Supplementary Materials to "Convex Banding of the Covariance Matrix"

Jacob Bien, Florentina Bunea, Luo Xiao

May 25, 2015

A.1 Dual problem

Define

$$L(\mathbf{\Sigma}, \mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(p-1)}) = \frac{1}{2}\|\mathbf{S} - \mathbf{\Sigma}\|_F^2 + \lambda \langle \sum_{\ell=1}^{p-1} \mathbf{W}^{(\ell)} * \mathbf{A}^{(\ell)}, \mathbf{\Sigma} \rangle.$$

Observe that

$$\|(\mathbf{W}^{(\ell)} * \mathbf{\Sigma})_{g_\ell}\|_2 = \max_{\mathbf{A}^{(\ell)}} \left\{ \langle \mathbf{W}^{(\ell)} * \mathbf{A}^{(\ell)}, \mathbf{\Sigma} \rangle \text{ s.t. } \|\mathbf{A}_{g_\ell}^{(\ell)}\|_2 \leq 1, \ \mathbf{A}_{g_\ell^c}^{(\ell)} = 0 \right\}.$$

It follows that (2.1) is equivalent to

$$\min_{\mathbf{\Sigma}} \left\{ \max_{\mathbf{A}^{(\ell)}} \left[L(\mathbf{\Sigma}, \mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(p-1)}) \text{ s.t. } \|\mathbf{A}_{g_\ell}^{(\ell)}\|_2 \leq 1, \mathbf{A}_{g_\ell{}^c}^{(\ell)} = 0 \right] \right\}.$$

We get the dual problem by interchanging the min and max. The inner minimization gives the primal-dual relation given in the theorem (by strong duality) and the following dual function:

$$\min_{\mathbf{\Sigma}} L(\mathbf{\Sigma}, \mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(p-1)}) = -\frac{1}{2}\|\mathbf{S} - \lambda \sum_{\ell=1}^{p-1} \mathbf{W}^{(\ell)} * \mathbf{A}^{(\ell)}\|_F^2 + \frac{1}{2}\|\mathbf{S}\|_F^2.$$

A.2 Ellipsoid projection

To update $\hat{\mathbf{A}}^{(\ell)}$ in Algorithm A.1, we must solve a problem of the form

$$\hat{\mathbf{A}}_{g_\ell}^{(\ell)} = \arg \min_{\mathbf{a} \in \mathbb{R}^{|g_\ell|}} \|\hat{\mathbf{R}}_{g_\ell}^{(\ell)} - \lambda \mathbf{D}^{(\ell)} \mathbf{a}\|_2^2 \text{ s.t. } \|\mathbf{a}\|_2^2 \leq 1,$$

which (in a change of coordinates) is the projection of a point onto an ellipsoid. Clearly, if $\|{\mathbf{D}^{(\ell)}}^{-1} \hat{\mathbf{R}}_{g_\ell}^{(\ell)}\| \leq \lambda$, then $\hat{\mathbf{A}}_{g_\ell}^{(\ell)} = {\mathbf{D}^{(\ell)}}^{-1} \hat{\mathbf{R}}_{g_\ell}^{(\ell)} / \lambda$. Otherwise, we use the method of Lagrange multipliers (to solve the problem with an equality constraint):

$$\mathcal{L}(\mathbf{a}, \nu) = \|\hat{\mathbf{R}}_{g_\ell}^{(\ell)} - \lambda \mathbf{D}^{(\ell)} \mathbf{a}\|_2^2 + \nu \lambda^2 (\|\mathbf{a}\|_2^2 - 1)$$

whence

$$0 = 2\lambda \mathbf{D}^{(\ell)}(\lambda \mathbf{D}^{(\ell)}\hat{\mathbf{a}} - \hat{\mathbf{R}}^{(\ell)}_{g_\ell}) + 2\nu\lambda^2 \mathbf{a} \implies \hat{\mathbf{a}} = (\mathbf{D}^{(\ell)2} + \nu \mathbf{I}_{|g_\ell|})^{-1}\lambda^{-1}\mathbf{D}^{(\ell)}\hat{\mathbf{R}}^{(\ell)}_{g_\ell}.$$

That is,

$$\hat{\mathbf{A}}^{(\ell)}_{s_m} = \frac{w_{\ell m}}{\lambda(w_{\ell m}^2 + \hat{\nu}_\ell)}\hat{\mathbf{R}}^{(\ell)}_{s_m}$$

where $\hat{\nu}_\ell$ is such that $\hat{\mathbf{A}}^{(\ell)}_{g_\ell}$ has unit norm, i.e. such that $h_\ell(\hat{\nu}_\ell) = \lambda^2$ where h_ℓ is defined in (3.2). We compute this root numerically. We can get limits within which $\hat{\nu}_\ell$ must lie. For example, replacing $w_{\ell m}$ by $w_\ell = \max_m w_{\ell m}$ makes the RHS smaller whereas replacing it by 0 makes the RHS larger:

$$\frac{\sum_{m=1}^{\ell} w_{\ell m}^2 \|\hat{\mathbf{R}}^{(\ell)}_{s_m}\|^2}{(w_\ell^2 + \nu)^2} =: h_\ell^L(\nu) \le h_\ell(\nu) \le h_\ell^U(\nu) := \frac{\sum_{m=1}^{\ell} w_{\ell m}^2 \|\hat{\mathbf{R}}^{(\ell)}_{s_m}\|^2}{(0 + \nu)^2}.$$

Now, since $h_\ell(\nu)$ is a decreasing function, we know that $\hat{\nu}_\ell^L \le \hat{\nu}_\ell \le \hat{\nu}_\ell^U$, where $h_\ell^L(\hat{\nu}_\ell^L) = \lambda^2 = h_\ell^U(\hat{\nu}_\ell^U)$. From this, it follows that

$$\left[\sqrt{\sum_{m=1}^{\ell} w_{\ell m}^2 \|\hat{\mathbf{R}}^{(\ell)}_{s_m}\|^2} - \lambda w_\ell^2\right]_+ \le \lambda\hat{\nu}_\ell \le \sqrt{\sum_{m=1}^{\ell} w_{\ell m}^2 \|\hat{\mathbf{R}}^{(\ell)}_{s_m}\|^2}.$$

Noting that $\sum_{m=1}^{\ell} w_{\ell m}^2 \|\hat{\mathbf{R}}^{(\ell)}_{s_m}\|^2 = \|\mathbf{D}^{(\ell)}\hat{\mathbf{R}}^{(\ell)}_{g_\ell}\|_2^2$, this simplifies to

$$\|\mathbf{D}^{(\ell)}\hat{\mathbf{R}}^{(\ell)}_{g_\ell}\|_2 - \lambda w_\ell^2 \le \lambda\hat{\nu}_\ell \le \|\mathbf{D}^{(\ell)}\hat{\mathbf{R}}^{(\ell)}_{g_\ell}\|_2.$$

To summarize, we have for $1 \le m \le \ell \le p-1$,

$$\hat{\mathbf{A}}^{(\ell)}_{s_m} = \begin{cases} \hat{\mathbf{R}}^{(\ell)}_{s_m}/(\lambda w_{\ell m}) & \text{if } \|\mathbf{D}^{(\ell)^{-1}}\hat{\mathbf{R}}^{(\ell)}_{g_\ell}\|_2 \le \lambda \\ \frac{w_{\ell m}}{\lambda(w_{\ell m}^2 + \hat{\nu}_\ell)}\hat{\mathbf{R}}^{(\ell)}_{s_m} & \text{otherwise} \end{cases}$$

and $\hat{\mathbf{A}}^{(\ell)}_{g_\ell^c} = 0$. This can be written more simply as $\frac{w_{\ell m}}{\lambda(w_{\ell m}^2 + [\hat{\nu}_\ell]_+)}\hat{\mathbf{R}}^{(\ell)}_{s_m}$ if we note that $\|\mathbf{D}^{(\ell)^{-1}}\hat{\mathbf{R}}^{(\ell)}_{g_\ell}\|_2 \le \lambda$ is equivalent to $h_\ell(0) \le \lambda^2$ and that in this case $\hat{\nu}_\ell \le 0$, since h_ℓ is nonincreasing. It turns out that often we will be able to get $\hat{\nu}_\ell$ in closed form. First, $h_1(\nu) = w_1^2\|\hat{\mathbf{R}}^{(1)}_{s_1}\|^2/(w_1^2 + \nu)^2$, so $\hat{\nu}_1 = w_1(\|\hat{\mathbf{R}}^{(1)}_{s_1}\|/\lambda - w_1)$. Furthermore, for $\ell \ge 1$, if $\hat{\nu}_\ell \le 0$ then for all $m \le \ell$, we have $\hat{\mathbf{R}}^{(\ell+1)}_{s_m} = 0$. This means that $h_{\ell+1}(\nu) = \frac{w_\ell^2}{(w_\ell^2+\nu)^2}\|\hat{\mathbf{R}}^{(\ell+1)}_{s_{\ell+1}}\|^2$ whence

$$\hat{\nu}_{\ell+1} = w_{\ell+1}\left(\|\hat{\mathbf{R}}^{(\ell+1)}_{s_{\ell+1}}\|/\lambda - w_{\ell+1}\right)$$

Thus, we only need to perform numerical root-finding when $\ell = p - 1 - \hat{K}$.

A.3 Proof of tapering theorem (Theorem 2)

Proof. By Proposition 5 in Jenatton et al. (2011), we can get $\hat{\mathbf{\Sigma}}$ by a single pass as in Algorithm 1. We begin with $\hat{\mathbf{R}}^{(1)} = \mathbf{S}$ and then for $\ell = 1, \ldots, p-1$, (and for each $m \leq \ell$), we have

$$\hat{\mathbf{R}}^{(\ell+1)}_{s_m} = \hat{\mathbf{R}}^{(\ell)}_{s_m} - \lambda w_{\ell m} \hat{\mathbf{A}}^{(\ell)}_{s_m} = \frac{[\hat{\nu}_\ell]_+}{w^2_{\ell m} + [\hat{\nu}_\ell]_+} \hat{\mathbf{R}}^{(\ell)}_{s_m}. \tag{A.3.1}$$

The optimality conditions give $\hat{\mathbf{\Sigma}} = \hat{\mathbf{R}}^{(p)}$, so that we have

$$\hat{\mathbf{\Sigma}}_{s_m} = \prod_{\ell=m}^{p-1} \frac{[\hat{\nu}_\ell]_+}{w^2_{\ell m} + [\hat{\nu}_\ell]_+} \cdot \mathbf{S}_{s_m} \tag{A.3.2}$$

which establishes this as an adaptively tapered estimator. □

A.4 Bounds on $\max_{ij} |\mathbf{S}_{ij} - \mathbf{\Sigma^*}_{ij}|$

Theorem A.1. *Assume* $\log p \leq \gamma n$ *for some constant* $\gamma > 0$ *and* $\max_j |\mathbf{\Sigma^*}_{jj}| \leq M$ *for some constant* M. *Let* $D = \max_{i,j} |\mathbf{S}_{ij} - \mathbf{\Sigma^*}_{ij}|$. *There exists some constant* $c > 0$ *such that for sufficiently large* $x > 0$,

$$\mathbb{P}\left(D > x\sqrt{\log(p \vee n)/n}\right) \leq \frac{c}{p \vee n} \tag{A.4.1}$$

and

$$\mathbb{E}D^2 \cdot 1_{\left\{D > x\sqrt{\log(p\vee n)/n}\right\}} \leq \frac{c}{n(p \vee n)} \tag{A.4.2}$$

