Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

SUPPLEMENT: ADAPTIVE ESTIMATION OF THE RANK OF THE COEFFICIENT MATRIX IN HIGH DIMENSIONAL MULTIVARIATE RESPONSE REGRESSION MODELS

By Xin Bing and Marten H. Wegkamp

Cornell University

APPENDIX A: PROOFS OF SECTIONS 2 & 3

A.1. Proof of Proposition 1. We first note that, since $(PY)_i = P(PY)_i$, see Giraud (2015) (page 124), and Pythagoras' identity

$$\begin{aligned}\|Y-(PY)_i\|^2 &= \|Y-PY\|^2+\|PY-(PY)_i\|^2 \\ &= \|Y-PY\|^2+\sum_{k=i+1}^{j} d_k^2(PY)+\sum_{k>j} d_k^2(PY) \\ &= \|Y-(PY)_j\|^2+\sum_{k=i+1}^{j} d_k^2(PY).\end{aligned} \tag{A.1}$$

Consequently, with $i<j$,

$$\begin{aligned}\widehat{\sigma}_i^2 \geq \widehat{\sigma}_j^2 &\iff \frac{\|Y-(PY)_i\|^2}{nm-\lambda i} \geq \frac{\|Y-(PY)_j\|^2}{nm-\lambda j} \\ &\iff \frac{\|Y-(PY)_j+\sum_{k=i+1}^{j} d_k^2(PY)}{(nm-\lambda j)+\lambda(j-i)} \geq \frac{\|Y-(PY)_j\|^2}{nm-\lambda j} \\ &\iff \frac{\sum_{k=i+1}^{j} d_k^2(PY)}{\lambda(j-i)} \geq \frac{\|Y-(PY)_j\|^2}{nm-\lambda j}=\widehat{\sigma}_j^2\end{aligned}$$

using the simple fact that $(a+b)/(c+d) \geq a/c \iff (b/d) \geq (a/c)$ for any positive numbers a,b,c,d. This proves (2.1). Claim (2.2) follows from (2.1) by taking $i=j-1$. Finally,

$$\frac{d_j^2(PY)}{\lambda} \leq \frac{\|Y-(PY)_j\|^2}{nm-\lambda j} = \frac{\|Y-(PY)_{j-1}\|^2-d_j^2(PY)}{nm-\lambda(j-1)-\lambda}$$

is equivalent with

$$\frac{d_j^2(PY)}{\lambda} \leq \frac{\|Y-(PY)_{j-1}\|^2}{nm-\lambda(j-1)}=\widehat{\sigma}_{j-1}^2$$

using the above elementary manipulation again and (2.3) follows. This completes our proof. □

A.2. Proof of Proposition 2. We first show (2.4). Suppose $\widehat{\sigma}_k^2 \le \widehat{\sigma}_{k-1}^2$. We observe

$$\begin{aligned}\frac{1}{k-\ell}\sum_{j=\ell+1}^{k} d_j^2(PY) &\ge d_k^2(PY) \quad \text{by } d_1(PY) \ge d_2(PY) \ge \cdots \\ &\ge \lambda \widehat{\sigma}_k^2 \qquad \text{by (2.2)}\end{aligned}$$

so that (2.1) implies $\widehat{\sigma}_k^2 \le \widehat{\sigma}_\ell^2$ for all $\ell \le k-1$. This proves the non-trivial direction of (2.4).

Next, we show (2.5). Suppose $\widehat{\sigma}_k^2 \ge \widehat{\sigma}_{k-1}^2$. Then, by (2.2) and $d_{k+1}(PY) \ge d_k(PY)$, we get

$$d_k^2(PY) \le \lambda \widehat{\sigma}_k^2 \quad \Longrightarrow \quad d_{k+1}^2(PY) \le \lambda \widehat{\sigma}_k^2 \quad \overset{(2.3)}{\Longleftrightarrow} \quad d_{k+1}^2(PY) \le \lambda \widehat{\sigma}_{k+1}^2.$$

Note that the last inequality further implies $\widehat{\sigma}_{k+1}^2 \ge \widehat{\sigma}_k^2$ by (2.2) again. Repeating the same reasoning completes our proof. □

A.3. Proof of Theorem 4. By Theorem 3, it suffices to show $d_1^2(PY) \le \lambda \widehat{\sigma}_1^2$. This is equivalent to $d_1^2(PY) \le \lambda \widehat{\sigma}_0^2$ by criterion (2.3) in Proposition 1. The latter, in turn, is equivalent to $d_1^2(PE) \le \lambda \widehat{\sigma}^2$ as $XA = 0$ implies $d_1^2(PE) = d_1^2(PY)$ and $\widehat{\sigma}_0^2 = \widehat{\sigma}^2$. □

A.4. Proof of Corollary 5. We can write $\lambda := C(\sqrt{m}+\sqrt{q})^2$ for some $C = (1+C_0)^2/(1-C_1) > 1$ with $C_0 > 0$ and $0 < C_1 < 1$. By Theorem 4, (3.2) and (3.3), we have

$$\begin{aligned}\mathbb{P}\{\widehat{k} \ne 0\} &\le \mathbb{P}\left\{d_1^2(PE) \ge \lambda \widehat{\sigma}^2\right\} \\ &\le \mathbb{P}\left\{d_1^2(PE) \ge (1+C_0)^2(\sqrt{m}+\sqrt{q})^2\sigma^2\right\} + \mathbb{P}\left\{\widehat{\sigma}^2 \le (1-C_1)\sigma^2\right\} \\ &\le \exp\left\{-C_0^2(\sqrt{m}+\sqrt{q})^2/2\right\} + \exp\left\{-C_1^2 nm/4\right\},\end{aligned}$$

which proves the claim. □

A.5. Proof of Theorem 6. By Proposition 1, for $\widehat{k} \le r$, we need to show that $d_{r+1}^2(PY) < \lambda\widehat{\sigma}_{r+1}^2$. We observe that

$$d_{r+1}^2(PY) < \lambda\widehat{\sigma}_{r+1}^2 \quad \Longleftrightarrow \quad \lambda > d_{r+1}^2(PY)/\widehat{\sigma}_r^2$$

by statement (2.3). Again, on the event (3.6), an application of Weyl's inequality and observing that $d_{r+1}(XA) = 0$ yield

