# SUPPLEMENT TO "OPTIMALITY OF SPECTRAL CLUSTERING IN THE GAUSSIAN MIXTURE MODEL"

BY Matthias Löffler, Anderson Y. Zhang, and Harrison H. Zhou

ETH Zürich, University of Pennsylvania, and Yale University

## APPENDIX A: CHARACTERISTICS OF THE POPULATION QUANTITIES

In this section, we present several propositions that characterize the population quantities defined in Section 4.1. We first define two matrices related to $z^*$. Let $D \in \mathbb{R}^{k\times k}$ be a diagonal matrix with

$$D_{j,j} = \left|\{i \in [n] : z_i^* = j\}\right|, \; j \in [k],$$

and let $Z^* \in \{0,1\}^{n\times k}$ be a matrix such that

$$Z^*_{i,j} = \mathbb{I}\{z_i^* = j\}, \; i \in [n], j \in [k]. \tag{40}$$

PROPOSITION A.1. *There exists an orthogonal matrix $W \in \mathbb{R}^{k\times k}$ such that*

$$V = Z^* D^{-\frac{1}{2}} W.$$

*Consequently, $V_{i,\cdot} = V_{j,\cdot}$ for all $i, j \in [n]$ such that $z_i^* = z_j^*$ In addition, we have that*

$$\sigma_1 \geq \sqrt{\frac{\beta n}{k}} \frac{\Delta}{2}.$$

PROOF. First note that

$$P = (\theta_1^*, \ldots, \theta_k^*)\, Z^{*T} = (\theta_1^*, \ldots, \theta_k^*)\, D^{\frac{1}{2}} D^{-\frac{1}{2}} Z^{*T} = (\theta_1^*, \ldots, \theta_k^*)\, D^{\frac{1}{2}} \left(Z^* D^{-\frac{1}{2}}\right)^T,$$

and observe that $Z^* D^{-\frac{1}{2}}$ has orthonormal columns. Now, we decompose $(\theta_1^*, \ldots, \theta_k^*)\, D^{\frac{1}{2}} = U\Lambda W^T$ into its SVD. Here $W$ is some orthonomal matrix $W \in \mathbb{R}^{k\times k}$. Then we have that

$$P = U\Lambda \left(Z^* D^{-\frac{1}{2}} W\right)^T,$$

with $Z^* D^{-\frac{1}{2}} W$ having orthonormal columns. Hence, we have that $\Sigma = \Lambda$ and $V = Z^* D^{-\frac{1}{2}} W$. The structure of $Z^*$ leads to the second statement presented in the proposition. Indeed, due to (2), the largest singular value of $(\theta_1^*, \ldots, \theta_k^*)$ must be greater than $\Delta/2$. Since $(\theta_1^*, \ldots, \theta_k^*)\, D^{\frac{1}{2}} = U\Sigma W^T$, we obtain that

$$\sigma_1 \geq \sqrt{\frac{\beta n}{k}} \frac{\Delta}{2}.$$

□

PROPOSITION A.2. *The matrix $V$ satisfies*

$$\max_{i\in[n]} \|V^T e_i\| \leq \sqrt{\frac{k}{\beta n}}.$$

PROOF. By Proposition A.1 we have that

$$\|V^T e_i\| = \|W^T D^{-1/2}(Z^*)^T e_i\| = \|D^{-1/2}(Z^*)^T e_i\|,$$

where we used that $W$ is orthogonal. Hence, we obtain that

$$\max_{i\in[n]} \|V^T e_i\| \leq \frac{1}{\min_{j\in[k]} D_{j,j}^{\frac{1}{2}}} \|(Z^*)^T e_i\| = \sqrt{\frac{k}{\beta n}}.$$

□

PROPOSITION A.3. *We have that*

$$\left|\left\langle u_l, \theta_j^* \right\rangle\right| \leq \sigma_l \sqrt{\frac{k}{\beta n}}, \ \forall j, l \in [k].$$

PROOF. Since $P = U\Sigma V^T$ and $P_{\cdot,i} = \theta^*_{z_i^*}$, $i \in [n]$, we have for any $u, l \in [k]$ that

$$\left\langle u_l, \theta_j^* \right\rangle = \sigma_l V_{i,l}, \text{ where } i \in [n] \text{ is any index such that } z_i^* = j.$$

The proof is completed by applying Proposition A.2. □

## APPENDIX B: AUXILIARY LEMMAS RELATED TO THE NOISE MATRIX $E$

In this section, we present three basic lemmas for the control of the noise term $E$ and empirical singular values and vectors, used in the proof of Theorem 2.1.

LEMMA B.1. *For a random matrix $E \in \mathbb{R}^{p\times n}$ with $\{E_{i,j}\} \overset{iid}{\sim} \mathcal{N}(0,1)$, define the event $\mathcal{F} = \{\|E\| \leq \sqrt{2}(\sqrt{n} + \sqrt{p})\}$. We have that*

$$\mathbb{P}\left(\|E\| \geq \sqrt{n} + \sqrt{p} + t\right) \leq e^{-t^2/2}$$

*and particularly*

$$\mathbb{P}(\mathcal{F}) \geq 1 - e^{-0.08n}.$$

PROOF. By Theorem 2.13 in [2] we have that $\mathbb{E}\|E\| \leq \sqrt{n} + \sqrt{p}$. Moreover, as $\|E\| = \sup_{\|u\|=\|v\|=1} \langle u, Ev \rangle$, we have by Borell's inequality (e.g. Theorem 2.2.7 in [3]) that $\mathbb{P}(\|E\| \geq \mathbb{E}\|E\| + t) \leq e^{-t^2/2}$. □

Weyl's inequality (e.g. Theorem 4.3.1 of [4]), the fact that $X = P + E$ and Lemma B.1 imply the following lemma.

LEMMA B.2. *Assume that the random event $\mathcal{F}$ holds. We have that*

$$\hat{\sigma}_j \leq \sigma_j + \sqrt{2}(\sqrt{n} + \sqrt{p}), \ \ \forall j \in [k].$$

The last lemma included in this section is the Davis-Kahan-Wedin $sin(\Theta)$ Theorem, which characterizes the distance between empirical and population singular vector spaces. We refer readers to Theorem 21 of [8] for its proof.

LEMMA B.3 (Davis-Kahan-Wedin $sin(\Theta)$ Theorem). *Consider any rank-$s$ matrices $W, \hat{W}$. Let $W = \sum_{i=1}^{s} \sigma_i u_i v_i^T$ be its SVD with $\sigma_1 \geq \ldots \geq \sigma_s$. Similarly, let $\hat{W} = \sum_{i=1}^{s} \hat{\sigma}_i \hat{u}_i \hat{v}_i^T$ be its SVD with $\hat{\sigma}_1 \geq \ldots \geq \hat{\sigma}_s$. For any $1 \leq j \leq l \leq s$, define $V = (v_j, \ldots, v_l)$ and $\hat{V} = (\hat{v}_j, \ldots, \hat{v}_l)$. Then, we have that*

$$\inf_{O:\text{ orthogonal matrix}} \left\| \hat{V} - VO \right\| \leq \sqrt{2} \left\| \hat{V}\hat{V}^T - VV^T \right\| \leq \frac{4\sqrt{2} \left\| \hat{W} - W \right\|}{\min\{\sigma_{j-1} - \sigma_j, \sigma_l - \sigma_{l+1}\}},$$

*where we denote $\sigma_0 = +\infty$ and $\sigma_{s+1} = 0$.*

## APPENDIX C: PROOFS OF KEY LEMMAS

In this section, we provide proofs of the lemmas stated in Section 4, except for the proof of of Lemma 4.3, which is deferred to Appendix F. Throughout this section, for any matrix $W$, we denote by span$(W)$ the space spanned by the columns of $W$.

