Supplement to “User-Friendly Covariance Estimation for Heavy-Tailed Distributions”

Yuan Ke[*], Stanislav Minsker[†], Zhao Ren[‡], Qiang Sun[§] and Wen-Xin Zhou[¶]

In Sections A–C, we provide proofs of all the theoretical results in the main text. In addition, we investigate robust covariance estimation and inference under factor models in Section D, which might be of independent interest.

A Proof of Proposition 2.1

Without loss of generality we assume $\boldsymbol{\mu} = \mathbf{0}$. We construct a random vector $\boldsymbol{X} \in \mathbb{R}^d$ that follows the distribution below:

$$\mathbb{P}\Big\{\boldsymbol{X} = (0, \ldots, 0, \underbrace{n\eta}_{j\text{th}}, 0, \ldots, 0)^\intercal\Big\} = \mathbb{P}\Big\{\boldsymbol{X} = (0, \ldots, 0, \underbrace{-n\eta}_{j\text{th}}, 0, \ldots, 0)^\intercal\Big\} = \frac{\sigma^2}{2n^2\eta^2}$$

for each $j = 1, \ldots, d$, and

$$\mathbb{P}(\boldsymbol{X} = \mathbf{0}) = 1 - \frac{d\sigma^2}{n^2\eta^2}.$$

Here we assume $\eta^2 > d\sigma^2/n^2$ so that $\mathbb{P}(\boldsymbol{X} = \mathbf{0}) > 0$. In other words, the number of non-zero elements of $\boldsymbol{X}$ is at most 1. It is easy to see that the mean and covariance matrix of $\boldsymbol{X}$ are $\mathbf{0}$ and $\sigma^2\mathbf{I}_d$, respectively.

Consider the empirical mean $\bar{\boldsymbol{X}} = (1/n)\sum_{i=1}^n \boldsymbol{X}_i$, where $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are i.i.d. from $\boldsymbol{X}$. It follows that

$$\begin{aligned}\mathbb{P}(\|\bar{\boldsymbol{X}}\|_\infty \geq \eta) &\geq \mathbb{P}\Big(\text{exactly one of the } n \text{ samples is not equal to } \mathbf{0}\Big)\\ &= \frac{d\sigma^2}{n\eta^2}\bigg(1 - \frac{d\sigma^2}{n^2\eta^2}\bigg)^{n-1}.\end{aligned}$$

Therefore, as long as $\delta < (2e)^{-1}$, the following bound

$$\|\bar{\boldsymbol{X}}\|_\infty \geq \sigma\sqrt{\frac{d}{n\delta}\bigg(1 - \frac{2e\delta}{n}\bigg)^{(n-1)/2}}$$

holds with probability at least δ, as claimed. □

[*] Department of Statistics, University of Georgia, Athens, GA 30602, USA. E-mail: yuan.ke@uga.edu.

[†] Department of Mathematics, University of Southern California, Los Angeles, CA 90089, USA. E-mail: minsker@usc.edu.

[‡] Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15260, USA. E-mail: zren@pitt.edu.

[§] Department of Statistical Sciences, University of Toronto, Toronto, ON M5S 3G3, Canada. E-mail: qsun@utstat.toronto.edu.

[¶] Department of Mathematics, University of California, San Diego, La Jolla, CA 92093, USA. E-mail: wez243@ucsd.edu.

B Proofs for Section 3

B.1 Proof of Theorem 3.1

For each $1 \le k \le \ell \le d$, note that $\widehat{\sigma}^{\mathcal{T}}_{1,k\ell}$ is a U-statistic with a bounded kernel of order two, say $\widehat{\sigma}^{\mathcal{T}}_{1,k\ell} = \binom{n}{2}^{-1} \sum_{1\le i<j\le n} h_{k\ell}(\boldsymbol{X}_i, \boldsymbol{X}_j)$. According to Hoeffding (1963), $\widehat{\sigma}^{\mathcal{T}}_{1,k\ell}$ can be represented as an average of (dependent) sums of independent random variables. Specifically, define

$$W(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) = \frac{h_{k\ell}(\boldsymbol{x}_1, \boldsymbol{x}_2) + h_{k\ell}(\boldsymbol{x}_3, \boldsymbol{x}_4) + \cdots + h_{k\ell}(\boldsymbol{x}_{2m-1}, \boldsymbol{x}_{2m})}{m}$$

for $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^d$. Let $\sum_{\mathcal{P}}$ denote the summation over all $n!$ permutations $(i_1, \dots, i_n)$ of $[n] := \{1, \dots, n\}$ and $\sum_C$ denote the summation over all $\binom{n}{2}$ pairs (i_1, i_2) $(i_1 < i_2)$ from $[n]$. Then we have $m \sum_{\mathcal{P}} W(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) = m2!(n-2)! \sum_C h_{k\ell}(\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2})$ and hence

$$\widehat{\sigma}^{\mathcal{T}}_{1,k\ell} = \frac{1}{n!} \sum_{\mathcal{P}} W(\boldsymbol{X}_{i_1}, \dots, \boldsymbol{X}_{i_n}). \tag{B.1}$$

Write $\tau = \tau_{k\ell}$ and $v = v_{k\ell}$ for simplicity. For any $y > 0$, by Markov's inequality, (B.1), convexity and independence, we derive that

$$\begin{aligned}
\mathbb{P}(\widehat{\sigma}^{\mathcal{T}}_{1,k\ell} - \sigma_{k\ell} \ge y) &\le e^{-(m/\tau)(y+\sigma_{k\ell})} \mathbb{E} e^{(m/\tau)\widehat{\sigma}^{\mathcal{T}}_{1,k\ell}} \\
&\le e^{-(m/\tau)(y+\sigma_{k\ell})} \frac{1}{n!} \sum_{\mathcal{P}} \mathbb{E} e^{(1/\tau)\sum_{j=1}^m h_{k\ell}(\boldsymbol{X}_{i_{2j-1}}, \boldsymbol{X}_{i_{2j}})} \\
&= e^{-(m/\tau)(y+\sigma_{k\ell})} \frac{1}{n!} \sum_{\mathcal{P}} \prod_{j=1}^m \mathbb{E} e^{(1/\tau) h_{k\ell}(\boldsymbol{X}_{i_{2j-1}}, \boldsymbol{X}_{i_{2j}})}.
\end{aligned}$$

Note that $h_{k\ell}(\boldsymbol{X}_{i_{2j-1}}, \boldsymbol{X}_{i_{2j}}) = \psi_\tau(Y_{\pi k} Y_{\pi \ell}/2) = \tau \psi_1(Y_{\pi k} Y_{\pi \ell}/(2\tau))$ for some $1 \le \pi \le N$. In addition, it is easy to verify the inequality that

$$-\log(1 - x + x^2) \le \psi_1(x) \le \log(1 + x + x^2) \ \text{ for all } x \in \mathbb{R}. \tag{B.2}$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E} e^{(1/\tau) h_{k\ell}(\boldsymbol{X}_{i_{2j-1}}, \boldsymbol{X}_{i_{2j}})} \le \mathbb{E}\{1 + Y_{\pi k} Y_{\pi \ell}/(2\tau) + (Y_{\pi k} Y_{\pi \ell})^2/(2\tau)^2\} \\
&= 1 + \sigma_{k\ell}/\tau + (1/\tau)^2 \mathbb{E}(Y_{\pi k} Y_{\pi \ell}/2)^2 \le e^{\sigma_{k\ell}/\tau + (v/\tau)^2}.
\end{aligned}$$

Combining the above calculations gives

$$\mathbb{P}(\widehat{\sigma}^{\mathcal{T}}_{1,k\ell} - \sigma_{k\ell} \ge y) \le e^{-(m/\tau)y + m(v/\tau)^2} = e^{-my^2/(4v^2)},$$

where the equality holds by taking $\tau = 2v^2/y$. Similarly, it can be shown that $\mathbb{P}(\widehat{\sigma}^{\mathcal{T}}_{1,k\ell} - \sigma_{k\ell} \le -y) \le e^{-my^2/(4v^2)}$. Consequently, for $\delta \in (0,1)$, taking $y = 2v\sqrt{(2\log d + \log \delta^{-1})/m}$, or equivalently, $\tau = v\sqrt{m/(2\log d + \log \delta^{-1})}$, we arrive at

$$\mathbb{P}\left(|\widehat{\sigma}^{\mathcal{T}}_{1,k\ell} - \sigma_{k\ell}| \ge 2v\sqrt{\frac{2\log d + \log \delta^{-1}}{m}}\right) \le \frac{2\delta}{d^2}.$$

From the union bound it follows that

$$\mathbb{P}\left(\|\widehat{\boldsymbol{\Sigma}}^{\mathcal{T}}_1 - \boldsymbol{\Sigma}\|_{\max} > 2 \max_{1\le k,\ell\le d} v_{k\ell} \sqrt{\frac{2\log d + \log \delta^{-1}}{n}}\right) \le (1 + d^{-1})\delta.$$

This proves (3.5). □

B.2 Proof of Theorem 3.2

To begin with, note that $\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}}$ can be written as a U-statistic of order 2. Define the index set $\mathcal{I} = \{(i,j) : 1 \le i < j \le n\}$ with cardinality $\binom{n}{2}$. Let $h(\boldsymbol{X}_i, \boldsymbol{X}_j) = (\boldsymbol{X}_i - \boldsymbol{X}_j)(\boldsymbol{X}_i - \boldsymbol{X}_j)^\intercal/2$ and $\mathbf{Z}_{i,j} = \tau^{-1}\psi_\tau(h(\boldsymbol{X}_i, \boldsymbol{X}_j)) = \psi_1(\tau^{-1}h(\boldsymbol{X}_i, \boldsymbol{X}_j))$, such that

$$\widetilde{\mathbf{\Sigma}} := \frac{1}{\tau}\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}} = \frac{1}{\binom{n}{2}}\sum_{(i,j)\in\mathcal{I}} \mathbf{Z}_{i,j}.$$

We now rewrite the U-statistic $\widetilde{\mathbf{\Sigma}}$ as a convex combination of sums of independent random matrices. As in the proof of Theorem 3.1, we define

$$\mathbf{W}_{(1,\ldots,n)} = m^{-1}(\mathbf{Z}_{1,2} + \mathbf{Z}_{3,4} + \ldots + \mathbf{Z}_{2m-1,2m}).$$

For every permutation $\pi = (i_1, \ldots, i_n)$, we adopt the notation $\mathbf{W}_\pi = \mathbf{W}_{(i_1,\ldots,i_n)}$ such that $\widetilde{\mathbf{\Sigma}}^\tau = (n!)^{-1}\sum_{\pi\in\mathcal{P}}\mathbf{W}_\pi$. Using the convexity of the mappings $\mathbf{A} \mapsto \lambda_{\max}(\mathbf{A})$ and $x \mapsto e^x$, we obtain that

$$\exp\{\lambda_{\max}(\widetilde{\mathbf{\Sigma}} - \mathbf{\Sigma}^\tau)\} \le \frac{1}{n!}\sum_{\pi\in\mathcal{P}} \exp\{\lambda_{\max}(\mathbf{W}_\pi - \mathbf{\Sigma}^\tau)\},$$

where $\mathbf{\Sigma}^\tau := \tau^{-1}\mathbf{\Sigma}$. Combined with Markov's inequality and the inequality $e^{\lambda_{\max}(\mathbf{A})} \le \operatorname{tr} e^{\mathbf{A}}$, this further implies

$$\begin{aligned}
&\mathbb{P}\{\sqrt{m}\,\lambda_{\max}(\widehat{\mathbf{\Sigma}}_2^{\mathcal{T}} - \mathbf{\Sigma}) \ge y\} = \mathbb{P}\Big\{e^{\lambda_{\max}(m\widetilde{\mathbf{\Sigma}} - m\mathbf{\Sigma}^\tau)} \ge e^{y\sqrt{m}/\tau}\Big\}\\
&\le e^{-y\sqrt{m}/\tau}\frac{1}{n!}\sum_{\pi\in\mathcal{P}}\mathbb{E}\exp\{\lambda_{\max}(m\mathbf{W}_\pi - m\mathbf{\Sigma}^\tau)\}\\
&\le e^{-y\sqrt{m}/\tau}\frac{1}{n!}\sum_{\pi\in\mathcal{P}}\mathbb{E}\operatorname{tr}\exp(m\mathbf{W}_\pi - m\mathbf{\Sigma}^\tau).
\end{aligned}$$

For every $\pi = (i_1, \ldots, i_n) \in \mathcal{P}$, define $\mathbf{Z}_{\pi,j} = \mathbf{Z}_{i_{2j-1},i_{2j}}$ and $\mathbf{H}_{\pi,j} = h(\boldsymbol{X}_{i_{2j-1}}, \boldsymbol{X}_{i_{2j}})$, such that $\mathbf{Z}_{\pi,1}, \ldots, \mathbf{Z}_{\pi,m}$ are independent and $\mathbb{E}\mathbf{H}_{\pi,j} = \mathbf{\Sigma}$. Then $\mathbf{W}_\pi$ can be written as $\mathbf{W}_\pi = m^{-1}(\mathbf{Z}_{\pi,1} + \ldots + \mathbf{Z}_{\pi,m})$. Recall that $\psi_\tau(x) = \tau\psi_1(x/\tau)$. In view of (B.2), we have the matrix inequality

$$-\log(\mathbf{I} - \tau^{-1}\mathbf{H}_{\pi,j} + \tau^{-2}\mathbf{H}_{\pi,j}^2) \preceq \mathbf{Z}_{\pi,j} \preceq \log(\mathbf{I} + \tau^{-1}\mathbf{H}_{\pi,j} + \tau^{-2}\mathbf{H}_{\pi,j}^2).$$

Then we can bound $\mathbb{E}\exp\operatorname{tr}(m\mathbf{W}_\pi - m\mathbf{\Sigma}^\tau)$ by

$$\begin{aligned}
&\mathbb{E}_{[m-1]}\mathbb{E}_m\operatorname{tr}\exp\left(\sum_{j=1}^{m-1}\mathbf{Z}_{\pi,j} - m\mathbf{\Sigma}^\tau + \mathbf{Z}_{\pi,m}\right)\\
&\qquad\leqslant \mathbb{E}_{[m-1]}\mathbb{E}_m\operatorname{tr}\exp\left\{\sum_{j=1}^{m-1}\mathbf{Z}_{\pi,j} - m\mathbf{\Sigma}^\tau + \log(\mathbf{I} + \tau^{-1}\mathbf{H}_{\pi,m} + \tau^{-2}\mathbf{H}_{\pi,m}^2)\right\},
\end{aligned}\tag{B.3}$$

where the expectation $\mathbb{E}_m$ is taken with respect to $\{\boldsymbol{X}_{i_{2m-1}}, \boldsymbol{X}_{i_{2m}}\}$ and the expectation $\mathbb{E}_{[m-1]}$ is taken with respect to $\{\boldsymbol{X}_{i_1}, ..., \boldsymbol{X}_{i_{2m-2}}\}$. To bound the right-hand side of (B.3), we follow a similar argument as in Minsker (2018). By Lieb's concavity theorem (see, e.g. Fact 2.5 in Minsker (2018)) and

Jensen's inequality, we arrive at

$$\begin{aligned}
&\mathbb{E}\operatorname{tr}\exp(m\mathbf{W}_\pi - m\boldsymbol{\Sigma}^\tau)\\
&\le \mathbb{E}\operatorname{tr}\exp\Big\{\sum_{j=1}^{m-1}\mathbf{Z}_{\pi,j} - m\boldsymbol{\Sigma}^\tau + \log(\mathbf{I} + \tau^{-1}\mathbb{E}\mathbf{H}_{\pi,m} + \tau^{-2}\mathbb{E}\mathbf{H}_{\pi,m}^2)\Big\}\\
&\le \operatorname{tr}\exp\Big\{\sum_{j=1}^{m}\log(\mathbf{I} + \tau^{-1}\mathbb{E}\mathbf{H}_{\pi,j} + \tau^{-2}\mathbb{E}\mathbf{H}_{\pi,j}^2) - m\boldsymbol{\Sigma}^\tau\Big\}\\
&\le \operatorname{tr}\exp\Big(\frac{1}{\tau^2}\sum_{j=1}^{m}\mathbb{E}\mathbf{H}_{\pi,j}^2\Big)\\
&\le d\exp(m\tau^{-2}\|\mathbb{E}\mathbf{H}_{\pi,1}^2\|_2)\\
&= d\exp(m\tau^{-2}v^2),
\end{aligned}$$

where we used the bound $\operatorname{tr} e^{\mathbf{A}} \le de^{\|\mathbf{A}\|}$ in the last inequality and the definition v^2 from (3.7) in the last equality.

