Supplement to
'Sparse recovery by thresholded non-negative least squares'

Martin Slawski and Matthias Hein
Department of Computer Science
Saarland University
Campus E 1.1, Saarbrücken, Germany
{ms,hein}@cs.uni-saarland.de

Abstract

We here provide additional proofs, definitions, lemmas and derivations omitted in the paper. Note that material contained in the latter are referred to by the captions used there (e.g. Theorem 1), whereas auxiliary statements contained exclusively in this supplement are preceded by a capital Roman letter (e.g. Theorem A.1).

A Sub-Gaussian random variables and concentration inequalities

A random variable Z is called sub-Gaussian if there exists a positive constant K such that $\mathbf{E}[|Z|^q]^{1/q} \leq K\sqrt{q}$. The smallest such K is called the sub-Gaussian norm $\|Z\|_{\psi_2}$ of Z. If $\mathbf{E}[Z] = 0$, which shall be assumed for the remainder of this paragraph, then the moment-generating function of Z satisfies $\mathbf{E}[\exp(tZ)] \leq \exp(-t^2/(2\sigma^2))$ for a parameter $\sigma > 0$ which is related to $\|Z\|_{\psi_2}$ by a multiplicative constant, cf. [1]. It follows that if $Z_1, \ldots, Z_n$ are i.i.d. copies of Z and $v \in \mathbb{R}^n$, then $\sum_{i=1}^n v_i Z_i$ is sub-Gaussian with parameter $\|v\|_2^2\,\sigma^2$. We have the well-known tail bound

$$\mathbf{P}(|Z| > z) \leq 2\exp\left(-\frac{z^2}{2\sigma^2}\right), \quad z \geq 0. \tag{A.1}$$

Combining the previous two facts and using a union bound, with $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$, it follows that for any collection of vectors $v_j \in \mathbb{R}^n$, $j = 1, \ldots, p$,

$$\mathbf{P}\left(\max_{1 \leq j \leq p} |v_j^\top \mathbf{Z}| > \sigma \max_{1 \leq j \leq p} \|v_j\|_2 \sqrt{2\log p} + \sigma z\right) \leq 2\exp\left(-\frac{1}{2}z^2\right), \quad z \geq 0. \tag{A.2}$$

A.1 Bernstein-type inequality for squared sub-Gaussian random variables

The following exponential inequality combines Lemma 14, Proposition 16 and Remark 18 in [1].

Lemma A. 1. *Let $Z_1, \ldots, Z_n$ be i.i.d. centered sub-Gaussian random variables with sub-Gaussian norm K. Then for every $a = (a_1, \ldots, a_n)^\top \in \mathbb{R}^n$ and every $z \geq 0$, one has*

$$\mathbf{P}\left(\left|\sum_{i=1}^n a_i (Z_i^2 - \mathbf{E}[Z_i^2])\right| > z\right) \leq 2\exp\left(-c\min\left(\frac{z^2}{K^4\|a\|_2^2}, \frac{z}{K^2\|a\|_\infty}\right)\right), \tag{A.3}$$

where $c > 0$ is an absolute constant.

A.2 Concentration of extreme singular values of sub-Gaussian random matrices

Denote by $s_{\min}(X)$ and $s_{\max}(X)$ the minimum and maximum singular value of a matrix X. The following statement is a special case covered by Theorem 39 in [1].

Theorem A. 1. *Let X be an $n \times s$ matrix with i.i.d. centered sub-Gaussian entries having unit variance and sub-Gaussian norm K. Then for every $z \geq 0$, with probability at least $1-2\exp(-cz^2)$, one has*

$$\sqrt{n} - C\sqrt{s} - z \leq s_{\min}(X) \leq s_{\max}(X) \leq \sqrt{n} + C\sqrt{s} + z, \text{ and} \tag{A.4}$$

$$s_{\max}\left(\frac{1}{n}X^\top X - I\right) \leq \max(\delta, \delta^2), \text{ where } \delta = C\sqrt{\frac{s}{n}} + \frac{z}{\sqrt{n}}, \tag{A.5}$$

with C, c depending only on K.

B Proof of Theorem 1

Self-regularizing property. We call a design self-regularizing with universal constant $\kappa \in (0,1]$ if

$$\beta^\top \Sigma \beta \geq \kappa(\mathbf{1}^\top \beta)^2 \quad \forall \beta \succeq 0. \tag{B.1}$$

Theorem 1 *Let Σ fulfill the self-regularizing property with constant κ. Then, with probability no less than 1 - $2/p$, the NNLS estimator obeys*

$$\frac{1}{n}\|X\beta^* - X\widehat{\beta}\|_2^2 \leq \frac{8\sigma}{\kappa}\sqrt{\frac{2\log p}{n}}\|\beta^*\|_1 + \frac{8\sigma^2}{\kappa}\frac{\log p}{n}.$$