PROOF OF LEMMA 4.1. Since $\hat{P}_{\cdot,i} = \hat{U}\hat{Y}_{\cdot,i} = \left(\hat{U}\hat{U}^T\right)\hat{U}\hat{Y}_{\cdot,i}$ lies in the column space span$(\hat{U})$ any $\{\theta_j\}_{j=1}^k$ that achieves the minimum of (11) must also lie in span$(\hat{U})$. In particular, we have that

$$\begin{aligned} \min_{z\in[k]^n, \{\theta_j\}_{j=1}^k \in \mathbb{R}^k} \sum_{i\in[n]} \left\| \hat{P}_{\cdot,i} - \theta_{z_i} \right\|^2 &= \min_{z\in[k]^n, \{c_j\}_{j=1}^k \in \mathbb{R}^k} \sum_{i\in[n]} \left\| \hat{U}\hat{Y}_{\cdot,i} - \hat{U}c_{z_i} \right\|^2 \\ &= \min_{z\in[k]^n, \{c_j\}_{j=1}^k \in \mathbb{R}^k} \sum_{i\in[n]} \left\| \hat{Y}_{\cdot,i} - c_{z_i} \right\|^2, \end{aligned}$$

where the last equation is due to the fact that $\hat{U}$ is an orthogonal matrix. □

PROOF OF LEMMA 4.2. Due to the fact that $\hat{P}$ is the best rank-$k$ approximation of $X$ in spectral norm and $P$ is also rank-$k$, we have that

$$\left\| \hat{P} - X \right\| \leq \|P - X\| = \|E\|.$$

This, the fact that both $\hat{P}$ and $P$ are at most rank $k$ and the fact that we work on the event $\mathcal{F}$ imply that,

$$\begin{aligned} \left\| \hat{P} - P \right\|_{\mathrm{F}} &\leq 2\sqrt{2k}\|P - X\| = 2\sqrt{2k}\|E\| \\ &\leq 4\sqrt{k}(\sqrt{n} + \sqrt{p}), \end{aligned} \tag{41}$$

where the last inequality is due to Lemma B.1. Now, denote by $\hat{\Theta}$ the center matrix after solving (11). That is, the $i$th column of $\hat{\Theta}$ is $\hat{\theta}_{\hat{z}'_i}$. Since $\hat{\Theta}$ is the solution to the $k$-means objective, we have that

$$\left\| \hat{\Theta} - \hat{P} \right\|_{\mathrm{F}} \leq \left\| \hat{P} - P \right\|_{\mathrm{F}}.$$

Hence, by the triangle inequality, we obtain that

$$\left\| \hat{\Theta} - P \right\|_{\mathrm{F}} \leq 2\left\| \hat{P} - P \right\|_{\mathrm{F}} \leq 8\sqrt{k}(\sqrt{n} + \sqrt{p}).$$

Now, define the set $S$ as

$$S = \left\{ i \in [n] : \left\| \hat{\theta}_{\hat{z}'_i} - \theta^*_{z^*_i} \right\| > \frac{\Delta}{2} \right\}.$$

Since $\left\{\hat{\theta}_{\hat{z}_i'} - \theta^*_{z_i^*}\right\}_{i\in[n]}$ are exactly the columns of $\hat{\Theta} - P$, we have that

$$|S| \leq \frac{\left\|\hat{\Theta} - P\right\|_{\mathrm{F}}^2}{(\Delta/2)^2} \leq \frac{256k\,(n+p)}{\Delta^2}.$$

Assuming that

$$\frac{\beta\Delta^2}{k^2\left(1+\frac{p}{n}\right)} \geq 512,$$

we have that

$$|S| \leq \frac{\beta n}{2k}.$$

We now show that all the data points in $S^C$ are correctly clustered. We define

$$C_j = \left\{i \in [n] : z_i^* = j, i \in S^C\right\},\ j \in [k].$$

The following holds:

- For each $j \in [k]$, $C_j$ cannot be empty, as $|C_j| \geq |\{i : z_i^* = j\}| - |S| > 0$.
- For each pair $j, l \in [k], j \neq l$, there cannot exist some $i \in C_j, i' \in C_l$ such that $\hat{z}_i' = \hat{z}_{i'}'$. Otherwise $\hat{\theta}_{\hat{z}_i'} = \hat{\theta}_{\hat{z}_{i'}'}$ which would imply

$$\begin{aligned}\left\|\theta_j^* - \theta_l^*\right\| &= \left\|\theta^*_{z_i^*} - \theta^*_{z_{i'}^*}\right\| \\ &\leq \left\|\theta^*_{z_i^*} - \hat{\theta}_{\hat{z}_i'}\right\| + \left\|\hat{\theta}_{\hat{z}_i'} - \hat{\theta}_{\hat{z}_{i'}'}\right\| + \left\|\hat{\theta}_{\hat{z}_{i'}'} - \theta^*_{z_{i'}^*}\right\| < \Delta,\end{aligned}$$

  contradicting (2).

Since $\hat{z}_i'$ can only take values in $[k]$, we conclude that the sets $\{\hat{z}_i' : i \in C_j\}$ are disjoint for all $j \in [k]$. That is, there exists a permutation $\phi \in \Phi$, such that

$$\hat{z}_i' = \phi(j),\ i \in C_j,\ j \in [k].$$

This implies that $\sum_{i\in S^C} \mathbb{I}\{\hat{z}_i \neq \phi(z_i^*)\} = 0$. Hence, we obtain that

$$\ell(\hat{z}, z^*) \leq |S| \leq \frac{256k\,(n+p)}{\Delta^2}.$$

When the ratio $\Delta^2/\left(k^2\,(n+p)\right)$ is large enough, an immediate implication is that $\min_{j\in[k]} |\{i \in [n] : \hat{z}_i = j\}| \geq \frac{\beta n}{k} - |S| \geq \frac{\beta n}{2k}$. Moreover, in this case we obtain that

$$\max_j \left\|\hat{\theta}_j - \theta^*_{\phi(j)}\right\|^2 \leq \frac{\left\|\hat{\Theta} - P\right\|_{\mathrm{F}}^2}{\frac{\beta n}{k} - |S|} \leq \frac{128k^2\,(n+p)}{\beta n}$$

□

PROOF OF LEMMA 4.4. Recall that $M$ has SVD $M = U\Sigma V^T$ where $U = (u_1, \ldots, u_k)$, $V = (v_1, \ldots, v_k)$ and $\Sigma = \mathrm{diag}\{\sigma_1, \ldots, \sigma_k\} \in \mathbb{R}^{k\times k}$ with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k \geq 0$. We denote $\mathbb{S} = \left\{x \in \mathrm{span}\left(I - VV^T\right) : \|x\| = 1\right\}$ to be the unit sphere in span $\left(I - VV^T\right)$. We also denote $\mathcal{O}$ to be the set of all orthonormal matrices in $\mathbb{R}^{n\times n}$ and furthermore

$$\mathcal{O}' = \{O \in \mathcal{O} : OV = V\}.$$

Let $V_\perp$ be an orthogonal extension of $V$ such that $(V, V_\perp) \in \mathcal{O}$. Then for any $O \in \mathcal{O}'$, due to the fact that $O(V, V_\perp) \in \mathcal{O}$ and $O(V, V_\perp) = (V, OV_\perp)$, we have that $OV_\perp$ is another orthogonal extension of $V$. This implies that

$$Ox \in \operatorname{span}\left(I - VV^T\right) \quad \forall x \in \operatorname{span}\left(I - VV^T\right). \tag{42}$$

Hence $\mathcal{O}'$ includes all rotation matrices in $\operatorname{span}\left(I - VV^T\right)$. In the following, we prove the three assertions of Lemma 4.4 one by one.

*Assertion (1).* Recall that $\hat{M} = M + E = U\Sigma V^T + E$ and $\hat{M} = \sum_{j=1}^{p\wedge n} \hat{\sigma}_j \hat{u}_j \hat{v}_j^T$ and denote by $\overset{d}{=}$ equality in distribution. For any $O \in \mathcal{O}'$, since $EO^T \overset{d}{=} E$, we have that $\hat{M}O^T = \left(U\Sigma V^T + E\right) O^T = U\Sigma V^T + EO^T \overset{d}{=} \hat{M}$. On the other hand, $\hat{M}O^T$ has SVD

$$\hat{M}O^T = \sum_{j=1}^{p\wedge n} \hat{\sigma}_j \hat{u}_j \left(O\hat{v}_j\right)^T.$$

Hence, for any $j \in [k]$, we have that $\hat{v}_j \overset{d}{=} O\hat{v}_j$.

