2 + \lambda_n \sum_{i=1}^p b_i - \lambda_n \sum_{i=1}^p \left(-z_i^T \beta_i + t_i\, b_i\right)$$
$$\text{s.t.} \qquad (z_i, t_i) \in \mathcal{K}, \quad 1 \leq i \leq p.$$

where Z is the matrix whose i^{th} row is z_i.

The advantage of an SOCP formulation is that it avoids manipulating the subdifferentials directly and replaces them conveniently with their counterparts arising from duality. In fact, the dual of $(1^\sharp)$ is also an SOCP, with conic variables $(Z_i, t_i) \in \mathbb{R}^K \times \mathbb{R}^+$ associated to each conic constraint. Moreover, as we show next, the variable Z_i coincides at optimality with an element of $[\partial \|B\|_{\ell_1/\ell_2}]_i$ which is characterized by the KKT conditions.

Indeed, since the original program is convex and strictly feasible, strong duality holds and any pair of primal $(B^\star, b^\star)$ and dual solutions $(Z^\star, t^\star)$ has to satisfy the Karush-Kuhn-Tucker conditions:

$$\|\beta_i^\star\|_2 \le b_i^\star, \quad 1 < i < p \tag{$2^\sharp$a}$$

$$\|z_i^\star\|_2 \le t_i^\star, \quad 1 < i < p \tag{$2^\sharp$b}$$

$$z_i^{\star T} \beta_i^\star - t_i^\star b_i^\star = 0, \quad 1 < i < p \tag{$2^\sharp$c}$$

$$\nabla_B \left[\frac{1}{2n} \, \| Y - XB \|_F^2 \right] \Big|_{B=B^\star} + \lambda_n Z^\star = 0 \tag{$2^\sharp$d}$$

$$\lambda_n (1 - t_i^\star) = 0 \tag{$2^\sharp$e}$$

Since equations ($2^\sharp$c) and ($2^\sharp$e) impose the constraints $t_i^\star = 1$ and $b_i^\star = \|\beta_i^\star\|_2$, a primal-dual solution to this conic program is determined by $(B^\star, Z^\star)$.

Any solution satisfying the conditions in Lemma 1 also satisfies these KKT conditions, since equation (6b) and the definition (6c) are equivalent to equation ($2^\sharp$d), and equation (6a) and the combination of conditions (6d) and (6c) imply that the complementary slackness equations ($2^\sharp$c) hold for each primal-dual conic pair (β_i, z_i).

Now consider some other primal solution $\widetilde{B}$; when combined with the optimal dual solution $\widehat{Z}$, the pair $(\widetilde{B}, \widehat{Z})$ must satisfy the KKT conditions [Ber95]. But since for $j \in S^c$, we have $\|\hat{z}_j\|_2 < 1$, then the complementary slackness condition ($2^\sharp$c) implies that for all $j \in S^c$, $\widetilde{\beta}_j = 0$. This fact in turn implies that the primal solution $\widetilde{B}$ must also be a solution to the restricted convex program (7), obtained by only considering the covariates in the set S or equivalently by setting $B_{S^c} = 0_{S^c}$. But since $s < n$ by assumption, the matrix $X_S^T X_S$ is strictly positive definite with probability one, and therefore the restricted convex program (7) has a unique solution $B_S^\star = \widehat{B}_S$. We have thus shown that a solution $(\widehat{B}, \widehat{Z})$ to the program (2) that satisfies the conditions of Lemma 1, if it exists, must be unique.

B Inequalities with block-matrix norms

In general, the two families of matrix norms that we have introduced, $\|\!| \cdot |\!\|_{p,q}$ and $\|\cdot\|_{\ell_a/\ell_b}$, are distinct, but they coincide in the following useful special case:

Lemma B.0.1. *For $1 \le p \le \infty$ and for r defined by $1/r + 1/p = 1$ we have*

$$\|\cdot\|_{\ell_\infty/\ell_p} \quad = \quad \|\!| \cdot |\!\|_{\infty, r}.$$

Proof. Indeed, if a_i^T denotes the i^{th} row of A, then

$$\|A\|_{\ell_\infty/\ell_p} = \max_i \|a_i\|_p = \max_i \max_{\|y_i\|_r \le 1} y_i^T a_i = \max_{\|y\|_r \le 1} \max_i |y^T a_i| = \max_{\|y\|_r \le 1} \|Ay\|_\infty = \|\!| A |\!\|_{\infty, r}.$$