Proof. For some vector $\delta \in \mathbb{R}^p$, set $P = \{j : \delta_j \geq 0\}$ and $N = \{j : \delta_j < 0\}$ and define $\widehat{\delta} = \beta^* - \widehat{\beta}$, where $\widehat{\beta}$ is a minimizer of the NNLS criterion. We will bound the ℓ_1-norm of $\widehat{\delta}$. Note that by feasibility of $\widehat{\beta}$, we have $\widehat{\delta} \preceq \beta^*$ and hence $\|\widehat{\delta}_P\|_1 \leq \|\beta^*\|_1$. By definition, $\widehat{\delta}$ minimizes

$$\frac{2}{n}\varepsilon^\top X\delta + \delta_P^\top \widehat{\Sigma}_{PP}\delta_P + 2\delta_P^\top \widehat{\Sigma}_{PN}\delta_N + \delta_N^\top \widehat{\Sigma}_{NN}\delta_N. \tag{B.2}$$

over all feasible $\delta \preceq \beta^*$. The ℓ_1-norm $\|\widehat{\delta}_N\|_1$ can be controlled by bounding the ℓ_1-norm of any minimizer $\widehat{d}$ of the problem

$$\min_{d \preceq 0} \frac{2}{n}\varepsilon^\top X_N d + 2\|\beta^*\|_1 \mathbf{1}^\top d + \kappa(\mathbf{1}^\top d)^2, \tag{B.3}$$

where (B.3) is obtained from (B.2) by omitting terms not depending on δ_N and replacing $\delta_P^\top \widehat{\Sigma}_{PN}\delta_N$ and $\delta_N^\top \widehat{\Sigma}_{NN}\delta_N$ by the lower bounds

$$\delta_P^\top \widehat{\Sigma}_{PN}\delta_N \geq -\|\beta^*\|_1 \mathbf{1}^\top \delta_N, \tag{B.4}$$

$$\delta_N^\top \widehat{\Sigma}_{NN}\delta_N \geq \kappa(\mathbf{1}^\top \delta_N)^2, \tag{B.5}$$

where (B.4) follows from the ℓ_1-bound on $\widehat{\delta}_P$ in combination with Hölder's inequality, and (B.5) is obtained by invoking the self-regularizing property (B.1). These replacements evidently ensure that $\|\widehat{\delta}_N\|_1 \leq \|\widehat{d}\|_1$, where $\widehat{d}$ is any minimizer of (B.3). The KKT optimality conditions of the quadratic program (B.3) read

$$\frac{1}{n}X_N^\top \varepsilon + \|\beta^*\|_1 \mathbf{1} + \kappa(\mathbf{1}^\top \widehat{d})\mathbf{1} + \widehat{\mu} = 0,$$
$$\widehat{d} \preceq 0, \quad \widehat{\mu} \succeq 0, \quad \widehat{\mu}_k \widehat{d}_k = 0, \; k = 1, \ldots, |N|,$$

where $\widehat{\mu}$ is a Langrangian multiplier. From the first equation, it follows that

$$\|\widehat{d}\|_1 \leq \frac{\|\beta^*\|_1 + A}{\kappa}, \quad A = \left\|\frac{X^\top \varepsilon}{n}\right\|_\infty.$$

Since β^* is feasible for the NNLS problem, using $\|\widehat{\delta}_N\|_1 \leq \|\widehat{d}\|_1$ and $\kappa < 1$, we have

$$\frac{1}{n}\|X\beta^* - X\widehat{\beta}\|_2^2 = \frac{1}{n}\|X\widehat{\delta}\|_2^2 \leq -\frac{2}{n}\varepsilon^\top X\widehat{\delta} \leq 2A(\|\widehat{\delta}_P\|_1 + \|\widehat{\delta}_N\|_1) \leq \frac{4A\|\beta^*\|_1 + 2A^2}{\kappa}.$$

Using the maximal inequality (A.2) for a finite collection of sub-Gaussian random variables, the event $\left\{A \leq 2\sigma\sqrt{\frac{2\log p}{n}}\right\}$ holds with probability no less than $1 - 2/p$. The result follows.

□

C Addendum for Definition 2

Lemma C. 1.

(i) $\widehat{\omega}(S) > 0 \Leftrightarrow \widehat{\tau}(S) > 0 \Leftrightarrow X_S \mathbb{R}_+^s$ *is a face of* $\mathcal{C}$.

(ii) $\widehat{\omega}(S) \leq 1$ *with equality if* $\{X_j\}_{j \in S}$ *and* $\{X_j\}_{j \in S^c}$ *are orthogonal and* $\frac{1}{n} X_{S^c}^\top X_{S^c}$ *is entry-wise non-negative.*

Proof. (i): We have

$$\widehat{\tau}^2(S) = \min_{\theta \in \mathbb{R}^s,\ \lambda \in T^{p-s-1}} \frac{1}{n} \left\| X_S \theta - X_{S^c} \lambda \right\|_2^2 = \min_{\lambda \in T^{p-s-1}} \frac{1}{n} \left\| Z\lambda \right\|_2^2, \text{ hence} \tag{C.1}$$

$$\exists \widehat{\lambda} \in T^{p-s-1} \text{ s.t. } Z\widehat{\lambda} = 0 \ \Rightarrow Z^\top Z \widehat{\lambda} = 0 \ \Rightarrow \| Z^\top Z \widehat{\lambda} \|_\infty = 0 \ \Rightarrow \widehat{\omega}(S) = 0.$$

On the other hand

$$\exists \widehat{v} \in \mathcal{V}(F) \text{ s.t. } \| Z_F^\top Z_F \widehat{v} \|_\infty = 0 \ \Rightarrow \ Z_F^\top Z_F \widehat{v} = 0 \ \Rightarrow \| Z_F \widehat{v} \|_2^2 \ \Rightarrow \widehat{\tau}(S) = 0.$$