Proof. By Lemma A.1 below, there exist constants $c_i, i = 1, 2, 3, 4$ such that, for any $0 < t < 2M$,

$$\mathbb{P}(D > t) \leq p^2 c_1 \exp\{-c_2 n t^2\} + p c_3 \exp\{-c_4 n t\}.$$

Hence,

$$\begin{aligned}\mathbb{P}\left(D > x\sqrt{\log(p \vee n)/n}\right) &\leq p^2 c_1 \exp\{-c_2 x^2 \log(p \vee n)\} + p c_3 \exp\{-c_4 x \log(p \vee n)\sqrt{n/\log(p \vee n)}\} \\ &\leq c_1 (p \vee n)^{2 - c_2 x^2} + c_3 (p \vee n)^{1 - c_4 x\sqrt{n/\log(p\vee n)}}.\end{aligned}$$

Next we derive that

$$\begin{aligned}&\mathbb{E}D^2 \cdot 1_{\left\{D > x\sqrt{\log(p\vee n)/n}\right\}} = \int_{x^2 \log(p\vee n)/n}^{\infty} \mathbb{P}(D^2 \geq y)\mathrm{d}y \\ &\leq \int_{x^2 \log(p\vee n)/n}^{\infty} \left(p^2 c_1 \exp\{-c_2 n t\} + p c_3 \exp\{-c_4 n\sqrt{t}\}\right) \mathrm{d}t \\ &= \frac{p^2 c_1 \exp\{-c_2 n x^2 \log(p \vee n)/n\}}{c_2 n} + \frac{p c_3 \exp\{-c_4 n x\sqrt{\log(p \vee n)/n}\}}{c_4 n}\left(x\sqrt{\log(p \vee n)/n} + \frac{1}{c_4 n}\right) \\ &\leq \frac{c_1 (p \vee n)^{2 - c_2 x^2}}{c_2 n} + \frac{c_3 (p \vee n)^{1 - c_4 x\sqrt{n/\log(p\vee n)}}}{c_4 n}\left(x\sqrt{\log(p \vee n)/n} + \frac{1}{c_4 n}\right).\end{aligned}$$

Hence if x is sufficiently large and by the assumption that $\log p \le \gamma n$, the inequalities in the lemma holds for some constant c. □

Lemma A.1. *There exist two constants c_1 and c_2 such that*

$$\mathbb{P}\left(\max_{ij}|\mathbf{S}_{ij}-\mathbf{\Sigma^*}_{ij}|>t\right)\le 2p^2\exp\left(-\frac{c_2nt^2}{\max_j \mathbf{\Sigma^*}_{jj}}\right)+8p\exp\left(-\frac{c_1nt}{\max_j \mathbf{\Sigma^*}_{jj}}\right).$$

for any $0<t<2\max_j \mathbf{\Sigma^*}_{jj}$.

Proof: Note that

$$\mathbf{S}_{ij}=n^{-1}\sum_{k=1}^n X_{ki}X_{kj}-\bar{X}_i\bar{X}_j.$$

and

$$|\mathbf{S}_{ij}-\mathbf{\Sigma^*}_{ij}|\le\left|n^{-1}\sum_{k=1}^n X_{ki}X_{kj}-\mathbf{\Sigma^*}_{ij}\right|+\left|\bar{X}_i\bar{X}_j\right|.$$

Hence

$$\begin{aligned}
&\mathbb{P}\left(\max_{ij}|\mathbf{S}_{ij}-\mathbf{\Sigma^*}_{ij}|>t\right)\\
\le&\mathbb{P}\left(\max_{ij}\left|n^{-1}\sum_{k=1}^n X_{ki}X_{kj}-\mathbf{\Sigma^*}_{ij}\right|>t/2\right)+\mathbb{P}\left(\max_{ij}\left|\bar{X}_i\bar{X}_j\right|>t/2\right)\\
\le&p^2\max_{ij}\mathbb{P}\left(\left|n^{-1}\sum_{k=1}^n X_{ki}X_{kj}-\mathbf{\Sigma^*}_{ij}\right|>t/2\right)+2p\max_j\mathbb{P}\left(\left|\bar{X}_j\right|>\sqrt{t/2}\right).
\end{aligned}$$

Let $I_{ij}=\mathbb{P}\left(|n^{-1}\sum_{k=1}^n X_{ki}X_{kj}-\mathbf{\Sigma^*}_{ij}|>t/2\right)$ and $I_j=\mathbb{P}\left(\left|\bar{X}_j\right|>\sqrt{t/2}\right)$. Then

$$\mathbb{P}\left(\max_{ij}|\mathbf{S}_{ij}-\mathbf{\Sigma^*}_{ij}|>t\right)\le p^2\max_{ij}I_{ij}+2p\max_j I_j. \tag{A.4.3}$$

We first consider I_j. $\bar{X}_j$ is sub-Gaussian with variance $\mathbf{\Sigma^*}_{jj}/n$ and

$$\begin{aligned}
\mathbb{E}\exp\left(t\bar{X}_j/\sqrt{\mathbf{\Sigma^*}_{jj}/n}\right)=&\prod_{k=1}^n\mathbb{E}\exp\left(tX_{kj}/\sqrt{n\mathbf{\Sigma^*}_{jj}}\right)\\
\le&\left\{\exp(Ct^2/n)\right\}^n\\
=&\exp(Ct^2).
\end{aligned}$$

By Lemma 5.5 in Vershynin (2011),

$$\mathbb{P}\left\{|\bar{X}_j|/\sqrt{\mathbf{\Sigma^*}_{jj}/n}>t\right\}\le\exp(1-t^2/K_1^2)$$

for some constant K_1 that does not depend on j. It follows that

$$I_j = \mathbb{P}\left(|\bar{X}_j| > \sqrt{t/2}\right) = \mathbb{P}\left\{|\bar{X}_j|/\sqrt{\mathbf{\Sigma}^*{}_{jj}/n} > \sqrt{tn/(2\mathbf{\Sigma}^*{}_{jj})}\right\} \leq \exp\left(1 - \frac{nt}{2K_1^2\mathbf{\Sigma}^*{}_{jj}}\right).$$

Therefore,

$$I_j \leq 4\exp\left(-\frac{c_1 nt}{\max_j \mathbf{\Sigma}^*{}_{jj}}\right) \tag{A.4.4}$$

for some constant c_1.

Now we consider I_{ij}. We shall find ν_{ij} and c_{ij} such that

$$\sum_{k=1}^{n} \mathbb{E}(X_{ki}^2 X_{kj}^2) \leq \nu_{ij} \tag{A.4.5}$$

and

$$\sum_{k=1}^{n} \mathbb{E}\left\{(X_{ki}X_{kj})_+^q\right\} \leq \frac{q!}{2} \cdot \nu_{ij} \cdot c_{ij}^{q-2} \tag{A.4.6}$$

for all integers $q \geq 3$. Then by Theorem 2.10 and Corollary 2.11 in Boucheron et al. (2013), for any $t > 0$,

$$\mathbb{P}\left(\left|\sum_{k=1}^{n}(X_{ki}X_{kj} - \mathbf{\Sigma}^*{}_{ij})\right| > t\right) \leq 2\exp\left\{-\frac{t^2}{2(\nu_{ij} + c_{ij}t)}\right\}. \tag{A.4.7}$$

To find ν_{ij} and c_{ij}, note that by Lemma 5.5 in Vershynin (2011), $\mathbb{E}|X_{ij}/\sqrt{\sigma_{jj}}|^q \leq K_2^q q^{q/2}$ for all $q \geq 1$ and some constant K_2 that does not depend on j. Hence

$$\sum_{k=1}^{n} \mathbb{E}(X_{ki}^2 X_{kj}^2) \leq \sum_{k=1}^{n} \sqrt{\mathbb{E}X_{ki}^4 \cdot \mathbb{E}X_{kj}^4} \leq 16n\mathbf{\Sigma}^*{}_{ii}\mathbf{\Sigma}^*{}_{jj}K_2^4.$$

Similarly,

$$\sum_{k=1}^{n} \mathbb{E}\left\{(X_{ki}X_{kj})_+^q\right\} \leq \sum_{k=1}^{n} \sqrt{\mathbb{E}X_{ki}^{2q} \cdot \mathbb{E}X_{kj}^{2q}} \leq n\mathbf{\Sigma}_{ii}^{*q/2}\mathbf{\Sigma}_{jj}^{*q/2}K_2^{2q}(2q)^q.$$

It is easy to show that (A.4.5) and (A.4.6) hold with $\nu_{ij} = K_3 n\mathbf{\Sigma}^*{}_{ii}\mathbf{\Sigma}^*{}_{jj}$ and $c_{ij} = K_3\sqrt{\mathbf{\Sigma}^*{}_{ii}\mathbf{\Sigma}^*{}_{jj}}$ for some constant K_3 that is sufficiently large and does not depend on i or j.

By (A.4.7), it follows that

$$\begin{aligned}
I_{ij} &\leq 2\exp\left\{-\frac{n^2t^2}{4(2\nu_{ij} + c_{ij}nt)}\right\} \\
&= 2\exp\left\{-\frac{n^2t^2}{4(2K_3 n\mathbf{\Sigma}^*{}_{ii}\mathbf{\Sigma}^*{}_{jj} + K_3\sqrt{\mathbf{\Sigma}^*{}_{ii}\mathbf{\Sigma}^*{}_{jj}}nt)}\right\} \\
&\leq 2\exp\left\{-\frac{nt^2}{4(2K_3\mathbf{\Sigma}^*{}_{ii}\mathbf{\Sigma}^*{}_{jj} + K_3\sqrt{\mathbf{\Sigma}^*{}_{ii}\mathbf{\Sigma}^*{}_{jj}}t)}\right\}.
\end{aligned}$$

If $t < 2\max_j {\boldsymbol{\Sigma}^*}_{jj}$, we have

$$I_{ij} \leq 2\exp\left(-\frac{c_2 n t^2}{\max_j {\boldsymbol{\Sigma}^*}_{jj}}\right), \tag{A.4.8}$$

where $c_2 = (16K_3)^{-1}$.