□

Two immediate consequences that we find useful in the case $p = r = 2$ are the following:

Corollary B.0.1. *For matrices $A \in \mathbb{R}^{m\times n}$ and $Z \in \mathbb{R}^{n\times r}$, we have*

$$\|AZ\|_{\ell_\infty/\ell_p} = |||AZ|||_{\infty,r} \le |||A|||_{\infty,\infty} |||Z|||_{\infty,r} = |||A|||_{\infty,\infty} \|Z\|_{\ell_\infty/\ell_p}. \quad (3^\sharp)$$

Corollary B.0.2.

$$|||A|||_r \le |||I_m|||_{r,\infty} |||A|||_{\infty,r} = s^{1/r} \|A\|_{\ell_\infty/\ell_p}.$$

C Analysis of $\mathcal{E}(U_S)$: proof of Lemma 2

This section is devoted to the analysis of the event $\mathcal{E}(U_S)$ from equation (9), and proves Lemma 2. We rewrite U_S as:

$$U_S = \widehat{\Sigma}_{SS}^{-\frac{1}{2}} \frac{\widetilde{W}}{\sqrt{n}} - \lambda_n (\widehat{\Sigma}_{SS})^{-1} \widehat{Z}_S, \qquad \text{with} \qquad \widetilde{W} := \frac{1}{\sqrt{n}} (\widehat{\Sigma}_{SS})^{-\frac{1}{2}} X_S^T W.$$

Using this representation and the triangle inequality, we get $\|U_S\|_{\ell_\infty/\ell_2} \le T_1 + T_2$ where $T_1 := \left\|(\widehat{\Sigma}_{SS})^{-\frac{1}{2}} \frac{\widetilde{W}}{\sqrt{n}}\right\|_{\ell_\infty/\ell_2}$ is a variance term due to the noise, and $T_2 := \lambda_n \left\|(\widehat{\Sigma}_{SS})^{-1} \widehat{Z}_S\right\|_{\ell_\infty/\ell_2}$ is a bias term coming from the regularization.

C.1 Bias term

Using inequality $(3^\sharp)$, we have $T_2 \le \lambda_n \left|\left|\left| (\widehat{\Sigma}_{SS})^{-1}\right|\right|\right|_\infty \left\|\widehat{Z}_S\right\|_{\ell_\infty/\ell_2} \le \lambda_n \left|\left|\left| (\widehat{\Sigma}_{SS})^{-1}\right|\right|\right|_\infty$ because, by construction, $\left\|\widehat{Z}_S\right\|_{\ell_\infty/\ell_2} \le 1$.

Therefore

$$\frac{T_2}{\lambda_n} \le |||(\Sigma_{SS})^{-1}|||_\infty + \left|\left|\left| (\widehat{\Sigma}_{SS})^{-1} - (\Sigma_{SS})^{-1}\right|\right|\right|_\infty \le D_{\max} + \sqrt{s} \left|\left|\left| (\widehat{\Sigma}_{SS})^{-1} - (\Sigma_{SS})^{-1}\right|\right|\right|_2$$

But the whitened random matrix $\widetilde{X}_S := \Sigma_{SS}^{-1/2} X_S$ has i.i.d. standard Gaussian entries and satisfies:

$$\left|\left|\left| (\widehat{\Sigma}_{SS})^{-1} - (\Sigma_{SS})^{-1}\right|\right|\right|_2 \le |||(\Sigma_{SS})^{-1}|||_2 \left|\left|\left| (\widetilde{X}_S^T \widetilde{X}_S/n)^{-1} - I_s\right|\right|\right|_2 \le \frac{1}{C_{\min}} \left|\left|\left| (\widetilde{X}_S^T \widetilde{X}_S/n)^{-1} - I_s\right|\right|\right|_2,$$

From concentration results in random matrix theory [DS01], for $s/n \to 0$, with probability $1 - \exp(-\Theta(n))$,we have

$$\left|\left|\left| (\widetilde{X}_S^T \widetilde{X}_S/n)^{-1} - I_s\right|\right|\right|_2 \le \mathcal{O}\left(\sqrt{\frac{s}{n}}\right) \quad \text{and therefore} \quad \frac{T_2}{\lambda_n} \le D_{\max} + \mathcal{O}\left(\frac{s}{\sqrt{n}}\right)$$