Letting $\tau = 2v^2\sqrt{m}/y$, we get

$$\mathbb{P}\{\sqrt{m}\,\lambda_{\max}(\widehat{\boldsymbol{\Sigma}}_2^{\mathcal{T}} - \boldsymbol{\Sigma}) \ge y\} \le d\exp\Big(-\frac{y\sqrt{m}}{\tau} + \frac{mv^2}{\tau^2}\Big) \le de^{-y^2/(4v^2)}.$$

Similarly, it can be shown that

$$\mathbb{P}\{\sqrt{m}\,\lambda_{\min}(\widehat{\boldsymbol{\Sigma}}_2^{\mathcal{T}} - \boldsymbol{\Sigma}) \le -y\} \le de^{-y^2/(4v^2)}.$$

Finally, taking $y = 2v\sqrt{\log(2d) + \log\delta^{-1}}$ in the last two displays proves (3.9). □

B.3 Proof of Theorem 3.3

Let $v_{\max} = \max_{1\le k,\ell\le d} v_{k\ell}$. By the union bound, for any $y > 0$ it holds

$$\begin{aligned}
&\mathbb{P}(\|\widehat{\boldsymbol{\Sigma}}_1^{\mathcal{H}} - \boldsymbol{\Sigma}\|_{\max} \ge v_{\max}\,y)\\
&\le \sum_{1\le k\le\ell\le d}\mathbb{P}(|\widehat{\sigma}_{1,k\ell}^{\mathcal{H}} - \sigma_{k\ell}| \ge v_{k\ell}\,y) \le \frac{d(d+1)}{2}\max_{1\le k\le\ell\le d}\mathbb{P}(|\widehat{\sigma}_{1,k\ell}^{\mathcal{H}} - \sigma_{k\ell}| \ge v_{k\ell}\,y).
\end{aligned}\tag{B.4}$$

In the rest of the proof, we fix $(k,\ell) \in [d]\times[d]$ and write $\tau = \tau_{k\ell}$ and $v = v_{k\ell}$ for simplicity. Moreover, define the index set $\mathcal{I} = \{(i,j) : 1 \le i < j \le n\}$, the collection $\{Z_{i,j} = (X_{ik} - X_{jk})(X_{i\ell} - X_{j\ell})/2 : (i,j) \in \mathcal{I}\}$ of random variables indexed by $\mathcal{I}$ and the loss function $\mathcal{L}(\theta) = \sum_{(i,j)\in\mathcal{I}}\ell_\tau(Z_{i,j} - \theta)$. With this notation, we have

$$\widehat{\sigma}_{1,k\ell}^{\mathcal{H}} = \widehat{\theta} := \operatorname*{argmin}_{\theta\in\mathbb{R}}\mathcal{L}(\theta).$$

Without loss of generality, we assume $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_d)^\intercal = \mathbf{0}$; otherwise, we can simply replace X_{ik} by $X_{ik} - \mu_k$ for all $i \in [n]$ and $k \in [d]$.

Note that $\widehat{\theta}$ is the unique solution of the equation

$$\Psi(\theta) := \frac{1}{\binom{n}{2}}\sum_{(i,j)\in\mathcal{I}}\psi_\tau(Z_{i,j} - \theta) = 0,$$

where $\psi_\tau(\cdot)$ is defined in (3.1). Similarly to the proof of Theorem 3.1, we define

$$w_{(1,\ldots,n)}(\theta) = \frac{1}{m\tau}\{\psi_\tau(Z_{1,2}-\theta)+\psi_\tau(Z_{3,4}-\theta)+\ldots+\psi_\tau(Z_{2m-1,2m}-\theta)\}.$$

Denote by $\mathcal{P}$ the class of all $n!$ permutations on $[n]$ and let $\pi = (i_1,\ldots,i_n)$ be a permutation, i.e., $\pi(j) = i_j$ for $j = 1,\ldots,n$. Put $w_\pi(\theta) = w_{(i_1,\ldots,i_n)}(\theta)$ for $\pi \in \mathcal{P}$, such that $\tau^{-1}m\Psi(\theta) = (n!)^{-1}\sum_{\pi\in\mathcal{P}} mw_\pi(\theta)$. By convexity, we have

$$\mathbb{E}\{e^{\tau^{-1}m\Psi(\theta)}\} \le \frac{1}{n!}\sum_{\pi\in\mathcal{P}}\mathbb{E}\{e^{mw_\pi(\theta)}\}.$$

Recall that $\mathbb{E}Z_{i,j} = \sigma_{k\ell}$ for any $(i,j)\in\mathcal{I}$. By (3.14),

$$v^2 = \operatorname{var}(Z_{1,2}) = \frac{1}{2}\{\mathbb{E}((X_k-\mu_k)^2(X_\ell-\mu_\ell)^2)+\sigma_{kk}\sigma_{\ell\ell}\}.$$

For $\pi = (1,\ldots,n)$, by (B.2) and the fact that $\tau^{-1}\psi_\tau(x) = \psi_1(x/\tau)$, we have

$$\begin{aligned}
\mathbb{E}\{e^{mw_\pi(\theta)}\} &= \prod_{j=1}^m \mathbb{E}\exp\{\psi_1((Z_{2j-1,2j}-\theta)/\tau)\}\\
&\le \prod_{j=1}^m \mathbb{E}\{1+\tau^{-1}(Z_{2j-1,2j}-\theta)+\tau^{-2}(Z_{2j-1,2j}-\theta)^2\}\\
&\le \prod_{j=1}^m [1+\tau^{-1}(\sigma_{k\ell}-\theta)+\tau^{-2}\{v^2+(\sigma_{k\ell}-\theta)^2\}]\\
&\le \exp[m\tau^{-1}(\sigma_{k\ell}-\theta)+m\tau^{-2}\{v^2+(\sigma_{k\ell}-\theta)^2\}]. \qquad \text{(B.5)}
\end{aligned}$$

Similarly, it can be shown that

$$\mathbb{E}\{-e^{mw_\pi(\theta)}\} \le \exp[-m\tau^{-1}(\sigma_{k\ell}-\theta)+m\tau^{-2}\{v^2+(\sigma_{k\ell}-\theta)^2\}]. \qquad \text{(B.6)}$$

Inequalities (B.5) and (B.6) hold for every permutation $\pi\in\mathcal{P}$. For $\eta\in(0,1)$, define

$$\begin{aligned}
B_+(\theta) &= \sigma_{k\ell}-\theta+\frac{v^2+(\sigma_{k\ell}-\theta)^2}{\tau}+\frac{\tau\log\eta^{-1}}{m},\\
B_-(\theta) &= \sigma_{k\ell}-\theta-\frac{v^2+(\sigma_{k\ell}-\theta)^2}{\tau}-\frac{\tau\log\eta^{-1}}{m}.
\end{aligned}$$

Together, (B.5), (B.6) and Markov's inequality imply

$$\begin{aligned}
&\mathbb{P}\{\Psi(\theta) > B_+(\theta)\} \le e^{-\tau^{-1}mB_+(\theta)}\mathbb{E}\{e^{\tau^{-1}m\Psi(\theta)}\} \le \eta,\\
\text{and }\ &\mathbb{P}\{\Psi(\theta) < B_-(\theta)\} \le e^{-\tau^{-1}mB_-(\theta)}\mathbb{E}\{-e^{\tau^{-1}m\Psi(\theta)}\} \le \eta.
\end{aligned}$$

Recall that $\Psi(\widehat{\theta}) = 0$. Let θ_+ be the smallest solution of the quadratic equation $B_+(\theta_+) = 0$, and θ_- be the largest solution of the equation $B_-(\theta_-) = 0$. Noting that $\Psi(\cdot)$ is decreasing, it follows from the last display that

$$\mathbb{P}(\theta_- \le \widehat{\theta} \le \theta_+) \ge 1-2\eta.$$

Similarly to the proof of Proposition 2.4 in Catoni (2012), it can be shown that with $\tau = \nu\sqrt{m/\log\eta^{-1}}$,

$$\theta_+ \le \sigma_{k\ell} + 2\Big(\frac{\nu^2}{\tau} + \frac{\tau\log\eta^{-1}}{m}\Big) \quad\text{and}\quad \theta_- \ge \sigma_{k\ell} - 2\Big(\frac{\nu^2}{\tau} + \frac{\tau\log\eta^{-1}}{m}\Big)$$

as long as $m \ge 8\log\eta^{-1}$. Consequently, we obtain that with probability at least $1-2\eta$, $|\widehat{\sigma}^{\mathcal{H}}_{1,k\ell} - \sigma_{k\ell}| \le 4\nu\sqrt{(\log\eta^{-1})/m}$.

Taking $y = 4\sqrt{(\log\eta^{-1})/m}$ in (B.4) yields $\|\widehat{\mathbf{\Sigma}}^{\mathcal{H}}_1 - \mathbf{\Sigma}\|_{\max} \le 4\nu_{\max}\sqrt{(\log\eta^{-1})/m}$ with probability at least $1 - d(d+1)\eta$. Finally, taking $\delta = d^2\eta$ proves (3.16). □

B.4 Proof of Theorem 3.5

We will use Theorem 1 and Lemma 1 in Minsker and Strawn (2017) that connect the performance of $\widehat{\sigma}^{\text{MOM}}_{\ell m}$ to the rate of convergence of $\widehat{\sigma}^{(1)}_{\ell m}$ to the normal law. It is well known that, whenever the 4th moments of the entries of $\boldsymbol{X}$ are finite, $\sqrt{|G_1|}\frac{\widehat{\sigma}^{(1)}_{\ell m} - \sigma_{\ell m}}{\Delta_{\ell m}}$ converges in distribution to the standard normal distribution. The rate of this convergence can be obtained via an analogue of the Berry-Esseen theorem for the sample covariance. Specifically, for any $1 \le \ell, m \le d$, we seek an upper bound on

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\bigg(\sqrt{|G_1|}\frac{\widehat{\sigma}^{(1)}_{\ell m} - \sigma_{\ell m}}{\Delta_{\ell m}} \le t\bigg) - \mathbb{P}(Z \le t)\right|,$$

where $Z \sim \mathcal{N}(0,1)$. To this end, we will use Theorem 2.9 in Pinelis and Molzon (2016). Using the notation therein, we take $\boldsymbol{V} = (X_\ell - \mathbb{E}X_\ell, X_m - \mathbb{E}X_m, X_\ell X_m - \mathbb{E}(X_\ell X_m))^\intercal$, $f(x_1, x_2, x_3) = x_3 - x_1\cdot x_2$, and deduce that

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\bigg(\sqrt{|G_1|}\frac{\widehat{\sigma}^{(1)}_{\ell m} - \sigma_{\ell m}}{\Delta_{\ell m}} \le t\bigg) - \mathbb{P}(Z \le t)\right| \le \frac{\mathfrak{C}_{\ell m}}{\sqrt{|G_1|}}, \tag{B.7}$$

where $\mathfrak{C}_{\ell m} > 0$ is a constant depending on $\Delta_{\ell m}$ and $\mathbb{E}|(X_\ell - \mathbb{E}X_\ell)(X_m - \mathbb{E}X_m)|^3$. Together with Theorem 1 and Lemma 1 of Minsker and Strawn (2017), (B.7) implies that

$$\big|\widehat{\sigma}^{\text{MOM}}_{\ell m} - \sigma_{\ell m}\big| \le 3\Delta_{\ell m}\sqrt{\frac{k}{n}}\left(\sqrt{\frac{s}{k}} + \mathfrak{C}_{\ell m}\sqrt{\frac{k}{n}}\right)$$

with probability at least $1 - 4e^{-2s}$ for all $s > 0$ satisfying

$$\sqrt{\frac{s}{k}} + \mathfrak{C}_{\ell m}\sqrt{\frac{k}{n}} \le 0.33. \tag{B.8}$$

Taking the union bound over all ℓ, m, we obtain that with probability at least $1 - 2d(d+1)e^{-2s}$,

$$\|\widehat{\mathbf{\Sigma}}^{\text{MOM}} - \mathbf{\Sigma}\|_{\max} \le 3\max_{\ell,m}\Delta_{\ell m}\left(\sqrt{\frac{s}{n}} + \max_{\ell,m}\mathfrak{C}_{\ell m}\frac{k}{n}\right)$$

for all $s > 0$ satisfying (B.8). The latter is equivalent to the statement of the theorem. □

B.5 Proof of Corollary 3.1

From the proof of Theorem 3.2, we find that

$$\|\mathbb{E}\{(\boldsymbol{X}_1-\boldsymbol{X}_2)(\boldsymbol{X}_1-\boldsymbol{X}_2)^\intercal\}^2\|_2 = 2\|\mathbb{E}\{(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^\intercal\}^2 + \mathrm{Tr}(\boldsymbol{\Sigma})\boldsymbol{\Sigma} + 2\boldsymbol{\Sigma}^2\|_2.$$

Under the bounded kurtosis condition that $K = \sup_{\boldsymbol{u}\in\mathbb{S}^{d-1}} \mathrm{kurt}(\boldsymbol{u}^\intercal\boldsymbol{X}) < \infty$, it follows from Lemma 4.1 in Minsker and Wei (2018) that

$$\|\mathbb{E}\{(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^\intercal\}^2\|_2 \le K\mathrm{Tr}(\boldsymbol{\Sigma})\|\boldsymbol{\Sigma}\|_2.$$

Together, the last two displays imply

$$\|\mathbb{E}\{(\boldsymbol{X}_1-\boldsymbol{X}_2)(\boldsymbol{X}_1-\boldsymbol{X}_2)^\intercal\}^2\|_2 \le 2\|\boldsymbol{\Sigma}\|_2\{(K+1)\mathrm{Tr}(\boldsymbol{\Sigma}) + 2\|\boldsymbol{\Sigma}\|_2\}.$$

Taking $v = \|\mathbb{E}\{(\boldsymbol{X}_1-\boldsymbol{X}_2)(\boldsymbol{X}_1-\boldsymbol{X}_2)^\intercal\}^2\|_2^{1/2}/2$ that scales with $\mathrm{Tr}(\boldsymbol{\Sigma})^{1/2}\|\boldsymbol{\Sigma}\|_2^{1/2} = \mathrm{r}(\boldsymbol{\Sigma})^{1/2}\|\boldsymbol{\Sigma}\|_2$, the resulting estimator satisfies

$$\|\widehat{\boldsymbol{\Sigma}}_2^{\mathcal{T}} - \boldsymbol{\Sigma}\|_2 \lesssim K^{1/2}\|\boldsymbol{\Sigma}\|_2\sqrt{\frac{\mathrm{r}(\boldsymbol{\Sigma})(\log d + t)}{n}} \tag{B.9}$$

with probability at least $1-e^{-t}$. □

C Proofs for Section 5

C.1 Proof of Theorem 5.1

Define each principal submatrix of $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma}^{(p,q)} = \mathbb{E}\boldsymbol{Z}_1^{(p,q)}\boldsymbol{Z}_1^{(p,q)\intercal}/2$, which is estimated by $\widehat{\boldsymbol{\Sigma}}_2^{(p,q),\mathcal{T}}$. As a result, we expect the final estimator $\widehat{\boldsymbol{\Sigma}}_q$ to be close to

$$\boldsymbol{\Sigma}_q = \sum_{j=-1}^{\lceil (d-1)/q\rceil} \mathbf{E}_{jq+1}^d(\boldsymbol{\Sigma}^{(jq+1,2q)}) - \sum_{j=0}^{\lceil (d-1)/q\rceil} \mathbf{E}_{jq+1}^d(\boldsymbol{\Sigma}^{(jq+1,q)}).$$

By the triangle inequality, we have $\|\widehat{\boldsymbol{\Sigma}}_q - \boldsymbol{\Sigma}\|_2 \le \|\widehat{\boldsymbol{\Sigma}}_q - \boldsymbol{\Sigma}_q\|_2 + \|\boldsymbol{\Sigma}_q - \boldsymbol{\Sigma}\|_2$. We first establish an upper bound for the bias term $\|\boldsymbol{\Sigma}_q - \boldsymbol{\Sigma}\|_2$. According to the decomposition illustrated by Figure 2, $\boldsymbol{\Sigma}_q$ is a banded version of the population covariance with bandwidth between q and $2q$. Therefore, we bound the spectral norm of $\boldsymbol{\Sigma}_q - \boldsymbol{\Sigma}$ by the $\|\cdot\|_{1,1}$ norm as follows:

$$\|\boldsymbol{\Sigma}_q - \boldsymbol{\Sigma}\|_2 \le \max_{1\le \ell\le d} \sum_{k:|k-\ell|>q} |\sigma_{k\ell}| \le \frac{M}{q^\alpha}.$$

It remains to control the estimation error $\|\widehat{\boldsymbol{\Sigma}}_q - \boldsymbol{\Sigma}_q\|_2$. Define $\mathbf{D}^{(p,q)} = \widehat{\boldsymbol{\Sigma}}_2^{(p,q),\mathcal{T}} - \boldsymbol{\Sigma}^{(p,q)}$,

$$\mathbf{S}_1 = \sum_{j=-1:j\text{ is odd}}^{\lceil (d-1)/q\rceil} \mathbf{E}_{jq+1}^d\{\mathbf{D}^{(jq+1,2q)}\}, \quad \mathbf{S}_2 = \sum_{j=0:j\text{ is even}}^{\lceil (d-1)/q\rceil} \mathbf{E}_{jq+1}^d\{\mathbf{D}^{(jq+1,2q)}\},$$

and $\mathbf{S}_3 = \sum_{j=0}^{\lceil (d-1)/q\rceil} \mathbf{E}_{jq+1}^d\{\mathbf{D}^{(jq+1,q)}\}$. Note that each $\mathbf{S}_i$ above is a sum of disjoint block diagonal matrices. Therefore,

$$\begin{aligned}\|\widehat{\boldsymbol{\Sigma}}_q - \boldsymbol{\Sigma}_q\|_2 &\le \|\mathbf{S}_1\|_2 + \|\mathbf{S}_3\|_2 + \|\mathbf{S}_3\|_2 \\ &\le 3\max_{j=-1}^{\lceil (d-1)/q\rceil}\{\|\mathbf{D}^{(jq+1,2q)}\|_2, \|\mathbf{D}^{(jq+1,q)}\|_2\}.\end{aligned} \tag{C.1}$$

Applying Theorem 3.2 to each principal submatrix with the choice of $\delta = (n^{c_0}d)^{-1}$ in τ, and by the union bound, we obtain that with probability at least $1 - 2d\delta = 1 - 2n^{-c_0}$,

$$\begin{aligned}
&\max_{j=-1}^{\lceil (d-1)/q\rceil}\{\|\mathbf{D}^{(jq+1,2q)}\|_2, \|\mathbf{D}^{(jq+1,q)}\|_2\}\\
&\le 2\|\boldsymbol{\Sigma}\|_2^{1/2}\{(M_1+1)q\|\boldsymbol{\Sigma}\|_2 + \|\boldsymbol{\Sigma}\|_2\}^{1/2}\sqrt{\frac{\log(4q)+\log\delta^{-1}}{m}}\\
&\le 2M_0\sqrt{1+(M_1+1)q}\sqrt{\frac{\log(4d)+c_0\log(nd)}{n}},
\end{aligned}$$

where we used the inequalities $\operatorname{tr}(\mathbf{D}^{(jq+1,2q)}) \le 2q\|\boldsymbol{\Sigma}\|_2$ and $\|\boldsymbol{\Sigma}\|_2 \le M_0$. Plugging this into (C.1), we obtain that with probability at least $1 - 2n^{-c_0}$,

$$\|\widehat{\boldsymbol{\Sigma}}_q - \boldsymbol{\Sigma}_q\|_2 \le 6M_0\sqrt{1+(M_1+1)q}\sqrt{\frac{\log(4d)+c_0\log(nd)}{n}}.$$

In view of the upper bounds on $\|\widehat{\boldsymbol{\Sigma}}_q - \boldsymbol{\Sigma}_q\|_2$ and $\|\boldsymbol{\Sigma}_q - \boldsymbol{\Sigma}\|_2$, the optimal bandwidth q is of order $\{n/\log(nd)\}^{1/(2\alpha+1)} \wedge d$, which leads to the desired result. □

C.2 Proof of Theorem 5.3

Define the symmetrized Bregman divergence for the loss function $\mathcal{L}(\boldsymbol{\Theta}) = \langle \boldsymbol{\Theta}^2, \widehat{\boldsymbol{\Sigma}}_1^{\mathcal{T}}\rangle - \operatorname{tr}(\boldsymbol{\Theta})$ as $D^{\mathrm{s}}_{\mathcal{L}}(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) = \langle \nabla\mathcal{L}(\boldsymbol{\Theta}_1) - \nabla\mathcal{L}(\boldsymbol{\Theta}_2), \boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\rangle$. We first need the following two lemmas.