A.5 Proof of bandwidth recovery

Proof of Theorem 3. Referring to the proof of Theorem 2, we have $\hat{\mathbf{R}}_{g_\ell}^{(\ell+1)} = 0$ if $\hat{\nu}_\ell \leq 0$ or equivalently if $h_\ell(0) \leq \lambda^2$, where h_ℓ is defined in (3.2). If $L = 0$, then $K = p-1$ and $\hat{K} \leq K$ holds automatically. Thus, assume that $L > 0$. We prove that $\hat{\mathbf{R}}_{g_L}^{(L+1)} = 0$ by induction on ℓ. For $\ell = 1$,

$$h_\ell(0) = 2\mathbf{S}_{1p}^2/w_1^2 \leq \max_{ij} |\mathbf{S}_{ij} - {\boldsymbol{\Sigma}^*}_{ij}|^2 \leq \lambda^2$$

on the set $\mathcal{A}_x$ defined in (4.1). Assume $\hat{\mathbf{R}}_{g_\ell}^{(\ell+1)} = 0$ for $\ell < L$. Then, since $\hat{\mathbf{R}}_{s_{\ell+1}}^{(\ell+1)} = \mathbf{S}_{s_{\ell+1}}$,

$$h_{\ell+1}(0) = \|\mathbf{S}_{s_{\ell+1}}\|^2/w_{\ell+1,\ell+1}^2 \leq \lambda^2$$

on $\mathcal{A}_x$. Therefore, $\hat{\mathbf{R}}_{g_L}^{(L+1)} = 0$ and so $\hat{\boldsymbol{\Sigma}}_{g_L} = 0$, i.e., $\hat{K} \leq K$. □

Proof of Theorems 4 and 5. In both theorems, we wish to show that $\hat{\boldsymbol{\Sigma}}_{s_{L+1}} \neq 0$ or equivalently that $h_\ell(0) > \lambda^2$ for each $\ell \geq L+1$, whence we get the condition

$$\min_{\ell \geq L+1} h_\ell(0) > \lambda^2. \tag{A.5.1}$$

Recalling that $h_\ell(0) = \sum_{m=1}^{\ell} \|\hat{\mathbf{R}}_{s_m}^{(\ell)}\|^2/w_{\ell m}^2$, we have

$$h_\ell(0) \geq \|\hat{\mathbf{R}}_{s_\ell}^{(\ell)}\|^2/w_\ell^2 = \|\mathbf{S}_{s_\ell}\|^2/w_\ell^2.$$

Now, being on the set $\mathcal{A}_x$ implies that for any ℓ,

$$\|\mathbf{S}_{s_\ell}\|_2 \geq \|{\boldsymbol{\Sigma}^*}_{s_\ell}\|_2 - \|\mathbf{S}_{s_\ell} - {\boldsymbol{\Sigma}^*}_{s_\ell}\|_2 \geq \|{\boldsymbol{\Sigma}^*}_{s_\ell}\|_2 - \sqrt{2\ell}\lambda = \|{\boldsymbol{\Sigma}^*}_{s_\ell}\|_2 - \lambda w_\ell. \tag{A.5.2}$$

We consider the two theorems separately:

1. (Theorem 4) By assumption, for $\ell \geq L+1$, (A.5.2) gives us $h_\ell(0) > \lambda^2$. Thus, (A.5.1) is satisfied, proving the first theorem.

2. (Theorem 5) By the same argument as above, we have $h_\ell(0) > \lambda^2$ for $\ell = L+1$ and for $\ell \geq L+3$. It remains to show that $h_{L+2}(0) > \lambda^2$. Since $h_{L+1}(0) > \lambda^2$, we have that $\hat{\nu}_{L+1} > 0$ and since $\hat{\nu}_L \leq 0$ (see appendix on ellipsoidal projection), $\hat{\nu}_{L+1} = w_{L+1}(\|\mathbf{S}_{s_{L+1}}\|/\lambda - w_{L+1}) > 0$. Thus,

$$\hat{\mathbf{R}}_{s_{L+1}}^{(L+2)} = \frac{\hat{\nu}_{L+1}\mathbf{S}_{s_{L+1}}}{w_{L+1}^2 + \hat{\nu}_{L+1}} = \left(\frac{\|\mathbf{S}_{s_{L+1}}\|_2 - \lambda w_{L+1}}{\|\mathbf{S}_{s_{L+1}}\|_2}\right)\mathbf{S}_{s_{L+1}}$$

and

$$\begin{aligned} h_{L+2}(0) &= \|\hat{\mathbf{R}}^{(L+2)}_{s_{L+2}}\|_2^2/w_{L+2}^2 + \|\hat{\mathbf{R}}^{(L+2)}_{s_{L+1}}\|_2^2/w_{L+2,L+1}^2 \\ &= \|\mathbf{S}_{s_{L+2}}\|_2^2/w_{L+2}^2 + (\|\mathbf{S}_{s_{L+1}}\|_2 - \lambda w_{L+1})^2/w_{L+2,L+1}^2 \\ &\geq (\|{\boldsymbol{\Sigma}^*}_{s_{L+2}}\|_2 - \lambda w_{L+2})_+^2/w_{L+2}^2 + (\|{\boldsymbol{\Sigma}^*}_{s_{L+1}}\|_2 - 2\lambda w_{L+1})^2/w_{L+2,L+1}^2 \\ &= (\|{\boldsymbol{\Sigma}^*}_{s_{L+2}}\|_2 - \lambda w_{L+2})_+^2/w_{L+2}^2 + \lambda^2\gamma^2 w_{L+1}^2/w_{L+2,L+1}^2 \\ &\geq (\|{\boldsymbol{\Sigma}^*}_{s_{L+2}}\|_2 - \lambda w_{L+2})_+^2/w_{L+2}^2 + \lambda^2\gamma^2 \end{aligned}$$

again applying (A.5.2), and using that $w_{L+1} = w_{L+1,L+1} \geq w_{L+2,L+1}$. Now, for this to exceed λ we have the following: If $\gamma \geq 1$, there is no requirement on ${\boldsymbol{\Sigma}^*}_{s_{L+2}}$; if $0 < \gamma < 1$, then

$$\|{\boldsymbol{\Sigma}^*}_{s_{L+2}}\|_2 > \lambda w_{L+2}\left(1 + \sqrt{1-\gamma^2}\right)$$

This establishes that $h_\ell(0) > \lambda^2$ for $\ell \geq L+1$, completing the proof of the second theorem.

□

A.6 Proof of convergence in Frobenius norm

Define

$$\|\boldsymbol{\Sigma}\|_{2,1} = \sum_{\ell=1}^{p-1} w_\ell \|\boldsymbol{\Sigma}_{s_\ell}\|_2 \quad \text{and} \quad \|\boldsymbol{\Sigma}\|_{2,\infty} = \max_{1\leq\ell\leq p-1} w_\ell^{-1}\|\boldsymbol{\Sigma}_{s_\ell}\|_2. \tag{A.6.1}$$

Recall that for any $\mathbf{B} \in \mathbb{R}^{p\times p}$, we define $L(\mathbf{B})$ to be such that $\mathbf{B}_{g_{L(\mathbf{B})}} = 0$ and $\mathbf{B}_{s_{L(\mathbf{B})+1}} \neq 0$, and $S(\mathbf{B}) = \{L(\mathbf{B})+1, \cdots, p-1\}$ with $K(\mathbf{B}) = |S(\mathbf{B})|$. Note then that $K(\mathbf{B}) = p-1-L(\mathbf{B})$. We first establish the following theorem.

A.6.1 Theorem A.2

Theorem A.2. *For any* $\mathbf{B} \in \mathbb{R}^{p\times p}$,

$$\begin{aligned} \|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\|_F^2 \leq& \|\mathbf{S}_{s_p} - {\boldsymbol{\Sigma}^*}_{s_p}\|_2^2 + \|\boldsymbol{\Sigma}^* - \mathbf{B}\|_F^2 + 4\lambda^2 K(\mathbf{B}) w_0(L(\mathbf{B})) \\ &+ 2\left(\|\mathbf{S} - \boldsymbol{\Sigma}^*\|_{2,\infty} - \lambda\right)\cdot 1_{\{\|\mathbf{S}-\boldsymbol{\Sigma}^*\|_{2,\infty}\geq\lambda\}} \cdot \sqrt{\sum_{\ell=1}^{p-1} w_\ell^2} \cdot \|\hat{\boldsymbol{\Sigma}} - \mathbf{B}\|_F, \end{aligned}$$

where $w_0(\ell) = \max_{\ell+1\leq m\leq p-1} \sum_{s=m}^{p-1} w_{sm}^2$.

Recalling that the subdiagonal s_m is included in g_ℓ for $m \leq \ell \leq p-1$, we see that $\sum_{\ell=m}^{p-1} w_{\ell m}^2$ measures the amount of "net weight" applied to the subdiagonal s_m and $w_0(L(\mathbf{B}))$ measures the largest amount of "net weight" applied to any subdiagonal in $m \in S(\mathbf{B})$.

To prove Theorem A.2, we will rely on the following proposition:

Proposition A.1. *For any* $\mathbf{B} \in \mathbb{R}^{p\times p}$,

$$\begin{aligned}\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_F^2 \leq & \|\mathbf{\Sigma}^* - \mathbf{B}\|_F^2 - \|\hat{\mathbf{\Sigma}} - \mathbf{B}\|_F^2 + 2\langle \mathbf{S}_{s_p} - {\mathbf{\Sigma}^*}_{s_p}, \hat{\mathbf{\Sigma}}_{s_p} - \mathbf{B}_{s_p}\rangle \\ & + 2\left(\|\mathbf{S} - \mathbf{\Sigma}^*\|_{2,\infty} - \lambda\right) \cdot \|\hat{\mathbf{\Sigma}} - \mathbf{B}\|_{2,1} + 4\lambda \|\hat{\mathbf{\Sigma}}_{S(\mathbf{B})} - \mathbf{B}\|_{2,1}^*.\end{aligned}$$

Proof. See Section A.13 of the supplementary materials. □

We are now ready to prove Theorem A.2.

Proof of Theorem A.2. By Proposition A.1,

$$\begin{aligned}\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_F^2 \leq & \|\mathbf{\Sigma}^* - \mathbf{B}\|_F^2 - \|\hat{\mathbf{\Sigma}} - \mathbf{B}\|_F^2 + \ 2\langle \mathbf{S}_{s_p} - {\mathbf{\Sigma}^*}_{s_p}, \hat{\mathbf{\Sigma}}_{s_p} - \mathbf{B}_{s_p}\rangle \\ & + 2\left(\|\mathbf{S} - \mathbf{\Sigma}^*\|_{2,\infty} - \lambda\right) \cdot \|\hat{\mathbf{\Sigma}} - \mathbf{B}\|_{2,1} + 4\lambda \|\hat{\mathbf{\Sigma}}_{S(\mathbf{B})} - \mathbf{B}\|_{2,1}^*.\end{aligned} \tag{A.6.2}$$

First we have

$$2\langle \mathbf{S}_{s_p} - {\mathbf{\Sigma}^*}_{s_p}, \hat{\mathbf{\Sigma}}_{s_p} - \mathbf{B}_{s_p}\rangle \leq \|\mathbf{S}_{s_p} - {\mathbf{\Sigma}^*}_{s_p}\|_2^2 + \|\hat{\mathbf{\Sigma}}_{s_p} - \mathbf{B}_{s_p}\|_2^2. \tag{A.6.3}$$

Next,

$$\begin{aligned}\|\hat{\mathbf{\Sigma}} - \mathbf{B}\|_{2,1} = \sum_{\ell=1}^{p-1} w_\ell \|\hat{\mathbf{\Sigma}}_{s_\ell} - \mathbf{B}_{s_\ell}\|_2 &\leq \sqrt{\sum_{\ell=1}^{p-1} w_\ell^2} \cdot \sqrt{\sum_{\ell=1}^{p-1} \|\hat{\mathbf{\Sigma}}_{s_\ell} - \mathbf{B}_{s_\ell}\|_2^2} \\ &\leq \sqrt{\sum_{\ell=1}^{p-1} w_\ell^2} \cdot \|\hat{\mathbf{\Sigma}} - \mathbf{B}\|_F.\end{aligned} \tag{A.6.4}$$

Finally, since $2\lambda b \leq a\lambda^2 + b^2/a$, for any $a > 0$, we obtain

$$\begin{aligned}2\lambda\|\hat{\mathbf{\Sigma}}_{S(\mathbf{B})} - \mathbf{B}\|_{2,1}^* &= 2\lambda \sum_{\ell=L(\mathbf{B})+1}^{p-1} \sqrt{\sum_{m=L(\mathbf{B})+1}^{\ell} w_{\ell m}^2 \|\hat{\mathbf{\Sigma}}_{s_m} - \mathbf{B}_{s_m}\|_2^2} \\ &\leq K(\mathbf{B})\lambda^2 a + \sum_{\ell=L(\mathbf{B})+1}^{p-1} \sum_{m=L(\mathbf{B})+1}^{\ell} w_{\ell m}^2 \|\hat{\mathbf{\Sigma}}_{s_m} - \mathbf{B}_{s_m}\|_2^2/a \\ &\leq K(\mathbf{B})\lambda^2 a + \sum_{m=L(\mathbf{B})+1}^{p-1} \left(\sum_{\ell=m}^{p-1} w_{\ell m}^2\right) \|\hat{\mathbf{\Sigma}}_{s_m} - \mathbf{B}_{s_m}\|_2^2/a.\end{aligned}$$