The second equivalence is by the definition of a face of a cone.
(ii) Consider all principal sub-matrices $\frac{1}{n} Z_F^\top Z_F$. By definition, $\widehat{\omega}(S)$ equals the maximum of the absolute values of the entries of $\frac{1}{n} Z_F^\top Z_F v$, where one minimizes over all v contained in the boundary of the unit cube in $[0,1]^{|F|}$. We may restrict our attention to matrices $\frac{1}{n} Z^\top Z$ which are entry-wise non-negative. To see this, assume that there exists a non-negative off-diagonal entry for a pair (j,k). Then pick $F_0 = \{j,k\}$ and set $\mathcal{V}(F_0) = \{v \in \mathbb{R}^2 : \ v \succeq 0, \|v\|_\infty = 1\}$ to obtain that

$$\begin{aligned}
\widehat{\omega}(S) \leq \min_{v \in \mathcal{V}(F_0)} \left\| \frac{1}{n} Z_{F_0}^\top Z_{F_0} v \right\|_\infty &\leq \max \Big\{ \frac{1}{n} (Z^\top Z)_{jj} + \frac{1}{n} (Z^\top Z)_{jk}, \\
&\qquad\qquad \frac{1}{n} (Z^\top Z)_{kk} + \frac{1}{n} (Z^\top Z)_{jk} \Big\} \\
&\leq \max \left\{ \frac{1}{n} (Z^\top Z)_{jj}, \frac{1}{n} (Z^\top Z)_{kk} \right\} \leq 1,
\end{aligned}$$

re-calling that $\|Z_j\|_2^2 = \left\| \Pi_S^\perp X_j \right\|_2^2 \leq \|X_j\|_2^2 = n$ for all j. If $Z^\top Z$ is entry-wise non-negative, a similar argument shows that $\widehat{\omega}(S)$ equals the minimum diagonal entry of $\frac{1}{n} Z^\top Z$, which is upper bounded by 1. Since

$$\frac{1}{n} Z^\top Z = \frac{1}{n} X_{S^c}^\top X_{S^c} - \frac{1}{n} X_{S^c}^\top X_S \left(\frac{1}{n} X_S^\top X_S \right)^{-1} X_S^\top X_{S^c},$$

orthogonality implies that $\frac{1}{n} Z^\top Z = \frac{1}{n} X_{S^c}^\top X_{S^c}$. Using entry-wise non-negativity of $\frac{1}{n} X_{S^c}^\top X_{S^c}$ together with $\|X_j\|_2^2 = n$, the assertion follows. □

D Proofs of Lemma 1 and Lemma 2

Lemma 1 *$\widehat{\beta}$ is a minimizer of the NNLS problem if and only if there exists $F \subseteq \{1, \ldots, p\}$ such that*

$$\frac{1}{n} X_j^\top (y - X\widehat{\beta}) = 0, \text{ and } \widehat{\beta}_j > 0, \ j \in F, \quad \frac{1}{n} X_j^\top (y - X\widehat{\beta}) \leq 0, \text{ and } \widehat{\beta}_j = 0, \ j \in F^c.$$

Proof. For $\mu, \beta \succeq 0$, the Lagrangian of the NNLS problem is given by

$$\mathcal{L}(\beta, \mu) = \frac{1}{n} \|y - X\beta\|_2^2 - \mu^\top \beta.$$

Lemma 1 is then immediately obtained from the resulting KKT optimality conditions. □

Lemma 2 *Consider the two non-negative least squares problems*

$$(P1): \min_{\beta^{(P1)} \succeq 0} \frac{1}{n} \|\Pi_S^{\perp}(\varepsilon - X_{S^c}\beta^{(P1)})\|_2^2 \quad (P2): \min_{\beta^{(P2)} \succeq 0} \frac{1}{n} \|\Pi_S y - X_S \beta^{(P2)} - \Pi_S X_{S^c} \widehat{\beta}^{(P1)}\|_2^2$$

with minimizers $\widehat{\beta}^{(P1)}$ of $(P1)$ and $\widehat{\beta}^{(P2)}$ of $(P2)$, respectively. If $\widehat{\beta}^{(P2)} \succ 0$, then setting $\widehat{\beta}_S = \widehat{\beta}^{(P2)}$ and $\widehat{\beta}_{S^c} = \widehat{\beta}^{(P1)}$ yields a minimizer $\widehat{\beta}$ of the non-negative least squares problem.

Proof. The NNLS objective is split into two parts in the following way:

$$\min_{\beta \succeq 0} \frac{1}{n} \|y - X\beta\|_2^2 = \min_{\beta \succeq 0} \frac{1}{n} \|\Pi_S y - X_S \beta_S - \Pi_S X_{S^c} \beta_{S^c}\|_2^2 + \frac{1}{n} \|\xi - Z\beta_{S^c}\|_2^2, \ \ \xi = \Pi_S^{\perp} \varepsilon. \tag{D.1}$$

Separate minimization of the second summand on the r.h.s. of (D.1) yields $\widehat{\beta}^{(P1)}$. Substituting $\widehat{\beta}^{(P1)}$ for β_{S^c} in the first summand, and minimizing the latter amounts to solving $(P2)$. In view of Lemma 1, if $\widehat{\beta}^{(P2)} \succ 0$, it coincides with the unconstrained least squares estimator (D.1) corresponding to problem $(P2)$. This implies that the optimal value of $(P2)$ must be zero, because the observation vector of the non-negative least squares problem $(P2)$ is contained in the column space of X_S. Since the second summand in (D.1) corresponding to $(P1)$ cannot be made smaller than by separate minimization, we have minimized the non-negative least squares objective. □