Lemma C.1. Provided $\lambda \ge 2\|\nabla\mathcal{L}(\boldsymbol{\Theta}^*)\|_{\max}$, $\widehat{\boldsymbol{\Theta}}$ falls in the ℓ_1-cone

$$\|\widehat{\boldsymbol{\Theta}}_{\mathcal{S}^{\mathrm{c}}} - \boldsymbol{\Theta}^*_{\mathcal{S}^{\mathrm{c}}}\|_{\ell_1} \le 3\|\widehat{\boldsymbol{\Theta}}_{\mathcal{S}} - \boldsymbol{\Theta}^*_{\mathcal{S}}\|_{\ell_1}.$$

Proof of Lemma C.1. Set $\widehat{\boldsymbol{\Gamma}} = (\widehat{\Gamma}_{k\ell})_{1\le k,\ell\le d} \in \mathbb{R}^{d\times d}$, where $\widehat{\Gamma}_{k\ell} = \partial|\widehat{\Theta}_{k\ell}| \in [1,1]$ whenever $k \ne \ell$, and $\widehat{\Gamma}_{k\ell} = 0$ whenever $k = \ell$. Here $\partial f(x_0)$ denotes the subdifferential of f at x_0. By the convexity of the loss function and the optimality condition, we have

$$\begin{aligned}
0 &\le \langle \nabla\mathcal{L}(\widehat{\boldsymbol{\Theta}}) - \nabla\mathcal{L}(\boldsymbol{\Theta}^*), \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\rangle\\
&= \langle -\lambda\widehat{\boldsymbol{\Gamma}} - \nabla\mathcal{L}(\boldsymbol{\Theta}^*), \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\rangle\\
&= -\langle \lambda\widehat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\rangle - \langle \nabla\mathcal{L}(\boldsymbol{\Theta}^*), \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\rangle\\
&\le -\lambda\|\widehat{\boldsymbol{\Theta}}_{\mathcal{S}^{\mathrm{c}}} - \boldsymbol{\Theta}^*_{\mathcal{S}^{\mathrm{c}}}\|_{\ell_1} + \lambda\|\widehat{\boldsymbol{\Theta}}_{\mathcal{S}} - \boldsymbol{\Theta}^*_{\mathcal{S}}\|_{\ell_1} + \frac{\lambda}{2}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\ell_1}.
\end{aligned}$$

Rearranging terms proves the stated result. □

Lemma C.2. Under the restricted eigenvalue condition, it holds

$$D^{\mathrm{s}}_{\mathcal{L}}(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}^*) \ge \kappa_-\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\mathrm{F}}^2.$$

Proof. We use $\operatorname{vec}(\mathbf{A})$ to denote the vectorized form of matrix $\mathbf{A}$. Let $\boldsymbol{\Delta} = \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*$. Then by the mean value theorem, there exists a $\gamma \in [0,1]$ such that

$$\begin{aligned}
D^{\mathrm{s}}_{\mathcal{L}}(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}^*) &= \langle \nabla\mathcal{L}(\widehat{\boldsymbol{\Theta}}) - \nabla\mathcal{L}(\boldsymbol{\Theta}^*), \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\rangle\\
&= \operatorname{vec}(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*)^{\intercal}\nabla^2\mathcal{L}(\widehat{\boldsymbol{\Theta}} + \gamma\boldsymbol{\Delta})\operatorname{vec}(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*)\\
&\ge \kappa_-\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2,
\end{aligned}$$

where the last step is due to the restricted eigenvalue condition and Lemma C.1. This completes the proof. □

Applying Lemma C.2 gives

$$\kappa_{-}\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathrm{F}}^2 \leq \langle \nabla \mathcal{L}(\widehat{\mathbf{\Theta}}) - \nabla \mathcal{L}(\mathbf{\Theta}^*), \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^* \rangle. \tag{C.2}$$

Next, note that the sub-differential of the norm $\|\cdot\|_{\ell_1}$ evaluated at $\mathbf{\Psi} = (\Psi_{k\ell})_{1\leq k,\ell\leq d}$ consists the set of all symmetric matrices $\mathbf{\Gamma} = (\Gamma_{k\ell})_{1\leq k,\ell\leq d}$ such that $\Gamma_{k\ell} = 0$ if $k = \ell$, $\Gamma_{k\ell} = \mathrm{sign}(\Psi_{k\ell})$ if $k \neq \ell$ and $\Psi_{k\ell} \neq 0$, $\Gamma_{k\ell} \in [-1, +1]$ if $k \neq \ell$ and $\Psi_{k\ell} = 0$. Then by the Karush-Kuhn-Tucker conditions, there exists some $\widehat{\mathbf{\Gamma}} \in \partial\|\widehat{\mathbf{\Theta}}\|_{\ell_1}$ such that

$$\nabla \mathcal{L}(\widehat{\mathbf{\Theta}}) + \lambda \widehat{\mathbf{\Gamma}} = \mathbf{0}.$$

Plugging the above equality into (C.2) and rearranging terms, we obtain

$$\kappa_{-}\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathrm{F}}^2 + \underbrace{\langle \nabla \mathcal{L}(\mathbf{\Theta}^*), \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^* \rangle}_{\mathrm{I}} + \underbrace{\langle \lambda \widehat{\mathbf{\Gamma}}, \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^* \rangle}_{\mathrm{II}} \leq 0. \tag{C.3}$$

We bound terms I and II separately, starting with I. Our first observation is

$$\nabla \mathcal{L}(\mathbf{\Theta}^*) = (\mathbf{\Theta}^* \widehat{\mathbf{\Sigma}}_1^{\mathcal{T}} - \mathbf{I})/2 + (\widehat{\mathbf{\Sigma}}_1^{\mathcal{T}} \mathbf{\Theta}^* - \mathbf{I})/2.$$

By Theorem 3.1, we obtain that with probability at least $1 - 2\delta$,

$$\|\nabla \mathcal{L}(\mathbf{\Theta}^*)\|_{\max} \leq \|\mathbf{\Theta}^*\|_{1,1} \|\widehat{\mathbf{\Sigma}}_1^{\mathcal{T}} - \mathbf{\Sigma}\|_{\max} \leq 2M\|\mathbf{V}\|_{\max} \sqrt{\frac{2\log d + \log \delta^{-1}}{\lfloor n/2 \rfloor}} \leq \lambda/2.$$

Let $\mathcal{S}$ be the support of nonzero elements of $\mathbf{\Theta}^*$ and $\mathcal{S}^{\mathrm{c}}$ be its complement with respect to the full index set $\{(k, \ell) : 1 \leq k, \ell \leq d\}$. For term I, separating the support of $\nabla \mathcal{L}(\mathbf{\Theta}^*)$ and $\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*$ to $\mathcal{S}$ and $\mathcal{S}^{\mathrm{c}}$ and applying the matrix Hölder inequality, we obtain

$$\begin{aligned}
\langle \nabla \mathcal{L}(\mathbf{\Theta}^*), \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^* \rangle &= \langle (\nabla \mathcal{L}(\mathbf{\Theta}^*))_{\mathcal{S}}, (\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)_{\mathcal{S}} \rangle + \langle (\nabla \mathcal{L}(\mathbf{\Theta}^*))_{\mathcal{S}^{\mathrm{c}}}, (\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)_{\mathcal{S}^{\mathrm{c}}} \rangle \\
&\geq -\|(\nabla \mathcal{L}(\mathbf{\Theta}^*))_{\mathcal{S}}\|_{\mathrm{F}} \|(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)_{\mathcal{S}}\|_{\mathrm{F}} - \|(\nabla \mathcal{L}(\mathbf{\Theta}^*))_{\mathcal{S}^{\mathrm{c}}}\|_{\mathrm{F}} \|(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)_{\mathcal{S}^{\mathrm{c}}}\|_{\mathrm{F}}.
\end{aligned}$$

For term II, separating the support of $\lambda \widehat{\mathbf{\Gamma}}$ and $\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*$ to $\mathcal{S}$ and $\mathcal{S}^{\mathrm{c}}$, we have

$$\langle \lambda \widehat{\mathbf{\Gamma}}, \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^* \rangle = \langle \lambda \widehat{\mathbf{\Gamma}}_{\mathcal{S}}, (\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)_{\mathcal{S}} \rangle + \langle \lambda \widehat{\mathbf{\Gamma}}_{\mathcal{S}^{\mathrm{c}}}, (\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)_{\mathcal{S}^{\mathrm{c}}} \rangle. \tag{C.4}$$

Let $1_{\mathcal{A}} \in \mathbb{R}^{d\times d}$ be a d-by-d matrix such that $1_{k\ell} = 1$ if $(k, \ell) \in \mathcal{A}$, $1_{k\ell} = 0$ otherwise. For the last term in the above equality, we have

$$\langle \lambda \widehat{\mathbf{\Gamma}}_{\mathcal{S}^{\mathrm{c}}}, (\widehat{\mathbf{\Theta}} - \mathbf{\Theta})_{\mathcal{S}^{\mathrm{c}}} \rangle = \langle \lambda \cdot 1_{\mathcal{S}^{\mathrm{c}}}, |\widehat{\mathbf{\Theta}}_{\mathcal{S}^{\mathrm{c}}}| \rangle = \langle \lambda \cdot 1_{\mathcal{S}^{\mathrm{c}}}, |(\widehat{\mathbf{\Theta}} - \mathbf{\Theta})_{\mathcal{S}^{\mathrm{c}}}| \rangle. \tag{C.5}$$

Plugging (C.5) into (C.4) and applying the matrix Hölder inequality yields

$$\begin{aligned}
\langle \lambda \widehat{\mathbf{\Gamma}}, \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^* \rangle &= \langle \lambda \widehat{\mathbf{\Gamma}}_{\mathcal{S}}, (\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)_{\mathcal{S}} \rangle + \langle \lambda \cdot 1_{\mathcal{S}^{\mathrm{c}}}, |(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)_{\mathcal{S}^{\mathrm{c}}}| \rangle \\
&= \langle \lambda \widehat{\mathbf{\Gamma}}_{\mathcal{S}}, (\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)_{\mathcal{S}} \rangle + \|\lambda \cdot 1_{\mathcal{S}^{\mathrm{c}}}\|_{\mathrm{F}} \|(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)_{\mathcal{S}^{\mathrm{c}}}\|_{\mathrm{F}} \\
&\geq -\|\lambda \cdot 1_{\mathcal{S}}\|_{\mathrm{F}} \|(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)_{\mathcal{S}}\|_{\mathrm{F}} + \lambda\sqrt{s} \|(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)_{\mathcal{S}^{\mathrm{c}}}\|_{\mathrm{F}}.
\end{aligned}$$

Plugging the bounds for I and II back into (C.3), we find

$$\begin{aligned}\kappa_-\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathrm{F}}^2 &+ (\|\lambda \cdot 1_{\mathcal{S}^c}\|_{\mathrm{F}} - \|(\nabla\mathcal{L}(\mathbf{\Theta}^*))_{\mathcal{S}^c}\|_{\mathrm{F}})\|(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*)_{\mathcal{S}^c}\|_{\mathrm{F}} \\ &\leq (\|(\nabla\mathcal{L}(\mathbf{\Theta}^*))_{\mathcal{S}}\|_{\mathrm{F}} + \|\lambda \cdot 1_{\mathcal{S}}\|_{\mathrm{F}})\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathrm{F}}.\end{aligned}$$

Since $\|(\nabla\mathcal{L}(\mathbf{\Theta}^*))_{\mathcal{S}^c}\|_{\mathrm{F}} \leq |\mathcal{S}^c|^{1/2}\|\nabla\mathcal{L}(\mathbf{\Theta}^*)\|_{\max} \leq |\mathcal{S}^c|^{1/2}\lambda = \|\lambda \cdot 1_{\mathcal{S}^c}\|_{\mathrm{F}}$, it follows that

$$\kappa_-\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathrm{F}}^2 \leq (\|(\nabla\mathcal{L}(\mathbf{\Theta}^*))_{\mathcal{S}}\|_{\mathrm{F}} + \|\lambda \cdot 1_{\mathcal{S}}\|_{\mathrm{F}})\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathrm{F}}.$$

Canceling $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathrm{F}}$ on both sides yields

$$\kappa_-\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\|_{\mathrm{F}} \leq \|\lambda \cdot 1_{\mathcal{S}}\|_{\mathrm{F}} + \|\nabla\mathcal{L}(\mathbf{\Theta}^*)_{\mathcal{S}}\|_{\mathrm{F}} \leq 3\lambda\sqrt{s}/2$$

under the scaling $\lambda \geq 2\|\nabla\mathcal{L}(\mathbf{\Theta}^*)\|_{\max}$. Plugging λ completes the proof. □

D Robust estimation and inference under factor models

As a complement to the three examples considered in the main text, in this section we discuss robust covariance estimation (Section D.1) and inference (Section D.2) under factor models, which might be of independent interest. In Section D.2, we provide a self-contained analysis to prove the consistency of estimating the false discovery proportion, while there is no such a theoretical guarantee in Fan et al. (2019) without using sample splitting.

D.1 Covariance estimation through factor models

Consider the approximate factor model of the form $\boldsymbol{X} = (X_1, \ldots, X_d)^\intercal = \boldsymbol{\mu} + \mathbf{B}\boldsymbol{f} + \boldsymbol{\varepsilon}$, from which we observe

$$\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id})^\intercal = \boldsymbol{\mu} + \mathbf{B}\boldsymbol{f}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n, \tag{D.1}$$

where $\boldsymbol{\mu}$ is a d-dimensional unknown mean vector, $\mathbf{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d)^\intercal \in \mathbb{R}^{d\times r}$ is the factor loading matrix, $\boldsymbol{f}_i \in \mathbb{R}^r$ is a vector of common factors to the ith observation and is independent of the idiosyncratic noise $\boldsymbol{\varepsilon}_i$. For more details about factor analysis, we refer the readers to Anderson and Rubin (1956), Chamberlain and Rothschild (1983), Bai and Li (2012) and Fan and Han (2017), among others. Factor pricing model has been widely used in financial economics, where X_{ik} is the excess return of fund/asset k at time i, $\boldsymbol{f}_i$'s are the systematic risk factors related to some specific linear pricing model, such as the capital asset pricing model (CAPM) (Sharpe, 1964), and the Fama-French three-factor model (Fama and French, 1993).

Under model (D.1), the covariance matrix of $\boldsymbol{X}$ can be written as

$$\mathbf{\Sigma} = (\sigma_{k\ell})_{1\leq k,\ell\leq d} = \mathbf{B}\mathrm{cov}(\boldsymbol{f})\mathbf{B}^\intercal + \mathbf{\Sigma}_\varepsilon, \tag{D.2}$$

where $\mathbf{\Sigma}_\varepsilon = (\sigma_{\varepsilon,k\ell})_{1\leq k,\ell\leq d}$ denotes the covariance matrix of $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots \varepsilon_d)^\intercal$, which is typically assumed to be sparse. When $\mathbf{\Sigma}_\varepsilon = \mathbf{I}_d$, model (D.1) is known as the strict factor model. To make the model identifiable, following Bai and Li (2012) we assume that $\mathrm{cov}(\boldsymbol{f}) = \mathbf{I}_r$ and that the columns of $\mathbf{B}$ are orthogonal.