Letting $a = 2w_0(L(\mathbf{B})) = 2\max_{L(\mathbf{B})+1\leq m\leq p-1} \sum_{\ell=m}^{p-1} w_{\ell m}^2$,

$$2\lambda\|\hat{\mathbf{\Sigma}}_{S(\mathbf{B})} - \mathbf{B}\|_{2,1}^* \leq 2K(\mathbf{B})\lambda^2 w_0(L(\mathbf{B})) + \frac{1}{2}\|\hat{\mathbf{\Sigma}}_{S(\mathbf{B})} - \mathbf{B}\|_F^2 - \frac{1}{2}\|\hat{\mathbf{\Sigma}}_{s_p} - \mathbf{B}_{s_p}\|_2^2. \tag{A.6.5}$$

Then combining (A.6.2), (A.6.3), (A.6.4) and (A.6.5),

$$\begin{aligned}\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_F^2 \leq & \|\mathbf{\Sigma}^* - \mathbf{B}\|_F^2 + \|\mathbf{S}_{s_p} - \mathbf{\Sigma}^*{}_{s_p}\|_2^2 + 4K(\mathbf{B})\lambda^2 w_0(L(\mathbf{B})) \\ & + 2\left(\|\mathbf{S} - \mathbf{\Sigma}^*\|_{2,\infty} - \lambda\right) \cdot 1_{\{\|\mathbf{S}-\mathbf{\Sigma}^*\|_{2,\infty} \geq \lambda\}} \cdot \sqrt{\sum_{\ell=1}^{p-1} w_\ell^2} \cdot \|\hat{\mathbf{\Sigma}} - \mathbf{B}\|_F,\end{aligned}$$

which concludes the proof. □

A.6.2 Proof of Theorem 6

We will use the following lemma.

Lemma A.2. *We have* $\|\mathbf{S} - \mathbf{\Sigma}^*\|_{2,\infty} \leq \max_{i,j} |\mathbf{S}_{ij} - \mathbf{\Sigma}^*{}_{ij}| \cdot \max_{1 \leq \ell \leq p-1} \sqrt{2\ell}/w_\ell$.

The proof follows immediately from the definition of the $\| \ \|_{2,\infty}$ norm given in (A.6.1).

Proof of Theorem 6

Proof. The first oracle inequality follows immediately from Theorem A.2, the choice of λ and $\mathcal{A}_x$, and the fact that $w_0(L(\mathbf{B})) \leq w_0(0) \leq 4p$ for the given weights. We now focus on the bound for $\mathbb{E}\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_F^2$. By Theorem A.2,

$$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_F^2 \leq R_1 + 2R_2(\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_F + \|\mathbf{\Sigma}^* - \mathbf{B}\|_F), \tag{A.6.6}$$

where

$$R_1 = \|\mathbf{\Sigma}^* - \mathbf{B}\|_F^2 + \|\mathbf{S}_{s_p} - \mathbf{\Sigma}^*{}_{s_p}\|_2^2 + 4\lambda^2 K(\mathbf{B}) w_0(L(\mathbf{B}))$$

and

$$R_2 = \left(\|\mathbf{S} - \mathbf{\Sigma}^*\|_{2,\infty} - \lambda\right) \cdot 1_{\{\|\mathbf{S}-\mathbf{\Sigma}^*\|_{2,\infty} \geq \lambda\}} \cdot \sqrt{\sum_{\ell=1}^{p-1} w_\ell^2}.$$

Using that $2R_2\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_F \leq 2R_2^2 + \frac{1}{2}\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_F^2$ and that $2R_2\|\mathbf{\Sigma}^* - \mathbf{B}\|_F \leq R_2^2 + \|\mathbf{\Sigma}^* - \mathbf{B}\|_F^2$, it follows from (A.6.6) that

$$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_F^2 \leq 6R_2^2 + 2\|\mathbf{\Sigma}^* - \mathbf{B}\|_F^2 + 2R_1. \tag{A.6.7}$$

With the given weights, $\sqrt{\sum_{\ell=1}^{p-1} w_\ell^2} \lesssim p$, hence

$$R_2 \lesssim p\left(\|\mathbf{S} - \mathbf{\Sigma}^*\|_{2,\infty} - \lambda\right) \cdot 1_{\{\|\mathbf{S}-\mathbf{\Sigma}^*\|_{2,\infty} \geq \lambda\}}.$$

By Lemma A.2 and with the given weights,$\|\mathbf{S} - \mathbf{\Sigma}^*\|_{2,\infty} \leq \max_{i,j} |\mathbf{S}_{ij} - \mathbf{\Sigma}^*{}_{ij}|$. Let $D = \max_{i,j} |\mathbf{S}_{ij} - \mathbf{\Sigma}^*{}_{ij}|$. Then by Theorem A.1 in Section A.4 and the given λ,

$$\mathbb{E}R_2^2 \lesssim p^2\mathbb{E}\left[(D - \lambda)^2 \cdot 1_{\{D>\lambda\}}\right] \lesssim p^2\mathbb{E}\left[T^2 \cdot 1_{\{D>\lambda\}}\right] \lesssim p/n.$$

Also it is easy to show that $\mathbb{E}\|\mathbf{S}_{s_p} - \mathbf{\Sigma^*}_{s_p}\|_2^2 = \sum_{j=1}^p \mathbb{E}(\mathbf{S}_{ij} - \mathbf{\Sigma^*}_{ij})^2 \lesssim p/n$. It follows by (A.6.7) that

$$\begin{aligned}\mathbb{E}\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_F^2 &\leq 6\mathbb{E}R_2^2 + 2\|\mathbf{\Sigma}^* - \mathbf{B}\|_F^2 + 2\mathbb{E}R_1 \\ &\lesssim \|\mathbf{\Sigma}^* - \mathbf{B}\|_F^2 + \frac{p}{n} + \lambda^2 K(\mathbf{B}) w_0(L(\mathbf{B})).\end{aligned}$$

Recalling that $w_0(L(B)) \lesssim p$ for the given weights, the theorem now follows. □

A.7 Proof of Frobenius norm lower bound

Proof of Theorem 7

Proof. Fix $0 < \alpha < 1/2$. Let $\mathbf{B}_{k,\ell} = \mathbf{e}_k\mathbf{e}_\ell^T + \mathbf{e}_\ell\mathbf{e}_k^T$ where $\mathbf{e}_k$ is the unit vector in $\mathbb{R}^p$ with the kth entry being 1 and $\mathbf{e}_\ell$ similarly defined. Let Ω be the subset of $\{0,1\}^{p(p-1)/2}$ such that if $\boldsymbol{\epsilon} = (\epsilon_{12}, \epsilon_{13}, \ldots, \epsilon_{1p}, \epsilon_{23}, \ldots, \epsilon_{2p}, \ldots, \epsilon_{p-1,p}) \in \Omega$, then $\epsilon_{k,\ell} = 0$ whenever $|k-\ell| > K$. Denote by N the number of entries in ϵ that are not fixed at 0, then $N = 2pK + o(pK)$. By Varshamov-Gilbert's bound (see Lemma 2.9 in Tsybakov 2009), there exists a subset Ω_0 of Ω such that: (i) $\mathbf{0} \in \Omega_0$; (ii) $\mathrm{Card}(\Omega_0) \geq 2^{N/8} + 1$; (iii) for any two distinct ϵ and ϵ' in Ω_0, the Hamming distance $\sum_{k,\ell} |\epsilon_{k,\ell} - \epsilon'_{k,\ell}| \geq N/8$.

Now for $\boldsymbol{\epsilon} \in \Omega_0$, define $\mathbf{\Sigma}_{\boldsymbol{\epsilon}} = \mathbf{I}_p + \frac{\alpha}{\sqrt{n}} \sum_{k<\ell} \epsilon_{k,\ell} \mathbf{B}_{k,\ell}$. Note that $\mathbf{\Sigma}_{\boldsymbol{\epsilon}}$ has bandwidth at most K. For any two distinct $\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon}'$ in Ω_0,

$$\|\mathbf{\Sigma}_{\boldsymbol{\epsilon}} - \mathbf{\Sigma}_{\boldsymbol{\epsilon}'}\|_F^2 = \frac{2\alpha^2}{n} \sum_{k<\ell} |\epsilon_{k,\ell} - \epsilon'_{k,\ell}| \geq \alpha^2 N/(4n). \tag{A.7.1}$$

It's easy to see that $\mathrm{tr}(\mathbf{\Sigma}_{\boldsymbol{\epsilon}}) = p$. Note that $\|\mathbf{\Sigma}_{\boldsymbol{\epsilon}} - \mathbf{I}_p\|_{op} \leq \frac{\alpha}{\sqrt{n}} 2K < 1$. Hence $\mathbf{\Sigma}_{\boldsymbol{\epsilon}}$ is positive definite.

With slight abuse of notation, let $\mathbb{P}_{\mathbf{\Sigma}}$ denote the joint probability distribution of $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and each $\mathbf{X}_i$ is from a multivariate normal distribution with mean zero and covariance $\mathbf{\Sigma}$. Let $\mathbb{K}(\mathbb{P}_{\mathbf{\Sigma}_{\boldsymbol{\epsilon}}}, \mathbb{P}_{\mathbf{I}_p}) = \int \log\left(\frac{d\mathbb{P}_{\mathbf{\Sigma}_{\boldsymbol{\epsilon}}}}{d\mathbb{P}_{\mathbf{I}_p}}\right) d\mathbb{P}_{\mathbf{\Sigma}_{\boldsymbol{\epsilon}}}$ be the Kullback-Leibler divergence. Then we can verify that

$$\begin{aligned}\mathbb{K}(\mathbb{P}_{\mathbf{\Sigma}_{\boldsymbol{\epsilon}}}, \mathbb{P}_{\mathbf{I}_p}) &= n\left\{-\frac{p}{2} + \frac{1}{2}\mathrm{tr}(\mathbf{\Sigma}_{\boldsymbol{\epsilon}}) - \frac{1}{2}\log\det(\mathbf{\Sigma}_{\boldsymbol{\epsilon}})\right\} \\ &= -\frac{n}{2}\log\det(\mathbf{\Sigma}_{\boldsymbol{\epsilon}}) = -\frac{n}{2}\sum_{k=1}^p \log\left\{1 + \lambda_k(\tilde{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}})\right\},\end{aligned}$$

where $\tilde{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}} = \mathbf{\Sigma}_{\boldsymbol{\epsilon}} - \mathbf{I}_p$. By the fact that $\log(1+x) \geq x - x^2/2$ for any $x \geq 0$ and that $\sum_{k=1}^p \lambda_k(\tilde{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}) = 0$, we obtain that

$$\mathbb{K}(\mathbb{P}_{\mathbf{\Sigma}_{\boldsymbol{\epsilon}}}, \mathbb{P}_{\mathbf{I}_p}) \leq \frac{n}{4}\sum_{k=1}^p \lambda_k^2(\tilde{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}) = \frac{n}{4}\|\tilde{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}\|_F^2 \leq \alpha^2 pK.$$