$$\lambda \;\ge\; d_1^2(PE)/\widehat{\sigma}_r^2 \;\ge\; d_{r+1}^2(PY)/\widehat{\sigma}_r^2$$

which is exactly what needed to be shown. □

A.6. Proof of Proposition 7. To show (3.8), on the one hand, we use the Eckhart-Young theorem and the fact that $r(XA) \le r$ to deduce $\|Y - (PY)_r\|^2 \le \|Y - XA\|^2 = \|E\|^2$. Hence

$$\widehat{\sigma}_r^2 \le \frac{nm}{nm - \lambda r}\widehat{\sigma}^2$$

On the other hand, Weyl's inequality shows that $d_{r+i}(PY) \ge d_{2r+i}(PE)$ for $0 \le i \le N - r$ by defining $d_k(PE) := 0$ for $k > N$, and we obtain

$$\begin{aligned}
\widehat{\sigma}_r^2 &= \frac{\|Y - (PY)_r\|^2}{nm - \lambda r} = \frac{\|Y - PY\|^2 + \sum_{j=r+1}^N d_j^2(PY)}{nm - \lambda r} \\
&\ge \frac{\|E - PE\|^2 + \sum_{j=2r+1}^N d_j^2(PE)}{nm - \lambda r} \\
&= \frac{\|E - (PE)_{2r\wedge N}\|^2}{nm - \lambda r}. \qquad \text{(A.2)}
\end{aligned}$$

If $2r \ge N$, then

$$\frac{\|E - (PE)_{2r\wedge N}\|^2}{nm - \lambda r} = \frac{\|E - PE\|^2}{nm - \lambda r} \ge \frac{\|E - (PE)_N\|^2}{nm - \lambda N/2}.$$

We conclude the proof by invoking (A.4) in Lemma 1. □

LEMMA 1. *For any given $1 \le k \le r$ and $2k \le N - 2$, if λ satisfies*

$$\lambda \;\ge\; \frac{nm}{\|E - (PE)_{2k}\|^2 / \left[d_{2k+1}^2(PE) + d_{2k+2}^2(PE)\right] + k},$$

then

$$\frac{\|E - (PE)_{2k}\|^2}{nm - \lambda k} \le \frac{\|E - (PE)_{2r}\|^2}{nm - \lambda r} \qquad \text{(A.3)}$$

In particular, on the event $\{\lambda\widehat{\sigma}^2 \geq d_1^2(PE) + d_2^2(PE)\}$, *we have*

$$\frac{\|E\|^2}{nm} \leq \frac{\|E-(PE)_2\|^2}{nm-\lambda} \leq \frac{\|E-(PE)_4\|^2}{nm-2\lambda} \leq \cdots \leq \frac{\|E-(PE)_N\|^2}{nm-\lambda N/2}. \tag{A.4}$$

Proof of Lemma 1. We first show (A.4) by using the same argument as in Propositions 1 and 2. For any $0 \leq k \leq (N/2-1)$, we define

$$e_k := \frac{\|E-(PE)_{(2k)}\|^2}{nm-\lambda k}. \tag{A.5}$$

Observe that

$$\begin{aligned} & e_k \leq e_{k+1} \\ & \quad \Longleftrightarrow \frac{\|E-(PE)_{(2k)}\|^2}{nm-\lambda k} \leq \frac{\|E-(PE)_{(2k+2)}\|^2}{nm-\lambda(k+1)} \\ & \quad \Longleftrightarrow \frac{\|E-(PE)_{(2k)}\|^2}{nm-\lambda k} \leq \frac{\|E-(PE)_{(2k)}\|^2 - d_{2k+1}^2(PE) - d_{2k+2}^2(PE)}{nm-\lambda k-\lambda} \\ & \quad \Longleftrightarrow \frac{d_{2k+1}^2(PE) + d_{2k+2}^2(PE)}{\lambda} \leq \frac{\|E-(PE)_{2k}\|^2}{nm-\lambda k} = e_k. \end{aligned} \tag{A.6}$$

For $k=0$, we find that $e_0 \leq e_1$ is equivalent with $d_1^2(PE) + d_2^2(PE) \leq \lambda\widehat{\sigma}^2$, which is precisely our condition on λ. From the decreasing property of singular values, (A.6) implies

$$\frac{d_{2k+3}^2(PE) + d_{2k+4}^2(PE)}{\lambda} \leq \frac{\|E-(PE)_{2k+2}\|^2}{nm-\lambda(k+1)} = e_{k+1},$$

which is, by the same argument above, equivalent with $e_{k+1} \leq e_{k+2}$. We conclude that $e_0 \leq e_1 \leq e_2 \leq e_3 \leq \ldots$, proving (A.4). *In fact, this proves the sequence of* $\{e_k\}$ *could only be either globally monotonic or decreasing first and then increasing.*

Finally, since $k \leq r$, (A.3) follows immediately from (A.6) and the monotone or two-sided monotone property of $\{e_k\}$. □

A.7. Proof of Theorem 8. From Theorem 6 and Proposition 7, it suffices to show $\{2d_1^2(PE) \leq \lambda\widehat{\sigma}^2\}$ holds with high probability. The proof follows exactly the same arguments as the proof of Corollary 5 by adapting to the constant 2.

A.8. Proof of Theorem 9. To show $\widehat{k} \geq s$, it suffices to prove $d_s^2(PY) \geq \lambda\widehat{\sigma}_s^2$. Identity (A.1) gives

$$\begin{aligned}\|Y-(PY)_s\|^2 &= \|Y-(PY)_r\|^2+\sum_{j=s+1}^{r} d_j^2(PY)\\ &\leq \|Y-(PY)_r\|^2+(r-s)d_{s+1}^2(PY).\end{aligned}$$

Consequently,

$$d_s^2(PY) \geq \lambda\cdot\frac{\|Y-(PY)_s\|^2}{nm-\lambda s}$$

is implied by

$$d_s^2(PY) \geq \lambda\cdot\frac{\|Y-(PY)_r\|^2+(r-s)d_s^2(PY)}{nm-\lambda s}$$

which in turn, after a little algebra, is seen to be equivalent to

$$d_s^2(PY) \geq \lambda\cdot\frac{\|Y-(PY)_r\|^2}{nm-\lambda r}=\lambda\widehat{\sigma}_r^2.$$