We consider the robust estimation of $\boldsymbol{\Sigma}$ based on independent observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from model (D.1). By (D.2) and the identifiability condition, $\boldsymbol{\Sigma}$ is comprised of two components: the low-rank component $\mathbf{B}\mathbf{B}^\intercal$ and the sparse component $\boldsymbol{\Sigma}_\varepsilon$. Using a pilot robust covariance estimator $\widehat{\boldsymbol{\Sigma}}_1^{\mathcal{T}}$ given in (3.3) or $\widehat{\boldsymbol{\Sigma}}_1^{\mathcal{H}}$ given in (3.13), we propose the following robust version of the principal orthogonal complement thresholding (POET) procedure (Fan, Liao and Mincheva, 2013):

(i) Let $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_r$ be the top r eigenvalues of $\widehat{\boldsymbol{\Sigma}}_1^{\mathcal{H}}$ (or $\widehat{\boldsymbol{\Sigma}}_1^{\mathcal{T}}$) with corresponding eigenvectors $\widehat{\boldsymbol{v}}_1, \widehat{\boldsymbol{v}}_2, \ldots, \widehat{\boldsymbol{v}}_r$. Compute the principal orthogonal complement

$$\widehat{\boldsymbol{\Sigma}}_\varepsilon = (\widehat{\sigma}_{\varepsilon,k\ell})_{1\leq k,\ell\leq d} = \widehat{\boldsymbol{\Sigma}}_1^{\mathcal{H}} - \widehat{\mathbf{V}}\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{V}}^\intercal, \tag{D.3}$$

where $\widehat{\mathbf{V}} = (\widehat{\boldsymbol{v}}_1, \ldots, \widehat{\boldsymbol{v}}_r)$ and $\widehat{\boldsymbol{\Lambda}} = \mathrm{diag}\,(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_r)$.

(ii) To achieve sparsity, apply the adaptive thresholding method (Rothman, Levina and Zhu, 2009; Cai and Liu, 2011) to $\widehat{\boldsymbol{\Sigma}}_\varepsilon$ and obtain $\widehat{\boldsymbol{\Sigma}}_\varepsilon^{\mathcal{T}} = (\widehat{\sigma}_{\varepsilon,k\ell}^{\mathcal{T}})_{1\leq k,\ell\leq d}$ such that

$$\widehat{\sigma}_{\varepsilon,k\ell}^{\mathcal{T}} = \begin{cases} \widehat{\sigma}_{\varepsilon,k\ell} & \text{if } k = \ell, \\ s_{k\ell}(\widehat{\sigma}_{\varepsilon,k\ell}) & \text{if } k \neq \ell, \end{cases} \tag{D.4}$$

where $s_{k\ell}(z) = \mathrm{sign}(z)(|z| - \lambda_{k\ell})$, $z \in \mathbb{R}$ is the soft thresholding function with $\lambda_{k\ell} = \lambda(\widehat{\sigma}_{\varepsilon,kk}\,\widehat{\sigma}_{\varepsilon,\ell\ell})^{1/2}$ and $\lambda > 0$ being a regularization parameter.

(iii) Obtain the final estimator of $\boldsymbol{\Sigma}$ as $\widehat{\boldsymbol{\Sigma}} = \widehat{\mathbf{V}}\widehat{\boldsymbol{\Lambda}}\widehat{\mathbf{V}}^\intercal + \widehat{\boldsymbol{\Sigma}}_\varepsilon^{\mathcal{T}}$.

Remark 1. The POET method (Fan, Liao and Mincheva, 2013) employs the sample covariance matrix as an initial estimator and has desirable properties for sub-Gaussian data. For elliptical distributions, Fan, Liu and Wang (2018) proposed to use the marginal Kendall's tau to estimate $\boldsymbol{\Sigma}$, and to use its top r eigenvalues and the spatial Kendall's tau to estimate the corresponding leading eigenvectors. In the above robust POET procedure, we only need to compute one initial estimator of $\boldsymbol{\Sigma}$ and moreover, optimal convergence rates can be achieved in high dimensions under finite fourth moment conditions; see Theorem D.1.

Condition D.1. Under model (D.1), the latent factor $\boldsymbol{f} \in \mathbb{R}^r$ and the idiosyncratic noise $\boldsymbol{\varepsilon} \in \mathbb{R}^d$ are independent. Moreover,

(i) (Identifiability) $\mathrm{cov}(\boldsymbol{f}) = \mathbf{I}_r$ and the columns of $\mathbf{B}$ are orthogonal;

(ii) (Pervasiveness) there exist positive constants c_l, c_u and C_1 such that

$$c_l \leq \min_{1\leq\ell\leq r}\{\lambda_\ell(\mathbf{B}^\intercal\mathbf{B}/d) - \lambda_{\ell+1}(\mathbf{B}^\intercal\mathbf{B}/d)\} \leq c_u \quad \text{with} \quad \lambda_{r+1}(\mathbf{B}^\intercal\mathbf{B}/d) = 0,$$

and $\max\{\|\mathbf{B}\|_{\max}, \|\boldsymbol{\Sigma}_\varepsilon\|_2\} \leq C_1$;

(iii) (Moment condition) $\max_{1\leq\ell\leq d}\mathrm{kurt}(X_\ell) \leq C_2$ for some constant $C_2 > 0$;

(iv) (Sparsity) $\boldsymbol{\Sigma}_\varepsilon$ is sparse in the sense that $s := \max_{1\leq k\leq d}\sum_{\ell=1}^d I(\sigma_{\varepsilon,k\ell} \neq 0)$ satisfies

$$s^2 \log d = o(n) \quad \text{and} \quad s^2 = o(d) \quad \text{as} \quad n, d \to \infty.$$

Theorem D.1. Under Condition D.1, the robust POET estimator with

$$\tau_{k\ell} \asymp \sqrt{n/(\log d)},\ 1 \le k, \ell \le d, \quad \text{and} \quad \lambda \asymp w_{n,d} := \sqrt{\log(d)/n} + d^{-1/2}$$

satisfies

$$\|\widehat{\mathbf{\Sigma}}_{\varepsilon}^{\mathcal{T}} - \mathbf{\Sigma}_{\varepsilon}\|_{\max} = O_{\mathbb{P}}(w_{n,d}), \quad \|\widehat{\mathbf{\Sigma}}_{\varepsilon}^{\mathcal{T}} - \mathbf{\Sigma}_{\varepsilon}\|_{2} = O_{\mathbb{P}}(s w_{n,d}), \tag{D.5}$$

$$\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{\max} = O_{\mathbb{P}}(w_{n,d}) \ \text{ and } \ \|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_{2} = O_{\mathbb{P}}(d w_{n,d}) \tag{D.6}$$

as $n, d \to \infty$.

D.2 Factor-adjusted multiple testing

Here we consider the problem of simultaneously testing the hypotheses

$$H_{0k} : \mu_k = 0 \ \text{ versus } \ H_{1k} : \mu_k \neq 0, \ \text{ for } k = 1, \ldots, d, \tag{D.7}$$

under model (D.1). Although the key implication from the multi-factor pricing theory is that the intercept μ_k should be zero, known as the "mean-variance efficiency" pricing, for any asset k, an important question is whether such a pricing theory can be validated by empirical data. In fact, a very small proportion of μ_k's might be nonzero according to the Berk and Green equilibrium (Berk and Green, 2004). Various statistical methods have been proposed to identify those positive μ_k's (Barras, Scaillet and Wermer, 2010; Fan and Han, 2017; Lan and Du, 2019). These works assume that both the factor and idiosyncratic noise follow multivariate normal distributions. To accommodate the heavy-tailed character of empirical data, we develop a robust multiple testing procedure that controls the overall false discovery rate or false discovery proportion.

For each $1 \le k \le d$, let T_k be a generic test statistic for testing the individual hypothesis $H_{0k} : \mu_k = 0$. For any threshold level $z > 0$, we reject the jth hypothesis whenever $|T_j| \ge z$. The numbers of total discoveries $R(z)$ and false discoveries $V(z)$ are defined by

$$R(z) = \sum_{k=1}^{d} I(|T_k| \ge z) \quad \text{and} \quad V(z) = \sum_{k \in \mathcal{H}_0} I(|T_k| \ge z), \tag{D.8}$$

respectively, where $\mathcal{H}_0 = \{1 \le k \le d : \mu_k = 0\}$. The main object of interest is the false discovery proportion (FDP), given by

$$\mathrm{FDP}(z) = V(z)/R(z).$$

Throughout we use the convention $0/0 = 0$. Note that $R(z)$ is observable given all the test statistics, while $V(z)$ is an unobservable random variable that needs to be estimated. For testing individual hypotheses H_{0k}, the standardized means Z_k, where $Z_k = n^{-1/2} \sum_{i=1}^{n} X_{ik}$, are sensitive to the tails of the sampling distributions. In particular, when the number of features d is large, stochastic outliers from the test statistics Z_k can be so large that they are mistakenly regarded as discoveries. Motivated by recent advances on robust estimation and inference (Catoni, 2012; Zhou et al., 2018), we consider the following robust M-estimator of μ_k:

$$\widehat{\mu}_k = \underset{\theta \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} \ell_{\tau_k}(X_{ik} - \theta) \ \text{ for some } \ \tau_k > 0. \tag{D.9}$$

The corresponding test statistic is then given by $T_k = \sqrt{n}\,\widehat{\mu}_k$ for $k = 1, \dots, d$.

Based on the law of large numbers, we define the approximate FDP by

$$\mathrm{FDP}_{\mathrm{A}}(z) = \frac{1}{R(z)} \sum_{k=1}^{d} \bigg\{ \Phi\bigg(\frac{-z + \sqrt{n}\,\boldsymbol{b}_k^{\intercal} \overline{\boldsymbol{f}}}{\sqrt{\sigma_{kk} - \|\boldsymbol{b}_k\|_2^2}} \bigg) + \Phi\bigg(\frac{-z - \sqrt{n}\,\boldsymbol{b}_k^{\intercal} \overline{\boldsymbol{f}}}{\sqrt{\sigma_{kk} - \|\boldsymbol{b}_k\|_2^2}} \bigg) \bigg\}, \tag{D.10}$$

where $\overline{\boldsymbol{f}} = (1/n) \sum_{i=1}^{n} \boldsymbol{f}_i$. It is shown in the appendix that the approximate FDP in (D.10) serves as a conservative surrogate for the true FDP.

Note that the approximate FDP defined in (D.10) depends on a number of unknown parameters, say $\{\boldsymbol{b}_k, \sigma_{kk}\}_{k=1}^{d}$ and $\overline{\boldsymbol{f}}$. In this section, we describe robust procedures to estimate these quantities using the only observations $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$.

(a) Compute the Huber-type covariance estimator $\widehat{\boldsymbol{\Sigma}}_1^{\mathcal{H}} = (\widehat{\sigma}_{1,k\ell}^{\mathcal{H}})_{1 \le k,\ell \le d}$ (or the truncated estimator $\widehat{\boldsymbol{\Sigma}}_1^{\mathcal{T}}$), and let $\widehat{\lambda}_1 \ge \cdots \ge \widehat{\lambda}_r$ and $\widehat{\boldsymbol{v}}_1, \dots, \widehat{\boldsymbol{v}}_r$ be its top r eigenvalues and the corresponding eigenvectors, respectively.

(b) Compute $\widehat{\mathbf{B}} = (\widehat{\lambda}_1^{1/2} \widehat{\boldsymbol{v}}_1, \dots, \widehat{\lambda}_r^{1/2} \widehat{\boldsymbol{v}}_r) \in \mathbb{R}^{d \times r}$ and $\widehat{\boldsymbol{u}} = \sqrt{n}\,(\widehat{\mathbf{B}}^{\intercal} \widehat{\mathbf{B}})^{-1} \widehat{\mathbf{B}}^{\intercal} \overline{\boldsymbol{X}} \in \mathbb{R}^r$, which serve as estimators of $\mathbf{B}$ and $\sqrt{n}\overline{\boldsymbol{f}}$, respectively. Here $\overline{\boldsymbol{X}} = (1/n) \sum_{i=1}^{n} \boldsymbol{X}_i$.

(c) Denote by $\widehat{\boldsymbol{b}}_1, \dots, \widehat{\boldsymbol{b}}_d$ the d rows of $\widehat{\mathbf{B}}$. For any $z \ge 0$, we estimate the approximate FDP $\mathrm{FDP}_{\mathrm{A}}(z)$ by

$$\widehat{\mathrm{FDP}}_{\mathrm{A}}(z) = \frac{1}{R(z)} \sum_{k=1}^{d} \bigg\{ \Phi\bigg(\frac{-z + \widehat{\boldsymbol{b}}_k^{\intercal} \widehat{\boldsymbol{u}}}{\sqrt{\widehat{\sigma}_{\varepsilon,kk}}} \bigg) + \Phi\bigg(\frac{-z - \widehat{\boldsymbol{b}}_k^{\intercal} \widehat{\boldsymbol{u}}}{\sqrt{\widehat{\sigma}_{\varepsilon,kk}}} \bigg) \bigg\}, \tag{D.11}$$

where $\widehat{\sigma}_{\varepsilon,kk} = \widehat{\sigma}_{1,kk}^{\mathcal{H}} - \|\widehat{\boldsymbol{b}}_k\|_2^2$ for $k = 1, \dots, d$.

The construction of $\widehat{\mathbf{B}}$ is based on the observation that principal component analysis and factor analysis are approximately equivalent under the pervasive assumption in high dimensions (Fan, Liao and Mincheva, 2013). To estimate $\overline{\boldsymbol{f}}$, note from model (D.1) that $\overline{\boldsymbol{X}} = \boldsymbol{\mu} + \mathbf{B}\overline{\boldsymbol{f}} + \overline{\boldsymbol{\varepsilon}}$, where $\boldsymbol{\mu}$ is assumed to be sparse and therefore is ignored for simplicity.

Theorem D.2. Under model (D.1), assume that $\boldsymbol{f}$ and $\boldsymbol{\varepsilon}$ are independent zero-mean random vectors and let $s_1 = \|\boldsymbol{\mu}\|_0$. Assume (i)–(iii) of Condition D.1 hold, and that (n, d, s_1) satisfies $\log d = o(n)$ and $n s_1 = o(d)$ as $n, d \to \infty$. Then for any $z \ge 0$,

$$\widehat{\mathrm{FDP}}_{\mathrm{A}}(z) / \mathrm{FDP}_{\mathrm{A}}(z) \xrightarrow{\mathbb{P}} 1 \quad \text{as} \quad n, d \to \infty. \tag{D.12}$$

D.3 Proof of Theorem D.1

The proof is based on Theorem 2.1 and (A.1) in Fan, Liu and Wang (2018), which provides high level results for the generic POET procedure. To that end, it suffices to show that with properly chosen $\tau_{k\ell}$,

$$\|\widehat{\boldsymbol{\Sigma}}_1^{\mathcal{H}} - \boldsymbol{\Sigma}\|_{\max} = O_{\mathbb{P}}\{n^{-1/2} (\log d)^{1/2}\}, \quad \max_{1 \le \ell \le r} |\widehat{\lambda}_\ell / \lambda_\ell - 1| = O_{\mathbb{P}}\{n^{-1/2} (\log d)^{1/2}\} \tag{D.13}$$

$$\text{and} \ \max_{1 \le \ell \le r} \|\widehat{\boldsymbol{v}}_\ell - \boldsymbol{v}_\ell\|_\infty = O_{\mathbb{P}}\{(nd)^{-1/2} (\log d)^{1/2}\}, \tag{D.14}$$

where $\lambda_1 \geq \cdots \geq \lambda_r$ are the top r eigenvalues of $\boldsymbol{\Sigma}$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r$ are the corresponding eigenvectors.

First, applying Theorem 3.3 with $\tau_{k\ell} \asymp \sqrt{n/(\log d)}$ implies that $\|\widehat{\boldsymbol{\Sigma}}_1^{\mathcal{H}} - \boldsymbol{\Sigma}\|_{\max} \lesssim n^{-1/2}(\log d)^{1/2}$ with probability at least $1-d^{-1}$. This verifies the first criterion in (D.13). Next, by Weyl's inequality and the inequality $\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_{1,1}$ for symmetric matrices, we have

$$\max_{1\leq\ell\leq r} |\widehat{\lambda}_\ell - \lambda_\ell| \leq \|\widehat{\boldsymbol{\Sigma}}_1^{\mathcal{H}} - \boldsymbol{\Sigma}\|_2 \leq \|\widehat{\boldsymbol{\Sigma}}_1^{\mathcal{H}} - \boldsymbol{\Sigma}\|_{1,1} \leq d\|\widehat{\boldsymbol{\Sigma}}_1^{\mathcal{H}} - \boldsymbol{\Sigma}\|_{\max}.$$

Let $\overline{\lambda}_1 \geq \cdots \geq \overline{\lambda}_r$ be the top r eigenvalues of $\mathbf{B}\mathbf{B}^\intercal$, and therefore of $\mathbf{B}^\intercal\mathbf{B}$. Note that, by Weyl's inequality, $\max_{1\leq\ell\leq r} |\lambda_\ell - \overline{\lambda}_\ell| \leq \|\boldsymbol{\Sigma}_\varepsilon\|_2$. It thus follows from Condition D.1 that

$$\min_{1\leq\ell\leq r-1} |\lambda_\ell - \lambda_{\ell+1}| \asymp d \ \text{ and } \ \lambda_r \asymp d \ \text{ as } d \to \infty.$$

Together, the last two displays imply $\max_{1\leq\ell\leq r} |\widehat{\lambda}_\ell/\lambda_\ell - 1| \lesssim n^{-1/2}(\log d)^{1/2}$ with probability at least $1 - d^{-1}$. Therefore, the second criterion in (D.13) is fulfilled.

For (D.14), applying Theorem 3 and Proposition 3 in Fan, Wang and Zhong (2018) we arrive at

$$\begin{aligned}&\max_{1\leq\ell\leq r} \|\widehat{\boldsymbol{v}}_\ell - \boldsymbol{v}_\ell\|_\infty \\ &\lesssim d^{-3/2}(r^4\|\widehat{\boldsymbol{\Sigma}}_1^{\mathcal{H}} - \boldsymbol{\Sigma}\|_{\infty,\infty} + r^{3/2}\|\widehat{\boldsymbol{\Sigma}}_1^{\mathcal{H}} - \boldsymbol{\Sigma}\|_2) \lesssim r^4 d^{-1/2}\|\widehat{\boldsymbol{\Sigma}}_1^{\mathcal{H}} - \boldsymbol{\Sigma}\|_{\max}.\end{aligned}$$

This validates (D.14).