Therefore, $\sum_{\mathbf{\Sigma}_\epsilon \in \Omega_0} \mathbb{K}(\mathbb{P}_{\mathbf{\Sigma}_\epsilon}, \mathbb{P}_{\mathbf{I}_p})/\mathrm{Card}(\Omega_0) \leq \alpha^2 pK$. Since $\log(\mathrm{Card}(\Omega_0)-1) \geq \log(2)N/8$ and $N = 2pK + o(pK)$, for any $0 < a < 1/8$, we can choose α small enough (depends only on a) such that

$$\frac{1}{\mathrm{Card}(\Omega_0)} \sum_{\mathbf{\Sigma}_\epsilon \in \Omega_0} \mathbb{K}(\mathbb{P}_{\mathbf{\Sigma}_\epsilon}, \mathbb{P}_{\mathbf{I}_p}) \leq a \log(\mathrm{Card}(\Omega_0) - 1). \tag{A.7.2}$$

With (A.7.1) and (A.7.2), by Theorem 2.5 in Tsybakov (2009), the theorem holds. □

A.8 Proof of convergence in operator norm

Proof of Theorem 9

Proof. The arguments given here hold on the set $\mathcal{A}_x$ defined in (4.1), with x as in Theorem 6. Since, under our assumptions, we have $\hat{K} = K$, with high probability, we further have:

$$\begin{aligned}
\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{op} \leq & \|\hat{\mathbf{\Sigma}}_S - \mathbf{S}_S\|_{op} + \|\mathbf{S}_S - \mathbf{\Sigma}^*{}_S\|_{op} + \|\hat{\mathbf{\Sigma}}_{S^c} - \mathbf{\Sigma}^*{}_{S^c}\|_{op} \\
= & \|\hat{\mathbf{\Sigma}}_S - \mathbf{S}_S\|_{op} + \|\mathbf{S}_S - \mathbf{\Sigma}^*{}_S\|_{op} + \|\mathbf{\Sigma}^*{}_{S^c}\|_{op} \\
\lesssim & \|\hat{\mathbf{\Sigma}}_S - \mathbf{S}_S\|_{1,1} + \|\mathbf{S}_S - \mathbf{\Sigma}^*{}_S\|_{1,1} + K\sqrt{\log p/n} \\
\lesssim & \max_i \sum_{|i-j| \leq K} |\hat{\mathbf{\Sigma}}_{ij} - \mathbf{S}_{ij}| + K\sqrt{\log p/n}.
\end{aligned}$$

We claim: *there exists a constant $c > 0$ such that*

$$|\hat{\mathbf{\Sigma}}_{ij} - \mathbf{S}_{ij}| \leq c\lambda \quad \textit{for all } |i-j| \leq K. \tag{A.8.1}$$

Then we have $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{op} \lesssim K\sqrt{\log p/n}$ and the proof is complete.

Next, we prove claim (A.8.1). By (A.3.1), we have for $\ell \geq L$ and $m \leq \ell$,

$$\hat{\mathbf{R}}^{(\ell+1)}_{s_m} = \frac{\hat{\nu}_\ell}{w^2_{\ell,m} + \hat{\nu}_\ell} \hat{\mathbf{R}}^{(\ell)}_{s_m} = \hat{\mathbf{R}}^{(\ell)}_{s_m} - \frac{w^2_{\ell,m}}{w^2_{\ell,m} + \hat{\nu}_\ell} \hat{\mathbf{R}}^{(\ell)}_{s_m},$$

where $\hat{\nu}_\ell$ satisfies

$$\sum_{m=1}^{\ell} \frac{w^2_{\ell,m}}{(w^2_{\ell,m} + \hat{\nu}_\ell)^2} \|\hat{\mathbf{R}}^{(\ell)}_{s_m}\|_2^2 = \lambda^2. \tag{A.8.2}$$

Let $h_{\ell,m} = \frac{w^2_{\ell,m}}{w^2_{\ell,m} + \hat{\nu}_\ell}$, then we have $\hat{\mathbf{R}}^{(\ell+1)}_{s_m} = (1 - h_{\ell,m})\hat{\mathbf{R}}^{(\ell)}_{s_m}$, and hence for $m \geq L+1$,

$$\hat{\mathbf{\Sigma}}_{s_m} = \mathbf{S}_{s_m} \prod_{\ell=m}^{p-1} (1 - h_{\ell,m}).$$

Let $h_m = \prod_{\ell=m}^{p-1}(1 - h_{\ell,m})$, then $h_m < 1$ and $h_m = 1 - \sum_{\ell=m}^{p-1} h_{\ell,m} + o\left\{\left(\sum_{\ell=m}^{p-1} h_{\ell,m}\right)^2\right\}$. Note that if we establish that

$$\sum_{\ell=m}^{p-1} h_{\ell,m} \leq C\lambda, \tag{A.8.3}$$

for some constant $C > 0$, then, for each $L+1 \le m \le p$ and each $(i,j) \in S_m$, we have $|\hat{\mathbf{\Sigma}}_{ij} - \mathbf{\Sigma^*}_{ij}| = |h_m \mathbf{S}_{ij} - h_m \mathbf{\Sigma^*}_{ij} + h_m \mathbf{\Sigma^*}_{ij} - \mathbf{\Sigma^*}_{ij}| \le c\lambda$
for some sufficiently large c that does not depend on i or j. Therefore to prove (A.8.1), it suffices to prove (A.8.3).

Now we focus on $\hat{\nu}_\ell$. By (A.8.2), if $\ell \ge L+1$, then $\sum_{m=L+1}^{\ell} \frac{w_{\ell,m}^2}{(w_{\ell,\ell}^2+\hat{\nu}_\ell)^2} \|\hat{\mathbf{R}}_{s_m}^{(\ell)}\|_2^2 \le \lambda^2$, which leads to

$$w_{\ell,\ell}^2 + \hat{\nu}_\ell \ge \frac{\sqrt{\sum_{m=L+1}^{\ell} w_{\ell,m}^2 \|\hat{\mathbf{R}}_{s_m}^{(\ell)}\|_2^2}}{\lambda} \ge \frac{w_{\ell\ell}\|\hat{\mathbf{R}}_{s_\ell}^{(\ell)}\|_2}{\lambda},$$

since $\max_m w_{\ell m} = w_{\ell\ell} =: w_\ell$. Note that, by (A.3.1), for every $m \ge L+1$, we have $\|\hat{\mathbf{R}}_{s_m}^{(\ell)}\|_2 = \|\mathbf{S}_{s_m}\|_2 \prod_{\ell'=m}^{\ell-1}(1-h_{\ell',m})$ and $\|\hat{\mathbf{R}}_{s_\ell}^{(\ell)}\|_2 = \|\mathbf{S}_{s_\ell}\|_2$. Then $\hat{\nu}_\ell \ge \frac{w_\ell \|\mathbf{S}_{s_\ell}\|_2}{\lambda} - w_\ell^2$. Since $h_{\ell,m} = \frac{w_{\ell,m}^2}{w_{\ell,m}^2+\hat{\nu}_\ell}$, we derive that

$$\sum_{\ell=m}^{p-1} h_{\ell m} \le \sum_{\ell=m}^{p-1} \frac{w_{\ell m}^2}{\hat{\nu}_\ell} \le \frac{\sum_{\ell=m}^{p-1} w_{\ell m}^2/(2\ell)}{\min_{\ell=L+1}^{p-1} \hat{\nu}_\ell/(2\ell)} \lesssim \frac{1}{\min_{\ell=L+1}^{p-1} \hat{\nu}_\ell/(2\ell)},$$

where the last inequality follows with the two sets of given weights.

Therefore, to prove (A.8.3), we just need to show that $\min_{L+1\le\ell\le p-1} \hat{\nu}_\ell/\ell \gtrsim \lambda^{-1}$, which follows immediately by the signal strength condition and by the fact that $\|\mathbf{S}_{s_\ell}\|_2 - \|\mathbf{\Sigma^*}_{s_\ell}\|_2 \gtrsim -\lambda\sqrt{2\ell}$ uniformly for all ℓ. □

A.9 Counterexample

The purpose of this section is to show that the signal strength condition of Theorem 9 of the main paper is necessary. In particular, we show that if the signal strength condition fails and $p > n$, then the estimator $\hat{\mathbf{\Sigma}}$ may not even be consistent in operator norm. The following example illustrates this. Let $\mathbf{\Sigma^*}$ be such that $\mathbf{\Sigma^*}_{ii} = 1$ for all i, $\mathbf{\Sigma^*}_{12} = \mathbf{\Sigma^*}_{21} = 0.5$, and $\mathbf{\Sigma^*}_{ij} = 0$ otherwise. Then $\|\mathbf{\Sigma^*}_{s_{p-1}}\|_2/\sqrt{2(p-1)} = o(1)$ and the signal strength condition cannot hold when p grows. For the above $\mathbf{\Sigma^*}$, $S = \{p-1, p\}$. For the estimator in Theorem 9 with $\lambda = 2x\sqrt{\log p/n}$, $\hat{S} \subseteq \{p-1, p\}$ by Theorem 3. Then, by (A.3.1) and (A.3.2) of Theorem 3 above, and recalling that $w_{p-1} = \sqrt{2(p-1)}$, we have $\hat{\mathbf{\Sigma}}_{s_{p-1}} = \frac{[\hat{\nu}_{p-1}]_+}{[\hat{\nu}_{p-1}]_+ + 2(p-1)} \mathbf{S}_{s_{p-1}}$, where $\hat{\nu}_{p-1}$ satisfies

$$\lambda^2 = \frac{2(p-1)}{(2(p-1)+\hat{\nu}_{p-1})^2} \|\mathbf{S}_{s_{p-1}}\|_2^2. \tag{A.9.1}$$

Notice that we have

$$\|\mathbf{S}_{s_{p-1}}\|_2 \le \|\mathbf{\Sigma^*}_{s_{p-1}}\|_2 + \|\mathbf{S}_{s_{p-1}} - \mathbf{\Sigma^*}_{s_{p-1}}\|_2 \le \sqrt{2 \times 0.5^2} + \|\mathbf{S}_{s_{p-1}} - \mathbf{\Sigma^*}_{s_{p-1}}\|_2 < \lambda\sqrt{2(p-1)}$$

with high probability, for large p. Then, (A.9.1) implies that $\hat{\nu}_{p-1} \le 0$, in which case $\hat{\mathbf{\Sigma}}_{s_{p-1}} = 0$, so $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma^*}\|_{op} \ge |\mathbf{\Sigma^*}_{12}| = 0.5$, and the estimator cannot be consistent.