By Weyl's inequality, on the event (3.9), we have

$$d_s(PY)\geq d_s(XA)-d_1(PE)\geq\sqrt{\lambda}\widehat{\sigma}_r,$$

which shows $d_s^2(PY)\geq\lambda\widehat{\sigma}_s^2$. □

A.9. Proof of Corollary 10. The proof follows immediately by noting the events in this corollary combined with (3.8) imply (3.6) and (3.9). □

A.10. Proof of Theorem 11. We define $\mathcal{C}=\{(1-C_1)\sigma^2\leq\widehat{\sigma}^2\leq(1+C_1)\sigma^2\}$ for any $0<C_1<1$. Choose $C'=(1+C_1)\sqrt{C}(1+\sqrt{2(1+\delta)})$ and $C_0>0$ such that $C=(1+C_0)^2/(1-C_1)$. Then we have

$$\begin{aligned}&\mathbb{P}\left\{\lambda\leq 2d_1^2(PE)/\widehat{\sigma}^2 \text{ or } d_s(XA)\leq\sqrt{\lambda}\widehat{\sigma}\left[\frac{1}{\sqrt{2}}+\sqrt{\frac{nm}{nm-\lambda r}}\,\right]\right\}\\ &\quad\leq\mathbb{P}\left\{d_1^2(PE)\geq(1+C_0)^2(\sqrt{m}+\sqrt{q})^2\sigma^2\right\}\\ &\qquad+\mathbb{P}\left\{d_s(XA)\leq C'\sigma(\sqrt{m}+\sqrt{q})\right\}+\mathbb{P}\{\mathcal{C}^c\}\\ &\quad=\mathbb{P}\left\{d_1^2(PE)\geq(1+C_0)^2(\sqrt{m}+\sqrt{q})^2\sigma^2\right\}+\mathbb{P}\{\mathcal{C}^c\}\\ &\quad\leq\exp\left\{-C_0^2(\sqrt{m}+\sqrt{q})^2/2\right\}+2\exp\left\{-3C_1^2nm/16\right\}\end{aligned}$$

using (3.2), (3.3) and (3.4). Invoking Corollary 10 concludes the proof. □

APPENDIX B: PROOFS OF SECTION 4

We first present several lemmas which are repeatedly used in the following proofs.

LEMMA 2. *Let c_i be some positive constants for $i = s, s+1, \ldots, t$. Then, for any $\varepsilon \in [0,1)$,*

$$\sum_{i=s}^{t} (c_i + \varepsilon S)^2 \le (1+\varepsilon)^2 \sum_{i=s}^{t} c_i^2,$$
$$\sum_{i=s}^{t} (c_i - \varepsilon S)^2 \ge (1-\varepsilon)^2 \sum_{i=s}^{t} c_i^2,$$

where $S = \sqrt{\sum_{i=s}^{t} c_i^2/(t-s)}$.

PROOF. Working out the square yields

$$\sum_{i=s}^{t} (c_i + \varepsilon S)^2 = \sum_{i=s}^{t} c_i^2 + \varepsilon^2 \sum_{i=s}^{t} c_i^2 + 2\varepsilon S \sum_{i=s}^{t} c_i \le (1 + \varepsilon^2 + 2\varepsilon) \sum_{i=s}^{t} c_i^2$$

using the Cauchy-Schwarz inequality. This proves the first statement. The second one follows by the same arguments. □

LEMMA 3. *(Interlacing inequality Horn and Johnson (2013)) Let A be a given $m \times n$ matrix, and let A_r denote a submatrix of A obtained by deleting a total of r rows and/or columns from A. Then*

$$d_k(A) \ge d_k(A_r) \ge d_{k+r}(A), \quad k = 1, \ldots, \min\{m,n\}$$

where for $M \in \mathbb{R}^{p\times q}$ we set $d_j(M) = 0$ if $j \ge \min\{p,q\}$.

LEMMA 4. *Let $n \times m$ matrix E have i.i.d. $N(0,1)$ entries and P be any $n \times n$ projection matrix with rank equal to q. Suppose $q \le m$. Then, one has*

$$\mathbb{E}[d_1^2(PE)] \ge m \tag{B.1}$$

and, for any $2 \le k \le q$,

$$\mathbb{E}[d_k(PE)] \le \sqrt{m} + \sqrt{q-k+1}, \quad \mathbb{E}[d_k(PE)] \ge \sqrt{m} - \sqrt{k}. \tag{B.2}$$

Moreover, for any $1 \le k \le q$,

$$\mathbb{P}\left\{\left|d_k(PE) - \mathbb{E}[d_k(PE)]\right| \ge t\right\} \le 2\exp(-t^2/2), \quad \forall\ t \ge 0. \tag{B.3}$$

For the case $m < q$, similar results hold for any $1 \le k \le m$, by switching q and m.

PROOF OF LEMMA 4. We only prove the case when $q \le m$. The complementary case $q > m$ can be derived using the fact $d_k(M) = d_k(M^T)$ for any matrix M.

We start by writing the eigenvalue decomposition of P as $P = U\Lambda U^T$ with orthogonal matrix $U \in \mathbb{R}^{n\times n}$ and diagonal matrix Λ whose first q diagonal elements are 1 and 0 elsewhere. Thus, for any $1 \le j \le q$, $d_j^2(PE) = \lambda_j(E^TPE) = d_j^2(\Lambda U^T E)$ where $\lambda_j(M)$ denotes the jth largest eigenvalue of M. Note that $Z = \Lambda U^T E \in \mathbb{R}^{n\times m}$ has $q \times m$ submatrix with i.i.d. $N(0,1)$ entries while the remaining $(n-q)\times m$ entries are all 0. For any $1 \le k \le q$, by Lemma 3, we have $d_k(Z) \le d_1(\overline{Z}_{q-k+1})$ and $d_k(Z) \ge d_k(\overline{Z}_k)$, where $\overline{Z}_j$ denotes the matrix made of the first j rows of Z. Then (B.2) follows immediately from Theorem 5.32 of Vershynin (2012).

Concentration inequality (B.3) follows from the Gaussian concentration inequality of (Giraud, 2015, page 221) and the fact that each singular value function is 1-Lipschitz with respect to the Frobenius norm.