In summary, (D.5) and the first bound in (D.6) follow from Theorem 2.1 and (A.1) in Fan, Liu and Wang (2018), and the second bound in (D.6) follows directly from the fact that $\|\widehat{\boldsymbol{\Sigma}}_1^{\mathcal{H}} - \boldsymbol{\Sigma}\|_2 \leq \|\widehat{\boldsymbol{\Sigma}}_1^{\mathcal{H}} - \boldsymbol{\Sigma}\|_{1,1} \leq d\|\widehat{\boldsymbol{\Sigma}}_1^{\mathcal{H}} - \boldsymbol{\Sigma}\|_{\max}$. □

D.4 Asymptotic property of FDP

In this section, we show that the approximate FDP in (D.10) serves as a conservative surrogate for the true FDP.

Condition D.2. Under model (D.1), $\boldsymbol{f}$ are $\boldsymbol{\varepsilon}$ are independent zero-mean random vectors. (i) $\mathrm{cov}(\boldsymbol{f}) = \mathbf{I}_r$ and $\|\boldsymbol{f}\|_{\psi_2} = \sup_{\boldsymbol{u}\in\mathbb{S}^{r-1}} \|\boldsymbol{u}^\intercal\boldsymbol{f}\|_{\psi_2} \leq C_f$ for some constant $C_f > 0$; (ii) the correlation matrix $\mathbf{R}_\varepsilon = (\varrho_{\varepsilon,k\ell})_{1\leq k,\ell\leq d}$ of $\boldsymbol{\varepsilon}$ satisfies $d^{-2}\sum_{1\leq k\neq\ell\leq d}\varrho_{\varepsilon,k\ell} \leq C_0 d^{-\delta_0}$ for some constants $C_0, \delta_0 > 0$; (iii) $d = d(n) \to \infty$ and $\log d = o(n^{1/2})$ as $n \to \infty$, and $\liminf_{n\to\infty} \frac{d_0}{d} > 0$, where $d_0 = \sum_{k=1}^d I(\mu_k = 0)$; (iv) $C_l \leq \sigma_{\varepsilon,kk} \leq v_k^{1/2} \vee \sigma_{kk} \leq C_u$ for all $1 \leq k \leq d$, where $v_k = \mathbb{E}(\varepsilon_k^4)$ and C_u, C_l are positive constants.

Theorem D.3. Assume that Condition D.2 holds. In (D.9), let $\tau_k = a_k\sqrt{n/\log(nd)}$ with $a_k \geq \sigma_{kk}^{1/2}$ for $k = 1, \ldots, d$. Then, as $n, d \to \infty$,

$$\begin{aligned}\frac{V(z)}{d_0} &= \frac{1}{d_0}\sum_{k\in\mathcal{H}_0}\bigg\{\Phi\bigg(\frac{-z + \sqrt{n}\boldsymbol{b}_k^\intercal\overline{\boldsymbol{f}}}{\sqrt{\sigma_{\varepsilon,kk}}}\bigg) + \Phi\bigg(\frac{z - \sqrt{n}\,\boldsymbol{b}_k^\intercal\overline{\boldsymbol{f}}}{\sqrt{\sigma_{\varepsilon,kk}}}\bigg)\bigg\} \\ &\quad + O_{\mathbb{P}}\bigg[\frac{1}{d^{(\delta_0\wedge 1)/2}} + \frac{\log(nd)}{\sqrt{n}} + \bigg\{\frac{\log(nd)}{n}\bigg\}^{1/4}\bigg] \qquad \text{(D.15)}\end{aligned}$$

and

$$\frac{R(z)}{d} = \frac{1}{d}\sum_{k=1}^{d}\bigg\{\Phi\bigg(\frac{-z+\sqrt{n}(\mu_k+\boldsymbol{b}_k^\intercal\overline{\boldsymbol{f}})}{\sqrt{\sigma_{\varepsilon,kk}}}\bigg)+\Phi\bigg(\frac{-z-\sqrt{n}(\mu_k+\boldsymbol{b}_k^\intercal\overline{\boldsymbol{f}})}{\sqrt{\sigma_{\varepsilon,kk}}}\bigg)\bigg\} + O_{\mathbb{P}}\bigg[\frac{1}{d^{(\delta_0\wedge 1)/2}}+\frac{\log(nd)}{\sqrt{n}}+\bigg\{\frac{\log(nd)}{n}\bigg\}^{1/4}\bigg] \tag{D.16}$$

uniformly over $z \geq 0$. In addition, for any $z \geq 0$ it holds

$$\mathrm{FDP}(z) = \mathrm{FDP}_{\mathrm{orc}}(z) + o_{\mathbb{P}}(1) \ \text{ as } n, d \to \infty, \tag{D.17}$$

where

$$\mathrm{FDP}_{\mathrm{orc}}(z) := \frac{1}{R(z)}\sum_{k\in\mathcal{H}_0}\bigg\{\Phi\bigg(\frac{-z+\sqrt{n}\,\boldsymbol{b}_k^\intercal\overline{\boldsymbol{f}}}{\sqrt{\sigma_{kk}-\|\boldsymbol{b}_k\|_2^2}}\bigg)+\Phi\bigg(\frac{-z-\sqrt{n}\,\boldsymbol{b}_k^\intercal\overline{\boldsymbol{f}}}{\sqrt{\sigma_{kk}-\|\boldsymbol{b}_k\|_2^2}}\bigg)\bigg\}.$$

D.4.1 Preliminaries

To prove Theorem D.3, we need the following results on the robust estimators μ_k's given in (D.9). Define $u_k = X_k - \mu_k = \boldsymbol{b}_k^\intercal\boldsymbol{f} + \varepsilon_k$ for $k = 1, \ldots, d$. Assume that $\mathbb{E}(\boldsymbol{f}) = \mathbf{0}$, $\mathbb{E}(\varepsilon_k) = 0$ and $\boldsymbol{f}$ are ε_k are independent. Then we have $\mathbb{E}(u_k) = 0$ and $\mathbb{E}(u_k^2) = \sigma_{kk} = \|\boldsymbol{b}_k\|_2^2 + \sigma_{\varepsilon,kk}$.

The first lemma is Theorem 5 in Fan, Li and Wang (2017) regarding the concentration of the robust mean estimator.

Lemma D.1. For every $1 \leq k \leq d$ and $t > 0$, the estimator $\widehat{\mu}_k$ in (D.9) with $\tau_k = a_k(n/t)^{1/2}$ for $a_k \geq \sigma_{kk}^{1/2}$ satisfies $|\widehat{\mu}_k - \mu_k| \leq 4a_k(t/n)^{1/2}$ with probability at least $1 - 2e^{-t}$ provided $n \geq 8t$.

The next result provides a nonasymptotic Bahadur representation for $\widehat{\mu}_k$, which follows directly from Lemma D.1 and Theorem 2.1 in Zhou et al. (2018). Let $u_{ik} = \boldsymbol{b}_k^\intercal\boldsymbol{f}_i + \varepsilon_{ik}$ for $i = 1, \ldots, n$ and $k = 1, \ldots, d$.

Lemma D.2. Under the conditions of Lemma D.1, it holds for every $1 \leq k \leq d$ that

$$\bigg|\sqrt{n}\,(\widehat{\mu}_k - \mu_k) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{\tau_k}(u_{ik})\bigg| \leq C\frac{a_k t}{\sqrt{n}} \tag{D.18}$$

with probability at least $1 - 3e^{-t}$ as long as $n \geq 8t$, where $C > 0$ is an absolute constant and $\psi_\tau(\cdot)$ is given in (3.1).

Let τ_k be as in Lemma D.1 and write

$$v_k = \mathbb{E}(\varepsilon_k^4), \quad \xi_k = \psi_k(u_k) \ \text{ for } \ k = 1, \ldots, d. \tag{D.19}$$

Here ξ_k are truncated versions of u_k. The next result shows that the differences between the first two (conditional) moments of u_k and ξ_k given $\boldsymbol{f}$ decay as τ_k grows.

Lemma D.3. Assume that $v_k < \infty$ for $k = 1, \ldots, d$.

1. On the event $\mathcal{G}_k := \{|\boldsymbol{b}_k^\top \boldsymbol{f}| \le \tau_k/2\}$, the following inequalities hold almost surely:

$$|\mathbb{E}_{\boldsymbol{f}}(\xi_k) - \boldsymbol{b}_k^\top \boldsymbol{f}| \le \min(2\tau_k^{-1}\sigma_{\varepsilon,kk}, 8\tau_k^{-3}v_k) \tag{D.20}$$

$$\text{and} \;\; \sigma_{\varepsilon,kk} - 4\tau_k^{-2}(v_k + \sigma_{\varepsilon,kk}^2) \le \mathrm{var}_{\boldsymbol{f}}(\xi_k) \le \sigma_{\varepsilon,kk}, \tag{D.21}$$

   where $\mathbb{E}_{\boldsymbol{f}}(\cdot)$ and $\mathrm{var}_{\boldsymbol{f}}(\cdot)$ denote the conditional mean and variance, separately.

2. On the event $\mathcal{G}_k \cap \mathcal{G}_\ell$, the following holds almost surely:

$$|\mathrm{cov}_{\boldsymbol{f}}(\xi_k, \xi_\ell) - \sigma_{\varepsilon,k\ell}| \le C\frac{v_k \vee v_\ell}{(\tau_k \wedge \tau_\ell)^2}, \tag{D.22}$$

   where $C > 0$ is an absolute constant.

Proof of Lemma D.3. First we prove (D.20) and (D.21). Fix k and let $\tau = \tau_k$ for simplicity. Since ε_k and $\boldsymbol{f}$ are independent, we have

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{f}}\xi_k - \boldsymbol{b}_k^\top \boldsymbol{f}\\ &= -\mathbb{E}_{\boldsymbol{f}}(\varepsilon_k + \boldsymbol{b}_k^\top \boldsymbol{f} - \tau)I(\varepsilon_k > \tau - \boldsymbol{b}_k^\top \boldsymbol{f}) + \mathbb{E}_{\boldsymbol{f}}(-\varepsilon_k - \boldsymbol{b}_k^\top \boldsymbol{f} - \tau)I(\varepsilon_k < -\tau - \boldsymbol{b}_k^\top \boldsymbol{f}).\end{aligned}$$

Therefore, on the event $\mathcal{G}_k$, it holds for any $2 \le q \le 4$ that

$$|\mathbb{E}_{\boldsymbol{f}}\xi_k - \boldsymbol{b}_k^\top \boldsymbol{f}| \le \mathbb{E}_{\boldsymbol{f}}\{|\varepsilon_k| I(|\varepsilon_k| > \tau - |\boldsymbol{b}_k^\top \boldsymbol{f}|)\} \le (\tau - |\boldsymbol{b}_k^\top \boldsymbol{f}|)^{1-q}\,\mathbb{E}(|\varepsilon_k|^q)$$

almost surely. This proves (D.20) by taking q to be 2 and 4. For the conditional variance, by (D.20) and the decomposition $\mathbb{E}_{\boldsymbol{f}}(\xi_k - \boldsymbol{b}_k^\top \boldsymbol{f})^2 = \mathrm{var}_{\boldsymbol{f}}(\xi_k) + (\mathbb{E}_{\boldsymbol{f}}\xi_k - \boldsymbol{b}_k^\top \boldsymbol{f})^2$, we have

$$\mathbb{E}_{\boldsymbol{f}}(\xi_k - \boldsymbol{b}_k^\top \boldsymbol{f})^2 - \frac{\sigma_{\varepsilon,kk}^2}{(\tau - |\boldsymbol{b}_k^\top \boldsymbol{f}|)^2} \le \mathrm{var}_{\boldsymbol{f}}(\xi_k) \le \mathbb{E}_{\boldsymbol{f}}(\xi_k - \boldsymbol{b}_k^\top \boldsymbol{f})^2. \tag{D.23}$$

Note that $\xi_k - \boldsymbol{b}_k^\top \boldsymbol{f}$ can be written as

$$\varepsilon_k I(|\boldsymbol{b}_k^\top \boldsymbol{f} + \varepsilon_k| \le \tau) + (\tau - \boldsymbol{b}_k^\top \boldsymbol{f})I(\boldsymbol{b}_k^\top \boldsymbol{f} + \varepsilon_k > \tau) - (\tau + \boldsymbol{b}_k^\top \boldsymbol{f})I(\boldsymbol{b}_k^\top \boldsymbol{f} + \varepsilon_k < -\tau).$$

It follows that

$$\begin{aligned}&(\xi_k - \boldsymbol{b}_k^\top \boldsymbol{f})^2\\ &= \varepsilon_k^2 I(|\boldsymbol{b}_k^\top \boldsymbol{f} + \varepsilon_k| \le \tau) + (\tau - \boldsymbol{b}_k^\top \boldsymbol{f})^2 I(\boldsymbol{b}_k^\top \boldsymbol{f} + \varepsilon_k > \tau) + (\tau + \boldsymbol{b}_k^\top \boldsymbol{f})^2 I(\boldsymbol{b}_k^\top \boldsymbol{f} + \varepsilon_k < -\tau).\end{aligned}$$

Taking conditional expectations on both sides gives

$$\begin{aligned}\mathbb{E}_{\boldsymbol{f}}(\xi_k - \boldsymbol{b}_k^\top \boldsymbol{f})^2 &= \mathbb{E}(\varepsilon_k^2) - \mathbb{E}_{\boldsymbol{f}}\{\varepsilon_k^2 I(|\boldsymbol{b}_k^\top \boldsymbol{f} + \varepsilon_k| > \tau)\}\\ &\quad + (\tau - \boldsymbol{b}_k^\top \boldsymbol{f})^2 \mathbb{P}_{\boldsymbol{f}}(\varepsilon_k > \tau - \boldsymbol{b}_k^\top \boldsymbol{f}) + (\tau + \boldsymbol{b}_k^\top \boldsymbol{f})^2 \mathbb{P}_{\boldsymbol{f}}(\varepsilon_k < -\tau - \boldsymbol{b}_k^\top \boldsymbol{f}).\end{aligned}$$

Using the identity that $u^2 = 2\int_0^u t\,dt$ for any $u > 0$, we have

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{f}}\{\varepsilon_k^2 I(\boldsymbol{b}_k^\top \boldsymbol{f} + \varepsilon_k > \tau)\}\\ &= 2\mathbb{E}_{\boldsymbol{f}}\int_0^\infty I(\varepsilon_k > t) I(\varepsilon_k > \tau - \boldsymbol{b}_k^\top \boldsymbol{f}) t\,dt\\ &= 2\mathbb{E}_{\boldsymbol{f}}\int_0^{\tau - \boldsymbol{b}_k^\top \boldsymbol{f}} I(\varepsilon_k > \tau - \boldsymbol{b}_k^\top \boldsymbol{f}) t\,dt + 2\mathbb{E}_{\boldsymbol{f}}\int_{\tau - \boldsymbol{b}_k^\top \boldsymbol{f}}^\infty I(\varepsilon_k > t) t\,dt\\ &= (\tau - \boldsymbol{b}_k^\top \boldsymbol{f})^2 \mathbb{P}_{\boldsymbol{f}}(\varepsilon_k > \tau - \boldsymbol{b}_k^\top \boldsymbol{f}) + 2\int_{\tau - \boldsymbol{b}_k^\top \boldsymbol{f}}^\infty \mathbb{P}(\varepsilon_k > t) t\,dt.\end{aligned}$$

It can be similarly shown that

$$\mathbb{E}_f\{\varepsilon_k^2 I(\boldsymbol{b}_k^\top \boldsymbol{f} + \varepsilon_k < -\tau)\} = (\tau + \boldsymbol{b}_k^\top \boldsymbol{f})^2 \mathbb{P}_f(\varepsilon_k < -\tau - \boldsymbol{b}_k^\top \boldsymbol{f}) + 2\int_{\tau+\boldsymbol{b}_k^\top \boldsymbol{f}}^\infty \mathbb{P}(-\varepsilon_k > t) t\, dt.$$

Together, the last three displays imply

$$\begin{aligned} 0 &\geq \mathbb{E}_f(\xi_k - \boldsymbol{b}_k^\top \boldsymbol{f})^2 - \mathbb{E}(\varepsilon_k^2) \\ &\geq -2\int_{\tau-|\boldsymbol{b}_k^\top \boldsymbol{f}|}^\infty \mathbb{P}(|\varepsilon_k| > t) t\, dt \geq -2v_k \int_{\tau-|\boldsymbol{b}_k^\top \boldsymbol{f}|}^\infty \frac{dt}{t^3} = -\frac{v_k}{(\tau - |\boldsymbol{b}_k^\top \boldsymbol{f}|)^2}. \end{aligned}$$

Combining this with (D.23) and (D.20) proves (D.21).