A.10 Positive definiteness

A.10.1 Proof of Theorem 10

Proof. Let $\mathbf{u}$ be the eigenvector of $\hat{\mathbf{\Sigma}}$ such that $\mathbf{u}^T\hat{\mathbf{\Sigma}}\mathbf{u} = \lambda_{\min}(\hat{\mathbf{\Sigma}})$. Then

$$\lambda_{\min}(\hat{\mathbf{\Sigma}}) = \mathbf{u}^T\mathbf{\Sigma}^*\mathbf{u} - \mathbf{u}^T(\mathbf{\Sigma}^* - \hat{\mathbf{\Sigma}})\mathbf{u} \geq \lambda_{\min}(\mathbf{\Sigma}^*) - \|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{op}.$$

Now, by Theorem 9, $\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{op} \leq C'K\sqrt{\frac{\log p}{n}}$, so the assumption on $\lambda_{\min}(\mathbf{\Sigma}^*)$ ensures that, whp, $\hat{\mathbf{\Sigma}} \succeq \delta\mathbf{I}_p$. Thus, the constraint in (2.1) may be dropped without changing the solution, meaning $\hat{\mathbf{\Sigma}} = \tilde{\mathbf{\Sigma}}$. □

A.10.2 Algorithm for $\tilde{\mathbf{\Sigma}}$ and its derivation

Theorem A.3. *A dual of* (2.3) *is given by*

$$\begin{aligned} &\text{Minimize}_{\mathbf{A}^{(\ell)},\mathbf{C}\in\mathbb{R}^{p\times p}} \frac{1}{2}\left\|\mathbf{S} - \lambda\sum_{\ell=1}^{p-1}\mathbf{W}^{(\ell)} * \mathbf{A}^{(\ell)} + \lambda\mathbf{C}\right\|_F^2 - \lambda\delta\cdot\mathrm{tr}(\mathbf{C}) \qquad \text{(A.10.1)} \\ &\text{s.t. } \mathbf{C}\succeq 0,\ \|\mathbf{A}^{(\ell)}_{g_\ell}\|_2 \leq 1,\ \mathbf{A}^{(\ell)}_{g_\ell{}^c} = 0 \text{ for } 1\leq\ell\leq p-1. \end{aligned}$$

In particular, given a solution to the dual, $(\hat{\mathbf{A}}^{(1)},\ldots,\hat{\mathbf{A}}^{(p-1)},\hat{\mathbf{C}})$*, the solution to* (2.3) *is given by*

$$\tilde{\mathbf{\Sigma}} = \mathbf{S} - \lambda\sum_{\ell=1}^{p-1}\mathbf{W}^{(\ell)} * \hat{\mathbf{A}}^{(\ell)} + \lambda\hat{\mathbf{C}}. \qquad \text{(A.10.2)}$$

Proof. Define

$$L(\mathbf{\Sigma},\mathbf{A}^{(1)},\ldots,\mathbf{A}^{(p-1)},\mathbf{C}) = \frac{1}{2}\|\mathbf{S}-\mathbf{\Sigma}\|_F^2 + \lambda\langle\sum_{\ell=1}^{p-1}W_\ell * \mathbf{A}^{(\ell)},\mathbf{\Sigma}\rangle - \lambda\langle\mathbf{C},\mathbf{\Sigma}-\delta\mathbf{I}_p\rangle.$$

Observe that

$$1_\infty\{\mathbf{\Sigma}\succeq\delta\mathbf{I}_p\} = \max_{\mathbf{C}\succeq 0} -\langle\mathbf{\Sigma}-\delta\mathbf{I}_p,\mathbf{C}\rangle$$

and (as before) that

$$\|(\mathbf{W}^{(\ell)} * \mathbf{\Sigma})_{g_\ell}\|_2 = \max_{\mathbf{A}^{(\ell)}}\left\{\langle\mathbf{W}^{(\ell)} * \mathbf{A}^{(\ell)},\mathbf{\Sigma}\rangle \text{ s.t. } \|\mathbf{A}^{(\ell)}_{g_\ell}\|_2\leq 1,\ \mathbf{A}^{(\ell)}_{g_\ell^c} = 0\right\}.$$

It follows that (2.1) is equivalent to

$$\min_{\mathbf{\Sigma}}\left\{\max_{\mathbf{A}^{(\ell)},\mathbf{C}}\left[L(\mathbf{\Sigma},\mathbf{A}^{(1)},\ldots,\mathbf{A}^{(p-1)},\mathbf{C}) \text{ s.t. } \|\mathbf{A}^{(\ell)}_{g_\ell}\|_2\leq 1, \mathbf{A}^{(\ell)}_{g_\ell{}^c} = 0, \mathbf{C}\succeq 0\right]\right\}.$$

Algorithm A.1 BCD on dual of Problem (2.3).

Inputs: $\mathbf{S}$, δ, λ, and weights matrices, $\mathbf{W}^{(\ell)}$. Initialize $\mathbf{A}^{(\ell)}, \mathbf{C}$.
Repeat until convergence:

- $\{\hat{\mathbf{A}}^{(\ell)}\} \leftarrow$ `threshold_subdiagonals`$\Big(\mathbf{S} + \lambda\hat{\mathbf{C}}, \{\hat{\mathbf{A}}^{(\ell)}\}, \lambda, \{w_{\ell m}\}\Big)$
- Let $\mathbf{U}\mathbf{D}\mathbf{U}^T = \lambda\sum_{\ell=1}^{p-1}\mathbf{W}^{(\ell)} * \hat{\mathbf{A}}^{(\ell)} - \mathbf{S}$ be the eigenvalue decomposition. Then,

$$\lambda\hat{\mathbf{C}} \leftarrow \mathbf{U}[\mathbf{D} + \delta\mathbf{I}_p]_+\mathbf{U}^T$$

  where the positive part, $[\cdot]_+$, is applied to each diagonal element.

Subroutine `threshold_subdiagonals`$\Big(\mathbf{R}, \{\hat{\mathbf{A}}^{(\ell)}\}, \lambda, \{w_{\ell m}\}\Big)$ For $\ell = 1, \ldots, p-1$:

- Compute $\hat{\mathbf{R}}^{(\ell)} \leftarrow \mathbf{R} - \lambda\sum_{\ell=1}^{p-1}\mathbf{W}^{(\ell)} * \hat{\mathbf{A}}^{(\ell)}$
- For $m \leq \ell$, set $\hat{\mathbf{A}}^{(\ell)}_{s_m} \leftarrow \frac{w_{\ell m}}{\lambda(w_{\ell m}^2 + \max\{\hat{\nu}_\ell, 0\})}\hat{\mathbf{R}}^{(\ell)}_{s_m}$ where $\hat{\nu}_\ell$ satisfies $\lambda^2 = h_\ell(\hat{\nu}_\ell)$, as in (3.2).

Return $\{\hat{\mathbf{A}}^{(\ell)}\}$.

We get the dual problem by interchanging the min and max. The inner minimization gives the primal-dual relation given in the theorem and the following dual function:

$$\min_{\mathbf{\Sigma}} L(\mathbf{\Sigma}, \mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(p-1)}, \mathbf{C}) = -\frac{1}{2}\|\mathbf{S} - \lambda\sum_{\ell=1}^{p-1}\mathbf{W}^{(\ell)} * \mathbf{A}^{(\ell)} + \lambda\mathbf{C}\|_F^2 + \frac{1}{2}\|\mathbf{S}\|_F^2 + \lambda\delta \cdot \mathrm{tr}(\mathbf{C}).$$

□

A BCD algorithm for solving (A.10.1) is given in Algorithm A.1. The update over C involves projecting a matrix onto the positive semidefinite cone. The other details are similar to those explained in Section 3.

A.11 Additional simulation results

In this section, we present additional simulation results. In Section 5.1.1, we showed that the empirical convergence rates in Frobenius norm correspond well to those predicted by theory. In Figure A.1 of this section we show (under identical simulation settings to those of Section 5.1.1) that the same is true for the convergence in operator norm. We scale the operator norm by $\sqrt{\log(p)}$ in accordance with the right-hand side of Theorem 9. In Figure A.2, we observe that the operator norm decays like $n^{-1/2}$ in agreement with Theorem 9.

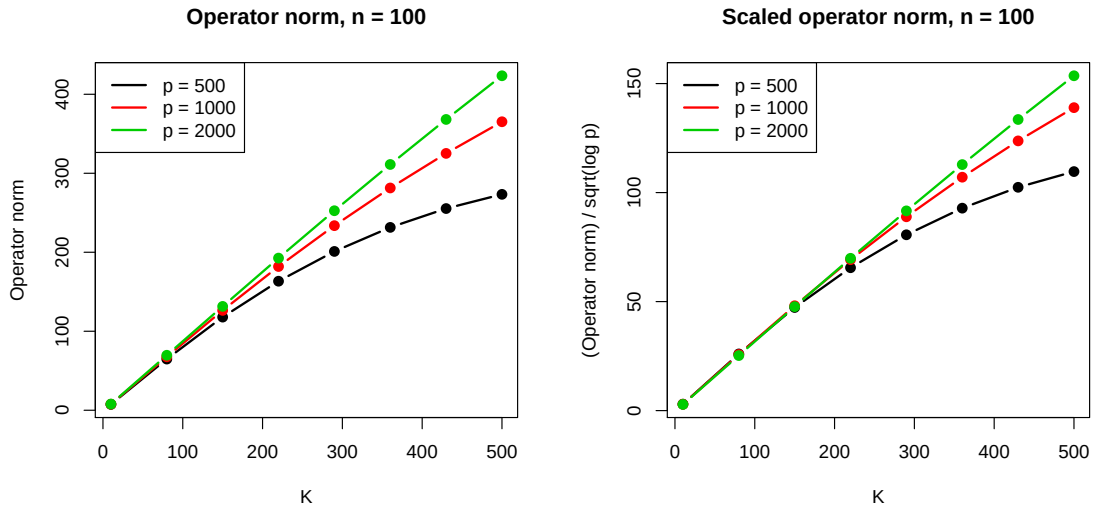

Figure A.1: *Convergence in operator norm. Monte Carlo estimate of (Left)* $E\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{op}$ *and (Right)* $E\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{op}/\sqrt{\log p}$ *as a function of* K*, the bandwidth of* $\mathbf{\Sigma}^*$*.*

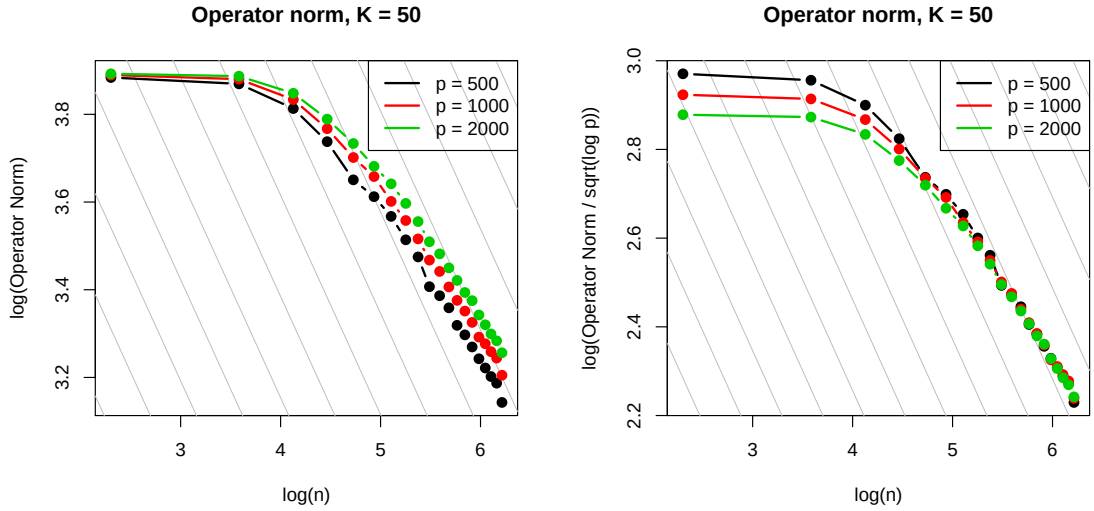

Figure A.2: *(Left:) For large* n*, expected* $E\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_{op}$ *is seen to decay like* $n^{-1/2}$ *as can be seen from* $-1/2$ *slope of* log-log *plot (gray lines are of slope* $-1/2$*). (Right:) Scaling this quantity by* $\sqrt{\log p}$ *aligns the curves for large* n*. Both of these phenomena are suggested by Theorem 9.*

A.12 Phoneme data sample covariance matrices

In Section 5.2 of the main paper, we state that inspection of the sample covariance matrices supports the notion that an approximate banded structure may be present in the phoneme data. See Figure A.3.