For any $x \in \mathbb{R}^n$, we define the mapping $f : \mathbb{R}^n \to \mathbb{S}$ as $f(x) = (I - VV^T)x/\|(I - VV^T)x\|$. Applying $f$ on both $\hat{v}_j$ and $O\hat{v}_j$, we obtain that

$$\frac{(I - VV^T)O\hat{v}_j}{\|(I - VV^T)O\hat{v}_j\|} \overset{d}{=} \frac{(I - VV^T)\hat{v}_j}{\|(I - VV^T)\hat{v}_j\|}.$$

Since $\hat{v}_j = VV^T\hat{v}_j + (I - VV^T)\hat{v}_j$, we have $O\hat{v}_j = VV^T\hat{v}_j + O(I - VV^T)\hat{v}_j$. By (42), we have that $O(I - VV^T)\hat{v}_j \in \operatorname{span}\left(I - VV^T\right)$. Hence, we obtain that

$$VV^T O\hat{v}_j = VV^T\hat{v}_j \tag{43}$$

$$(I - VV^T)O\hat{v}_j = (I - VV^T)O(I - VV^T)\hat{v}_j = O(I - VV^T)\hat{v}_j. \tag{44}$$

As a consequence of (44), we obtain that

$$O\frac{(I - VV^T)\hat{v}_j}{\|(I - VV^T)\hat{v}_j\|} = \frac{O(I - VV^T)\hat{v}_j}{\|O(I - VV^T)\hat{v}_j\|} \overset{d}{=} \frac{(I - VV^T)\hat{v}_j}{\|(I - VV^T)\hat{v}_j\|} \quad \forall\, O \in \mathcal{O}'. \tag{45}$$

In particular, $(I - VV^T)\hat{v}_j/\|(I - VV^T)\hat{v}_j\|$ is contained in $\mathbb{S}$ and is rotation-invariant. Hence, we obtain that $(I - VV^T)\hat{v}_j/\|(I - VV^T)\hat{v}_j\|$ is uniformly distributed on $\mathbb{S}$.

*Assertion (2).* For any $x \in \mathbb{R}^n$, we define another mapping $g : \mathbb{R}^n \to \mathbb{R}^n$ as $g(x) = ((VV^Tx)^T, ((I - VV^T)x)^T/\|(I - VV^T)x\|)^T$. Recall that $\hat{v}_j \overset{d}{=} O\hat{v}_j\ \forall O \in \mathcal{O}'$. Applying $g$ on both $\hat{v}_j$ and $O\hat{v}_j$ and using (43), (44) and (45), we obtain that

$$\begin{pmatrix} VV^T\hat{v}_j \\ \frac{(I - VV^T)\hat{v}_j}{\|(I - VV^T)\hat{v}_j\|} \end{pmatrix} \overset{d}{=} \begin{pmatrix} VV^T\hat{v}_j \\ O\frac{(I - VV^T)\hat{v}_j}{\|(I - VV^T)\hat{v}_j\|} \end{pmatrix}. \tag{46}$$

Let $\mathcal{A}$ be a Borel subset of $\operatorname{span}(VV^T)$ and $\mathcal{B}$ a Borel subset of $\mathbb{S}$. By (46) we have for any $O \in \mathcal{O}'$ that

$$\mathbb{P}\left(\frac{(I - VV^T)\hat{v}_j}{\|(I - VV^T)\hat{v}_j\|} \in \mathcal{B} \middle| VV^T\hat{v}_j \in \mathcal{A}\right) = \mathbb{P}\left(O\frac{(I - VV^T)\hat{v}_j}{\|(I - VV^T)\hat{v}_j\|} \in \mathcal{B} \middle| VV^T\hat{v}_j \in \mathcal{A}\right).$$

Hence, we obtain that $\frac{(I - VV^T)\hat{v}_j}{\|(I - VV^T)\hat{v}_j\|}\big|VV^T\hat{v}_j$ is also uniformly distributed on $\mathbb{S}$, invariant to the value of $VV^T\hat{v}_j$. This implies that $\frac{(I - VV^T)\hat{v}_j}{\|(I - VV^T)\hat{v}_j\|}$ is independent of $VV^T\hat{v}_j$.

*Assertion (3).* Since $\|(I - VV^T)\hat{v}_j\| = \sqrt{1 - \|VV^T\hat{v}_j\|^2}$ is a function of only $VV^T\hat{v}_j$, this is an immediate consequence of the second assertion.

□

## APPENDIX D: EXTENSION OF PROPOSITION 2.1

In this appendix, we provide an extension of Proposition 2.1.

PROPOSITION D.1. *Assume the observations $\{X_i\}_{i\in[n]}$ are generated as follows:*

$$X_i = \theta^*_{z^*_i} + \epsilon_i.$$

*Denote $E := (\epsilon_1, \ldots, \epsilon_n)$. Assume that $\Delta/(\beta^{-0.5}kn^{-0.5}\|E\|) \geq C$ for some large enough constant $C > 0$. Then the output of Algorithm 1, $\hat{z}$, satisfies for another constant $C' > 0$*

$$\ell(\hat{z}, z^*) \leq \frac{C'k\|E\|^2}{n\Delta^2}. \tag{47}$$

*In particular, if $\{\epsilon_i\}_{i=1}^n \overset{iid}{\sim} \mathcal{N}(0,\Sigma)$ and assuming that $\Delta/(\beta^{-0.5}k(\|\Sigma\| + (trace(\Sigma) + \|\Sigma\|\log(p+n))/n)^{0.5}) \geq C$, we have for another constant $C'' > 0$ with probability at least $1 - \exp(-0.08n)$ that*

$$\ell(\hat{z}, z^*) \leq \frac{C''k\left(\|\Sigma\| + (trace(\Sigma) + \|\Sigma\|\log(p+n))/n\right)^2}{\Delta^2}. \tag{48}$$

*Moreover, if $\{\epsilon_i\}_{i=1}^n \overset{iid}{\sim} subG(\sigma^2)$ (i.e., sub-Gaussian with variance proxy $\sigma^2$) and assuming that $\Delta/(\beta^{-0.5}kn^{0.5}\sigma(1+p/n)^{-0.5}) \geq C$, we have with probability at least $1-\exp(-0.08n)$ that*

$$\ell(\hat{z}, z^*) \leq \frac{C''k\sigma^2\left(1+\frac{p}{n}\right)}{\Delta^2}. \tag{49}$$

PROOF. Following the proof of Proposition 2.1 line by line (47) immediately follows. To obtain (48) and (49), we provide upper bounds for $\|E\|$.

When the errors $\{\epsilon_i\}_{i=1}^n$ are independent Gaussians with covariance matrix $\Sigma$, we bound $\|E\|$ by applying Corollary 3.11 in [1]. More precisely, assume that $\Sigma$ has eigendecomposition $\Sigma = \Gamma\Lambda\Gamma^T$ with $\Lambda$ being a diagonal matrix and $\Gamma$ an orthogonal matrix. Denote by $\tilde{E}$ a $p \times n$ matrix with i.i.d. standard Gaussian entries. Then, by rotation invariance of isotropic Gaussian random variables, we have that $E \overset{d}{=} \Gamma\Lambda^{1/2}\Gamma^T\tilde{E} \overset{d}{=} \Gamma\Lambda^{1/2}\tilde{E}$. Hence, $\|E\| \overset{d}{=} \|\Gamma\Lambda^{1/2}\tilde{E}\| \leq \|\Lambda^{1/2}\tilde{E}\|$. The entries of $\Lambda^{1/2}\tilde{E}$ are independent and hence we can now apply Corollary 3.11 in [1] and (48) follows.

When the errors are sub-Gaussian distributed, we bound $\|E\|$ by a net argument, see for instance Theorem 5.39 in [9]. □

## APPENDIX E: PROOF OF THEOREM 2.2

To prove Theorem 2.2, we first note that

$$\begin{aligned}\sum_{i\in[n]}\|\hat{Y}_{\cdot,i} - \tilde{c}_{\tilde{z}_i}\|^2 &\leq \sum_{i\in[n]}\|\hat{Y}_{\cdot,i} - \check{c}_{\check{z}_i}\|^2 \\ &\leq (1+\varepsilon)\inf_{\{c_j\}_{j=1}^k\in\mathbb{R}^k}\sum_{i\in[n]}\min_{j\in[k]}\|\hat{Y}_{\cdot,i} - c_j\|^2,\end{aligned}$$

as each iteration of Lloyd's algorithm is guaranteed to not increase the value of the objective function. By the same analysis as for $\hat{z}$ in Proposition 2.1 , $\tilde{z}$ and $\check{z}$ satisfy (13) with an additional factor of $(1+\varepsilon)$ on the right hand side of the inequality and the centres $\{\tilde{\theta}_j\}_{j=1}^k = \{\hat{U}\tilde{c}_j\}_{j=1}^k$ satisfy (14) with an additional factor of $\sqrt{1+\varepsilon}$ on the right hand side of the inequality. Then the exponential bound (10) follows similarly as the proof of Theorem 2.1 and we line out the necessary modifications below.