C.2 Noise term

On the other hand, conditionally on X_S, the other term, T_1, is a maximum of χ-distributed random variables, and using concentration results for χ^2 random variables and for spectral matrix norms, we have

Lemma C.2.1. *With probability* $1 - \mathcal{O}\left(\exp(-\Theta(\log s))\right)$, $\quad {T_1}^2 \geq \frac{8K}{C_{\min}} \frac{\log s}{n}$

Proof. Note that conditioned on X_S, we have $(\mathrm{vec}(\widetilde{W}) \mid X_S) \sim N(\vec{0}_{s\times K}, I_s \otimes I_K)$ where $\mathrm{vec}(A)$ denotes the vectorization of matrix A . Using this fact and the definition of the block ℓ_∞/ℓ_2 norm,

$$\begin{aligned} T_1 &= \max_{i\in S} \left\| e_i^T (\widehat{\Sigma}_{SS})^{-\frac{1}{2}} \frac{\widetilde{W}}{\sqrt{n}} \right\|_2 \\ &\leq \left\| (\widehat{\Sigma}_{SS})^{-1} \right\|_2^{1/2} \left[\frac{1}{n} \max_{i\in S} \zeta_i^2 \right]^{1/2}, \end{aligned}$$

which defines ζ_i^2 as independent χ^2 variates with K degrees of freedom. Using the tail bound in Lemma E.0.1 with $t = 2K \log s > K$, we have

$$\mathbb{P}\left[\frac{1}{n} \max_{i\in S} \zeta_i^2 \geq \frac{4K \log s}{n} \right] \leq \exp\left(-2K \log s \left(1 - 2\sqrt{\frac{1}{2\log s}} \right) \right) \to 0.$$

Defining the event $\mathcal{T} := \left\{ \left\| (\widehat{\Sigma}_{SS})^{-1} \right\|_2 \leq \frac{2}{C_{\min}} \right\}$, we have $\mathbb{P}[\mathcal{T}] \geq 1 - 2\exp(-\Theta(n))$, again using concentration results from random matrix theory [DS01]. Therefore,

$$\begin{aligned} \mathbb{P}\left[T_1 \geq \sqrt{\frac{8K \log s}{C_{\min} n}} \right] &\leq \mathbb{P}\left[T_1 \geq \sqrt{\frac{8K \log s}{C_{\min} n}} \,\middle|\, \mathcal{T} \right] + \mathbb{P}[\mathcal{T}^c] \\ &\leq \mathbb{P}\left[\frac{1}{n} \max_{i\in S} \zeta_i^2 \geq \frac{4K \log s}{n} \right] + 2\exp(-\Theta(n)) \\ &= \mathcal{O}\left(\exp(-\Theta(\log s))\right) \to 0. \end{aligned}$$

□

Combining noise and bias terms yields that, under assumption $A3$ and conditions (5) of Theorem 1, with probability $1 - \exp(-\Theta(\log s))$, we have

$$\|U_S\|_{\ell_\infty/\ell_2} \leq \mathcal{O}\left(\sqrt{\frac{(\log s)}{n}} \right) + \lambda_n \left(D_{\max} + \mathcal{O}\left(\sqrt{\frac{s^2}{n}} \right) \right).$$

which proves lemma 2.

D Analysis of $\mathcal{E}(V_{S^c})$: proofs.

By definition of the model (1) and by construction of the primal-dual pair $(\widehat{B}, \widehat{Z})$, the following conditional independence hold, and play a key role in the following analysis.

$$W \perp\!\!\!\perp X_{S^c} \mid X_S, \qquad \widehat{Z}_S \perp\!\!\!\perp X_{S^c} \mid X_S, \quad \text{and} \quad \widehat{Z}_S \perp\!\!\!\perp X_{S^c} \mid \{X_S, W\}.$$

D.1 Proof of Lemma 3

Statement of lemma 3:

1. *Under assumption A2, $T_1' \leq 1 - \gamma$.*
2. *Under conditions (5) of Theorem 1, $T_2' = o_p(1)$.*