Next, we study the covariance $\mathrm{cov}_f(\xi_k, \xi_\ell)$ for $k \neq \ell$, which can be written as

$$\begin{aligned} \mathrm{cov}_f(\xi_k, \xi_\ell) &= \mathbb{E}_f(\xi_k - \boldsymbol{b}_k^\top \boldsymbol{f} + \boldsymbol{b}_k^\top \boldsymbol{f} - \mathbb{E}_f \xi_k)(\xi_\ell - \boldsymbol{b}_\ell^\top \boldsymbol{f} + \boldsymbol{b}_\ell^\top \boldsymbol{f} - \mathbb{E}_f \xi_\ell) \\ &= \underbrace{\mathbb{E}_f(\xi_k - \boldsymbol{b}_k^\top \boldsymbol{f})(\xi_\ell - \boldsymbol{b}_\ell^\top \boldsymbol{f})}_{\Pi_1} - \underbrace{(\mathbb{E}_f \xi_k - \boldsymbol{b}_k^\top \boldsymbol{f})(\mathbb{E}_f \xi_\ell - \boldsymbol{b}_\ell^\top \boldsymbol{f})}_{\Pi_2}. \end{aligned}$$

For Π_2, it follows immediately from (D.20) that $|\Pi_2| \lesssim (\tau_k \tau_\ell)^{-1} \sigma_{\varepsilon,kk} \sigma_{\varepsilon,\ell\ell}$ almost surely on the event $\mathcal{G}_{k\ell} := \{|\boldsymbol{b}_k^\top \boldsymbol{f}| \leq \tau_k/2\} \cap \{|\boldsymbol{b}_\ell^\top \boldsymbol{f}| \leq \tau_\ell/2\}$. It remains to consider Π_1. Recall that $\xi_k - \boldsymbol{b}_k^\top \boldsymbol{f} = \varepsilon_k I(|u_k| \leq \tau_k) + (\tau_k - \boldsymbol{b}_k^\top \boldsymbol{f}) I(u_k > \tau_k) - (\tau_k + \boldsymbol{b}_k^\top \boldsymbol{f}) I(u_k < -\tau_k)$, where $u_k = \boldsymbol{b}_k^\top \boldsymbol{f} + \varepsilon_k$. Then, Π_1 can be written as

$$\begin{aligned} &\mathbb{E}_f \varepsilon_k \varepsilon_\ell I(|u_k| \leq \tau_k, |u_\ell| \leq \tau_\ell) + (\tau_\ell - \boldsymbol{b}_\ell^\top \boldsymbol{f}) \mathbb{E}_f \varepsilon_k I(|u_k| \leq \tau, u_\ell > \tau) \\ &\quad - (\tau_\ell + \boldsymbol{b}_\ell^\top \boldsymbol{f}) \mathbb{E}_f \varepsilon_k I(|u_k| \leq \tau_k, u_\ell < -\tau_\ell) + (\tau_k - \boldsymbol{b}_k^\top \boldsymbol{f}) \mathbb{E}_f \varepsilon_\ell I(u_k > \tau_k, |u_\ell| \leq \tau_\ell) \\ &\quad + (\tau_k - \boldsymbol{b}_k^\top \boldsymbol{f})(\tau_\ell - \boldsymbol{b}_\ell^\top \boldsymbol{f}) \mathbb{E}_f I(u_k > \tau_k, u_\ell > \tau_\ell) - (\tau_k - \boldsymbol{b}_k^\top \boldsymbol{f})(\tau_\ell + \boldsymbol{b}_\ell^\top \boldsymbol{f}) \mathbb{E}_f I(u_k > \tau_k, u_\ell < -\tau_\ell) \\ &\quad - (\tau_k + \boldsymbol{b}_k^\top \boldsymbol{f}) \mathbb{E}_f \varepsilon_\ell I(u_k < -\tau_k, |u_\ell| \leq \tau_\ell) - (\tau_k + \boldsymbol{b}_k^\top \boldsymbol{f})(\tau_\ell - \boldsymbol{b}_\ell^\top \boldsymbol{f}) \mathbb{E}_f I(u_k < -\tau_k, u_\ell > \tau_\ell) \\ &\quad + (\tau_k + \boldsymbol{b}_k^\top \boldsymbol{f})(\tau_\ell + \boldsymbol{b}_\ell^\top \boldsymbol{f}) \mathbb{E}_f I(u_k < -\tau_k, u_\ell < -\tau_\ell). \end{aligned} \tag{D.24}$$

For the first term in (D.24), note that

$$\begin{aligned} &\mathbb{E}_f \varepsilon_k \varepsilon_\ell I(|u_k| \leq \tau_k, |u_\ell| \leq \tau_\ell) \\ &\quad = \mathrm{cov}(\varepsilon_k, \varepsilon_\ell) - \mathbb{E}_f \varepsilon_k \varepsilon_\ell I(|u_k| > \tau_k) - \mathbb{E}_f \varepsilon_k \varepsilon_\ell I(|u_\ell| > \tau_\ell) + \mathbb{E}_f \varepsilon_k \varepsilon_\ell I(|u_k| > \tau_k, |u_\ell| > \tau_\ell), \end{aligned}$$

where $|\mathbb{E}_f \varepsilon_k \varepsilon_\ell I(|u_k| > \tau_k)| \leq (\tau_k - |\boldsymbol{b}_k^\top \boldsymbol{f}|)^{-2} \mathbb{E}(|\varepsilon_k|^3 |\varepsilon_\ell|) \leq 4\tau_k^{-2} v_k^{3/4} v_\ell^{1/4}$ and

$$|\mathbb{E}_f \varepsilon_k \varepsilon_\ell I(|u_k| > \tau_k, |u_\ell| > \tau_\ell)| \leq (\tau_k - |\boldsymbol{b}_k^\top \boldsymbol{f}|)^{-1} (\tau_\ell - |\boldsymbol{b}_\ell^\top \boldsymbol{f}|)^{-1} \mathbb{E}(\varepsilon_k^2 \varepsilon_\ell^2) \leq 4\tau_k^{-1} \tau_\ell^{-1} v_k^{1/2} v_\ell^{1/2}$$

almost surely on $\mathcal{G}_{k\ell}$. Hence,

$$|\mathbb{E}_f \varepsilon_k \varepsilon_\ell I(|u_k| \leq \tau_k, |u_\ell| \leq \tau_\ell) - \mathrm{cov}(\varepsilon_k, \varepsilon_\ell)| \lesssim (\tau_k \wedge \tau_\ell)^{-2}$$

holds almost surely on the same event. For the remaining terms in (D.24), it can be similarly derived that, almost surely on the same event,

$$\begin{aligned} |\mathbb{E}_f \varepsilon_k I(|u_k| \leq \tau_k, u_\ell > \tau_\ell)| &\leq |\tau_\ell - \boldsymbol{b}_\ell^\top \boldsymbol{f}|^{-3} \mathbb{E}(|\varepsilon_k| |\varepsilon_\ell|^3), \\ |\mathbb{E}_f \varepsilon_k I(|u_k| \leq \tau_k, u_\ell < -\tau_\ell)| &\leq |\tau_\ell + \boldsymbol{b}_\ell^\top \boldsymbol{f}|^{-3} \mathbb{E}(|\varepsilon_k| |\varepsilon_\ell|^3), \\ \text{and } \mathbb{E}_f I(u_k > \tau_k, \xi_\ell < -\tau_\ell) &\leq |\tau_k - \boldsymbol{b}_k^\top \boldsymbol{f}|^{-2} |\tau_\ell + \boldsymbol{b}_\ell^\top \boldsymbol{f}|^{-2} \mathbb{E}(\varepsilon_k^2 \varepsilon_\ell^2). \end{aligned}$$

Putting the pieces together, we arrive at $|\Pi_1 - \mathrm{cov}(\varepsilon_k, \varepsilon_\ell)| \lesssim (\tau_k \wedge \tau_\ell)^{-2} (v_k \vee v_\ell)$ almost surely on $\mathcal{G}_{k\ell}$. This proves the stated result (D.22). □

The next lemma provides several concentration results regarding the factors $\boldsymbol{f}_i$'s and their functionals.

Lemma D.4. Assume that (i) of Condition D.2 holds. Then, for any $t > 0$,

$$\mathbb{P}\Big\{\max_{1\le i\le n} \|\boldsymbol{f}_i\|_2 > C_1 C_f (r + \log n + t)^{1/2}\Big\} \le e^{-t}, \tag{D.25}$$

$$\mathbb{P}\{\|\sqrt{n}\,\overline{\boldsymbol{f}}\|_2 > C_2 C_f (r + t)^{1/2}\} \le e^{-t}, \tag{D.26}$$

$$\text{and } \mathbb{P}[\|\widehat{\boldsymbol{\Sigma}}_f - \mathbf{I}_r\|_2 > \max\{C_3 C_f^2 n^{-1/2}(r+t)^{1/2}, C_3^2 C_f^4 n^{-1}(r+t)\}] \le 2e^{-t}, \tag{D.27}$$

where $\overline{\boldsymbol{f}} = (1/n)\sum_{i=1}^n \boldsymbol{f}_i$, $\widehat{\boldsymbol{\Sigma}}_f = (1/n)\sum_{i=1}^n \boldsymbol{f}_i \boldsymbol{f}_i^\intercal$ and C_1–C_3 are absolute constants.

Proof of Lemma D.4. Under (i) of Condition D.2, it holds for any $\boldsymbol{u} \in \mathbb{R}^r$ that

$$\mathbb{E}e^{\boldsymbol{u}^\intercal \boldsymbol{f}_i} \le e^{CC_f^2\|\boldsymbol{u}\|_2^2} \text{ for } i = 1,\ldots,n,$$

$$\text{and } \mathbb{E}e^{\sqrt{n}\,\boldsymbol{u}^\intercal \overline{\boldsymbol{f}}} = \prod_{i=1}^n \mathbb{E}e^{n^{-1/2}\boldsymbol{u}^\intercal \boldsymbol{f}_i} \le \prod_{i=1}^n e^{CC_f^2 n^{-1}\|\boldsymbol{u}\|_2^2} \le e^{CC_f^2\|\boldsymbol{u}\|_2^2},$$

where $C > 0$ is an absolute constant. Using Theorem 2.1 in Hsu, Kakade and Zhang (2012), we derive that for any $t > 0$,

$$\mathbb{P}\{\|\boldsymbol{f}_i\|_2^2 > 2CC_f^2(r + 2\sqrt{rt} + 2t)\} \le e^{-t} \text{ and } \mathbb{P}\{\|\sqrt{n}\,\overline{\boldsymbol{f}}\|_2^2 > 2CC_f^2(r + 2\sqrt{rt} + 2t)\} \le e^{-t}.$$

Then, (D.25) follows from the first inequality and the union bound, and the second inequality leads to (D.26) immediately. Finally, applying Theorem 5.39 in Vershynin (2012) gives (D.27). □

D.4.2 Proof of Theorem D.3

First we introduce the following notations:

$$v_k = \mathbb{E}(\varepsilon_k^4), \quad \kappa_{\varepsilon,k} = v_k/\sigma_{\varepsilon,kk}^2, \quad u_{ik} = \boldsymbol{b}_k^\intercal \boldsymbol{f}_i + \varepsilon_{ik}, \quad k = 1,\ldots,d,\ i = 1,\ldots,n.$$

Let $t \ge 1$ and set $\tau_k = a_k(n/t)^{1/2}$ with $a_k \ge \sigma_{kk}^{1/2}$ for $k = 1,\ldots,d$. In view of Lemma D.2, define the event

$$\mathcal{E}_1(t) = \bigcap_{k=1}^d \bigg\{\bigg|\sqrt{n}\,(\widehat{\mu}_k - \mu_k) - \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_{\tau_k}(u_{ik})\bigg| \le C\frac{a_k t}{\sqrt{n}}\bigg\}, \tag{D.28}$$

such that $\mathbb{P}\{\mathcal{E}_1(t)^{\mathrm{c}}\} \le 3de^{-t}$. Moreover, by Lemma D.4, let $\mathcal{E}_2(t)$ be the event that the following hold:

$$\max_{1\le i\le n} \|\boldsymbol{f}_i\|_2 \le C_1 C_f (r + \log n + t)^{1/2}, \quad \|\sqrt{n}\,\overline{\boldsymbol{f}}\|_2 \le C_2 C_f (r+t)^{1/2},$$
$$\text{and } \|\widehat{\boldsymbol{\Sigma}}_f - \mathbf{I}_r\|_2 \le \max\{C_3 C_f^2 n^{-1/2}(r+t)^{1/2}, C_3^2 C_f^4 n^{-1}(r+t)\}. \tag{D.29}$$

By the union bound, $\mathbb{P}\{\mathcal{E}_2(t)^{\mathrm{c}}\} \le 4e^{-t}$.

Now we are ready to prove (D.15). The proof of (D.16) follows the same argument, and thus is omitted. For $k = 1,\ldots,d$, define

$$B_k = \sqrt{n}\,\boldsymbol{b}_k^\intercal \overline{\boldsymbol{f}}, \quad V_k = \frac{1}{\sqrt{n}}\sum_{i=1}^n V_{ik} := \frac{1}{\sqrt{n}}\sum_{i=1}^n \{\psi_{\tau_k}(u_{ik}) - \mathbb{E}_{\boldsymbol{f}_i}\psi_{\tau_k}(u_{ik})\}, \tag{D.30}$$

and $R_k = n^{-1/2} \sum_{i=1}^n \{\mathbb{E}_{\boldsymbol{f}_i} \psi_{\tau_k}(u_{ik}) - \boldsymbol{b}_k^\top \boldsymbol{f}_i\}$, where $\mathbb{E}_{\boldsymbol{f}_i}(\cdot) := \mathbb{E}(\cdot|\boldsymbol{f}_i)$. On the event $\mathcal{E}_1(t)$,

$$|T_k - (V_k + B_k + R_k)| \le C a_k n^{-1/2} t \ \text{ for all } 1 \le k \le d. \tag{D.31}$$

On $\mathcal{E}_2(t)$, it holds $\max_{1\le i\le n} |\boldsymbol{b}_k^\top \boldsymbol{f}_i| \le C_1 C_f \|\boldsymbol{b}_k\|_2 (r + \log n + t)^{1/2} \le C_1 C_f \sigma_{kk}^{1/2} (r + \log n + t)^{1/2}$, which further implies

$$\max_{1\le i\le n} |\boldsymbol{b}_k^\top \boldsymbol{f}_i| \le \tau_k/2 \ \text{ for all } 1 \le k \le d \tag{D.32}$$

as long as $n \ge 4(C_1 C_f)^2 (r + \log n + t)t$. Then, it follows from Lemma D.3 that

$$|R_k| \le \sqrt{n} \max_{1\le i\le n} |\mathbb{E}_{\boldsymbol{f}_i} \psi_{\tau_k}(u_{ik}) - \boldsymbol{b}_k^\top \boldsymbol{f}_i| \le 8\sqrt{n}\, \tau_k^{-3} v_k \le 8\sigma_{kk}^{-3/2} v_k n^{-1} t^{3/2} \tag{D.33}$$

holds almost surely on $\mathcal{E}_2(t)$ for all $1 \le k \le d$. Together, (D.31) and (D.33) imply that for any $z \ge 0$,

$$\begin{aligned} &\sum_{k\in\mathcal{H}_0} I\Big(|V_k + B_k| \ge z + C a_k n^{-1/2} t + 8\kappa_{\varepsilon,k} \sigma_{\varepsilon,kk}^{1/2} n^{-1} t^{3/2}\Big) \\ &\le V(z) \le \sum_{k\in\mathcal{H}_0} I\Big(|V_k + B_k| \ge z - C a_k n^{-1/2} t - 8\kappa_{\varepsilon,k} \sigma_{\varepsilon,kk}^{1/2} n^{-1} t^{3/2}\Big) \end{aligned} \tag{D.34}$$

holds almost surely on $\mathcal{E}_1(t) \cap \mathcal{E}_2(t)$. In view of (D.34), we will instead deal with $\widetilde{V}_+(z)$ and $\widetilde{V}_-(z)$, where

$$\widetilde{V}_+(z) := \sum_{k\in\mathcal{H}_0} I(V_k \ge z - B_k) \ \text{ and } \ \widetilde{V}_-(z) := \sum_{k\in\mathcal{H}_0} I(V_k \le -z - B_k)$$

are such that $\widetilde{V}_+(z) + \widetilde{V}_-(z) = \sum_{k\in\mathcal{H}_0} I(|V_k + B_k| \ge z)$.