In particular, the local optimality guarantee $\|\hat{Y}_i - \tilde{c}_{\tilde{z}_i}\| \le \|\hat{Y}_i - \tilde{c}_j\|, \forall i \in [n], j \neq \tilde{z}_i$ ensures that the equality in (17) holds. Moreover, by definition of the centres $\tilde{\theta}_j$ we have, similarly as in (21), that

$$\tilde{\theta}_j = \frac{\sum_{\check{z}_i = j} \hat{P}_{\cdot,i}}{\sum_{\check{z}_i = j} 1} = \frac{\hat{\sigma}_l}{\sqrt{|\{i \in [n] : \check{z}_i = j\}|}}$$

and that

$$|\langle \hat{u}_l, \tilde{\theta}_j \rangle| \le \frac{\hat{\sigma}_l}{\sqrt{|\{i \in [n] : \check{z}_i = j\}|}}.$$

Finally, since $\check{z}$ fulfills (13) with an additional factor of $(1+\varepsilon)$ on the right hand side, we further have that $|\{i \in [n] : \check{z}_i = j\}| \ge \frac{\beta n}{2k}$ and thus we obtain, as in (23), that

$$\max_{j\in[k]} \max_{r+1\le l\le k} \left|\left\langle \hat{u}_l, \tilde{\theta}_j \right\rangle\right| \mathbb{I}(\mathcal{F}) \le (k\rho + 4)\sqrt{\frac{2k}{\beta}\left(1+\frac{p}{n}\right)}.$$

With these modifications, the rest of the proof is the same as the proof of Theorem 2.1.

## APPENDIX F: SPECTRAL PROJECTION MATRIX PERTURBATION THEORY

In this section, we give the proof of Lemma 4.3. Before that, we first introduce two lemmas used in the proof of Lemma 4.3.

The following lemma gives an upper bound on the operator norm of $\|S_{a:b}\|$. The setting considered here is slightly more general than that in Lemma 4.3, as $E$ is not necessarily a Gaussian noise matrix. The proof of Lemma F.1 mainly follows that of Lemma 2 in [6]. It is included in the later part of this section for completeness.

LEMMA F.1. *Consider any rank-$k$ matrix $M \in \mathbb{R}^{p\times n}$ with SVD $M = \sum_{j=1}^k \sigma_j u_j v_j^T$ where $\sigma_1 \ge \sigma_2 \ldots \ge \sigma_k > 0$. Define $\sigma_0 = \sigma_{k+1} = 0$.*

*Consider any matrix $E \in \mathbb{R}^{p\times n}$. Define $\hat{M} = M + E$. Let the SVD of $\hat{M}$ be $\sum_{j=1}^{p\wedge n} \hat{\sigma}_j \hat{u}_j \hat{v}_j^T$ where $\hat{\sigma}_1 \ge \hat{\sigma}_2 \ge \ldots \ge \hat{\sigma}_{p\wedge n}$.*

*For any two indexes $a, b$ such that $1 \le a \le b \le k$, define $V_{a:b} = (v_a, \ldots, v_b)$, $\hat{V}_{a:b} = (\hat{v}_a, \ldots, \hat{v}_b)$ and $V := (v_1, \ldots, v_k)$. Define the singular value gap $g_{a:b} = \min\{\sigma_{a-1} - \sigma_a, \sigma_b - \sigma_{b+1}\}$. Define*

$$S_{a:b} = \left(I - VV^T\right)\left(\hat{V}_{a:b}\hat{V}_{a:b}^T - V_{a:b}V_{a:b}^T\right)V_{a:b} - \sum_{a\le j\le b} \frac{1}{\sigma_j}\left(I - VV^T\right)E^T u_j v_j^T V_{a:b}. \tag{50}$$

*Then, we have that*

$$\|S_{a:b}\| \le \left(\frac{32(\sigma_a - \sigma_b)}{\pi g_{a:b}} + 16\right)\frac{\|E\|^2}{g_{a:b}^2}.$$

$S_{a:b}$ in Lemma 4.3 and Lemma F.1 depends on $E$. It can be written as $S_{a:b}(E)$ with $S_{a:b}(\cdot)$ treated as a function of the noise matrix. Lemma F.2 studies the Lipschitz continuity of $S_{a:b}(\cdot)$. It slightly generalizes Lemma 2.4 in [7] and its proof follows along the same arguments. Its proof will be given in the later part of this section for completeness.

LEMMA F.2. *Consider the same setting as in Lemma F.1. Define $S_{a:b}(E)$ as in (50). Consider another matrix $E' \in \mathbb{R}^{p\times n}$ and define $\hat{M}' := M + E'$. Define $S_{a:b}(E')$ analogously. Assuming that* $\max\{\|E\|, \|E'\|\} \le g_{a:b}/4$, *we have that*

$$\|S_{a:b}(E) - S_{a:b}(E')\| \le 1024\left(1 + \frac{\sigma_a - \sigma_b}{g_{a:b}}\right)\frac{\max\{\|E\|, \|E'\|\}}{g_{a:b}^2}\|E - E'\|. \tag{51}$$

Applying Lemma F.1 and Lemma F.2, we are able to prove Lemma 4.3. It generalizes Theorem 1.1 in [7], and its proof follows the same argument.

PROOF OF LEMMA 4.3. Define $\phi$ as follows

$$\phi(s) = \begin{cases} 1, & s \le 1 \\ 3 - 2s, & 1 < s < 3/2 \\ 0, & s \ge 3/2 \end{cases}$$

and note that $\phi$ is Lipschitz with Lipschitz constant 2. As we mention earlier in this section, we write $S_{a:b}(E)$ and treat $S_{a:b}(\cdot)$ as a matrix valued function.

*Step 1.* Define a function

$$h_\delta(E) = \langle S_{a:b}(E), W\rangle \phi\left(\frac{6\|E\|}{\delta}\right).$$

We are going to show that $h_\delta$ is also Lipschitz for any $\delta \le g_{a:b}/4$. We use the notation $\|\cdot\|_*$ for the nuclear norm of a matrix.

- First suppose that $\max\{\|E\|, \|E'\|\} \le \delta$. Then, by Lemma F.1, Lemma F.2 and the fact that $\phi$ is Lipschitz, we obtain that

$$\begin{aligned}
&|h_\delta(E) - h_\delta(E')| \\
&\le \left|\langle S_{a:b}(E) - S_{a:b}(E'), W\rangle\right| \phi\left(\frac{6\|E\|}{\delta}\right) + \left|\langle S_{a:b}(E'), W\rangle\right| \left|\phi\left(\frac{6\|E\|}{\delta}\right) - \phi\left(\frac{6\|E'\|}{\delta}\right)\right| \\
&\le \|S_{a:b}(E) - S_{a:b}(E')\|\|W\|_*\phi\left(\frac{6\|E\|}{\delta}\right) + \|S_{a:b}(E')\|\|W\|_* \left|\phi\left(\frac{6\|E\|}{\delta}\right) - \phi\left(\frac{6\|E'\|}{\delta}\right)\right| \\
&\le 1024\left(1 + \frac{\sigma_a - \sigma_b}{g_{a:b}}\right)\frac{\max\{\|E\|, \|E'\|\}}{g_{a:b}^2}\|E - E'\|\|W\|_* \\
&\quad + 16\left(1 + \frac{\sigma_a - \sigma_b}{g_{a:b}}\right)\frac{\|E'\|^2}{g_{a:b}^2}\|W\|_* \frac{12\,|\|E\| - \|E'\||}{\delta} \\
&\le C_1\left(1 + \frac{\sigma_a - \sigma_b}{g_{a:b}}\right)\frac{\delta}{g_{a:b}^2}\|E - E'\|\|W\|_*,
\end{aligned}$$

  for some constant $C_1 > 0$ that is independent of $E, E'$.
- If $\min\{\|E\|, \|E'\|\} \ge \delta$ then $h(E) = h(E') = 0$ and the above inequality trivially holds.
- Finally, if $\|E\| < \delta \le \|E'\|$, by a similar argument as above, we obtain that

$$\begin{aligned}
|h_\delta(E) - h_\delta(E')| = |h_\delta(E)| &= \left|\langle S_{a:b}(E), W\rangle \left(\phi\left(\frac{6\|E\|}{\delta}\right) - \phi\left(\frac{6\|E'\|}{\delta}\right)\right)\right| \\
&\le \|S_{a:b}(E)\|\|W\|_* \left|\phi\left(\frac{6\|E\|}{\delta}\right) - \phi\left(\frac{6\|E'\|}{\delta}\right)\right|
\end{aligned}$$

$$\leq C_1 \left(1+\frac{\sigma_a-\sigma_b}{g_{a:b}}\right)\frac{\delta}{g_{a:b}^2}\|E-E'\|\|W\|_*,$$

and the same bound holds if we switch the places of $E$ and $E'$ in the last case.