Finally, $\mathbb{E}[d_1^2(PE)] \ge \mathbb{E}[d_1^2(\overline{Z}_1)]$ and

$$\mathbb{E}\left[d_1^2(\overline{Z}_1)\right] = \mathbb{E}\left[\overline{Z}_1^T\overline{Z}_1\right] = \sum_{i=1}^{m}\mathbb{E}\left[\overline{Z}_{1i}^2\right] = m$$

implying (B.1). This completes the proof. □

The next lemma proves one-side concentration of $\|PE - (PE)_k\|$ around its mean.

LEMMA 5. *Let $n \times m$ matrix E have i.i.d. $N(0,1)$ entries and P be any $n \times n$ projection matrix with rank equal to q. Assume $q \le m$. Then, for any $1 \le k < q$ and any $\varepsilon \in (0,1)$, there exists a constant $C = C(\varepsilon) > 0$, such that*

$$\mathbb{P}\left\{\|PE - (PE)_k\|^2 \le (1-\varepsilon)\mathbb{E}[\|PE - (PE)_k\|^2]\right\} \le e^{-Cm}.$$

Similar results hold by switching q and m when $m < q$ for $1 \le k < m$.

PROOF OF LEMMA 5. Fix $1 \le k < q$. Note that

$$\|PE - (PE)_k\|^2 = \|PE\|^2 - \sum_{j=1}^{k} d_j^2(PE).$$

We first study $\sum_{j=1}^{k} d_j^2(PE)$. From (B.3) in Lemma 4, for any $k \le j \le q$, we have

$$\mathbb{P}\left\{d_j^2(PE) \ge (\mathbb{E}[d_j(PE)] + t)^2\right\} \le \exp(-t^2/2), \quad \forall\, t \ge 0.$$

This further implies

$$\mathbb{P}\left\{\sum_{j=1}^{k} d_j^2(PE) \le \sum_{j=1}^{k}(\mathbb{E}[d_j(PE)]+t)^2\right\} \ge 1-k\exp(-t^2/2).$$

Let $c>0$ be sufficiently small and choose $t^2=(c^2/k)\sum_{j=1}^{k}(\mathbb{E}[d_j(PE)])^2$. Invoking Lemma 2 and using $\mathbb{E}[d_j^2(PE)] \ge (\mathbb{E}[d_j(PE)])^2$ yield

$$\begin{aligned}\mathbb{P}&\left\{\sum_{j=1}^{k} d_j^2(PE) \le (1+c)^2\sum_{j=1}^{k}\mathbb{E}[d_j^2(PE)]\right\}\\ &\ge \mathbb{P}\left\{\sum_{j=1}^{k} d_j^2(PE) \le (1+c)^2\sum_{j=1}^{k}(\mathbb{E}[d_j(PE)])^2\right\}\\ &\ge 1-k\exp\left[-\frac{c^2\sum_{j=1}^{k}(\mathbb{E}[d_j(PE)])^2}{2k}\right].\end{aligned}$$

From $(\mathbb{E}[d_j(PE)])^2 \ge \mathbb{E}[d_j^2(PE)]-1$ in the proof of Lemma 1 of Giraud (2011), observe that

$$\begin{aligned}\sum_{j=1}^{k}(\mathbb{E}[d_j(PE)])^2 &\ge \frac{k}{q}\sum_{j=1}^{q}(\mathbb{E}[d_j(PE)])^2 \ge \frac{k}{q}\sum_{j=1}^{q}(\mathbb{E}[d_j^2(PE)]-1)\\ &= \frac{k}{q}\cdot\mathbb{E}[\|PE\|^2]-k = k(m-1).\end{aligned}$$

In the last equality, we use the fact that $\|PE\|^2$ has a central χ^2_{qm} distribution. This further implies

$$\mathbb{P}\left\{\sum_{j=1}^{k} d_j^2(PE) \le (1+c)^2\sum_{j=1}^{k}\mathbb{E}[d_j^2(PE)]\right\} \ge 1-e^{-Cm}$$

for some constant $C=C(c)>0$. On the other hand, since $\|PE\|^2 \sim \chi^2_{qm}$, using (3.3) yields

$$\mathbb{P}\left\{\|PE\|^2 \ge (1-c)^2\mathbb{E}\left[\|PE\|^2\right]\right\} \ge 1-\exp(-c^2qm/4).$$

Combining these two displays concludes our proof. □

B.1. Proof of Proposition 12. We first show that $\widehat{\lambda}_{t+1} := \widehat{\lambda}_{\widehat{k}_t}$ defined in (4.4) is decreasing in $\widehat{k}_t$. To simplify notation, we write k for $\widehat{k}_t$. Since it is immediate to verify the decreasing property when $2k \geq N$, we consider $1 \leq k < N/2$ only. It suffices to show $(1-\varepsilon)\widehat{R}_t/\widehat{U}_t + k$ is increasing. Recall that $S_j = \mathbb{E}[d_j^2(Z)]$ for any $1 \leq j \leq N$ and $\widehat{R}_t$ and $\widehat{U}_t$ are defined as

$$\widehat{R}_t \;:=\; (n-q)m + \sum_{j=2k+1}^{N} S_j, \qquad \widehat{U}_t \;:=\; S_1 \vee (S_{2k+1} + S_{2k+2})\,, \tag{B.4}$$

by using k in lieu of $\widehat{k}_t$. First, we show

$$a_k := \frac{(1-\varepsilon)\left[(n-q)m + \sum_{j=2k+1}^{N} S_j\right]}{S_{2k+1} + S_{2k+2}} + k \tag{B.5}$$

is increasing. This follows from

$$\frac{a_k - a_{k-1}}{1-\varepsilon} \geq \frac{(n-q)m + \sum_{j=2k+1}^{N} S_j}{S_{2k+1} + S_{2k+2}} - \frac{(n-q)m + \sum_{j=2k+1}^{N} S_j}{S_{2k-1} + S_{2k}} > 0.$$

It remains to show, if $S_1 \geq S_{2k+1} + S_{2k+2}$, the sequence

$$b_k := \frac{(1-\varepsilon)\left[(n-q)m + \sum_{j=2k+1}^{N} S_j\right]}{S_1} + k \tag{B.6}$$

is increasing in k. This is guaranteed by the fact that b_k starts increasing after k such that $S_1 \geq S_{2k+1} + S_{2k+2}$. This proves $\widehat{\lambda}_{t+1}$ is decreasing in $\widehat{k}_t$. We conclude the proof of $\widehat{\lambda}_{t+1} < \widehat{\lambda}_t$ for any $t \geq 0$ by noting that $\widehat{\lambda}_0$ is greater than the λ obtained from (4.4) by using $\widehat{k}_t = 0$.