E Addendum for Examples 1 and Examples 2

E.1 Example 1

The Gram matrix $\Sigma = \frac{1}{n} X^{\top} X$ can be identified with a covariance matrix of a set of zero-mean, unit variance random variables $\{R_j\}_{j=1}^p$. Correspondingly, for any $S \subset \{1, \dots, p\}$, the matrix

$$\frac{1}{n} Z^{\top} Z = \frac{1}{n} X_{S^c}^{\top} (I - \Pi_S) X_{S^c} = \Sigma_{S^c S^c} - \Sigma_{S^c S} \Sigma_{SS}^{-1} \Sigma_{S S^c} \tag{E.1}$$

can be interpreted as the *conditional* covariance matrix of the random variables $\{R_j\}_{j \in S^c}$ conditional on $\{R_j\}_{j \in S}$. The power decay structure of the matrix Σ induces a Markov random field (see [2]) so that the conditional covariances satisfy $\mathrm{Cov}(R_k, R_l | \{R_j\}_{j \in S}) \geq 0$, with equality if S contains an index j such that $k \wedge l < j < k \vee l$. The minimum diagonal entry of $\frac{1}{n} Z^{\top} Z$ used to lower bound $\widehat{\omega}(S)$ can be obtained from the following consideration.

$$\frac{1}{n} (Z^{\top} Z)_{jj} = \mathrm{Var}(R_j | \{R_k\}_{k \in S}) \geq \mathrm{Var}(R_j | \{R_{j-1}, R_{j+1}\}) = \sigma_{jj} - \Sigma_{j\mathcal{N}} (\Sigma_{\mathcal{N}\mathcal{N}})^{-1} \Sigma_{\mathcal{N} j}, \tag{E.2}$$

with $\mathcal{N} = \{j-1, j+1\}$ and

$$\Sigma_{j\mathcal{N}} = [\rho \ \rho], \quad \Sigma_{\mathcal{N}\mathcal{N}} = \begin{bmatrix} 1 & \rho^2 \\ \rho^2 & 1 \end{bmatrix}.$$

Explicit computation of the r.h.s. of (E.2) then yields that $\frac{1}{n}(Z^{\top} Z)_{jj} \geq 1 - \frac{2\rho^2}{1+\rho^2}$.
Moreover, it is well-known (see again [2]) that the off-diagonal entries of the *inverse* of a covariance matrix are – up to a change in sign and a multiplicative factor – equal to the conditional covariances after conditioning on all remaining variables, i.e. for $j \neq k$, $(\Sigma^{-1})_{jk} \propto -\mathrm{Cov}(R_j, R_k | \{R_l\}_{l \notin \{j,k\}})$. In view of the Markov random field structure under consideration, which implies that all R_j are conditionally independent of all remaining variables given $\{R_{j-1}, R_{j+1}\}$, it thus follows that Σ^{-1} as well as the inverses of sub-matrices Σ_{SS}^{-1} have at most two non-zero off-diagonal entries per row. Hence, $K(S) = \max_{v: \, \|v\|_{\infty} = 1} \left\|\Sigma_{SS}^{-1} v\right\|_{\infty}$ and $\phi_{\min}(S) = \min_{v: \, \|v\|_2 = 1} \|\Sigma_{SS} v\|_2$ are necessarily upper and lower bounded by constants depending on ρ only, but not on s.

E.2 Example 2

One computes that

$$(\Sigma_{SS}^{-1})_{jk} = \frac{1}{(1-\rho)(1+(s-1)\rho)} \begin{cases} 1 + (s-2)\rho & j = k, \\ -\rho & j \neq k. \end{cases} \tag{E.3}$$

and consequently, using (E.1),

$$\left(\frac{1}{n}Z^\top Z\right)_{jk} = \begin{cases} 1-\rho^2 s/(1+(s-1)\rho) & j=k, \\ \rho-\rho^2 s/(1+(s-1)\rho) & j\neq k. \end{cases} \tag{E.4}$$

From (C.1), $\widehat{\tau}^2(S) = \min_{\lambda\in T^{p-s-1}} \lambda^\top \frac{1}{n} Z^\top Z\lambda$. In view of the simple structure (E.4), one verifies that the minimum is attained for $\lambda = \mathbf{1}/(p-s)$, which yields that

$$\widehat{\tau}^2(S) = \frac{(1-\rho)\rho}{(s-1)\rho+1} + \frac{1-\rho}{p-s} = O(s^{-1}), \tag{E.5}$$

and, with high probability,

$$\|\widehat{\beta}_{S^c}\|_1 \leq \frac{2\sigma\sqrt{2\log(p)/n}}{\widehat{\tau}^2(S)} \leq \frac{((s-1)\rho+1)2\sigma\sqrt{2\log(p)/n}}{(1-\rho)\rho}, \tag{E.6}$$

as given in the paper. Given the closed form expression (E.3), the bound (14) in the paper, which, for some vector v, reads

$$\|\Sigma_{SS}^{-1}\Sigma_{SS^c}v\|_\infty \leq \underbrace{\max_{v:\ \|v\|_\infty=1}\|\Sigma_{SS}^{-1}v\|_\infty}_{K(S)} \underbrace{\max_{j\in S, k\in S^c}|\sigma_{jk}|}_{\mu(S)} \|v\|_1$$

is replaced by

$$\|\Sigma_{SS}^{-1}\Sigma_{SS^c}v\|_\infty \leq \|\Sigma_{SS}^{-1}\mathbf{1}\|_\infty \|v\|_1 = \frac{\rho}{1+(s-1)\rho}\|v\|_1,$$

using the fact that all off-diagonal entries of Σ are equal to ρ. Applying the previous bound to $v = \widehat{\beta}_{S^c}$ together with (E.6) and $\phi_{\min}(S) = 1-\rho$ and following Step 3 in the proof of Theorem 2 in the paper, one obtains that with high probability,

$$\left\|\widehat{\beta}_S - \beta^*_S\right\|_\infty \leq \frac{4\sigma}{1-\rho}\sqrt{\frac{2\log p}{n}},$$

provided $\beta_{\min}(S)$ exceeds the right hand side. Moreover, (E.4) implies that $\widehat{\omega}(S) = 1-\rho^2 s/(1+(s-1)\rho)$, since all entries of $\frac{1}{n}Z^\top Z$ are non-negative. Consequently, choosing the threshold as $\lambda = \frac{2\sigma}{\widehat{\omega}(S)}\sqrt{\frac{2\log p}{n}}$ with $\widehat{\omega}(S)$ as above,

$$\left\|\widehat{\beta}(\lambda) - \beta^*\right\|_\infty \leq \frac{4\sigma}{1-\rho} + 2\sigma\left(1-\frac{\rho^2 s}{1+(s-1)\rho}\right)\sqrt{\frac{2\log p}{n}}.$$