Combining the above cases together, we have shown that for any $\delta$ such that $\delta \leq g_{a:b}/4$, $h_\delta$ is a Lipschitz function with Lipschitz constant bounded by

$$C_1 \left(1+\frac{\sigma_a-\sigma_b}{g_{a:b}}\right)\frac{\delta}{g_{a:b}^2}\|W\|_*.$$

*Step 2.* In the following, for any two sequences $\{x_n\}, \{y_n\}$, we adopt the notation $x_n \lesssim y_n$ meaning there exists some constant $c > 0$ independent of $n$, such that $x_n \leq c y_n$.

By lemma B.1, we have that for all $t > 0$,

$$\mathbb{P}\left(|\|E\| - \mathbb{E}\|E\|| \geq \sqrt{2t}\right) \leq \exp(-t).$$

Set $\delta = \delta(t) = \mathbb{E}\|E\| + \sqrt{2t}$. We consider the following two scenarios depending on the values of $t$.

- We first consider the case where $\sqrt{2t} \leq g_{a:b}/24$, which implies $\delta(t) \leq g_{a:b}/6$. By the definition of $h_\delta(\cdot)$, we have that $h_\delta(E) = \langle S_{a:b}(E), W\rangle$. Denoting by $m$ the median of $\langle S_{a:b}(E), W\rangle$ we have that

$$\begin{aligned}\mathbb{P}(h_\delta(E) \geq m) &\geq \mathbb{P}(h_\delta(E) \geq m, \|E\| \leq \delta(t)) = \mathbb{P}(\langle S_{a:b}(E), W\rangle \geq m, \|E\| \leq \delta(t))\\ &\geq \mathbb{P}(\langle S_{a:b}, W\rangle \geq m) - \mathbb{P}(\|E\| > \delta(t)) \geq \frac{1}{2} - \frac{1}{2}e^{-t} \geq \frac{1}{4},\end{aligned}$$

  and likewise $\mathbb{P}(h_\delta(E) \leq m) \geq 1/4$. Hence, since $h_\delta$ is Lipschitz, we can apply Lemma 2.6 in [7], which is a corollary to the the Gaussian isoperimetric inequality, to show that with probability at least $1 - e^{-t}$ that

$$|h_\delta(E) - m| \lesssim \sqrt{t}\left(1+\frac{\sigma_a-\sigma_b}{g_{a:b}}\right)\frac{\delta(t)}{g_{a:b}^2}\|W\|_*. \tag{52}$$

  By Lemma B.1, we have that $\mathbb{E}\|E\| \lesssim \sqrt{n+p}$. Thus, we obtain that

$$|h_\delta(E) - m| \lesssim \left(1+\frac{\sigma_a-\sigma_b}{g_{a:b}}\right)\frac{\sqrt{t}}{g_{a:b}}\left(\frac{\sqrt{n+p}+\sqrt{t}}{g_{a:b}}\right)\|W\|_*.$$

  Moreover, the event where $\|E\| \leq \delta(t)$ occurs with probability at least $1 - e^{-t}$ and on this event $h_\delta$ coincides with $\langle S_{a:b}(E), W\rangle$. Hence, with probability at least $1 - 2e^{-t}$

$$|\langle S_{a:b}(E), W\rangle - m| \lesssim \left(1+\frac{\sigma_a-\sigma_b}{g_{a:b}}\right)\frac{\sqrt{t}}{g_{a:b}}\left(\frac{\sqrt{n+p}+\sqrt{t}}{g_{a:b}}\right)\|W\|_*. \tag{53}$$

- We need to prove a similar inequality in the case $\sqrt{2t} > g_{a:b}/24$. In this case we have that $\mathbb{E}\|E\| \lesssim \sqrt{t}$ as by assumption $\mathbb{E}\|E\| \leq g_{a:b}/8$. Hence, applying lemma F.1, we have that

$$|\langle S_{a:b}(E), W\rangle| \leq \|S_{a:b}(E)\|\,\|W\|_* \lesssim \left(1+\frac{\sigma_a-\sigma_b}{g_{a:b}}\right)\frac{t}{g_{a:b}^2}\|W\|_*. \tag{54}$$

  Hence, since $t \geq \log(4)$ and $e^{-t} \leq 1/4$, we conclude that we can bound

$$|m| \lesssim \left(1+\frac{\sigma_a-\sigma_b}{g_{a:b}}\right)\frac{t}{g_{a:b}^2}\|W\|_*. \tag{55}$$

  (54) and (55) together immediately imply that the inequality in (53) also holds for $\sqrt{2t} > g_{a:b}/24$.

So far we have proved that (53) holds for all $t > \log 4$. Integrating out the tails in the inequality in (53) we obtain that

$$|\mathbb{E}\langle S_{a:b}(E), W\rangle - m| \le \mathbb{E}|\langle S_{a:b}(E), W\rangle - m| \lesssim \left(1 + \frac{\sigma_a - \sigma_b}{g_{a:b}}\right) \frac{\sqrt{n+p}}{g_{a:b}^2} \|W\|_*,$$

and hence we can substitute the median by the mean in the concentration inequality (53).

□

The last two things left are the proofs of Lemma F.1 and Lemma F.2.

PROOF OF LEMMA F.1. As in the proof of Lemma 4.4, we use self-adjoint dilation. As before, we define for any matrix $W$

$$D(W) = \begin{pmatrix} 0 & W \\ W^T & 0 \end{pmatrix}.$$

Since $D(M)$ is symmetric and because of its relation to $M$ it has eigendecomposition

$$D(M) = \sum_{1 \le |i| \le k} \sigma_i P_i, \tag{56}$$

where for $i \in [k]$,

$$\sigma_{-i} = -\sigma_i,\ P_i = \frac{1}{2}\begin{pmatrix} u_i u_i^T & u_i v_i^T \\ v_i u_i^T & v_i v_i^T \end{pmatrix},\ P_{-i} = \frac{1}{2}\begin{pmatrix} u_i u_i^T & -u_i v_i^T \\ -v_i u_i^T & v_i v_i^T \end{pmatrix} \tag{57}$$

Similarly, we have that

$$D(\hat{M}) = \sum_{1 \le |i| \le p \wedge n} \hat{\sigma}_i \hat{P}_i.,$$

where for each $i \in [k]$, $\hat{\sigma}_{-i}, \hat{P}_i$ and $\hat{P}_{-i}$ are defined analogously. Denote

$$P = \sum_{|i| \in \{a,\dots,b\}} P_i, \quad \text{and } \hat{P} = \sum_{|i| \in \{a,\dots,b\}} \hat{P}_i. \tag{58}$$

Using this notation, we have that

$$\left(I - VV^T\right)\left(\hat{V}_{a:b}\hat{V}_{a:b}^T - V_{a:b}V_{a:b}^T\right) V_{a:b} = \left(O_{n\times p}\ \left(I - VV^T\right)\right)\left(\hat{P} - P\right)\begin{pmatrix} O_{n\times p} \\ V_{a:b} \end{pmatrix}, \tag{59}$$

where $O_{n\times p}$ denotes a $n \times p$-matrix consisting of 0's. We divide the following part of the proof into three steps.

*Step 1.* In this step, we decompose $\left(I - VV^T\right)\left(\hat{V}_{a:b}\hat{V}_{a:b}^T - V_{a:b}V_{a:b}^T\right) V_{a:b}$. Denote by $[\sigma_a, \sigma_b]$ the corresponding interval on the real axis of the complex plane $\mathbb{C}$. Define $\gamma^+$ to be a contour $\mathbb{C}$ around the intervals $[\sigma_a, \sigma_b]$ with distance equal to $g_{a:b}/2$, i.e.,

$$\gamma^+ = \left\{\eta \in \mathbb{C} : \text{dist}(\eta, [\sigma_a, \sigma_b]) = \frac{g_{a:b}}{2}\right\}, \tag{60}$$

where for any point $\eta \in \mathbb{C}$ and interval $B \in \mathbb{C}$, $\text{dist}(\eta, B) = \min_{\eta' \in \mathbb{B}} \|\eta - \eta'\|$. Likewise we define $\gamma^-$ as

$$\gamma^- = \left\{\eta \in \mathbb{C} : \text{dist}(\eta, [\sigma_{-b}, \sigma_{-a}]) = \frac{g_{a:b}}{2}\right\}. \tag{61}$$

This way, among the singular values of $D(M)$, only those with index in $\{a,\ldots,b\}$ and $\{-b,\ldots,-a\}$ are included in $\gamma^+$ and $\gamma^-$ respectively, and the remaining ones lie outside of the contours. By the Riesz representation Theorem for spectral projectors (c.f. page 39 of [5]), we have that

$$\hat{P} = -\frac{1}{2\pi i}\oint_{\gamma^+}(D(\hat{M})-\eta I)^{-1}d\eta - \frac{1}{2\pi i}\oint_{\gamma^-}(D(\hat{M})-\eta I)^{-1}d\eta. \tag{62}$$