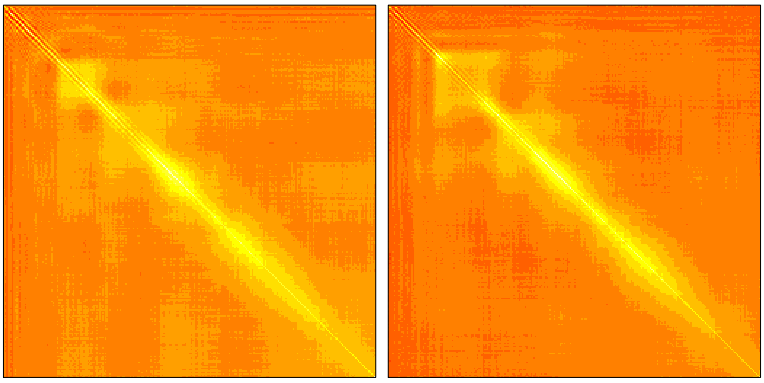

Figure A.3: *Within-class sample covariance matrices for phoneme data*

A.13 Proof of Proposition A.1

In Section A.6.1 of the supplementary materials, a key result for proving Theorem A.2 came from Proposition A.1, which (to remind the reader) states that, for any $B \in \mathbb{R}^{p\times p}$,

$$\begin{aligned}\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\|_F^2 \leq & \|\boldsymbol{\Sigma}^* - \mathbf{B}\|_F^2 - \|\hat{\boldsymbol{\Sigma}} - \mathbf{B}\|_F^2 + 2\langle \mathbf{S}_{s_p} - \boldsymbol{\Sigma}^*{}_{s_p}, \hat{\boldsymbol{\Sigma}}_{s_p} - \mathbf{B}_{s_p}\rangle \\ & + 2\left(\|\mathbf{S} - \boldsymbol{\Sigma}^*\|_{2,\infty} - \lambda\right) \cdot \|\hat{\boldsymbol{\Sigma}} - \mathbf{B}\|_{2,1} + 4\lambda \|\hat{\boldsymbol{\Sigma}}_{S(\mathbf{B})} - \mathbf{B}\|_{2,1}^*.\end{aligned}$$

We prove this result in this section. We begin by stating and proving two lemmas and a proposition that will be instrumental in the proof of Proposition A.1. The first lemma provides bounds on the inner product of two matrices in terms of the newly introduced norms in (A.6.1). We directly bound the inner product $\langle \mathbf{A}, \mathbf{B}\rangle^-$ in which we leave out the contribution of the diagonals,

$$\langle \mathbf{A}, \mathbf{B}\rangle^- = \langle \mathbf{A}, \mathbf{B}\rangle - \langle \mathbf{A}_{s_p}, \mathbf{B}_{s_p}\rangle = \sum_{j\neq k} A_{jk}B_{jk}.$$

We treat the main diagonal differently from the rest because it does not appear in the penalty term $\|\boldsymbol{\Sigma}\|_{2,1}^*$ of (2.2).

Lemma A.3. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be two arbitrary* $p \times p$ *matrices, then* $\langle \mathbf{A}, \mathbf{B}\rangle^- \leq \|\mathbf{A}\|_{2,1} \cdot \|\mathbf{B}\|_{2,\infty} \leq \|\mathbf{A}\|_{2,1}^* \cdot \|\mathbf{B}\|_{2,\infty}$.

Proof. $\langle \mathbf{A}, \mathbf{B}\rangle^- = \sum_{\ell=1}^{p-1} \langle \mathbf{A}_{s_\ell}, \mathbf{B}_{s_\ell}\rangle \leq \sum_{\ell=1}^{p-1} \|\mathbf{A}_{s_\ell}\|_2 \cdot \|\mathbf{B}_{s_\ell}\|_2 \leq \|\mathbf{A}\|_{2,1} \cdot \|\mathbf{B}\|_{2,\infty}$. The second inequality follows from the fact that $\|\mathbf{A}\|_{2,1} \leq \|\mathbf{A}\|_{2,1}^*$. □

For any matrix $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ and set $S \subseteq \{1,\ldots,p-1\}$, let $\mathbf{\Sigma}_S$ denote the $p \times p$ matrix such that $[\mathbf{\Sigma}_S]_{ij} = \mathbf{\Sigma}_{ij} 1\{p - |i-j| \in S\}$ and let $\mathbf{\Sigma}_{S^c} = \mathbf{\Sigma} - \mathbf{\Sigma}_S$.

Lemma A.4. *Let $S = \{L+1,\ldots,p-1\}$ for some L. For any $p \times p$ matrix $\mathbf{\Sigma}$,*

$$\begin{aligned} (i) \quad \|\mathbf{\Sigma}\|_{2,1} &= \|\mathbf{\Sigma}_S\|_{2,1} + \|\mathbf{\Sigma}_{S^c}\|_{2,1}, \\ (ii) \quad \|\mathbf{\Sigma}\|^*_{2,1} &\leq \|\mathbf{\Sigma}_S\|^*_{2,1} + \|\mathbf{\Sigma}_{S^c}\|^*_{2,1}, \\ (iii) \quad \|\mathbf{\Sigma}\|^*_{2,1} &\geq \|\mathbf{\Sigma}_S\|^*_{2,1} + \|\mathbf{\Sigma}_{S^c}\|_{2,1}. \end{aligned}$$

Proof. We have

$$\begin{aligned} \|\mathbf{\Sigma}\|_{2,1} &= \sum_{\ell=1}^{p-1} w_\ell \|\mathbf{\Sigma}_{s_\ell}\|_2 = \sum_{\ell=1}^{p-1} w_\ell \left(\|1_{\{\ell\in S\}}\mathbf{\Sigma}_{s_\ell}\|_2 + \|1_{\{\ell\notin S\}}\mathbf{\Sigma}_{s_\ell}\|_2 \right) \\ &= \|\mathbf{\Sigma}_S\|_{2,1} + \|\mathbf{\Sigma}_{S^c}\|_{2,1}. \end{aligned}$$

Similarly,

$$\begin{aligned} \|\mathbf{\Sigma}\|^*_{2,1} &= \sum_{\ell=1}^{p-1} \sqrt{\sum_{m=1}^{\ell} w_{\ell m}^2 \|\mathbf{\Sigma}_{s_m}\|_2^2} \\ &\leq \sum_{\ell=1}^{p-1} \left\{ \sqrt{\sum_{m=1}^{\ell} w_{\ell m}^2 \|1_{\{m\in S\}}\mathbf{\Sigma}_{s_m}\|_2^2} + \sqrt{\sum_{m=1}^{\ell} w_{\ell m}^2 \|1_{\{m\notin S\}}\mathbf{\Sigma}_{s_m}\|_2^2} \right\} \\ &= \|\mathbf{\Sigma}_S\|^*_{2,1} + \|\mathbf{\Sigma}_{S^c}\|^*_{2,1}. \end{aligned}$$

Finally,

$$\begin{aligned} \|\mathbf{\Sigma}\|^*_{2,1} &= \sum_{\ell=1}^{p-1} \sqrt{\sum_{m=1}^{\ell} w_{\ell m}^2 \|\mathbf{\Sigma}_{s_m}\|_2^2} \\ &\geq \sum_{\ell=1}^{p-1} \left\{ \sqrt{\sum_{m=1}^{\ell} w_{\ell m}^2 \|1_{\{m\in S\}}\mathbf{\Sigma}_{s_m}\|_2^2} + w_{\ell\ell}\|1_{\{\ell\notin S\}}\mathbf{\Sigma}_{s_\ell}\|_2 \right\} \\ &= \|\mathbf{\Sigma}_S\|^*_{2,1} + \|\mathbf{\Sigma}_{S^c}\|_{2,1}. \end{aligned}$$

□

Let $w_{\ell\cdot} \in \mathbb{R}^\ell$ denote the weights on the ℓth triangle and let the weight matrix $\mathbf{W}^{(\ell)} \in \mathbb{R}^{p\times p}$ be defined as: $\mathbf{W}^{(\ell)}_{s_m} = w_{\ell m}\mathbf{1}_{2m}$ for $1 \leq m \leq \ell$ and $\mathbf{W}^{(\ell)}_{s_m} = 0$ if $m > \ell$. Here $\mathbf{1}_{2m}$ is a length-$2m$ vector of 1's. Observe that the penalty term (2.2) can be equivalently written as $\|\mathbf{\Sigma}\|^*_{2,1} = \sum_{\ell=1}^{p-1} \|(\mathbf{W}^{(\ell)} * \mathbf{\Sigma})_{g_\ell}\|_2$, where $*$ denotes elementwise multiplication. Define $f_\ell(\mathbf{B}) := \|(\mathbf{W}^{(\ell)} * \mathbf{B})_{g_\ell}\|_2$. Recall the definitions of the new norms in (A.6.1).

Proposition A.2. *For any* $\mathbf{B} \in \mathbb{R}^{p\times p}$ *and* $\mathbf{W}^{(\ell)} * \mathbf{A}^{(\ell)} \in \partial f_\ell(\mathbf{B}), 1 \le \ell \le p-1$,

$$\begin{aligned}\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_F^2 \le & \|\mathbf{\Sigma}^* - \mathbf{B}\|_F^2 - \|\hat{\mathbf{\Sigma}} - \mathbf{B}\|_F^2 + 2\langle \mathbf{S}_{s_p} - \mathbf{\Sigma}^*{}_{s_p}, \hat{\mathbf{\Sigma}}_{s_p} - \mathbf{B}_{s_p}\rangle \\ & + 2\|\mathbf{S} - \mathbf{\Sigma}^*\|_{2,\infty} \cdot \|\hat{\mathbf{\Sigma}} - \mathbf{B}\|_{2,1} - 2\lambda \left\langle \sum_{\ell=1}^{p-1} \mathbf{W}^{(\ell)} * \mathbf{A}^{(\ell)}, \hat{\mathbf{\Sigma}} - \mathbf{B} \right\rangle.\end{aligned}$$

Proof. $f_\ell(\mathbf{B})$ is convex and its sub-differential is

$$\begin{aligned}\partial f_\ell(\mathbf{B}) = \Big\{ & \mathbf{W}^{(\ell)} * \mathbf{A}^{(\ell)} \in \mathbb{R}^{p\times p} : \|\mathbf{A}^{(\ell)}_{g_\ell}\|_2 \le 1, \mathbf{A}^{(\ell)}_{g_\ell^c} = 0 \text{ and} \\ & \big\langle (\mathbf{W}^{(\ell)} * \mathbf{B})_{g_\ell}, \mathbf{A}^{(\ell)}_{g_\ell} \big\rangle = \|(\mathbf{W}^{(\ell)} * \mathbf{B})_{g_\ell}\|_2 \cdot \|\mathbf{A}^{(\ell)}_{g_\ell}\|_2 \Big\}.\end{aligned} \tag{A.13.1}$$

Let $\mathbf{W}^{(\ell)} * \hat{\mathbf{A}}^{(\ell)} \in \partial f_\ell(\hat{\mathbf{\Sigma}})$. For an arbitrary $\mathbf{B} \in \mathbb{R}^{p\times p}$, let $\mathbf{W}^{(\ell)} * \mathbf{A}^{(\ell)} \in \partial f_\ell(\mathbf{B})$. Since the sub-gradient of a convex function is monotone, we have

$$\begin{aligned}\langle \mathbf{W}^{(\ell)} * \hat{\mathbf{A}}^{(\ell)}, \hat{\mathbf{\Sigma}} - \mathbf{B}\rangle &= \left\langle (\mathbf{W}^{(\ell)} * \hat{\mathbf{A}}^{(\ell)})_{g_\ell}, (\hat{\mathbf{\Sigma}} - \mathbf{B})_{g_\ell} \right\rangle \\ &\ge \left\langle (\mathbf{W}^{(\ell)} * \mathbf{A}^{(\ell)})_{g_\ell}, (\hat{\mathbf{\Sigma}} - \mathbf{B})_{g_\ell} \right\rangle = \langle \mathbf{W}^{(\ell)} * \mathbf{A}^{(\ell)}, \hat{\mathbf{\Sigma}} - \mathbf{B}\rangle.\end{aligned}$$