F Proof of Theorem 3

Consider the following ensemble of random matrices

Ens$_+ = \{X = (x_{ij}), \{x_{ij},\ 1\leq i\leq n,\ 1\leq j\leq p\}$ i.i.d. from a sub-Gaussian distribution on $\mathbb{R}_+\}$.

Theorem 3 *Let X be a random matrix from Ens$_+$, scaled s.t. $\mathbf{E}\left[\frac{1}{n}X^\top X\right] = \rho I + (1-\rho)\mathbf{1}\mathbf{1}^\top$ for some $\rho\in(0,1)$. Fix an $S\subset\{1,\ldots,p\}$, $|S|\leq s$. Then there exists constants $c, c_1, c_2, c_3, C, C' > 0$ such that for all $n \geq C\log(p)s^2$,*

$$\widehat{\tau}^2(S) \geq cs^{-1} - C'\sqrt{\log(p)/n}$$

with probability no less than $1 - 3/p - \exp(-c_1 n) - 2\exp(-c_2\log p) - \exp(-c_3\log^{1/2}(p)s)$.

We state and prove three basic concentration results first.

Lemma F. 1. *Let $Z_1,\ldots,Z_n$ be i.i.d. centered, unit variance sub-Gaussian random variables with sub-Gaussian norm K. Then for all $z\geq 0$*

$$\mathbf{P}\left(\sum_{i=1}^n Z_i^2 > n + zn\right) \leq \exp(-c\min(\frac{z^2}{K^4}, \frac{z}{K^2})n). \tag{F.1}$$

Proof. Noting that $\mathbf{E}[\sum_{i=1}^n Z_i^2] = n$ and re-arranging, the result follows from Lemma A.1 with $a = (1, \dots, 1)^\top$. $\square$

In the sequel, we denote by Σ^* the population covariance $\mathbf{E}[\frac{1}{n} X^\top X] = (1-\rho) I_p + \rho \mathbf{1}\mathbf{1}^\top$, where $\rho \in (0,1)$ depends on the specific distribution for the entries (x_{ij}).

Lemma F. 2. *If X is a random matrix from Ens$_+$, then for all $t \geq 0$ and any $S \subseteq \{1, \dots, p\}$, $|S| \leq s$, with probability at least $1 - 2\exp(-c_1 t^2) - \exp\left(-c_2 \min\left(t^2, t\right) s\right)$*

$$s_{\max}\left(\frac{1}{n} X_S^\top X_S - \Sigma_{SS}^*\right) \leq \max(\delta, \delta^2) + C_1 \sqrt{\frac{s^2(1+t)}{n}}, \; \delta = C_2 \sqrt{\frac{s}{n}} + \frac{t}{\sqrt{n}}, \tag{F.2}$$

where $C, C_1, C_2, c, c_1, c_2 > 0$ are universal constants.

Proof. We decompose $X_S^i = \widetilde{X}_S^i + \mu \mathbf{1}$, where $\mu > 0$ is the mean of the entries, $i = 1, \dots, n$. We have

$$\begin{aligned}
s_{\max}\left(\frac{1}{n} X_S^\top X_S - \Sigma_{SS}^*\right) &= \sup_{v\colon \|v\|_2 = 1} \left| \frac{1}{n} \sum_{i=1}^n \left(\langle \widetilde{X}_S^i + \mu \mathbf{1}, v \rangle^2 - \mathbf{E}[\langle \widetilde{X}_S^i + \mu \mathbf{1}, v \rangle^2] \right) \right|, \\
&= \sup_{v\colon \|v\|_2 = 1} \left| \frac{1}{n} \sum_{i=1}^n \left(\langle \widetilde{X}_S^i, v \rangle^2 - \mathbf{E}[\langle \widetilde{X}_S^i, v \rangle^2] + 2 \langle \mu \mathbf{1}, v \rangle \langle \widetilde{X}_S^i, v \rangle \right) \right| \\
&\leq \sup_{v\colon \|v\|_2 = 1} \left| \frac{1}{n} \sum_{i=1}^n \left(\langle \widetilde{X}_S^i, v \rangle^2 - \mathbf{E}[\langle \widetilde{X}_S^i, v \rangle^2] \right) \right| + 2 \sup_{v\colon \|v\|_2 = 1} \left| \langle \mu \mathbf{1}, v \rangle \frac{1}{n} \sum_{i=1}^n \langle \widetilde{X}_S^i, v \rangle \right|
\end{aligned}$$

The first summand is handled by an application of Theorem A.1. For the second summand, we have

$$2 \sup_{v\colon \|v\|_2 = 1} \left| \langle \mu \mathbf{1}, v \rangle \frac{1}{n} \sum_{i=1}^n \langle \widetilde{X}_S^i, v \rangle \right| \leq 2 \left| \mu \sqrt{s} \left\| \frac{1}{n} \sum_{i=1}^n \widetilde{X}_S^i \right\|_2 \right|.$$