For any matrix $W$ and any $\eta\in\mathbb{C}$, define the resolvent operator

$$R_W(\eta) = (D(W)-\eta I)^{-1}.$$

Then (62) can be written as

$$\hat{P} = -\frac{1}{2\pi i}\oint_{\gamma^+}R_{\hat{M}}(\eta)d\eta - \frac{1}{2\pi i}\oint_{\gamma^-}R_{\hat{M}}(\eta)d\eta.$$

Note that $D(\hat{M}) = D(M)+D(E)$ and that $R_M(\eta) = (D(M)-\eta I)^{-1}$. We expand $R_{\hat{M}}(\eta)$ into its Neumann series:

$$\begin{aligned}R_{\hat{M}}(\eta) &= (D(M)-\eta I + D(E))^{-1} = ((D(M)-\eta I)(I+R_M(\eta)D(E)))^{-1}\\ &= (I+R_M(\eta)D(E))^{-1}R_M(\eta) = \sum_{j=0}^{\infty}(-1)^j[R_M(\eta)D(E)]^jR_M(\eta)\\ &= R_M(\eta) - R_M(\eta)D(E)R_M(\eta) + \sum_{j=2}^{\infty}(-1)^j[R_M(\eta)D(E)]^jR_M(\eta). \end{aligned} \tag{63}$$

Applying the Riesz representation Theorem on $P$, we have that

$$\begin{aligned}P &= -\frac{1}{2\pi i}\oint_{\gamma^+}(D(M)-\eta I)^{-1}d\eta - \frac{1}{2\pi i}\oint_{\gamma^-}(D(M)-\eta I)^{-1}d\eta\\ &= -\frac{1}{2\pi i}\oint_{\gamma^+}R_M(\eta)d\eta - \frac{1}{2\pi i}\oint_{\gamma^-}R_M(\eta)d\eta.\end{aligned}$$

As a result, we have the decomposition

$$\hat{P} - P = L(E) + S(E) \tag{64}$$

where $L(E)$ and $S(E)$ are operators on $E$, defined as

$$L(E) = \frac{1}{2\pi i}\oint_{\gamma^+}R_M(\eta)D(E)R_M(\eta)d\eta + \frac{1}{2\pi i}\oint_{\gamma^-}R_M(\eta)D(E)R_M(\eta)d\eta. \tag{65}$$

and

$$\begin{aligned}S(E) = &-\frac{1}{2\pi i}\oint_{\gamma^+}\sum_{j=2}^{\infty}(-1)^j[R_M(\eta)D(E)]^jR_M(\eta)d\eta\\ &-\frac{1}{2\pi i}\oint_{\gamma^-}\sum_{j=2}^{\infty}(-1)^j[R_M(\eta)D(E)]^jR_M(\eta)d\eta.\end{aligned} \tag{66}$$

By (59), we have that

$$\begin{aligned}\left(I-VV^T\right)\left(\hat{V}_{a:b}\hat{V}_{a:b}^T - V_{a:b}V_{a:b}^T\right)V_{a:b} &= \left(O_{n\times p}\ \left(I-VV^T\right)\right)L(E)\begin{pmatrix}O_{n\times p}\\ V_{a:b}\end{pmatrix}\\ &\quad + \left(O_{n\times p}\ \left(I-VV^T\right)\right)S(E)\begin{pmatrix}O_{n\times p}\\ V_{a:b}\end{pmatrix}.\end{aligned}$$

*Step 2.* In the following, we show that the first term on the right hand side of the above formula equals$\sum_j \frac{1}{\sigma_j}\left(I - VV^T\right)E^T u_j e_j^T$, which implies that the second term equals $S_{a:b}$.

Define

$$L_{a:b} = \left(O_{n\times p}\ \left(I - VV^T\right)\right) L(E) \begin{pmatrix} O_{n\times p} \\ V_{a:b} \end{pmatrix}. \tag{67}$$

We first simplify $L(E)$. Recalling (57), for any $i$ such that $|i| \le k$, we have that $P_i = \theta_i\theta_i^T$, where $\theta_i = \frac{1}{\sqrt{2}}(u_i^T, v_i^T)^T, \theta_{-i} = \frac{1}{\sqrt{2}}(u_i^T, -v_i^T)^T$. We expand this into an orthonormal basis of $\mathbb{R}^{p+n}$, $\{\theta_i, \theta_{-i}\}_{i\in[k]} \cup \{\theta_j\}_{k+1\le j\le p+n-k}$. This implies the following:

- For $k+1 \le j \le p+n-k$, we define $P_j = \theta_j\theta_j^T$ and decomlpose the identity matrix as

$$I = \sum_{i\in\{1,\ldots,p+n-k\}\cup\{-k,\ldots,-1\}} P_i$$

  In the rest of the proof, by default we treat $\{1,\ldots,p+n-k\}\cup\{-k,\ldots,-1\}$ to be the whole set for the index $i$. We drop it when there is no ambiguity. For instance, the above equation can be simply written as $I = \sum_i P_i$.
- We define

$$\sigma_j = 0, \forall k+1 \le j \le p+n-k. \tag{68}$$

  Then (56) can be expressed as

$$D(M) = \sum_i \sigma_i P_i.$$

- For $k+1 \le j \le p+n-k$, $\theta_j$ is orthogonal to $\theta_i - \theta_{-i}, \forall i \in [k]$. This implies that the second part of $\theta_j$ (i.e., from the $(p+1)$th coordinate to the $(p+n)$th coordinate) is 0, or orthogonal to span$(v_1,\ldots,v_k)$. Thus,

$$\left(O_{n\times p}\ \left(I - VV^T\right)\right) P_i = O, \forall i \text{ s.t. } |i| \le k, \tag{69}$$

$$P_i \begin{pmatrix} O_{n\times p} \\ V_{a:b} \end{pmatrix} = O, \forall i \text{ s.t. } |i| \notin \{a,\ldots,b\}. \tag{70}$$

  and

$$\left(O_{n\times p}\ \left(I - VV^T\right)\right) \sum_{i>k} P_i \begin{pmatrix} O_{p\times n} \\ O_{I_{n\times n}} \end{pmatrix} = I - VV^T. \tag{71}$$

By (56), we have that

$$\begin{aligned} R_M(\eta) &= (D(M) - \eta I)^{-1} = \left(\sum_i \sigma_i P_i - \eta I\right)^{-1} = \left(\sum_i (\sigma_i - \eta) P_i\right)^{-1} \\ &= \sum_i \frac{1}{\sigma_i - \eta} P_i = \sum_{i\in\{a,\ldots,b\}} \frac{1}{\sigma_i - \eta} P_i + \sum_{i\notin\{a,\ldots,b\}} \frac{1}{\sigma_i - \eta} P_i, \end{aligned} \tag{72}$$

defined as $R_1^+(\eta)$ and $R_2^+(\eta)$ respectively. With this, for the first term of $L(E)$ in (65), we have that

$$\frac{1}{2\pi i}\oint_{\gamma^+} R_M(\eta) D(E) R_M(\eta) d\eta = \frac{1}{2\pi i}\oint_{\gamma^+} \left(R_1^+(\eta) + R_2^+(\eta)\right) D(E) \left(R_1^+(\eta) + R_2^+(\eta)\right) d\eta.$$

Observe that by the Cauchy-Goursat Theorem,

$$\oint_{\gamma^+} R_1^+(\eta)D(E)R_1^+(\eta)d\eta$$

$$=\sum_{i\in\{a,\dots,b\}} P_iD(E)P_i \oint_{\gamma^+} \frac{1}{(\sigma_i-\eta)^2}d\eta + \sum_{i\neq j,\ i,j\in\{a,\dots,b\}} P_iD(E)P_j \oint_{\gamma^+} \frac{1}{(\sigma_i-\eta)(\sigma_j-\eta)}d\eta$$

$$=0,$$

since there is no singularity inside $\gamma^+$. The identical result holds for $\oint_{\gamma^+} R_2^+(\eta)D(E)R_2^+(\eta)d\eta$. Using the Cauchy integral formula, we obtain that

$$\frac{1}{2\pi i}\oint_{\gamma^+} R_1^+(\eta)D(E)R_2^+(\eta)d\eta = \sum_{i\in\{a,\dots,b\}}\sum_{j\notin\{a,\dots,b\}} \frac{1}{2\pi i}\oint_{\gamma^+} \frac{d\eta}{(\sigma_i-\eta)(\sigma_j-\eta)} P_iEP_j$$

$$=\sum_{i\in\{a,\dots,b\}}\sum_{j\notin\{a,\dots,b\}} \frac{P_iEP_j}{\sigma_i-\sigma_j}.$$