Both terms T_1' and T_2' rely on the matrix expectations $\mathbb{E}[V \mid X_S]$ and $\mathbb{E}[V \mid X_S, W]$ which lemmas D.2.1 and D.2.2 show to be respectively:

$$\mathbb{E}[V \mid X_S] = -\lambda_n \Sigma_{S^cS}\Sigma_{SS}^{-1}\mathbb{E}[\widehat{Z}_S|X_S] \qquad \text{and} \qquad \mathbb{E}[V \mid X_S, W] = -\frac{\lambda_n}{n}\Sigma_{S^cS}\Sigma_{SS}^{-1}\widehat{Z}_S.$$

For $T_1' = \|\mathbb{E}[V \mid X_S]\|_{\ell_\infty/\ell_2}$, using the matrix-norm inequality ($3^\sharp$) and then Jensen's inequality yields the announced result:

$$T_1' \;=\; \|\Sigma_{S^cS}\Sigma_{SS}^{-1}\mathbb{E}[\widehat{Z}_S|X_S]\|_{\ell_\infty/\ell_2} \leq |\!|\!| \Sigma_{S^cS}\Sigma_{SS}^{-1} |\!|\!|_\infty \mathbb{E}[\|\widehat{Z}_S\|_{\ell_\infty/\ell_2}|X_S] \leq (1-\gamma).$$

For $T_2' = \|\mathbb{E}[V \mid X_S] - \mathbb{E}[V \mid X_S, W]\|_{\ell_\infty/\ell_2}$, using again inequality ($3^\sharp$), we have

$$\begin{aligned} T_2' &\leq |\!|\!| \Sigma_{S^cS}(\Sigma_{SS})^{-1} |\!|\!|_\infty \|\widehat{Z}_S - \mathbb{E}[\widehat{Z}_S|X_S]\|_{\ell_\infty/\ell_2} \\ &\leq (1-\gamma)\,\mathbb{E}\left[\left\|\widehat{Z}_S - Z_S^*\right\|_{\ell_\infty/\ell_2}\right] + (1-\gamma)\left\|\widehat{Z}_S - Z_S^*\right\|_{\ell_\infty/\ell_2} \end{aligned}$$

But Lemma D.2.3, which relates the consistency of the primal variables to the consistency of dual variables, shows that $\left\|\widehat{Z}_S - Z_S^*\right\|_{\ell_\infty/\ell_2} = o_p(1)$, so that the (sub)gradients of the regularization are consistent on the support S. This shows that $T_2' = o_p(1)$.

D.2 Technical lemmas

Lemma D.2.1. $\mathbb{E}[V \mid X_S] = -\lambda_n \Sigma_{S^cS}\Sigma_{SS}^{-1}\mathbb{E}[\widehat{Z}_S|X_S]$.

Proof. Using the conditional independencies $W \perp\!\!\!\perp X_{S^c} \mid X_S$ and $\widehat{Z}_S \perp\!\!\!\perp X_{S^c} \mid X_S$, we have

$$\mathbb{E}[V \mid X_S] \;=\; \mathbb{E}[X_{S^c}^T|X_S]\left([\Pi_S - I_n]\frac{\mathbb{E}[W|X_S]}{n} - \lambda_n \frac{X_S}{n}(\widehat{\Sigma}_{SS})^{-1}\mathbb{E}[\widehat{Z}_S|X_S]\right).$$

Since $\mathbb{E}[W|X_S] = 0$, the first term vanishes, and using $\mathbb{E}[X_{S^c}^T|X_S] = \Sigma_{S^cS}\Sigma_{SS}^{-1}X_S^T$, we obtain the announced expression. □

Lemma D.2.2. $\mathbb{E}\,[V \mid X_S, W] = -\frac{\lambda_n}{n}\Sigma_{S^cS}\Sigma_{SS}^{-1}\widehat{Z}_S$.