Re-writing the norm as

$$\left\| \frac{1}{n} \sum_{i=1}^n \widetilde{X}_S^i \right\|_2 = \left(\frac{1}{n} \sum_{j \in S} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \widetilde{x}_{ij} \right)^2 \right)^{1/2} = \left(\frac{1}{n} \sum_{j=1}^s Z_j^2 \right)^{1/2}, \; Z_j = \frac{1}{\sqrt{n}} \sum_{i=1}^n \widetilde{x}_{ij}.$$

and noting that, as explained in Appendix A, the sub-Gaussian norm of the $\{Z_j\}$ is uniformly bounded by an absolute constant, say L, we invoke (F.1), which yields for all $t \geq 0$

$$\mathbf{P}\left(\sum_{j=1}^s Z_j^2 > s + ts \right) \leq \exp\left(-c \min\left(\frac{t^2}{L^4}, \frac{t}{L^2} \right) s \right).$$

The claim follows by taking roots and back-substituting. $\square$

Lemma F. 3.

$$\max_{1 \leq j,k \leq p} \left| \left(\frac{1}{n} X^\top X - \Sigma^* \right)_{jk} \right| \leq C \sqrt{\frac{\log p}{n}},$$

with probability at least $1 - 3/p - \exp(-cn)$, where $C, c > 0$ are universal constants.

Proof. Write $\widetilde{X}_j = X_j - \mu \mathbf{1}$, $j = 1, \dots, p$, for the column vectors obtained by centering the columns of X. We have

$$\frac{1}{n} \left(\langle X_j, X_k \rangle - \mathbf{E}[\langle X_j, X_k \rangle] \right) = \frac{1}{n} \langle \widetilde{X}_j, \widetilde{X}_k \rangle - \mu \left(\frac{1}{n} \langle \widetilde{X}_j, \mathbf{1} \rangle + \frac{1}{n} \langle \widetilde{X}_k, \mathbf{1} \rangle \right). \tag{F.3}$$

For the second term in (F.3), we have, in view of the properties of sub-Gaussian random variables in Appendix A

$$\mathbf{P}\left(\left| \frac{\mu}{n} \langle \widetilde{X}_j + \widetilde{X}_k, \mathbf{1} \rangle \right| > \sqrt{2} \mu z \right) \leq 2 \exp(-c_0 n z^2). \tag{F.4}$$

For the first term in (F.3), let us first consider the case $j \neq k$. Fix any $j \in \{1, \ldots, p\}$. It follows from Lemma F.1 that the event $\mathcal{E}_j = \{\|X_j\|_2^2 \leq 2n\}$ holds with probability at least $1 - \exp(-c_1 n)$. Conditional on $\mathcal{E}_j$, $\langle \widetilde{X}_j, \widetilde{X}_k \rangle$ is a sub-Gaussian random variable with sub-Gaussian norm bounded by $L\sqrt{n}$, for some universal constant $L > 0$. It follows that

$$\begin{aligned}
\mathbf{P}\left(\left|\frac{1}{n}\langle \widetilde{X}_j, \widetilde{X}_k \rangle\right| > z\right) &\leq \mathbf{P}\left(\left|\frac{1}{n}\langle \widetilde{X}_j, \widetilde{X}_k \rangle\right| > z \middle| \mathcal{E}_j\right) + \mathbf{P}(\mathcal{E}_j^c) \\
&\leq 2\exp(-c_2 n z^2 / L^2) + \exp(-c_1 n) \leq 2\exp(-c_3 n z^2) + \exp(-c_1 n).
\end{aligned} \tag{F.5}$$

Let now $j = k$. With the aim to control the first term in (F.3), an application of Lemma A.1 yields $\forall z \geq 0$

$$\mathbf{P}\left(\left|\frac{1}{n}\sum_{j=1}^{n}\left(\widetilde{x}_{ij}^2 - \mathbf{E}[\widetilde{x}_{ij}^2]\right)\right| > z\right) \leq 2\exp(-c_4 \min(z, z^2) n). \tag{F.6}$$

Combining (F.4), (F.5) and (F.6), with a union bound over all p^2 entries of $\frac{1}{n}X^\top X$ and setting $z = 2/\sqrt{\min\{c_0, c_3, c_4\}}\sqrt{\frac{\log p}{n}}$, we obtain

$$\mathbf{P}\left(\left|\left(\frac{1}{n}X^\top X - \Sigma^*\right)_{jk}\right| > C\sqrt{\frac{\log p}{n}}\right) \leq \frac{3}{p} + \exp(-c_1 n + \log p).$$

□

Equipped with these auxiliary results, we turn to the actual proof of the Theorem. We analyze the random scaling of $\widehat{\tau}^2(S)$ using the dual formulation (C.1). In the following, denote by $\mathbb{S}^{s-1} = \{u \in \mathbb{R}^s : \|u\|_2 = 1\}$ the unit sphere in $\mathbb{R}^s$. Expanding the square in (C.1), we have