A similar result holds for $\frac{1}{2\pi i}\oint_{\gamma^+} R_2^+(\eta)D(E)R_1^+(\eta)d\eta$. Hence, we obain that

$$\frac{1}{2\pi i}\oint_{\gamma^+} R_M(\eta)D(E)R_M(\eta)d\eta = \sum_{i\in\{a,\dots,b\}}\sum_{j\notin\{a,\dots,b\}} \frac{P_iD(E)P_j+P_jD(E)P_i}{\sigma_i-\sigma_j}.$$

In the same manner, splitting

$$R_M(\eta) = R_1^-(\eta) + R_2^-(\eta) \triangleq \sum_{i\in\{-b,\dots,-a\}} \frac{P_i}{\sigma_i-\eta} + \sum_{i\notin\{-b,\dots,-a\}} \frac{P_i}{\sigma_i-\eta}, \tag{73}$$

we also obtain that

$$\frac{1}{2\pi i}\oint_{\gamma^-} R_M(\eta)D(E)R_M(\eta)d\eta = \sum_{i\in\{-b,\dots,-a\}}\sum_{j\notin\{-b,\dots,-a\}} \frac{P_iD(E)P_j+P_jD(E)P_i}{\sigma_i-\sigma_j}. \tag{74}$$

Hence, we have that

$$L(E) = \left(\sum_{i\in\{a,\dots,b\}}\sum_{j\notin\{a,\dots,b\}} + \sum_{i\in\{-b,\dots,-a\}}\sum_{j\notin\{-b,\dots,-a\}}\right) \frac{P_iD(E)P_j+P_jD(E)P_i}{\sigma_i-\sigma_j}. \tag{75}$$

Note that for any $i$ such $|i|\in\{a,\dots,b\}$ and any $|j|\notin\{a,\dots,b\}$, (69) and (70) imply

$$\left(O_{n\times p}\ \left(I-VV^T\right)\right) P_iD(E)P_j \begin{pmatrix} O_{n\times p} \\ V_{a:b}\end{pmatrix} = 0.$$

Together with (67), this implies that

$$L_{a:b} = \left(\sum_{i\in\{a,\dots,b\}}\sum_{j\notin\{a,\dots,b\}} + \sum_{i\in\{-b,\dots,-a\}}\sum_{j\notin\{-b,\dots,-a\}}\right) \left(O_{n\times p}\ \left(I-VV^T\right)\right) \frac{P_jD(E)P_i}{\sigma_i-\sigma_j} \begin{pmatrix} O_{n\times p} \\ V_{a:b}\end{pmatrix}$$

$$= \left(\sum_{i\in\{a,\dots,b\}} + \sum_{i\in\{-b,\dots,-a\}}\right) \sum_{j>k} \left(O_{n\times p}\ \left(I-VV^T\right)\right) \frac{P_jD(E)P_i}{\sigma_i} \begin{pmatrix} O_{n\times p} \\ V_{a:b}\end{pmatrix}.$$

Recall that for all $i \leq k$, $\sigma_{-i} = -\sigma_i$. This yields

$$
\begin{aligned}
L_{a:b} &= \sum_{i \in \{a,\dots,b\}} \frac{1}{\sigma_i} \left(O_{n\times p}\ \left(I - VV^T\right)\right) \left(\sum_{j>k} P_j\right) D(E)\left(P_i - P_{-i}\right) \begin{pmatrix} O_{n\times p} \\ V_{a:b} \end{pmatrix} \\
&= \sum_{i \in \{a,\dots,b\}} \frac{1}{\sigma_i} \left(O_{n\times p}\ \left(I - VV^T\right)\right) \left(\sum_{j>k} P_j\right) \begin{pmatrix} O & E \\ E^T & O \end{pmatrix} \begin{pmatrix} O & u_i v_i^T \\ v_i^T u_i & O \end{pmatrix} \begin{pmatrix} O_{n\times p} \\ V_{a:b} \end{pmatrix} \\
&= \sum_{i \in \{a,\dots,b\}} \frac{1}{\sigma_i} \left(O_{n\times p}\ \left(I - VV^T\right)\right) \left(\sum_{j>k} P_j\right) \begin{pmatrix} O_{p\times n} \\ I_{n\times n} \end{pmatrix} E^T u_i v_i^T V_{a:b} \\
&= \sum_{i \in \{a,\dots,b\}} \frac{1}{\sigma_i} \left(I - VV^T\right) E^T u_i v_i^T V_{a:b},
\end{aligned}
$$

where the last equation is due to (71). This implies

$$
\begin{aligned}
S_{a:b} &= \left(I - VV^T\right)\left(\hat{V}_{a:b}\hat{V}_{a:b}^T - V_{a:b}V_{a:b}^T\right) V_{a:b} - L_{a:b} \\
&= \left(O_{n\times p}\ \left(I - VV^T\right)\right) S(E) \begin{pmatrix} O_{n\times p} \\ V_{a:b} \end{pmatrix}. \qquad (76)
\end{aligned}
$$

*Step 3.* In the final step, we upper bound $\|S_{a:b}\|$ by using the formula above. By (72), for any $\eta \in \gamma^+$ or $\eta \in \gamma^-$, we have that

$$
\|R_M(\eta)\| \leq \frac{2}{g_{a:b}}. \qquad (77)
$$

Moreover, we have that

$$
|\gamma^+| = |\gamma^-| \leq 2(\sigma_a - \sigma_b) + \pi g_{a:b}.
$$

Recall the definition of $S(E)$ in (66). Note that $\|D(E)\| = \|E\|$.

- Under the assumption that $\|E\| \leq g_{a:b}/4$, we have that

$$
\begin{aligned}
\|S_{a:b}\| \leq \|S(E)\| &\leq \frac{|\gamma^+| + |\gamma^-|}{2\pi} \sum_{j=2}^{\infty} \|R_M(\eta)\|^{j+1} \|D(E)\|^j \\
&\leq \frac{2(\sigma_a - \sigma_b) + \pi g_{a:b}}{\pi} \|E\|^2 \left(\frac{2}{g_{a:b}}\right)^3 \sum_{j=0}^{\infty} \|E\|^j \left(\frac{2}{g_{a:b}}\right)^j \\
&\leq \left(\frac{16(\sigma_a - \sigma_b)}{\pi g_{a:b}} + 8\right) \frac{\|E\|^2}{g_{a:b}^2} \sum_{j=0}^{\infty} \|E\|^j \left(\frac{2}{g_{a:b}}\right)^j \\
&\leq \left(\frac{32(\sigma_a - \sigma_b)}{\pi g_{a:b}} + 16\right) \frac{\|E\|^2}{g_{a:b}^2}. \qquad (78)
\end{aligned}
$$

- If $\|E\| > g_{a:b}/4$, by (76) we have that

$$
\|S_{a:b}\| \leq \left\|\hat{V}_{a:b}\hat{V}_{a:b}^T - V_{a:b}V_{a:b}^T\right\| + \|L_{a:b}\|.
$$

The first term is bounded by 2. By the definition of $\|L_{a:b}\|$, the second term can be bounded as follows

$$\|L_{a:b}\| = \left\| \left(I - VV^T\right) E^T \left( \sum_{i\in\{a,\ldots,b\}} \frac{1}{\sigma_i} u_i v_i^T \right) V_{a:b} \right\| \le \frac{\|E\|}{\min_{i\in\{a,\ldots,b\}} \sigma_i} \le \frac{\|E\|}{g_{a:b}}.$$

Hence, we finally obtain that

$$\|S_{a:b}\| \le 2 + \frac{\|E\|}{g_{a:b}} \le 16 \frac{\|E\|^2}{g_{a:b}^2}.$$

□

Finally, we prove Lemma F.2.