Next, we show that, for given $\widehat{\lambda}_t > \widehat{\lambda}_{t+1}$, we have $\widehat{k}_t \leq \widehat{k}_{t+1}$. Suppose $\widehat{k}_t > \widehat{k}_{t+1}$, we obtain

$$\frac{\|Y - (PY)_{\widehat{k}_{t+1}}\|^2}{nm - \widehat{\lambda}_{t+1}\widehat{k}_{t+1}} \leq \frac{\|Y - (PY)_{\widehat{k}_t}\|^2}{nm - \widehat{\lambda}_{t+1}\widehat{k}_t} \iff \frac{\sum_{j=\widehat{k}_{t+1}}^{\widehat{k}_t} d_j^2(PY)}{\widehat{k}_t - \widehat{k}_{t+1}} \leq \frac{\|Y - (PY)_{\widehat{k}_t}\|^2}{nm/\widehat{\lambda}_{t+1} - \widehat{k}_t}$$

Similarly,

$$\frac{\|Y - (PY)_{\widehat{k}_{t+1}}\|^2}{nm - \widehat{\lambda}_{t}\widehat{k}_{t+1}} \geq \frac{\|Y - (PY)_{\widehat{k}_t}\|^2}{nm - \widehat{\lambda}_{t}\widehat{k}_t} \iff \frac{\sum_{j=\widehat{k}_{t+1}}^{\widehat{k}_t} d_j^2(PY)}{\widehat{k}_t - \widehat{k}_{t+1}} \geq \frac{\|Y - (PY)_{\widehat{k}_t}\|^2}{nm/\widehat{\lambda}_{t} - \widehat{k}_t}$$

which is a contradiction as $\widehat{\lambda}_t > \widehat{\lambda}_{t+1}$.

Finally, we show $\widehat{k}_t \le r$ for all $0 \le t \le T$. We start defining the event

$$\mathcal{E} := \left\{ \bigcap_{k=1}^{N} \mathcal{E}_k \right\} \bigcap \left\{ 2d_1^2(PE) \le \widehat{\lambda}_0 \widehat{\sigma}^2 \right\}$$

$$\mathcal{E}_k := \left\{ \frac{\|E - (PE)_{2k\wedge N}\|^2}{d_1^2(PE) \vee \left[d_{2k+1}^2(PE) + d_{2k+2}^2(PE) \right]} \ge \frac{(1-\varepsilon)\widehat{R}_t}{\widehat{U}_t} \right\} \tag{B.7}$$

with $\widehat{\lambda}_0$ chosen as (4.2) and $\widehat{R}_t$ and $\widehat{U}_t$ defined in (4.6). We will work on this event in the remainder of the proof. From Theorem 6 and Proposition 7, we know $1 \le \widehat{k}_0 \le r$ where $\widehat{k}_0$ is the selected rank from (4.3). Let $\widehat{\lambda}_1$ be updated via (4.4) using $\widehat{k}_0$. In order to guarantee $\widehat{k}_1 \le r$, from (4.1) and (A.2), it suffices to show

$$d_1^2(PE) \le \frac{\|E - (PE)_{(2r)\wedge N}\|^2}{nm/\widehat{\lambda}_1 - r}. \tag{B.8}$$

On the one hand, the choice of $\widehat{\lambda}_1$ satisfies

$$\begin{aligned} \widehat{\lambda}_1 &= \frac{nm}{(1-\varepsilon)\widehat{R}_t/\widehat{U}_t + \widehat{k}_0} \\ &\overset{\mathcal{E}}{\ge} \frac{nm}{\|E - (PE)_{(2\widehat{k}_0)\wedge N}\|^2 / \left(d_{2\widehat{k}_0+1}^2(PE) + d_{2\widehat{k}_0+2}^2(PE) \right) + \widehat{k}_0} \end{aligned}$$

by which (A.3) in Lemma 1 guarantees

$$\frac{\|E - (PE)_{(2\widehat{k}_0)\wedge N}\|^2}{nm - \widehat{\lambda}_1 \widehat{k}_0} \le \frac{\|E - (PE)_{(2r)\wedge N}\|^2}{nm - \widehat{\lambda}_1 r} \tag{B.9}$$

provided that $\widehat{k}_0 \le r$. On the other hand,

$$\widehat{\lambda}_1 \overset{\mathcal{E}}{\ge} \frac{nm}{\|E - (PE)_{(2\widehat{k}_0)\wedge N}\|^2 / d_1^2(PE) + \widehat{k}_0}$$

which is equivalent with

$$d_1^2(PE) \le \frac{\|E - (PE)_{(2\widehat{k}_0)\wedge N}\|^2}{nm/\widehat{\lambda}_1 - \widehat{k}_0}. \tag{B.10}$$

Combining (B.9) with (B.10) proves (B.8). After repeating this argument, on the event $\mathcal{E}$, we find

$$d_1^2(PE) \le \frac{\|E - (PE)_{(2r)\wedge N}\|^2}{nm/\widehat{\lambda}_t - r} \le \frac{\|Y - (PY)_r\|^2}{nm/\widehat{\lambda}_t - r} \tag{B.11}$$

for any $t \geq 1$, which implies $\widehat{k}_t \leq r$.