In the following, we will focus on $\widetilde{V}_+(z)$ ($\widetilde{V}_-(z)$ can be dealt with in the same way). Observe that, conditional on $\mathcal{F}_n := \{\boldsymbol{f}_i\}_{i=1}^n$, $I(V_1 \ge z - B_1), \ldots, I(V_d \ge z - B_d)$ are weakly correlated random variables. Define $Y_k = I(V_k \ge z - B_k)$ and $P_k = \mathbb{E}(Y_k|\mathcal{F}_n)$ for $k = 1, \ldots, d$. To prove the consistency of $\widetilde{V}_+(z)$, we calculate its variance:

$$\begin{aligned} \mathrm{var}\bigg(\frac{1}{d_0} \sum_{k\in\mathcal{H}_0} Y_k \bigg| \mathcal{F}_n\bigg) &= \mathbb{E}\bigg[\bigg\{\frac{1}{d_0} \widetilde{V}_+(z) - \frac{1}{d_0} \sum_{k\in\mathcal{H}_0} P_k\bigg\}^2 \bigg| \mathcal{F}_n\bigg] \\ &= \frac{1}{d_0^2} \sum_{k\in\mathcal{H}_0} \mathrm{var}(Y_k|\mathcal{F}_n) + \frac{1}{d_0^2} \sum_{k,\ell\in\mathcal{H}_0:k\ne\ell} \mathrm{cov}(Y_k, Y_\ell|\mathcal{F}_n) \\ &\le \frac{1}{4d_0} + \frac{1}{d_0^2} \sum_{k,\ell\in\mathcal{H}_0:k\ne\ell} \{\mathbb{E}(Y_k Y_\ell|\mathcal{F}_n) - P_k P_\ell\} \end{aligned} \tag{D.35}$$

almost surely. In what follows, we study P_k and $\mathbb{E}(Y_k Y_\ell|\mathcal{F}_n)$ separately, starting with the former. For each k, V_k is a sum of conditionally independent zero-mean random variables given $\mathcal{F}_n$. Define

$$\nu_k^2 = \mathrm{var}(V_k|\mathcal{F}_n) = \frac{1}{n} \sum_{i=1}^n \nu_{ik}^2 \ \text{ with } \ \nu_{ik}^2 = \mathrm{var}(V_{ik}|\mathcal{F}_n).$$

Then, it follows from the Berry-Esseen theorem that

$$\begin{aligned} &\sup_{z\in\mathbb{R}} |\mathbb{P}(V_k \le \nu_k x|\mathcal{F}_n) - \Phi(x)| \\ &\lesssim \frac{1}{\nu_k^3 n^{3/2}} \sum_{i=1}^n \mathbb{E}\{|\psi_{\tau_k}(u_{ik})|^3|\mathcal{F}_n\} \lesssim \frac{1}{\nu_k^3 n^{3/2}} \sum_{i=1}^n (|\boldsymbol{b}_k^\top \boldsymbol{f}_i|^3 + \mathbb{E}|\varepsilon_{ik}|^3) \end{aligned} \tag{D.36}$$

almost surely. By (D.32) and (D.21), it holds

$$1 - 4(1 + \kappa_{\varepsilon,k})n^{-1}t \le \sigma_{\varepsilon,kk}^{-1}\nu_k^2 \le 1 \tag{D.37}$$

almost surely on $\mathcal{E}_2(t)$ for all $1 \le k \le d$. Combining this with (D.29) and (D.36) gives

$$\left|P_k - \Phi\left(\frac{-z + B_k}{\nu_k}\right)\right| \lesssim \kappa_{\varepsilon,k}^{3/4}\frac{1}{\sqrt{n}} + \sigma_{\varepsilon,kk}^{-3/2}\|\boldsymbol{b}_k\|_2^3\sqrt{\frac{r + \log n + t}{n}}$$

almost surely on $\mathcal{E}_2(t)$ for all $1 \le k \le d$ as long as $n \ge 2C_3^2C_f^4(r + t) \vee 8(1 + \kappa_{\varepsilon,\max})t$, where $\kappa_{\varepsilon,\max} = \max_{1\le\ell\le d}\kappa_{\varepsilon,\ell}$. By the mean value theorem, there exists some $\eta_k \in [\sigma_{\varepsilon,kk}^{-1/2}, \nu_k^{-1}]$ such that

$$\left|\Phi\left(\frac{-z + B_k}{\nu_k}\right) - \Phi\left(\frac{-z + B_k}{\sqrt{\sigma_{\varepsilon,kk}}}\right)\right| = \phi(\eta_k|z - B_k|)\frac{\eta_k|z - B_k|}{\eta_k}\left|\frac{1}{\nu_k} - \frac{1}{\sqrt{\sigma_{\varepsilon,kk}}}\right| \lesssim \frac{\kappa_{\varepsilon,k}t}{n}.$$

Together, the last two displays imply that almost surely on $\mathcal{E}_2(t)$,

$$\left|P_k - \Phi\left(\frac{-z + B_k}{\sqrt{\sigma_{\varepsilon,kk}}}\right)\right| \lesssim \kappa_{\varepsilon,k}\left(\frac{1}{\sqrt{n}} + \frac{t}{n}\right) + \sigma_{\varepsilon,kk}^{-3/2}\|\boldsymbol{b}_k\|_2^3\sqrt{\frac{r + \log n + t}{n}} \tag{D.38}$$

uniformly for all $1 \le k \le d$ and $z \ge 0$.

Next we consider $\mathbb{E}(Y_kY_\ell|\mathcal{F}_n) = \mathbb{P}(V_k \ge z - B_k, V_\ell \ge z - B_\ell|\mathcal{F}_n)$. Define bivariate random vectors $\boldsymbol{V}_i = (\nu_k^{-1}V_{ik}, \nu_\ell^{-1}V_{i\ell})^\intercal$ for $i = 1, \dots, n$, where $V_{ik}, V_{i\ell}$ are as in (D.30). Observe that $\boldsymbol{V}_1, \dots, \boldsymbol{V}_n$ are conditionally independent random vectors given $\mathcal{F}_n$. Denote by $\boldsymbol{\Theta} = (\theta_{uv})_{1\le u,v\le 2}$ the conditional covariance matrix of $n^{-1/2}\sum_{i=1}^n \boldsymbol{V}_i = (\nu_k^{-1}V_k, \nu_\ell^{-1}V_\ell)^\intercal$ given $\mathcal{F}_n$, such that $\theta_{11} = \theta_{22} = 1$ and $\theta_{12} = \theta_{21} = (\nu_k\nu_\ell n)^{-1}\sum_{i=1}^n \mathrm{cov}_{f_i}(V_{ik}, V_{i\ell})$. By (D.22), (D.32) and (D.37),

$$|\theta_{12} - r_{\varepsilon,k\ell}| \lesssim (\kappa_{\varepsilon,k} \vee \kappa_{\varepsilon,\ell})n^{-1}t \tag{D.39}$$

holds almost surely on $\mathcal{E}_2(t)$ for all $1 \le k \ne \ell \le d$ and sufficient large n, say $n \gtrsim \kappa_{\varepsilon,\max}t$. Let $\boldsymbol{G} = (G_1, G_2)^\intercal$ be a Gaussian random vector with $\mathbb{E}(\boldsymbol{G}) = \boldsymbol{0}$ and $\mathrm{cov}(\boldsymbol{G}) = \boldsymbol{\Theta}$. Applying Theorem 1.1 in Bentkus (2005) and (D.29), we have

$$\begin{aligned}
&\sup_{x,y\in\mathbb{R}} |\mathbb{P}(V_k \ge \nu_kx, V_\ell \ge \nu_\ell y|\mathcal{F}_n) - \mathbb{P}(G_1 \ge x, G_2 \ge y)| \\
&\lesssim \frac{1}{n^{3/2}}\sum_{i=1}^n \mathbb{E}\|\boldsymbol{\Theta}^{-1/2}\boldsymbol{V}_i\|_2^3 \\
&\lesssim \frac{1}{(\sigma_{\varepsilon,kk}n)^{3/2}}\sum_{i=1}^n(\mathbb{E}|\varepsilon_{ik}|^3 + |\boldsymbol{b}_k^\intercal\boldsymbol{f}_i|^3) + \frac{1}{(\sigma_{\varepsilon,\ell\ell}n)^{3/2}}\sum_{i=1}^n(\mathbb{E}|\varepsilon_{i\ell}|^3 + |\boldsymbol{b}_\ell^\intercal\boldsymbol{f}_i|^3) \\
&\lesssim \frac{\kappa_{\varepsilon,k} + \kappa_{\varepsilon,\ell}}{\sqrt{n}} + (\sigma_{\varepsilon,kk}^{-3/2}\|\boldsymbol{b}_k\|_2^3 + \sigma_{\varepsilon,\ell\ell}^{-3/2}\|\boldsymbol{b}_\ell\|_2^3)\sqrt{\frac{r + \log n + t}{n}}
\end{aligned}$$

almost surely on $\mathcal{E}_2(t)$ for all $1 \le k \ne \ell \le d$. In particular, taking $x = \nu_k^{-1}(z - B_k)$ and $y = \nu_\ell^{-1}(z - B_\ell)$ gives

$$\begin{aligned}
&|\mathbb{E}(Y_kY_\ell|\mathcal{F}_n) - \mathbb{P}\{G_1 \ge \nu_k^{-1}(z - B_k), G_2 \ge \nu_\ell^{-1}(z - B_\ell)\}| \\
&\lesssim \frac{\kappa_{\varepsilon,k} + \kappa_{\varepsilon,\ell}}{\sqrt{n}} + (\sigma_{\varepsilon,kk}^{-3/2}\|\boldsymbol{b}_k\|_2^3 + \sigma_{\varepsilon,\ell\ell}^{-3/2}\|\boldsymbol{b}_\ell\|_2^3)\sqrt{\frac{r + \log n + t}{n}}
\end{aligned} \tag{D.40}$$

almost surely on $\mathcal{E}_2(t)$. In addition, it follows from Corollary 2.1 in Li and Shao (2002) that

$$|\mathbb{P}(G_1 \geq x, G_2 \geq y) - \{1 - \Phi(x)\}\{1 - \Phi(y)\}| \leq \frac{|\theta_{12}|}{4} e^{-(x^2+y^2)/(2+2|\theta_{12}|)} \leq \frac{|\theta_{12}|}{4} \tag{D.41}$$

for all $x, y \in \mathbb{R}$.

Substituting the bounds (D.38), (D.39), (D.40) and (D.41) into (D.35), we obtain

$$\mathbb{E}\bigg[\bigg\{\frac{1}{d_0}\widetilde{V}_+(z) - \frac{1}{d_0}\sum_{k\in\mathcal{H}_0} P_k\bigg\}^2 \bigg| \mathcal{F}_n\bigg]$$
$$\lesssim \frac{1}{d_0^2}\sum_{k,\ell\in\mathcal{H}_0:k\neq\ell} \varrho_{\varepsilon,k\ell} + \frac{1}{d_0} + \frac{\kappa_{\varepsilon,\max}}{\sqrt{n}} + \frac{\kappa_{\varepsilon,\max} t}{n} + \max_{1\leq k\leq d} \sigma_{\varepsilon,kk}^{-3/2}\|\boldsymbol{b}_k\|_2^3 \sqrt{\frac{r + \log n + t}{n}}$$

and

$$\bigg|\frac{1}{d_0}\sum_{k\in\mathcal{H}_0} P_k - \frac{1}{d_0}\sum_{k\in\mathcal{H}_0} \Phi\bigg(\frac{-z + B_k}{\sqrt{\sigma_{\varepsilon,kk}}}\bigg)\bigg|$$
$$\lesssim \frac{\kappa_{\varepsilon,\max}}{\sqrt{n}} + \frac{\kappa_{\varepsilon,\max} t}{n} + \max_{1\leq k\leq d} \sigma_{\varepsilon,kk}^{-3/2}\|\boldsymbol{b}_k\|_2^3 \sqrt{\frac{r + \log n + t}{n}}$$

almost surely on $\mathcal{E}_2(t)$. Similar bounds can be derived for

$$\mathrm{var}\bigg(\frac{1}{d_0}\widetilde{V}_-(z)\bigg|\mathcal{F}_n\bigg) \text{ and } \frac{1}{d_0}\mathbb{E}\{\widetilde{V}_-(z)|\mathcal{F}_n\}.$$

Taking $t = \log(nd)$ so that $\mathbb{P}\{\mathcal{E}_1(t)^{\mathrm{c}}\} \leq 3n^{-1}$ and $\mathbb{P}\{\mathcal{E}_2(t)^{\mathrm{c}}\} \leq 4(nd)^{-1}$. Under Condition D.2, it follows that

$$\frac{1}{d_0}\widetilde{V}_+(z) = \frac{1}{d_0}\sum_{k\in\mathcal{H}_0}\Phi\bigg(\frac{-z + B_k}{\sqrt{\sigma_{\varepsilon,kk}}}\bigg) + O_{\mathbb{P}}[d^{-(1\wedge\delta_0)/2} + n^{-1/4}\{\log(nd)\}^{1/4}],$$
$$\frac{1}{d_0}\widetilde{V}_-(z) = \frac{1}{d_0}\sum_{k\in\mathcal{H}_0}\Phi\bigg(\frac{-z - B_k}{\sqrt{\sigma_{\varepsilon,kk}}}\bigg) + O_{\mathbb{P}}[d^{-(1\wedge\delta_0)/2} + n^{-1/4}\{\log(nd)\}^{1/4}]$$

uniformly over all $z \geq 0$. This, together with (D.34) and the fact that $|\Phi(z_1) - \Phi(z_2)| \leq (2\pi)^{-1/2}|z_1 - z_2|$, proves (D.15). □

D.5 Proof of Theorem D.2

Let $\boldsymbol{b}^{(1)}, \ldots, \boldsymbol{b}^{(r)} \in \mathbb{R}^d$ be the columns of $\mathbf{B}$. Without loss of generality, assume that $\|\boldsymbol{b}^{(1)}\|_2 \geq \cdots \geq \|\boldsymbol{b}^{(r)}\|_2$. Under (i) of Condition D.1, $\mathbf{B}\mathbf{B}^\intercal$ has non-vanishing eigenvalues $\{\bar{\lambda}_\ell := \|\boldsymbol{b}^{(\ell)}\|_2^2\}_{\ell=1}^r$ with eigenvectors $\{\bar{\boldsymbol{v}}_\ell := \boldsymbol{b}^{(\ell)}/\|\boldsymbol{b}^{(\ell)}\|_2\}_{\ell=1}^r$, and $\mathbf{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_d)^\intercal = (\bar{\lambda}_1^{1/2}\bar{\boldsymbol{v}}_1, \ldots, \bar{\lambda}_r^{1/2}\bar{\boldsymbol{v}}_r)$. Moreover, write $\widehat{\mathbf{B}} = (\widehat{\boldsymbol{b}}_1, \ldots, \widehat{\boldsymbol{b}}_d)^\intercal = (\widehat{\lambda}_1^{1/2}\widehat{\boldsymbol{v}}_1, \ldots, \widehat{\lambda}_r^{1/2}\widehat{\boldsymbol{v}}_r)$ with $\widehat{\boldsymbol{v}}_\ell = (\widehat{v}_{\ell 1}, \ldots, \widehat{v}_{\ell d})^\intercal$ for $\ell = 1, \ldots, r$ and $\widehat{\boldsymbol{b}}_k = (\widehat{\lambda}_1^{1/2}\widehat{v}_{1k}, \ldots, \widehat{\lambda}_r^{1/2}\widehat{v}_{rk})^\intercal$ for $k = 1, \ldots, d$.

A key step in proving (D.12) is the derivation of an upper bound on the estimation error $\Delta_d := \max_{1\leq k\leq d} |\|\widehat{\boldsymbol{b}}_k\|_2 - \|\boldsymbol{b}_k\|_2|$. By Weyl's inequality and the decomposition that $\widehat{\boldsymbol{\Sigma}}_1^{\mathcal{H}} = \mathbf{B}\mathbf{B}^\intercal + (\widehat{\boldsymbol{\Sigma}}_1^{\mathcal{H}} - \boldsymbol{\Sigma}) + \boldsymbol{\Sigma}_\varepsilon$,

$$\max_{1\leq\ell\leq r} |\widehat{\lambda}_\ell - \bar{\lambda}_\ell| \leq \|\widehat{\boldsymbol{\Sigma}}_1^{\mathcal{H}} - \boldsymbol{\Sigma}\|_2 + \|\boldsymbol{\Sigma}_\varepsilon\|_2. \tag{D.42}$$

Applying Corollary 1 in Yu, Wang and Samworth (2015) yields that, for every $1 \le \ell \le r$,

$$\|\widehat{\boldsymbol{v}}_\ell - \overline{\boldsymbol{v}}_\ell\|_2 \le \frac{2^{3/2}(\|\widehat{\boldsymbol{\Sigma}}_1^{\mathcal{H}} - \boldsymbol{\Sigma}\|_2 + \|\boldsymbol{\Sigma}_\varepsilon\|_2)}{\min(\overline{\lambda}_{\ell-1} - \overline{\lambda}_\ell, \overline{\lambda}_\ell - \overline{\lambda}_{\ell+1})},$$

where we set $\overline{\lambda}_0 = \infty$ and $\overline{\lambda}_{r+1} = 0$. Under (ii) of Condition D.1, it follows that

$$\max_{1 \le \ell \le r} \|\widehat{\boldsymbol{v}}_\ell - \overline{\boldsymbol{v}}_\ell\|_2 \lesssim d^{-1}(\|\widehat{\boldsymbol{\Sigma}}_1^{\mathcal{H}} - \boldsymbol{\Sigma}\|_2 + \|\boldsymbol{\Sigma}_\varepsilon\|_2). \tag{D.43}$$

Moreover, apply Theorem 3 and Proposition 3 in Fan, Wang and Zhong (2018) to reach

$$\max_{1 \le \ell \le r} \|\widehat{\boldsymbol{v}}_\ell - \overline{\boldsymbol{v}}_\ell\|_\infty \lesssim r^4(d^{-1/2}\|\widehat{\boldsymbol{\Sigma}}_1^{\mathcal{H}} - \boldsymbol{\Sigma}\|_{\max} + d^{-1}\|\boldsymbol{\Sigma}_\varepsilon\|_2). \tag{D.44}$$

Note that, under (ii) of Condition D.1,

$$\|\overline{\boldsymbol{v}}_\ell\|_\infty = \|\boldsymbol{b}^{(\ell)}\|_\infty / \|\boldsymbol{b}^{(\ell)}\|_2 \le \|\mathbf{B}\|_{\max} / \|\boldsymbol{b}^{(\ell)}\|_2 \lesssim d^{-1/2} \text{ for all } \ell = 1, \ldots, r. \tag{D.45}$$

Define $\widetilde{\boldsymbol{b}}_k = (\overline{\lambda}_1^{1/2}\widehat{v}_{1k}, \ldots, \overline{\lambda}_r^{1/2}\widehat{v}_{rk})^\intercal$. By the triangular inequality,

$$\begin{aligned}
\|\widehat{\boldsymbol{b}}_k - \boldsymbol{b}_k\|_2 &\le \|\widehat{\boldsymbol{b}}_k - \widetilde{\boldsymbol{b}}_k\|_2 + \|\widetilde{\boldsymbol{b}}_k - \boldsymbol{b}_k\|_2 \\
&= \bigg\{ \sum_{\ell=1}^r (\widehat{\lambda}_\ell^{1/2} - \overline{\lambda}_\ell^{1/2})^2 \widehat{v}_{\ell k}^2 \bigg\}^{1/2} + \bigg\{ \sum_{\ell=1}^r \overline{\lambda}_\ell (\widehat{v}_{\ell k} - \overline{v}_{\ell k})^2 \bigg\}^{1/2} \\
&\le r^{1/2} \Big(\max_{1 \le \ell \le r} |\widehat{\lambda}_\ell^{1/2} - \overline{\lambda}_\ell^{1/2}| \|\widehat{\boldsymbol{v}}_\ell\|_\infty + \max_{1 \le \ell \le r} \overline{\lambda}_\ell^{1/2} \|\widehat{\boldsymbol{v}}_\ell - \overline{\boldsymbol{v}}_\ell\|_\infty \Big).
\end{aligned}$$