$$\begin{aligned}
\widehat{\tau}^2(S) &= \min_{\theta \in \mathbb{R}^s,\, \lambda \in T^{p-s-1}} \theta^\top \frac{1}{n} X_S^\top X_S \theta - 2\theta^\top \frac{1}{n} X_S^\top X_{S^c}\lambda + \lambda^\top \frac{1}{n} X_{S^c}^\top X_{S^c}\lambda \\
&\geq \min_{r>0,\, u \in \mathbb{S}^{s-1},\, \lambda \in T^{p-s-1}} r^2 u^\top \Sigma_{SS}^* u - r^2 s_{\max}\left(\frac{1}{n}X_S^\top X_S - \Sigma_{SS}^*\right) - \\
&\quad - 2ru^\top \frac{1}{n} X_S^\top X_{S^c}\lambda + \lambda^\top \frac{1}{n} X_{S^c}^\top X_{S^c}\lambda \\
&\geq \min_{r>0,\, u \in \mathbb{S}^{s-1},\, \lambda \in T^{p-s-1}} r^2 u^\top \Sigma_{SS}^* u - r^2 s_{\max}\left(\frac{1}{n}X_S^\top X_S - \Sigma_{SS}^*\right) \\
&\quad - 2\rho r u^\top \mathbf{1} - 2ru^\top(\frac{1}{n}X_S^\top X_{S^c} - \Sigma_{SS^c}^*)\lambda + \rho + \frac{1-\rho}{p-s} - \\
&\quad - \sup_{\lambda \in T^{p-s-1}} \left|\lambda^\top(\frac{1}{n}X_{S^c}^\top X_{S^c} - \Sigma_{S^cS^c}^*)\lambda\right|.
\end{aligned} \tag{F.7}$$

For the last inequality, we have used that $\min_{\lambda \in T^{p-s-1}} \lambda^\top \Sigma_{S^cS^c}^* \lambda = \rho + \frac{1-\rho}{p-s}$ by setting $\lambda = \mathbf{1}/(p-s)$. We further set $\Delta = s_{\max}\left(\frac{1}{n}X_S^\top X_S - \Sigma_{SS}^*\right)$ and $\delta = \sup_{u \in \mathbb{S}^{s-1}, \lambda \in T^{p-s-1}} \left|u^\top\left(\frac{1}{n}X_{S^c}^\top X_{S^c} - \Sigma_{S^cS^c}^*\right)\lambda\right|$. The random deviation terms Δ and δ will be controlled uniformly over $u \in \mathbb{S}^{s-1}$ and $\lambda \in T^{p-s-1}$ by means of the two preceding lemmas, and are hence subsequently treated as constants. This approach allows us to minimize the lower bound in (F.7) w.r.t. u and r separately from λ. The minimization problem involving u and r reads

$$\min_{r>0,\, u \in \mathbb{S}^{s-1}} r^2 u^\top \Sigma_{SS}^* u - 2\rho r u^\top \mathbf{1} - r^2\Delta - 2r\delta. \tag{F.8}$$

We first derive an expression for

$$\phi(r) = \min_{u \in \mathbb{S}^{s-1}} r^2 u^\top \Sigma_{SS}^* u - 2\rho r u^\top \mathbf{1}. \tag{F.9}$$

We decompose $u = u^\| + u^\perp$, where $u^\| = \left\langle \frac{\mathbf{1}}{\sqrt{s}}, u \right\rangle \frac{\mathbf{1}}{\sqrt{s}}$ is the projection of u on the unit vector $\mathbf{1}/\sqrt{s}$, which is the eigenvector of Σ_{SS}^* associated with its largest eigenvalue $1 + \rho(s-1)$. By

Parseval's identity, we have $\|u^{\|}\|_2^2 = \gamma$, $\|u^{\perp}\|_2^2 = (1-\gamma)$ for some $\gamma \in [0,1]$. Inserting this decomposition and noting that the remaining eigenvalues of Σ_{SS}^* are all equal to $(1-\rho)$, we obtain the following expression to be minimized w.r.t. $\gamma \in [0,1]$

$$r^2\gamma \underbrace{(1+(s-1)\rho)}_{s_{\max}(\Sigma_{SS}^*)} + r^2(1-\gamma) \underbrace{(1-\rho)}_{s_{\min}(\Sigma_{SS}^*)} - 2\rho r\sqrt{\gamma}\sqrt{s}, \tag{F.10}$$

where we have used that $\langle u^{\perp}, \mathbf{1}\rangle = 0$ and that all potential minimizers must satisfy $\langle u^{\|}, \mathbf{1}\rangle > 0$. Let us put aside the constraint $\gamma \in [0,1]$ for a moment. The expression (F.10) is a convex function of γ, hence we may find an (unconstrained) minimizer $\widetilde{\gamma}$ by differentiating and setting the derivative equal to zero. This yields $\widetilde{\gamma} = \frac{1}{r^2 s}$, which coincides with the constrained minimizer if and only if $r \geq \frac{1}{\sqrt{s}}$. Now observe that the minimizer of the problem $\min_{r>0, u \in \mathbb{S}^{s-1}} r^2 u^\top \Sigma_{SS}^* u - 2\rho r u^\top \mathbf{1}$ with r being unfixed equals the minimizer $\widehat{\theta}$ of the problem $\min_{\theta \in \mathbb{R}^s} \theta^\top \Sigma_{SS}^* \theta - 2\rho\theta^\top \mathbf{1}$, which is given by $\widehat{\theta} = \frac{\rho \mathbf{1}}{1+(s-1)\rho} = \frac{\mathbf{1}}{\sqrt{s}} \cdot \frac{\sqrt{s}\rho}{1+(s-1)\rho}$, a unit vector satisfying $\gamma = 1$ times a radius less than $1/\sqrt{s}$. We conclude that for all $r < 1/\sqrt{s}$, the minimum is attained for $\gamma = 1$, hence the function $\phi(r)$ (F.9) is given by

$$\phi(r) = \begin{cases} r^2 s_{\max}(\Sigma_{SS}^*) - 2r\rho\sqrt{s} & r < 1/\sqrt{s}, \\ r^2(1-\rho) - \rho & \text{otherwise,} \end{cases} \tag{F.11}$$

where the second line is obtained by inserting $\widetilde{\gamma} = \frac{1}{r^2 s}$ for γ in (F.10). The minimization problem (F.8) to be considered eventually reads

$$\min_{r>0} \psi(r), \quad \psi(r) = \phi(r) - r^2\Delta - 2r\delta. \tag{F.12}$$