PROOF OF LEMMA F.2. We follow the same decomposition and notation as in the proof of Lemma F.1. Recall the definition of $\hat{P}$ and $P$ in (58). In particular, due to (64), we have that

$$\hat{P} - P = L(E) + S(E),$$

where $L(E)$ and $S(E)$ are defined in (65) and (66), respectively. Define $\hat{P}', L(E'), S(E')$ in the same manner for $M'$. Then we have that

$$S(E') - S(E) = \hat{P}' - \hat{P} - \left(L(E') - L(E)\right).$$

As a consequence, due to (76), we obtain that

$$\begin{aligned} S_{a:b}(E') - S_{a:b}(E) &= \left(O_{n\times p} \; \left(I - VV^T\right)\right) \left(S(E') - S(E)\right) \begin{pmatrix} O_{n\times p} \\ V_{a:b} \end{pmatrix} \\ &= \left(O_{n\times p} \; \left(I - VV^T\right)\right) \left(\hat{P}' - \hat{P}\right) \begin{pmatrix} O_{n\times p} \\ V_{a:b} \end{pmatrix} \\ &\quad - \left(O_{n\times p} \; \left(I - VV^T\right)\right) \left(L(E') - L(E)\right) \begin{pmatrix} O_{n\times p} \\ V_{a:b} \end{pmatrix}. \end{aligned}$$

In the proof of Lemma F.1, we analyze the difference between $\hat{P}$ and $P$. By the exactly the same argument, we analyze the difference between $\hat{P}'$ and $\hat{P}$. As in (64), we have that

$$\hat{P}' - \hat{P} = \hat{L}(E' - E) + \hat{S}(E' - E),$$

where

$$\begin{aligned} \hat{L}(E' - E) &= \frac{1}{2\pi i} \oint_{\gamma^+} R_{\hat{M}}(\eta) D(E' - E) R_{\hat{M}}(\eta) d\eta \\ &\quad + \frac{1}{2\pi i} \oint_{\gamma^-} R_{\hat{M}}(\eta) D(E' - E) R_{\hat{M}}(\eta) d\eta. \end{aligned} \tag{79}$$

and

$$\begin{aligned} \hat{S}(E' - E) &= -\frac{1}{2\pi i} \oint_{\gamma^+} \sum_{j=2}^{\infty} (-1)^j [R_{\hat{M}}(\eta) D(E' - E)]^j R_{\hat{M}}(\eta) d\eta \\ &\quad - \frac{1}{2\pi i} \oint_{\gamma^-} \sum_{j=2}^{\infty} (-1)^j [R_{\hat{M}}(\eta) D(E' - E)]^j R_{\hat{M}}(\eta) d\eta, \end{aligned}$$

with $\gamma^+, \gamma^-$ defined in (60) and (61). Hence, we have that

$$S_{a:b}(E') - S_{a:b}(E) = \left(O_{n\times p}\ \left(I - VV^T\right)\right) \hat{S}(E'-E) \begin{pmatrix} O_{n\times p} \\ V_{a:b} \end{pmatrix}$$
$$+ \left(O_{n\times p}\ \left(I - VV^T\right)\right) \left(\hat{L}(E'-E) - \left(L(E') - L(E)\right)\right) \begin{pmatrix} O_{n\times p} \\ V_{a:b} \end{pmatrix},$$

which implies

$$\left\|S_{a:b}(E') - S_{a:b}(E)\right\| \le \left\|\hat{S}(E'-E)\right\| + \left\|\hat{L}(E'-E) - \left(L(E') - L(E)\right)\right\|. \tag{80}$$

We are going to establish upper bounds on the two terms individually.

*Step 1.* We first bound the second term above. Due to (65), (79) and the fact that $D(E'-E) = D(E') - D(E)$, we have that

$$\begin{aligned}
&\hat{L}(E'-E) - \left(L(E') - L(E)\right)\\
&= \frac{1}{2\pi i}\oint_{\gamma^+} \left(R_{\hat{M}}(\eta)D(E'-E)R_{\hat{M}}(\eta)d\eta - R_M(\eta)D(E'-E)R_M(\eta)\right) d\eta\\
&\quad + \frac{1}{2\pi i}\oint_{\gamma^-} \left(R_{\hat{M}}(\eta)D(E'-E)R_{\hat{M}}(\eta)d\eta - R_M(\eta)D(E'-E)R_M(\eta)\right) d\eta.
\end{aligned}$$

By Weyl's inequality (Theorem 4.3.1 of [4]), we have $|\hat{\sigma}_i - \sigma_i| \le \|E\|, \forall i \in [p\wedge n]$. Assuming that $\|E\| \le g_{a:b}/4$, the minimum distance between $\gamma^+, \gamma^-$ to the points $\{(\hat{\sigma}_i, 0)\}$ is at least $g_{a:b}/2 - \|E\| \ge g_{a:b}/4$, for all $i \in [p\wedge n]$. Similarly as (77), we obtain that

$$\left\|R_{\hat{M}}(\eta)\right\| \le \frac{4}{g_{a:b}}, \forall \eta \in \gamma^+, \gamma^-.$$

Hence, together with the fact that $\|D(E'-E)\| = \|E'-E\|$, we have that

$$\begin{aligned}
&\left\|\oint_{\gamma^+} \left(R_{\hat{M}}(\eta)D(E'-E)R_{\hat{M}}(\eta)d\eta - R_M(\eta)D(E'-E)R_M(\eta)\right) d\eta\right\|\\
&\le \left\|\oint_{\gamma^+} R_{\hat{M}}(\eta)D(E'-E)(R_{\hat{M}}(\eta) - R_M(\eta))d\eta\right\| + \left\|\oint_{\gamma^+} (R_{\hat{M}}(\eta) - R_M(\eta))D(E'-E)R_M(\eta)d\eta\right\|\\
&\le \frac{8|\gamma^+|}{g_{a_b}}\|E'-E\| \sup_{\eta\in\gamma^+} \|R_{\hat{M}}(\eta) - R_M(\eta)\|.
\end{aligned}$$

Moreover, by the expansion of the resolvent into a Neumann series in (63), we have that

$$\begin{aligned}
\|R_{\hat{M}}(\eta) - R_M(\eta)\| &\le \sum_{j=1}^{\infty} (\|R_M(\eta)\|\|E\|)^j \|R_M(\eta)\| \le \|R_M(\eta)\|^2 \|E\| \sum_{j=0}^{\infty} (\|R_M(\eta)\|\|E\|)^j\\
&\le \frac{8\|E\|}{g_{a:b}^2},
\end{aligned}$$

where the last inequality is due to (77). Hence, as $|\gamma^+| \le \pi g_{a:b} + 2(\sigma_a - \sigma_b)$, we have that

$$\begin{aligned}
&\left\|\oint_{\gamma^+} \left(R_{\hat{M}}(\eta)D(E'-E)R_{\hat{M}}(\eta)d\eta - R_M(\eta)D(E'-E)R_M(\eta)\right) d\eta\right\|\\
&\le 64\left(\pi + \frac{2(\sigma_a - \sigma_b)}{g_{a:b}}\right) \frac{\|E\|\,\|E'-E\|}{g_{a:b}^2}.
\end{aligned}$$

The same result holds for the other integral over $\gamma^-$. Hence, we obtain that

$$\left\|\hat{L}(E'-E)-\big(L(E')-L(E)\big)\right\| \leq 64\left(1+\frac{2(\sigma_a-\sigma_b)}{\pi g_{a:b}}\right)\frac{\|E\|\,\|E'-E\|}{g_{a:b}^2}.$$

*Step 2.* For the term related to $\hat{S}$, we bound it analogously as in the proof of Lemma F.1. Following (78), we have that

$$\left\|\hat{S}(E'-E)\right\| \leq 64\left(\frac{32(\sigma_a-\sigma_b)}{\pi g_{a:b}}+16\right)\frac{\|E'-E\|^2}{g_{a:b}^2}.$$

Combining the above result with (80), we obtain that

$$\left\|S_{a:b}(E')-S_{a:b}(E)\right\| \leq 1024\left(1+\frac{\sigma_a-\sigma_b}{g_{a:b}}\right)\frac{\max\{\|E\|,\|E'\|\}}{g_{a:b}^2}\|E-E'\|.$$

□

## REFERENCES


[1] BANDEIRA, A. S. and VAN HANDEL, R. (2016). Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Ann. Probab.* **44** 2479-2506.

[2] DAVISON, K. R. and SZAREK, S. J. (2001). Local operator theory, random matrices and Banach spaces. In *Handbook of the geometry of Banach spaces*, **1** 317-366. North-Holland, Amsterdam.

[3] GINÉ, E. and NICKL, R. (2016). *Mathematical Foundations of infinite-Dimensional Statistical Methods*. Cambridge University Press.

[4] HORN, R. A. and JOHNSON, C. R. (2012). *Matrix analysis*. Cambridge university press.

[5] KATO, T. (2013). *Perturbation theory for linear operators* **132**. Springer Science & Business Media.

[6] KOLTCHINSKII, V. and LOUNICI, K. (2016). Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* **52** 1976–2013.

[7] KOLTCHINSKII, V. and XIA, D. (2016). Perturbation of linear forms of singular vectors under Gaussian noise. In *High Dimensional Probability VII* 397–423. Springer.

[8] O'ROURKE, S., VU, V. and WANG, K. (2018). Random perturbation of low rank matrices: Improving classical bounds. *Linear Algebra Appl.* **540** 26–59.

[9] VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed sensing* 210-268. Cambridge Univ. Press, Cambridge.