To conclude our proof, we show that $\mathcal{E}$ holds with probability tending to 1. Again, we only prove for $q \leq m$ since the case of $q > m$ can be obtained similarly. The proof of Lemma 4 shows that there exists a $q \times m$ matrix Z with i.i.d. standard normal entries such that $d_k(PE) = \sigma d_k(Z)$ for any $1 \leq k \leq q$. Without loss of generality, we assume $\sigma = 1$. We first consider the event $\{2d_1^2(Z) \leq \widehat{\lambda}_0 \widehat{\sigma}^2\}$. From (B.3) and using $\mathbb{E}[d_1^2(Z)] \geq m$ in Lemma 4, we have

$$\mathbb{P}\Big\{ d_1^2(Z) \leq (1 + C_1)^2 \mathbb{E}\left[d_1^2(Z)\right] \Big\} \geq 1 - \exp\left(-C_1^2 m/2\right). \tag{B.12}$$

Then, using (3.3), we have

$$\begin{aligned}
\mathbb{P}\left\{2d_1^2(Z) \leq \widehat{\lambda}_0 \widehat{\sigma}^2\right\} &\geq 1 - \mathbb{P}\Big\{ d_1^2(Z) \geq (1 + C_1)^2 \mathbb{E}\left[d_1^2(Z)\right] \Big\} \\
&\quad - \mathbb{P}\{\widehat{\sigma}^2 \leq (1 - C_2)\} \\
&\geq 1 - \exp\left(-C_1^2 m^2/2\right) - \exp\left(-C_2^2 nm/4\right),
\end{aligned}$$

for any $0 < C_1, C_2 < 1$ satisfying $(1 + C_1)^2/(1 - C_2) = 1 + \varepsilon$.

Next we quantify the event $\mathcal{E}_k$ in (B.7) which is equivalent to

$$\mathcal{E}_k := \left\{ \frac{\|E - (PE)_{2k \wedge N}\|^2}{d_1^2(Z) \vee \left[d_{2k+1}^2(Z) + d_{2k+2}^2(Z)\right]} \geq \frac{(1 - \varepsilon)\widehat{R}_t}{\widehat{U}_t} \right\}.$$

We will proceed to to show, that each of the random quantities concentrate around their means. We shall only consider when $n > q$ since $n = q$ is easy to obtain by using the same arguments and the fact $\|E - PE\| = 0$ for $n = q$.

Fix $1 \leq k \leq r$ and consider two cases:

(1) When $2k < q$, observe that

$$\|E - (PE)_{2k}\|^2 = \|E - PE\|^2 + \|PE - (PE)_{2k}\|^2.$$

Since $\|E - PE\|^2$ has a central $\chi^2_{(n-q)m}$ distribution, inequality (3.3) gives

$$\mathbb{P}\left\{\|E - PE\|^2 \leq (1 - C_3)(n - q)m\right\} \leq \exp\left(-C_3^2 (n - q)m/4\right) \tag{B.13}$$

for any $C_3 \in (0, 1)$. This bound together with Lemma 5 yield

$$\mathbb{P}\Big\{\|E - (PE)_{2k}\|^2 \leq (1 - C_3)\widehat{R}_t \sigma^2\Big\} \leq \exp\left(-C_3^2 (n - q)m/4\right) + \exp\left(-C_3' m\right)$$

with some constant $C_3' > 0$.

On the other hand, recall that $\widehat{U}_t = S_1 \vee (S_{2k+1} + S_{2k+2})$ from (B.4). If $d_1^2(Z) \le d_{2k+1}^2(Z) + d_{2k+2}^2(Z)$ such that $\widehat{U}_t = S_{2k+1} + S_{2k+2}$, the concentration inequality (B.3) in Lemma 4 gives

$$\mathbb{P}\Big\{ d_{2k+1}^2(Z) + d_{2k+2}^2(Z) \le \sum_{j=2k+1}^{2k+2} \left(\mathbb{E}[d_j(Z)] + t\right)^2 \Big\} \ge 1 - 2\exp(-t^2/2),$$

for any $t \ge 0$. Take $t^2 = (C_4^2/2)\sum_{j=2k+1}^{2k+2}(\mathbb{E}[d_j(Z)])^2$ with arbitrary $C_4 \in (0,1)$ and invoke Lemma 2 to obtain

$$\mathbb{P}\Big\{ d_{2k+1}^2(Z) + d_{2k+2}^2(Z) \le (1+C_4)^2 \widehat{U}_t \Big\} \ge 1 - 2\exp\left(-C_4^2 m/2\right).$$

For the exponent in the probability tail, we use $(\mathbb{E}[d_j(Z)])^2 \ge \mathbb{E}[d_j^2(Z)] - 1 = S_j - 1$ from the proof of Lemma 1 in Giraud (2011) and $S_1 \ge m$ from Lemma 4 to obtain

$$\sum_{j=2k+1}^{2k+2} (\mathbb{E}[d_j(Z)])^2 \ge S_{2k+1} + S_{2k+2} - 2 \ge S_1 - 2 \ge m - 2.$$

If $d_1^2(Z) \ge d_{2k+1}^2(Z) + d_{2k+2}^2(Z)$, the concentration inequality (B.12) gives a similar result as above. Choose $1 - \varepsilon = (1 - C_3)/(1 + C_4)^2$ and conclude that

$$\mathbb{P}(\mathcal{E}_k^c) \le 2\exp\left(-C_4^2 m/2\right) + \exp\left(-C_3^2(n-q)m/4\right) + \exp\left(-C_3' m\right),$$

for any $2k < q$.

(2) If $2k \ge q$, we immediately have $\widehat{R}_t = (n-q)m$ and $\widehat{U}_t = S_1$. Therefore, (B.12) and (B.13) give

$$\begin{aligned}\mathbb{P}\{\mathcal{E}_k^c\} &= \mathbb{P}\left\{ \frac{\|E - PE\|^2}{d_1^2(Z)} \le \frac{(1-C_3)\widehat{R}_t}{(1+C_4)^2 \widehat{U}_t} \right\} \\ &\le \exp\left(-C_4^2 m/2\right) + \exp\left(-C_3^2(n-q)m/4\right).\end{aligned}$$

Taking the union bound over all $1 \le k \le r$ concludes our proof. □

B.2. Proof of Theorem 13. The choice $\widehat{\lambda}_0 = C(\sqrt{m} + \sqrt{q})^2$ is feasible by using

$$\mathbb{E}[d_1^2(Z)] \le (\mathbb{E}[d_1(Z)])^2 + 1 \le (\sqrt{m} + \sqrt{q})^2 + 1$$

and choosing ε such that

$$2(1+\varepsilon)[(\sqrt{m}+\sqrt{q})^2+1] \le C(\sqrt{m}+\sqrt{q})^2.$$

This implies $\widetilde{k} \le \widehat{k}$ from the proof of Proposition 12. Invoking Proposition 12 once again concludes the proof of Theorem 13. □