Proof. Appealing to the conditional independence $\widehat{Z}_S \perp\!\!\!\perp X_{S^c} \mid \{X_S, W\}$, we have

$$\mathbb{E}\,[V \mid X_S, W] \;=\; \mathbb{E}\,[X_{S^c}^T \mid X_S, W]\left([\Pi_S - I_n]\frac{W}{n} - \lambda_n \frac{X_S}{n}\,(\widehat{\Sigma}_{SS})^{-1}\mathbb{E}\,[\widehat{Z}_S | X_S, W]\right).$$

Observe that $\mathbb{E}\,[\widehat{Z}_S|X_S, W] = \widehat{Z}_S$ because (X_S, W) uniquely specifies $\widehat{B}_S$ through the convex program (7), and the triple $(X_S, W, \widehat{B}_S)$ defines $\widehat{Z}_S$ through equation (6b). Moreover, the noise term disappears because the kernel of the orthogonal projection matrix $(I_n - \Pi_S)$ is the same as the range space of X_S, and

$$\begin{aligned}\mathbb{E}\,[X_{S^c}^T \mid X_S, W][\Pi_S - I_n] &= \mathbb{E}\,[X_{S^c}^T \mid X_S][\Pi_S - I_n] \\ &= \Sigma_{S^cS}\Sigma_{SS}^{-1}X_S^T[\Pi_S - I_n] \;=\; 0.\end{aligned}$$

The result follows from the fact that $\mathbb{E}\,[X_{S^c}^T \mid X_S, W] = \mathbb{E}\,[X_{S^c}^T \mid X_S] = \Sigma_{S^cS}\Sigma_{SS}^{-1}X_S^T$. □

Lemma D.2.3. *Define the matrix $\Delta \in \mathbb{R}^{s\times K}$ with rows $\Delta_i := U_i/\|\beta_i^*\|_2$. As long as $\|\Delta_i\|_2 \le 1/2$ for all row indices $i \in S$, we have*

$$\left\|\widehat{Z}_S - \zeta(B_S^*)\right\|_{\ell_\infty/\ell_2} \;\le\; 4\,\|\Delta\|_{\ell_\infty/\ell_2}.$$

Hence $\|\Delta\|_{\ell_\infty/\ell_2} = o_p(1)$ (shown in Sec. 3.1) implies that $\left\|\widehat{Z}_S - \zeta(B_S^*)\right\|_{\ell_\infty/\ell_2} = o_p(1)$.

Proof. From lemma 2 , the condition $\|\Delta_i\|_2 \le 1/2$ implies that $\widehat{\beta}_i \ne \vec{0}$ and hence $\widehat{Z}_i = \widehat{\beta}_i/\|\widehat{\beta}_i\|_2$ for all rows $i \in S$. Therefore, using the notation $Z_i^* = \beta_i^*/\|\beta_i^*\|_2$ we have

$$\begin{aligned}\widehat{Z}_i - Z_i^* &= \frac{\widehat{\beta}_i}{\|\widehat{\beta}_i\|_2} - Z_i^* = \frac{Z_i^* + \Delta_i}{\|Z_i^* + \Delta_i\|_2} - Z_i^* \\ &= Z_i^*\left(\frac{1}{\|Z_i^* + \Delta_i\|_2} - 1\right) + \frac{\Delta_i}{\|Z_i^* + \Delta_i\|_2}.\end{aligned}$$

Note that, for $z \ne 0$, $g(z,\delta) = \frac{1}{\|z+\delta\|_2}$ is differentiable with respect to δ, with gradient $\nabla_\delta\, g(z,\delta) = -\frac{z+\delta}{2\|z+\delta\|_2^3}$. By the mean-value theorem, there exists $h \in [0,1]$ such that

$$\frac{1}{\|z+\delta\|_2} - 1 = g(z,\delta) - g(z,0) = \nabla_\delta\, g(z, h\delta)^T\delta = -\frac{(z+h\delta)^T\delta}{2\|z+h\delta\|_2^3},$$

which implies that there exists $h_i \in [0,1]$ such that

$$\begin{aligned}\|\widehat{Z}_i - Z_i^*\|_2 &\leq \|Z_i^*\|_2 \frac{|(Z_i^* + h_i\Delta_i)^T \Delta_i|}{2\|Z_i^* + h_i\Delta_i\|_2^3} + \frac{\|\Delta_i\|_2}{\|Z_i^* + \Delta_i\|_2} \\ &\leq \frac{\|\Delta_i\|_2}{2\|Z_i^* + h_i\Delta_i\|_2^2} + \frac{\|\Delta_i\|_2}{\|Z_i^* + \Delta_i\|_2}. \qquad (4^\sharp)\end{aligned}$$