This, together with (D.42)–(D.45) and Theorem 3.3, implies

$$\Delta_d \le \max_{1 \le k \le d} \|\widehat{\boldsymbol{b}}_k - \boldsymbol{b}_k\|_2 = O_{\mathbb{P}}(w_{n,d}). \tag{D.46}$$

With the above preparations, now we are ready to prove (D.12). To that end, define $\widetilde{\boldsymbol{u}} = \sqrt{n}\,(\mathbf{B}^\intercal\mathbf{B})^{-1}\mathbf{B}^\intercal\overline{\boldsymbol{X}}$, so that for every $1 \le k \le d$,

$$\boldsymbol{b}_k^\intercal \widetilde{\boldsymbol{u}} = \sqrt{n}\,\boldsymbol{b}_k^\intercal \overline{\boldsymbol{f}} + \sqrt{n}\,\boldsymbol{b}_k^\intercal (\mathbf{B}^\intercal\mathbf{B})^{-1}\mathbf{B}^\intercal\boldsymbol{\mu} + \sqrt{n}\,\boldsymbol{b}_k^\intercal (\mathbf{B}^\intercal\mathbf{B})^{-1}\mathbf{B}^\intercal\overline{\boldsymbol{\varepsilon}}.$$

Consider the decomposition

$$\begin{aligned}
&\bigg|\Phi\Big(\frac{-z + \widehat{\boldsymbol{b}}_k^\intercal \widehat{\boldsymbol{u}}}{\sqrt{\widehat{\sigma}_{\varepsilon,kk}}}\Big) - \Phi\Big(\frac{-z + \sqrt{n}\,\boldsymbol{b}_k^\intercal \overline{\boldsymbol{f}}}{\sqrt{\sigma_{\varepsilon,kk}}}\Big)\bigg| \\
&\le \bigg|\Phi\Big(\frac{-z + \widehat{\boldsymbol{b}}_k^\intercal \widehat{\boldsymbol{u}}}{\sqrt{\widehat{\sigma}_{\varepsilon,kk}}}\Big) - \Phi\Big(\frac{-z + \boldsymbol{b}_k^\intercal \widetilde{\boldsymbol{u}}}{\sqrt{\sigma_{\varepsilon,kk}}}\Big)\bigg| + \bigg|\Phi\Big(\frac{-z + \boldsymbol{b}_k^\intercal \widetilde{\boldsymbol{u}}}{\sqrt{\sigma_{\varepsilon,kk}}}\Big) - \Phi\Big(\frac{-z + \sqrt{n}\,\boldsymbol{b}_k^\intercal \overline{\boldsymbol{f}}}{\sqrt{\sigma_{\varepsilon,kk}}}\Big)\bigg| \\
&\le \bigg|\Phi\Big(\frac{-z + \boldsymbol{b}_k^\intercal \widetilde{\boldsymbol{u}}}{\sqrt{\widehat{\sigma}_{\varepsilon,kk}}}\Big) - \Phi\Big(\frac{-z + \boldsymbol{b}_k^\intercal \widetilde{\boldsymbol{u}}}{\sqrt{\sigma_{\varepsilon,kk}}}\Big)\bigg| \\
&\qquad + \bigg|\Phi\Big(\frac{-z + \widehat{\boldsymbol{b}}_k^\intercal \widehat{\boldsymbol{u}}}{\sqrt{\widehat{\sigma}_{\varepsilon,kk}}}\Big) - \Phi\Big(\frac{-z + \boldsymbol{b}_k^\intercal \widetilde{\boldsymbol{u}}}{\sqrt{\widehat{\sigma}_{\varepsilon,kk}}}\Big)\bigg| + \bigg|\Phi\Big(\frac{-z + \boldsymbol{b}_k^\intercal \widetilde{\boldsymbol{u}}}{\sqrt{\sigma_{\varepsilon,kk}}}\Big) - \Phi\Big(\frac{-z + \sqrt{n}\,\boldsymbol{b}_k^\intercal \overline{\boldsymbol{f}}}{\sqrt{\sigma_{\varepsilon,kk}}}\Big)\bigg| \\
&:= \Delta_{k1} + \Delta_{k2} + \Delta_{k3}.
\end{aligned} \tag{D.47}$$

In the following, we deal with Δ_{k1}, Δ_{k2} and Δ_{k3} separately.

By the mean value theorem, there exists some ξ_k between $\widehat{\sigma}_{\varepsilon,kk}^{-1/2}$ and $\sigma_{\varepsilon,kk}^{-1/2}$ such that

$$\Delta_{k1} = \phi(\xi_k|z - \boldsymbol{b}_k^\top \widetilde{\boldsymbol{u}}|)|z - \boldsymbol{b}_k^\top \widetilde{\boldsymbol{u}}| \left| \frac{1}{\sqrt{\widehat{\sigma}_{\varepsilon,kk}}} - \frac{1}{\sqrt{\sigma_{\varepsilon,kk}}} \right|,$$

where $\phi(\cdot) = \Phi'(\cdot)$. With $\tau_{kk} \asymp \sqrt{n/\log(nd)}$ for $k = 1, \ldots, d$, it follows that the event

$$\mathcal{E}_0 = \left\{ \max_{1 \le k \le d} |\widehat{\sigma}_{kk} - \sigma_{kk}| \lesssim \sqrt{\log(d)/n} \right\}$$

satisfies $\mathbb{P}(\mathcal{E}_0^{\mathrm{c}}) \lesssim n^{-1}$. On $\mathcal{E}_0$, it holds $\widehat{\sigma}_{\varepsilon,kk}^{-1} \ge (2\sigma_{kk})^{-1}$, $\sigma_{\varepsilon,kk}^{-1} \ge \sigma_{kk}^{-1}$ and therefore $\xi_k \ge (2\sigma_{kk})^{-1/2}$ uniformly for all $1 \le k \le d$ as long as $n \gtrsim \log d$. This further implies $\max_{1\le k\le d} \max_{z\ge 0} \phi(\xi_k|z - \boldsymbol{b}_k^\top \widetilde{\boldsymbol{u}}|)|z - \boldsymbol{b}_k^\top \widetilde{\boldsymbol{u}}| = O_{\mathbb{P}}(1)$. By (D.46),

$$\left| \frac{1}{\sqrt{\widehat{\sigma}_{\varepsilon,kk}}} - \frac{1}{\sqrt{\sigma_{\varepsilon,kk}}} \right| = O_{\mathbb{P}}(|\widehat{\sigma}_{kk} - \sigma_{kk}| + \|\widehat{\boldsymbol{b}}_k - \boldsymbol{b}_k\|_2) = O_{\mathbb{P}}(w_{n,d}) \tag{D.48}$$

uniformly over $k = 1, \ldots, d$, where $w_{n,d} = \sqrt{\log(d)/n} + d^{-1/2}$. Putting the above calculations together, we arrive at

$$\frac{1}{d} \sum_{k=1}^{d} \Delta_{k1} = O_{\mathbb{P}}(w_{n,d}). \tag{D.49}$$

Turning to Δ_{k2}, again by the mean value theorem, there exists some η_k between $\widehat{\boldsymbol{b}}_k^\top \widehat{\boldsymbol{u}}$ and $\boldsymbol{b}_k^\top \widetilde{\boldsymbol{u}}$ such that

$$\Delta_{k2} = \phi\bigg(\frac{-z + \eta_k}{\sqrt{\widehat{\sigma}_{\varepsilon,kk}}} \bigg) \frac{\widehat{\boldsymbol{b}}_k^\top \widehat{\boldsymbol{u}} - \boldsymbol{b}_k^\top \widetilde{\boldsymbol{u}}}{\sqrt{\widehat{\sigma}_{\varepsilon,kk}}}.$$

In view of (D.48),

$$\max_{1\le k\le d} \phi\bigg(\frac{-z + \eta_k}{\sqrt{\widehat{\sigma}_{\varepsilon,kk}}} \bigg) \frac{1}{\sqrt{\widehat{\sigma}_{\varepsilon,kk}}} = O_{\mathbb{P}}(1).$$

Observe that $\widehat{\mathbf{B}}\widehat{\boldsymbol{u}} = (\sum_{\ell=1}^r \widehat{\boldsymbol{v}}_\ell \widehat{\boldsymbol{v}}_\ell^\top)\boldsymbol{Z}$ and $\mathbf{B}\widetilde{\boldsymbol{u}} = (\sum_{\ell=1}^r \overline{\boldsymbol{v}}_\ell \overline{\boldsymbol{v}}_\ell^\top)\boldsymbol{Z}$, where $\boldsymbol{Z} = \sqrt{n}\overline{\boldsymbol{X}}$. By the Cauchy-Schwarz inequality,

$$\begin{aligned} \sum_{k=1}^{d} |\widehat{\boldsymbol{b}}_k^\top \widehat{\boldsymbol{u}} - \boldsymbol{b}_k^\top \widetilde{\boldsymbol{u}}| &\le d^{1/2} \|\widehat{\mathbf{B}}\widehat{\boldsymbol{u}} - \mathbf{B}\widetilde{\boldsymbol{u}}\|_2 \\ &\le d^{1/2} \bigg\| \sum_{\ell=1}^{r} (\widehat{\boldsymbol{v}}_\ell \widehat{\boldsymbol{v}}_\ell^\top - \overline{\boldsymbol{v}}_\ell \overline{\boldsymbol{v}}_\ell^\top) \bigg\|_2 \|\boldsymbol{Z}\|_2 \le 2rd^{1/2} \max_{1\le \ell\le r} \|\widehat{\boldsymbol{v}}_\ell - \overline{\boldsymbol{v}}_\ell\|_2 \, \|\boldsymbol{Z}\|_2. \end{aligned} \tag{D.50}$$

For $\|\boldsymbol{Z}\|_2$, we calculate $\mathbb{E}\|\boldsymbol{Z}\|_2^2 = n\|\boldsymbol{\mu}\|_2^2 + \sum_{k=1}^d \sigma_{kk}$, indicating that $\|\boldsymbol{Z}\|_2 = O_{\mathbb{P}}(\sqrt{n}\,\|\boldsymbol{\mu}\|_2 + d^{1/2})$. Combining this with (D.43), (D.50) and Theorem 3.3, we conclude that

$$\frac{1}{d} \sum_{k=1}^{d} \Delta_{k2} = O_{\mathbb{P}}\{(d^{-1/2}\sqrt{n}\,\|\boldsymbol{\mu}\|_2 + 1)w_{n,d}\}. \tag{D.51}$$

For Δ_{k3}, following the same arguments as above, it suffices to consider

$$\frac{\sqrt{n}}{d} \sum_{k=1}^{d} |\boldsymbol{b}_k^\top (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top \boldsymbol{\mu} + \boldsymbol{b}_k^\top (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top \overline{\boldsymbol{\varepsilon}}|,$$

which, by the Cauchy-Schwarz inequality, is bounded by

$$\sqrt{\frac{n}{d}} \Big\| \sum_{\ell=1}^{r} \bar{\boldsymbol{v}}_\ell \bar{\boldsymbol{v}}_\ell^\top \Big\|_2 \|\boldsymbol{\mu}\|_2 + \max_{1\le k\le d} \|\boldsymbol{b}_k\|_2 \|\boldsymbol{u}\|_2 \le \sqrt{\frac{n}{d}} \|\boldsymbol{\mu}\|_2 + \max_{1\le k\le d} \sqrt{\sigma_{kk}} \|\boldsymbol{u}\|_2,$$

where $\boldsymbol{u} = \sqrt{n}\,(\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\bar{\boldsymbol{\varepsilon}} \in \mathbb{R}^r$ is a zero-mean random vector with covariance matrix $\boldsymbol{\Sigma}_u = (\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top\boldsymbol{\Sigma}_\varepsilon\mathbf{B}(\mathbf{B}^\top\mathbf{B})^{-1}$. Recall that $\mathbf{B}^\top\mathbf{B} \in \mathbb{R}^{r\times r}$ has non-increasing eigenvalues $\bar{\lambda}_1 \ge \cdots \ge \bar{\lambda}_r$. Under (ii) of Condition D.1, it holds $\mathbb{E}\|\boldsymbol{u}\|_2^2 = \mathrm{Tr}(\boldsymbol{\Sigma}_u) \le \|\boldsymbol{\Sigma}_\varepsilon\| \sum_{\ell=1}^r \bar{\lambda}_\ell^{-1} \lesssim rd^{-1}$ and thus $\|\boldsymbol{u}\|_2 = O_{\mathbb{P}}(d^{-1/2})$. Putting the pieces together, we get

$$\frac{1}{d}\sum_{k=1}^{d} \Delta_{k3} = O_{\mathbb{P}}(d^{-1/2}\sqrt{n}\,\|\boldsymbol{\mu}\|_2 + d^{-1/2}). \tag{D.52}$$

Combining (D.47), (D.49), (D.51) and (D.52), we reach

$$\frac{1}{d}\sum_{k=1}^{d} \Phi\bigg(\frac{-z + \widehat{\boldsymbol{b}}_k^\top\widehat{\boldsymbol{u}}}{\sqrt{\widehat{\sigma}_{\varepsilon,kk}}}\bigg) = \frac{1}{d}\sum_{k=1}^{d} \Phi\bigg(\frac{-z + \sqrt{n}\,\boldsymbol{b}_k^\top\bar{\boldsymbol{f}}}{\sqrt{\sigma_{\varepsilon,kk}}}\bigg) + O_{\mathbb{P}}(w_{n,d} + d^{-1/2}\sqrt{n}\,\|\boldsymbol{\mu}\|_2).$$

Using the same argument, it can similarly derived that

$$\frac{1}{d}\sum_{k=1}^{d} \Phi\bigg(\frac{-z - \widehat{\boldsymbol{b}}_k^\top\widehat{\boldsymbol{u}}}{\sqrt{\widehat{\sigma}_{\varepsilon,kk}}}\bigg) = \frac{1}{d}\sum_{k=1}^{d} \Phi\bigg(\frac{-z - \sqrt{n}\,\boldsymbol{b}_k^\top\bar{\boldsymbol{f}}}{\sqrt{\sigma_{\varepsilon,kk}}}\bigg) + O_{\mathbb{P}}(w_{n,d} + d^{-1/2}\sqrt{n}\,\|\boldsymbol{\mu}\|_2).$$

Together, the last two displays lead to the stated result (D.12). □

References

Anderson, T. W. and Rubin, H. (1956). Statistical inference for factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. V* 111–150. Univ. California Press, Berkeley.

Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension. *Ann. Statist.* **40** 436–465.

Barras, L., Scaillet, O. and Wermers, R. (2010). False discoveries in mutual fund performance: Measuring luck in estimated alphas. *J. Finance* **65** 179–216.

Berk, J. B. and Green, R. C. (2004). Mutual fund flows and performance in rational markets. *J. Polit. Econ.* **112** 1269–1295.

Bentkus, V. (2005). A Lyapunov-type bound in R^d. *Theory Probab. Appl.* **49** 311–323.

Cai, T. T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106** 672–684.

Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure and mean-variance analysis in large asset markets. *Econometrica* **51** 1305–1324.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *J. Financ. Econ.* **33** 3–56.

Fan, J. and Han, X. (2017). Estimation of the false discovery proportion with unknown dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1143–1164.

Fan, J., Ke, Y., Sun, Q. and Zhou, W.-X. (2019). FarmTest: Factor-adjusted robust multiple testing with approximate false discovery control. *J. Amer. Statist. Assoc.* To appear.

Fan, J., Li, Q. and Wang, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 247–265.

Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 603–680.

Fan, J., Liu, H. and Wang, W. (2018). Large covariance estimation through elliptical factor models. *Ann. Statist.* **46** 1383–1414.

Fan, J., Wang, W. and Zhong, Y. (2018). An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *J. Mach. Learn. Res.* **18**(207) 1–42.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

Hsu, D., Kakade, S. M. and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.* **17**(52) 1–6.

Lan, W. and Du, L. (2019). A factor-adjusted multiple testing procedure with application to mutual fund selection. *J. Bus. Econom. Statist.* **37** 147–157.

Li, W. V. and Shao, Q.-M. (2002). A normal comparison inequality and its applications. *Probab. Theory Relat. Fields* **122** 494–508.

Minsker, S. (2018). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Ann. Statist.* **46** 2871–2903.

Minsker, S. and Strawn, N. (2017). Distributed statistical estimation and rates of convergence in normal approximation. *arXiv preprint arXiv:1704.02658*.

Minsker, S. and Wei, X. (2018). Robust modifications of U-statistics and applications to covariance estimation problems. *arXiv preprint arXiv:1801.05565*.

Pinelis, I. and Molzon, R. (2016). Optimal-order bounds on the rate of convergence to normality in the multivariate delta method. *Electron. J. Statist.* **10** 1001–1063.

Rothman, A. J., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* **104** 177–186.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *J. Finance* **19** 425–442.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge.

Yu, Y., Wang, T. and Samworth, R. J. (2015). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika* **102** 315–323.

Zhou, W.-X., Bose, K., Fan, J. and Liu, H. (2018). A new perspective on robust M-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *Ann. Statist.* **46** 1904–1931.