B.3. Proof of Theorem 14. We work on the event

$$\mathcal{E}' = \{\widehat{\sigma} \le (1+\varepsilon)\sigma\} \cap \mathcal{E}$$

where $\mathcal{E}$ is defined in (B.7). From (3.4) and the proof of Proposition 12, $\mathcal{E}'$ holds with probability tending to 1.
Let $\widehat{k}_0, \widehat{k}_1, \ldots$ be the sequence of selected ranks from (4.3) and (4.5). Our assumption

$$\begin{aligned} d_{k_0}(XA) &\ge C''\sigma\sqrt{\lambda_0} \ge (1+\varepsilon)\sigma\sqrt{\lambda_0}\left(\frac{\sqrt{2}}{2}+\sqrt{\frac{nm}{nm-\lambda_0 r}}\right) \\ &\overset{\mathcal{E}'}{\ge} d_1(PE) + \frac{\|E\|}{\sqrt{nm/\lambda_0 - r}} \ge d_1(PE) + \frac{\|Y-(PY)_r\|}{\sqrt{nm/\lambda_0 - r}} \end{aligned}$$

implies that $\widehat{k}_0 \in [k_0, r]$ by Theorem 9 under the rank constraint (3.10). This, in turn, from the decreasing property of $\widehat{\lambda}_t$ in Proposition 12, yields $\widehat{\lambda}_1 \le \lambda_1$, where $\widehat{\lambda}_1$ is computed from (4.4) by using $\widehat{k}_0$. Moreover, $\lambda_1 \le \lambda_0$ implies r also satisfies (3.10) for λ_1. Hence,

$$\begin{aligned} d_{k_1}(XA) &\ge C''\sigma\sqrt{\lambda_1} \ge (1+\varepsilon)\sigma\sqrt{\lambda_1}\left(\frac{\sqrt{2}}{2}+\sqrt{\frac{nm}{nm-\lambda_1 r}}\right) \\ &\ge (1+\varepsilon)\sigma\sqrt{\widehat{\lambda}_1}\left(\frac{\sqrt{2}}{2}+\sqrt{\frac{nm}{nm-\widehat{\lambda}_1 r}}\right) \end{aligned}$$

and by the same argument above, we have $\widehat{k}_1 \in [k_1, r]$. Now the proof follows by repeating these arguments and the fact that $k_T = r$. □

B.4. Proof of Theorem 15. We work on the event

$$\mathcal{Z} = \{d_1(Z) = \sqrt{m}+\sqrt{q}\}$$

which holds almost surely as $N = q \wedge m \to \infty$ from the celebrated result of Bai and Yin (1993).

We only prove the result for $q \le m$ since the case $q > m$ can be derived in a similar way. From the first signal condition (4.8), Theorem 14 immediately implies $\widehat{k}_0 \ge q/2$. Without loss of generality, we assume $\widehat{k}_0 = q/2$, and q is even for notational simplicity. Step (4.4) implies the following update

$$\lambda_1 = \frac{nm}{(1-c)(n-q)m/(\sqrt{m}+\sqrt{q})^2 + q/2}. \tag{B.14}$$

To show $\widehat{k}_1 = r$, from (9), it suffices to show

$$d_r(XA) \ge d_1(PE) + \sqrt{\lambda_1}\widehat{\sigma}_r. \tag{B.15}$$

Observe that, with high probability,

(B.16)

$$\lambda_1\widehat{\sigma}_r^2 = \frac{\|Y-(PY)_r\|^2}{nm/\lambda_1 - r} \le \frac{\|E\|^2}{nm/\lambda_1 - r} \le \frac{(1+\varepsilon)nm(\sqrt{m}+\sqrt{q})^2\sigma^2}{7(n-q)m/8 - (r-q/2)(\sqrt{m}+\sqrt{q})^2}$$

for any $\varepsilon \in (0, 1/4)$, by choosing $c = 7/8$ and using (3.4) in the last inequality. We consider two cases:

Case 1:

$$\frac{nm}{\lambda_0} \ge \frac{1+\delta}{\delta}q \quad \Longleftrightarrow \quad nm \ge \frac{1+\delta}{\delta}2C(\sqrt{m}+\sqrt{q})^2 q. \tag{B.17}$$

This implies $r \le q$ from the rank constraint (3.10). Hence,

$$\lambda_1\widehat{\sigma}_r^2 \le \frac{(1+\varepsilon)nm(\sqrt{m}+\sqrt{q})^2\sigma^2}{7(n-q)m/8 - (q/2)(\sqrt{m}+\sqrt{q})^2} \le \frac{C(1+\varepsilon)(1+\delta)}{1+5\delta/16}(\sqrt{m}+\sqrt{q})^2\sigma^2$$

by using (B.17), $m/(\sqrt{m}+\sqrt{q})^2 \le 1$ and $C \ge 8/7$. This further implies

$$d_1(PE) + \sqrt{\lambda_1}\widehat{\sigma}_r \le C'\left[1 + \sqrt{\frac{1+\delta}{1+5\delta/16}}\right](\sqrt{m}+\sqrt{q})\sigma$$

by taking $C' = \sqrt{C(1+\varepsilon)}$. Thus, (B.15) holds under the assumed signal condition (4.9).

Case 2:

$$\frac{nm}{\lambda_0} \le \frac{1+\delta}{\delta}q \quad \Longleftrightarrow \quad nm \le \frac{1+\delta}{\delta}\cdot 2C(\sqrt{m}+\sqrt{q})^2 q. \tag{B.18}$$

It follows that

$$r \le \frac{\delta}{1+\delta}\frac{nm}{\lambda_0} = \frac{\delta}{1+\delta}\cdot\frac{nm}{2C(\sqrt{m}+\sqrt{q})^2}.$$

Plugging the above upper bound of r into (B.16) yields

$$\lambda_1\widehat{\sigma}_r^2 \le \frac{(1+\varepsilon)(\sqrt{m}+\sqrt{q})^2\sigma^2}{7/8-\delta/(2C(1+\delta))-(q/n)[7/8-(\sqrt{m}+\sqrt{q})^2/(2m)]}. \tag{B.19}$$