We note that $\|Z_i^*\|_2 = 1$ and $\|\Delta_i\|_2 \leq \frac{1}{2}$ imply that $\|Z_i^* + h_i\Delta_i\|_2 \geq \frac{1}{2}$. Combined with inequality $(4^\sharp)$, we obtain $\|\widehat{Z}_i - Z_i^*\|_2 \leq 4\|\Delta_i\|_2$, which proves the lemma. □

D.3 Proof of Lemma 4

We begin by noting that conditionally on X_S and W, each vector $V_j \in \mathbb{R}^K$ is normally distributed. Since $\mathrm{Cov}(X^{(j)} \mid X_S, W) = (\Sigma_{S^c|S})_{jj}\, I_n$, we have

$$\mathrm{Cov}(V_j \mid X_S, W) \;=\; M_n\, (\Sigma_{S^c|S})_{jj}$$

where the $K \times K$ matrix $M_n = M_n(X_S, W)$ is given by

$$M_n \;:=\; \frac{\lambda_n^2}{n} \widehat{Z}_S^T (\widehat{\Sigma}_{SS})^{-1} \widehat{Z}_S + \frac{1}{n^2} W^T (\Pi_S - I_n) W. \qquad (5^\sharp)$$

In the expression of M_n, the cross terms of the form $W^T(\Pi_S - I_n)(\widehat{\Sigma}_{SS})^{-1}\widehat{Z}_S$ vanish in the previous expression because of the same orthogonality arguments as in the proof of lemma D.2.2. Conditionally on W and X_S, the matrix M_n is fixed, and we have

$$\left(\|V_j - \mathbb{E}\left[V_j \mid X_S,\, W\right]\|_2^2 \,|\, W, X_S \right) \;\overset{d}{=}\; \left(\Sigma_{S^c\,|\,S}\right)_{jj} \xi_j^T M_n \xi_j.$$

where $\xi_j \sim N(\vec{0}_K, I_K)$.

D.4 Proof of Lemma 5

Statement of lemma 5:

Under the conditions (5) of Theorem 1, $\|\!| M_n - M^* |\!\|_2 = o_p(\|\!| M^* |\!\|_2)$ *where*

$$M^* = \frac{\lambda_n^2}{n} (Z_S^*)^T (\Sigma_{SS})^{-1} Z_S^*, \quad \text{so that} \quad \|\!| M^* |\!\|_2 = \lambda_n^2 \, \frac{\psi(B^*)}{n}.$$

Consequently, for any $\delta > 0$ *the following event* $\mathcal{T}(\delta)$ *has probability converging to 1.*

$$\mathcal{T}(\delta) \;:=\; \left\{ \|\!| M_n |\!\|_2 \leq \lambda_n^2 \, \frac{\psi(B^*)}{n} \, (1+\delta) \right\}.$$

With $Z_S^* = \zeta(B_S^*)$, define the $K \times K$ random matrix

$$M_n^* \;:=\; \frac{\lambda_n^2}{n} (Z_S^*)^T (\widehat{\Sigma}_{SS})^{-1} Z_S^* + \frac{1}{n^2} W^T (I_n - \Pi_S) W$$

and note that (using standard results on Wishart matrices [And84])

$$\mathbb{E}\left[M_n^*\right] \quad = \quad \frac{\lambda_n^2}{n-s-1}(Z_S^*)^T(\Sigma_{SS})^{-1}Z_S^* + \sigma^2\frac{n-s}{n^2}I_K.$$

To bound M_n from M^* in spectral norm, we use the triangle inequality:

$$|\!|\!| M_n - M^* |\!|\!|_2 \;\leq\; |\!|\!| M_n - M_n^* |\!|\!|_2 + |\!|\!| M_n^* - \mathbb{E}\left[M_n^*\right] |\!|\!|_2 + |\!|\!| \mathbb{E}\left[M_n^*\right] - M^* |\!|\!|_2 \quad (6^\sharp)$$