It follows that

$$\left\langle \sum_{\ell=1}^{p-1} \mathbf{W}^{(\ell)} * \hat{\mathbf{A}}^{(\ell)}, \hat{\mathbf{\Sigma}} - \mathbf{B} \right\rangle \ge \left\langle \sum_{\ell=1}^{p-1} \mathbf{W}^{(\ell)} * \mathbf{A}^{(\ell)}, \hat{\mathbf{\Sigma}} - \mathbf{B} \right\rangle. \tag{A.13.2}$$

Using the primal-dual relation given in Theorem 1 of the main paper and the fact that $\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^* = \hat{\mathbf{\Sigma}} - \mathbf{S} + \mathbf{S} - \mathbf{\Sigma}^*$, we have

$$\langle \hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*, \hat{\mathbf{\Sigma}} - \mathbf{B}\rangle = \langle \mathbf{S} - \mathbf{\Sigma}^*, \hat{\mathbf{\Sigma}} - \mathbf{B}\rangle - \lambda \left\langle \sum_{\ell=1}^{p-1} \mathbf{W}^{(\ell)} * \hat{\mathbf{A}}^{(\ell)}, \hat{\mathbf{\Sigma}} - \mathbf{B} \right\rangle.$$

Combining this with (A.13.2), we derive that

$$\langle \hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*, \hat{\mathbf{\Sigma}} - \mathbf{B}\rangle \le \langle \mathbf{S} - \mathbf{\Sigma}^*, \hat{\mathbf{\Sigma}} - \mathbf{B}\rangle - \lambda \left\langle \sum_{\ell=1}^{p-1} \mathbf{W}^{(\ell)} * \mathbf{A}^{(\ell)}, \hat{\mathbf{\Sigma}} - \mathbf{B} \right\rangle. \tag{A.13.3}$$

By the cosine formula, $2\langle \hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*, \hat{\mathbf{\Sigma}} - \mathbf{B}\rangle = \|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_F^2 + \|\hat{\mathbf{\Sigma}} - \mathbf{B}\|_F^2 - \|\mathbf{\Sigma}^* - \mathbf{B}\|_F^2$. Therefore, we can rewrite (A.13.3) as

$$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^*\|_F^2 + \|\hat{\mathbf{\Sigma}} - \mathbf{B}\|_F^2 \le \|\mathbf{\Sigma}^* - \mathbf{B}\|_F^2 + 2\langle \mathbf{S} - \mathbf{\Sigma}^*, \hat{\mathbf{\Sigma}} - \mathbf{B}\rangle - 2\lambda \left\langle \sum_{\ell=1}^{p-1} \mathbf{W}^{(\ell)} * \mathbf{A}^{(\ell)}, \hat{\mathbf{\Sigma}} - \mathbf{B} \right\rangle$$

and the proposition follows since, by Lemma A.3,

$$\langle \mathbf{S} - \mathbf{\Sigma}^*, \hat{\mathbf{\Sigma}} - \mathbf{B}\rangle \le \langle \mathbf{S}_{s_p} - \mathbf{\Sigma}^*{}_{s_p}, \hat{\mathbf{\Sigma}}_{s_p} - \mathbf{B}_{s_p}\rangle + \|\mathbf{S} - \mathbf{\Sigma}^*\|_{2,\infty} \cdot \|\hat{\mathbf{\Sigma}} - \mathbf{B}\|_{2,1}.$$

□

We are now prepared to prove Proposition A.1. For simplicity, let $S = S(\mathbf{B})$ and $L = L(\mathbf{B})$. The focus of this proof is on the term $\left\langle \sum_{\ell=1}^{p-1} \mathbf{W}^{(\ell)} * \mathbf{A}^{(\ell)}, \hat{\mathbf{\Sigma}} - \mathbf{B} \right\rangle$ in Proposition A.2. For $1 \le \ell \le L$, the constraints on $\mathbf{A}^{(\ell)}$ are $\|\mathbf{A}^{(\ell)}_{g_\ell^c}\|_2 = 0$ and $\|\mathbf{A}^{(\ell)}_{g_\ell}\|_2 \le 1$ (the third constraint holds automatically since $\mathbf{B}_{g_\ell} = 0$ for $\ell \le L$). We let $\mathbf{A}^{(\ell)}_{s_m} = w_{\ell m}\hat{\mathbf{\Sigma}}_{s_m}/f_\ell(\hat{\mathbf{\Sigma}})$ if $f_\ell(\hat{\mathbf{\Sigma}}) \neq 0$ and 0 otherwise, for $1 \le m \le \ell, 1 \le \ell \le L$. Then for $\ell \le L$,

$$\begin{aligned}\left\langle \mathbf{W}^{(\ell)} * \mathbf{A}^{(\ell)}, \hat{\mathbf{\Sigma}} - \mathbf{B} \right\rangle &= \sum_{m=1}^{\ell} \langle w_{\ell m}\mathbf{A}^{(\ell)}_{s_m}, \hat{\mathbf{\Sigma}}_{s_m} - \mathbf{B}_{s_m}\rangle = \sum_{m=1}^{\ell} \langle w_{\ell m}^2 \hat{\mathbf{\Sigma}}_{s_m}/f_\ell(\hat{\mathbf{\Sigma}}), \hat{\mathbf{\Sigma}}_{s_m}\rangle \\ &= \sum_{m=1}^{\ell} w_{\ell m}^2 \|\hat{\mathbf{\Sigma}}_{s_m}\|_2^2/f_\ell(\hat{\mathbf{\Sigma}}) = f_\ell(\hat{\mathbf{\Sigma}}).\end{aligned}$$

It follows that $\left\langle \sum_{\ell=1}^{L} \mathbf{W}^{(\ell)} * \mathbf{A}^{(\ell)}, \hat{\mathbf{\Sigma}} - \mathbf{B} \right\rangle = \sum_{\ell=1}^{L} f_\ell(\hat{\mathbf{\Sigma}}) \ge \|\hat{\mathbf{\Sigma}}_{S^c}\|_{2,1}$, by Lemma A.4 (iii).

Next, fix $\ell \ge L+1$. By the definition of subgradient in (A.13.1), $\mathbf{A}^{(\ell)}_{g_L}$ can be chosen to have arbitrary values (as long as $\|\mathbf{A}^{(\ell)}_{g_\ell}\|_2 \le 1$), and we take $\mathbf{A}^{(\ell)}_{g_L} = 0$ because of the equality $\left\langle (\mathbf{W}^{(\ell)} * \mathbf{B})_{g_\ell}, \mathbf{A}^{(\ell)}_{g_\ell} \right\rangle = \|(\mathbf{W}^{(\ell)} * \mathbf{B})_{g_\ell}\|_2 \cdot \|\mathbf{A}^{(\ell)}_{g_\ell}\|_2$. Then

$$\begin{aligned}-\left\langle \mathbf{W}^{(\ell)} * \mathbf{A}^{(\ell)}, \hat{\mathbf{\Sigma}} - \mathbf{B} \right\rangle &= -\sum_{m=L+1}^{\ell} \langle w_{\ell m}\mathbf{A}^{(\ell)}_{s_m}, \hat{\mathbf{\Sigma}}_{s_m} - \mathbf{B}_{s_m}\rangle \\ &\le \sum_{m=L+1}^{\ell} w_{\ell m}\|\mathbf{A}^{(\ell)}_{s_m}\|_2 \cdot \|\hat{\mathbf{\Sigma}}_{s_m} - \mathbf{B}_{s_m}\|_2 \\ &\le \sqrt{\sum_{m=L+1}^{\ell} w_{\ell m}^2 \|\hat{\mathbf{\Sigma}}_{s_m} - \mathbf{B}_{s_m}\|_2^2} \cdot \sqrt{\sum_{m=L+1}^{\ell} \|\mathbf{A}^{(\ell)}_{s_m}\|_2^2} \\ &\le \sqrt{\sum_{m=L+1}^{\ell} w_{\ell m}^2 \|\hat{\mathbf{\Sigma}}_{s_m} - \mathbf{B}_{s_m}\|_2^2}.\end{aligned}$$

In the above we used the fact that $\mathbf{A}^{(\ell)}_{g_\ell^c} = 0$ and $\|\mathbf{A}^{(\ell)}_{g_\ell}\|_2 \le 1$. It follows that

$$\begin{aligned}-\left\langle \sum_{\ell=L+1}^{p-1} \mathbf{W}^{(\ell)} * \mathbf{A}^{(\ell)}, \hat{\mathbf{\Sigma}} - \mathbf{B} \right\rangle &\le \sum_{\ell=L+1}^{p-1} \sqrt{\sum_{m=L+1}^{\ell} w_{\ell m}^2 \|\hat{\mathbf{\Sigma}}_{s_m} - \mathbf{B}_{s_m}\|_2^2} \\ &= \|\hat{\mathbf{\Sigma}}_S - \mathbf{B}_S\|_{2,1}^*.\end{aligned}$$

Therefore

$$-\left\langle \sum_{\ell=1}^{p-1} \mathbf{W}^{(\ell)} * \mathbf{A}^{(\ell)}, \hat{\mathbf{\Sigma}} - \mathbf{B} \right\rangle \le -\|\hat{\mathbf{\Sigma}}_{S^c}\|_{2,1} + \|\hat{\mathbf{\Sigma}}_S - \mathbf{B}_S\|_{2,1}^*,$$

and, by Lemma A.4 (i),

$$\begin{aligned}
&\|\hat{\mathbf{\Sigma}} - \mathbf{B}\|_{2,1} - \left\langle \sum_{\ell=1}^{p-1} \mathbf{W}^{(\ell)} * \mathbf{A}^{(\ell)}, \hat{\mathbf{\Sigma}} - \mathbf{B} \right\rangle \\
\leq &\|\hat{\mathbf{\Sigma}}_S - \mathbf{B}_S\|_{2,1} + \|\hat{\mathbf{\Sigma}}_{S^c} - \mathbf{B}_{S^c}\|_{2,1} - \|\hat{\mathbf{\Sigma}}_{S^c}\|_{2,1} + \|\hat{\mathbf{\Sigma}}_S - \mathbf{B}_S\|_{2,1}^* \\
\leq &2\|\hat{\mathbf{\Sigma}}_S - \mathbf{B}_S\|_{2,1}^*.
\end{aligned}$$

Here we have used that $\mathbf{B}_{S^c} = 0$. The proposition follows by noting that $\mathbf{B}$ has all zero in S^c.

References

Boucheron, S., Lugosi, G. & Massart, P. (2013), *Concentration inequalities: a nonasymptotic theory of independence*, Oxford University Press, Oxford.

Jenatton, R., Audibert, J. & Bach, F. (2011), 'Structured variable selection with sparsity-inducing norms', *The Journal of Machine Learning Research* **12**, 2777–2824.

Tsybakov, A. (2009), *Introduction to nonparametric estimation*, Springer, New York. Revised and extended from the 2004 French original, Translated by Valdimir Zaiats.

Vershynin, R. (2011), Introduction to the non-asymptotic analysis of random matrices. Available from the link `http://arxiv.org/abs/1201.0708v3`.