We argue that it suffices to consider the case $r < 1/\sqrt{s}$ in (F.11) provided

$$((1-\rho) - \Delta)^2 > \delta^2 s, \tag{F.13}$$

a condition we will comment on below. If this condition is met, differentiating shows that ψ is increasing on $[\frac{1}{\sqrt{s}}, \infty)$. In fact, for all r in that ray,

$$\frac{d}{dr}\psi(r) = 2r(1-\rho) - 2r\Delta - 2\delta, \text{ and thus}$$

$$\frac{d}{dr}\psi(r) > 0 \text{ for all } r \in \left[\frac{1}{\sqrt{s}}, \infty\right) \Leftrightarrow \frac{1}{\sqrt{s}}((1-\rho) - \Delta) > \delta \Leftrightarrow ((1-\rho)-\Delta)^2 > s\delta^2.$$

Considering the case $r < 1/\sqrt{s}$, we observe that $\psi(r)$ is convex provided

$$s_{\max}(\Sigma_{SS}^*) > \Delta, \tag{F.14}$$

a condition we shall comment on below as well. Provided (F.13) and (F.14) hold true, the minimizer $\widehat{r}$ of (F.12) is given by $(\rho\sqrt{s} + \delta)/(s_{\max}(\Sigma_{SS}^*) - \Delta)$. Substituting this result back into (F.12) and in turn into the lower bound (F.7), one obtains after collecting terms

$$\widehat{\tau}^2(S) \geq \rho \frac{(1-\rho) - \Delta}{(1-\rho) + s\rho - \Delta} - \frac{2\rho\sqrt{s}\delta + \delta^2}{s_{\max}(\Sigma_{SS}^*) - \Delta} + \frac{1-\rho}{p-s} - \sup_{\lambda \in T^{p-s-1}} \left| \lambda^\top \left(\frac{1}{n} X_{S^c}^\top X_{S^c} - \Sigma_{S^c S^c}^*\right)\lambda \right|. \tag{F.15}$$

Consider the two events

$$\mathcal{A} = \left\{ \Delta \leq C_1 \left(\sqrt{\frac{s^2 \log^{1/2} p}{n}} + \sqrt{\frac{\log p}{n}} \right) \right\}, \mathcal{B} = \left\{ \max_{j,k} \left| \left(\frac{1}{n} X^\top X - \Sigma^* \right)_{jk} \right| \leq C_2 \sqrt{\frac{\log p}{n}} \right\},$$

for universal constants $C_1, C_2 > 0$. Conditional on $\mathcal{A} \cap \mathcal{B}$, bounding

$$\delta \leq \sup_{u \in \mathbb{S}^{s-1}} \|u\|_1 \sup_{\lambda \in T^{p-s-1}} \left\| \left(\frac{1}{n} X_{S^c}^\top X_{S^c} - \Sigma_{S^c S^c}^* \right) \lambda \right\|_\infty \leq \sqrt{s} C_2 \sqrt{\frac{\log p}{n}},$$

and inserting the scaling for Δ under $\mathcal{A}$, there exists a sufficiently large constant $\widehat{C} > 0$ such that the two conditions (F.13) and (F.14) supposed to be fulfilled previously indeed hold given that

$n \geq \widehat{C}\log(p)s^2$. We may re-write (F.15) as

$$\widehat{\tau}^2(S) \geq \frac{\rho(1-\Delta/(1-\rho))}{(1-\Delta/(1-\rho)) + s\frac{\rho}{1-\rho}} + \frac{2\rho\frac{\sqrt{s}}{1+(s-1)\rho}\delta}{1-\Delta/(1+(s-1)\rho)} - \frac{\delta^2/(1+(s-1)\rho)}{1-\Delta/(1+(s-1)\rho)} - \\ - \sup_{\lambda \in T^{p-s-1}} \left| \lambda^\top (\frac{1}{n} X_{S^c}^\top X_{S^c} - \Sigma^*_{S^c S^c})\lambda \right|. \tag{F.16}$$

Conditional on $\mathcal{A} \cap \mathcal{B}$, there exists again a sufficiently large constant $\widetilde{C} > 0$ such that if $n \geq \widetilde{C}\log(p)s^2$

$$c_1\frac{1}{s} - C_3\sqrt{\frac{\log p}{n}} - C_4\frac{\log p}{n} - C_2\sqrt{\frac{\log p}{n}} = c_1\frac{1}{s} - C_5\sqrt{\frac{\log p}{n}} \tag{F.17}$$

by inserting the resulting scalings separately for each summand in (F.16), where $c_1, C_3, C_4, C_5 > 0$ are universal constants. We conclude that if $n \geq \max(\widehat{C}, \widetilde{C})\log(p)s^2$, (F.17) holds with probability no less than $1 - \mathbf{P}(\mathcal{A}) - \mathbf{P}(\mathcal{B})$. Using Lemmas F.2 and F.3 to control $\mathbf{P}(\mathcal{A})$ and $\mathbf{P}(\mathcal{B})$, the result follows.

References

[1] R. Vershynin. *In: Compressed Sensing: Theory and Applications, Y. Eldar and G. Kutyniok (eds.)*, chapter 'Introduction to the non-asymptotic analysis of random matrices'. Cambridge University Press, 2012.

[2] H. Rue and L. Held. *Gaussian Markov Random Fields*. Chapman and Hall/CRC, Boca Raton, 2001.