First, we have $|\!|\!| M_n^* - \mathbb{E}\left[M_n^*\right] |\!|\!|_2 \leq T_1^\dagger + T_2^\dagger$ where

$$T_1^\dagger \quad = \quad \frac{\lambda_n^2}{n}\, |\!|\!| Z_S^* |\!|\!|_2^2 \left|\!\left|\!\left| \frac{n}{n-s-1}(\Sigma_{SS})^{-1} - (\widehat{\Sigma}_{SS})^{-1} \right|\!\right|\!\right|_2 \;=\; o_p\left(\frac{\lambda_n^2 s}{n}\right),$$

since $|\!|\!| Z_S^* |\!|\!|_2^2 \leq s$, and $\left|\!\left|\!\left| \frac{n}{n-s-1}(\Sigma_{SS})^{-1} - (\widehat{\Sigma}_{SS})^{-1} \right|\!\right|\!\right|_2 = o_p(1)$, and

$$T_2^\dagger \quad := \quad \frac{1}{n^2}\; |\!|\!| W^T(I_n - \Pi_S)W - \sigma^2(n-s)I_K |\!|\!|_2 \;=\; \mathcal{O}_p\left(\frac{1}{n}\right) \;=\; o_p\left(\frac{\lambda_n^2 s}{n}\right),$$

since $\lambda_n^2 s \to +\infty$. Overall, we conclude that

$$|\!|\!| M_n^* - \mathbb{E}\left[M_n^*\right] |\!|\!|_2 \quad = \quad o_p\left(\frac{\lambda_n^2 s}{n}\right). \qquad (7^\sharp)$$

Then considering the first term in decomposition $(6^\sharp)$, we have

$$\begin{aligned}
|\!|\!| M_n^* - M_n |\!|\!|_2 \quad &= \quad \frac{\lambda_n^2}{n}\left|\!\left|\!\left| Z_S^*\widehat{\Sigma}_{SS}^{-1}Z_S^* - \widehat{Z}_S\widehat{\Sigma}_{SS}^{-1}\widehat{Z}_S \right|\!\right|\!\right|_2 \\
&= \quad \frac{\lambda_n^2}{n}\left|\!\left|\!\left| Z_S^*\widehat{\Sigma}_{SS}^{-1}(Z_S^* - \widehat{Z}_S) + (Z_S^* - \widehat{Z}_S)\widehat{\Sigma}_{SS}^{-1}(Z_S^* + (\widehat{Z}_S - Z_S^*)) \right|\!\right|\!\right|_2 \\
&\leq \quad \frac{\lambda_n^2}{n}\left|\!\left|\!\left| \widehat{\Sigma}_{SS}^{-1} \right|\!\right|\!\right|_2 \left|\!\left|\!\left| Z_S^* - \widehat{Z}_S \right|\!\right|\!\right|_2 \left(2\, |\!|\!| Z_S^* |\!|\!|_2 + \left|\!\left|\!\left| Z_S^* - \widehat{Z}_S \right|\!\right|\!\right|_2\right)
\end{aligned}$$

Moreover, since $\left|\!\left|\!\left| \widehat{\Sigma}_{SS}^{-1} \right|\!\right|\!\right|_2 = \mathcal{O}_p(1)$, $|\!|\!| Z_S^* |\!|\!|_2 = \mathcal{O}_p(\sqrt{s})$, $\left|\!\left|\!\left| Z_S^* - \widehat{Z}_S \right|\!\right|\!\right|_2 \leq \sqrt{s}\left\| Z_S^* - \widehat{Z}_S \right\|_{\ell_\infty/\ell_2}$ from Corollary B.0.2 and $\left\| Z_S^* - \widehat{Z}_S \right\|_{\ell_\infty/\ell_2} = o_p(1)$ from Lemma D.2.3, we conclude that

$$|\!|\!| M_n^* - M_n |\!|\!|_2 = o_p\left(\frac{\lambda_n^2 s}{n}\right). \qquad (8^\sharp)$$

For the matrix M^*, we have

$$|\!|\!| M^* |\!|\!|_2 = \frac{\lambda_n^2}{n-s-1}\,\psi(B^*) + \frac{\sigma^2}{n}\left(1 - \frac{s}{n}\right) = (1 + o(1))\left[\frac{\lambda_n^2 \psi(B^*)}{n}\right]. \qquad (9^\sharp)$$

Therefore $|\!|\!| M^* |\!|\!|_2 = \Theta(\lambda_n^2 s/n)$. Moreover, since

$$\left(\frac{1}{n} - \frac{1}{n-s-1}\right)\lambda_n^2\,\psi(B^*) = o\left(\frac{\lambda_n^2\, s}{n}\right), \quad \text{and} \quad \frac{\sigma^2}{n}\left(1 - \frac{s}{n}\right) = o\left(\frac{\lambda_n^2\, s}{n}\right)$$

using the first condition (5) on λ_n, we have

$$|\!|\!| M^* - \mathbb{E}[M_n^*] |\!|\!|_2 = o\left(\frac{\lambda_n^2 s}{n}\right) \tag{$10^\sharp$}$$

Combining bounds ($7^\sharp$), ($8^\sharp$), ($10^\sharp$) in the decomposition ($6^\sharp$) and ($9^\sharp$) shows that $|\!|\!| M_n - M^* |\!|\!|_2 = o_p(|\!|\!| M^* |\!|\!|_2)$ so that we can conclude that for any $\delta > 0$ the event

$$\mathcal{T}(\delta) := \left\{ |\!|\!| M_n |\!|\!|_2 \leq \lambda_n^2 \frac{\psi(B^*)}{n} (1+\delta) \right\}$$

has probability converging to 1.

D.5 Proof of Lemma 6

Statement of lemma 6:
If there exists $\nu > 0$, such that $t^*(n, B^*) > (1+\nu)\log(p-s)$, then

$$\mathbb{P}\left[\max_{j \in S^c} \|\xi_j\|_2^2 \geq 2t^*(n, B^*)\right] \to 0 \,.$$

Note that $t^* \to +\infty$ under the specified scaling of (n, p, s). By applying Lemma E.0.1 from Appendix E on large deviations for χ^2 variates with $t = t^*(n, B^*)$, we obtain

$$\mathbb{P}[T_3' \geq \gamma \mid \mathcal{T}(\delta)] \quad \leq \quad (p-s)\, \exp\left(-t^* \left[1 - 2\sqrt{\frac{K}{t^*}}\right]\right) \leq (p-s)\, \exp\left(-t^*\,(1-\delta)\right),$$

for (n, p, s) sufficiently large. Thus, the bound ($11^\sharp$) tends to zero as long as there exists $\nu > 0$ such that we have $(1-\delta)\, t^*(n, B^*) > (1+\nu)\log(p-s)$, or equivalently and as claimed

$$n \quad > \quad (1+\nu) \frac{(1+\delta)}{(1-\delta)} \frac{C_{\max}}{\gamma^2} \, [2\, \psi(B^*)\, \log(p-s)] \,.$$

E Large deviations for χ^2-variates

Lemma E.0.1. *Let $Z_1, \ldots, Z_m$ be i.i.d. χ^2-variates with d degrees of freedom. Then for all $t > d$, we have*

$$\mathbb{P}[\max_{i=1,\ldots,m} Z_i \geq 2t] \quad \leq \quad m\, \exp\Big(-t\Big[1 - 2\sqrt{\frac{d}{t}}\Big]\Big). \tag{$11^\sharp$}$$

Proof. Given a central χ^2-variate X with d degrees of freedom, Laurent and Massart [LM98] prove that $\mathbb{P}[X - d \geq 2\sqrt{d\,x} + 2x] \leq \exp(-x)$, or equivalently

$$\mathbb{P}\left[X \geq x + (\sqrt{x} + \sqrt{d})^2\right] \quad \leq \quad \exp(-x),$$

valid for all $x > 0$. Setting $\sqrt{x} + \sqrt{d} = \sqrt{t}$, we have

$$\begin{aligned} \mathbb{P}[X \geq 2t] \stackrel{(a)}{\leq} \mathbb{P}\left[X \geq (\sqrt{t} - \sqrt{d})^2 + t\right] &\leq \exp(-(\sqrt{t} - \sqrt{d})^2) \\ &\leq \exp(-t + 2\sqrt{td}) \\ &= \exp\left(-t\left[1 - 2\sqrt{\frac{d}{t}}\right]\right), \end{aligned}$$

where inequality (a) follows since $\sqrt{t} \geq \sqrt{d}$ by assumption. Thus, the claim ($11^\sharp$) follows by the union bound. □

References

[And84] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley, New York, 1984.

[Ber95] D.P. Bertsekas. *Nonlinear programming.* Athena Scientific, Belmont, MA, 1995.

[BV04] S. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge University Press, Cambridge, UK, 2004.

[DS01] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices, and Banach spaces. In *Handbook of Banach Spaces*, volume 1, pages 317–336. Elsevier, Amsterdan, NL, 2001.

[LM98] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1303–1338, 1998.