If $7/8 \le (\sqrt{m}+\sqrt{q})^2/(2m)$, we further have

$$\lambda_1\widehat{\sigma}_r^2 \le \frac{(1+\varepsilon)(\sqrt{m}+\sqrt{q})^2\sigma^2}{7/8-\delta/(2C(1+\delta))} \le \frac{C(1+\varepsilon)(1+\delta)}{1+3\delta/8}(\sqrt{m}+\sqrt{q})^2\sigma^2$$

by using $C \ge 8/7$. By repeating the same arguments before, (B.15) holds. Finally, we show (B.15) still holds when $7/8 \ge (\sqrt{m}+\sqrt{q})^2/(2m)$. Recall that $\widehat{k}_0 = q/2$ which implies

$$\frac{q}{2} \le \frac{\delta}{1+\delta}\frac{nm}{\lambda_0} \quad \Longleftrightarrow \quad \frac{q}{n} \le \frac{\delta m}{C(1+\delta)(\sqrt{m}+\sqrt{q})^2}.$$

Combining this with (B.19) gives

$$\lambda_1\widehat{\sigma}_r^2 \le \frac{C(1+\varepsilon)(1+\delta)}{1+\delta/8}(\sqrt{m}+\sqrt{q})^2\sigma^2.$$

Repeating the same arguments proves (B.15), hence concludes the proof. □

B.5. Proof of Proposition 16. Again, we work on the event $\mathcal{Z}$, defined in the proof of Theorem 14 and we only prove the case $q \le m$ since the complementary case follows by the same arguments.

Since $\widehat{k}_0 \ge N/2$, we assume $\widehat{k}_0 = N/2$ so this is the most difficult case. We take N even for simplicity. Thus, step (4.4) implies the update of λ as

$$\widehat{\lambda}_1 = \frac{nm}{(1-\varepsilon)(n-q)m/(\sqrt{m}+\sqrt{q})^2+q/2} = \frac{nm}{(7/8)(n-q)m/(\sqrt{m}+\sqrt{q})^2+q/2}$$

by taking $\varepsilon = 1/8$. On the one hand, a little algebra shows that

$$\widehat{K}_1 \;=\; \frac{nm}{\widehat{\lambda}_1} \;=\; \frac{7(n-q)m}{8(\sqrt{m}+\sqrt{q})^2}+\frac{q}{2} \;\ge\; \frac{9}{8}q$$

by using our assumption

$$\frac{nm}{\lambda_0} = \frac{nm}{2C(\sqrt{m}+\sqrt{q})^2} \ge \frac{3}{4}q. \tag{B.20}$$

with $C = 8/7$ and $m/(\sqrt{m}+\sqrt{q})^2 \le 1$. Hence $\widehat{k}_2$ is selected from $[q/2,\ q]$ according to (4.5). On the other hand, similar as (B.16) and by using $r \le q$, we have

$$\widehat{\lambda}_1\widehat{\sigma}_r^2 \le \frac{(1+\varepsilon')nm}{7(n-q)m/8-(q/2)(\sqrt{m}+\sqrt{q})^2}(\sqrt{m}+\sqrt{q})^2\sigma^2$$

with high probability for any $\varepsilon' \in (0, 1/4)$. Using (B.20), $m/(\sqrt{m}+\sqrt{q})^2 \le 1$ and $C > 8/7$ yields

$$\widehat{\lambda}_1\widehat{\sigma}_r^2 \le 12C(1+\varepsilon')(\sqrt{m}+\sqrt{q})^2\sigma^2.$$

Taking $C' = \sqrt{C(1+\varepsilon')}$ concludes

$$d_1(PE) + \sqrt{\widehat{\lambda}_1}\widehat{\sigma}_r \le C'(1+2\sqrt{3})(\sqrt{m}+\sqrt{q})\sigma,$$

which, by using our signal condition and invoking Theorem 9, completes the proof. □

APPENDIX C: PROOFS OF SECTION 5

The following lemmas are critical for extending the previous results to general errors with heavy tail distributions.

LEMMA 6. *Let $E \in \mathbb{R}^{n\times m}$ have independent entries with mean zero, variance σ^2 and fourth moment $\gamma < \infty$. For any $\varepsilon \in (0,1)$, we have*

$$\mathbb{P}\Big\{\big|\|E\|^2 - nm\sigma^2\big| > \varepsilon nm\sigma^2\Big\} \ \le\ \frac{\gamma/\sigma^4 - 1}{\varepsilon^2 nm}.$$

PROOF OF LEMMA 6. Since $\|E\|^2 = \sum_{i=1}^n\sum_{j=1}^m E_{ij}^2$ with $\mathbb{E}[\|E\|^2] = nm\sigma^2$ and $\mathrm{Var}(\|E\|^2) = nm(\gamma-\sigma^4)$, the result follows from the Bienaymé-Chebyshev inequality. □

LEMMA 7. *Let E has independent entries with mean zero and unit variance. Assume $n = O(m^\alpha)$ for some $\alpha \in [0,1)$. Then, for any $\varepsilon \in (0,1)$, one has*

$$(1-\varepsilon)\sqrt{m} \le d_k(E) \le (1+\varepsilon)\sqrt{m}, \qquad \text{for all } 1 \le k \le n \tag{C.1}$$

with probability at least $1 - 2\exp(-c\varepsilon^2 m^{1-\alpha})$ where $c > 0$ is some absolute constant. Moreover, for any $\varepsilon \in (0,1)$, we have

$$\|E - (E)_k\|^2 \ge (1-\varepsilon)^2(n-k)m, \qquad \text{for all } 1 \le k \le n \tag{C.2}$$

with probability converging to 1 as $m \to \infty$. Similar results hold for $m = O(n^\alpha)$ with m and n switched.

PROOF OF LEMMA 7. Let us first consider $n = O(m^\alpha)$ for some $0 \le \alpha < 1$. Fix any $1 \le k \le n$. Lemma 3 implies $d_k(\overline{E}_k) \le d_k(E) \le d_1(\overline{E}_{n-k+1})$ where $\overline{E}_j$ is the matrix made of the first j rows of E for any $1 \le j \le n$. For notational simplicity, we write $F^j = (\overline{E}_j)^T \in \mathbb{R}^{m\times j}$. Notice that, F^j still has independent entries, each row $F^j_{i\cdot}$ of F^j has $j \times j$ identity covariance matrix and

$$\|F^j_{i\cdot}\|_2 = \sum_{\ell=1}^{j} \left(F^j_{i\ell}\right)^2 = \sum_{\ell=1}^{j} E^2_{\ell i} = j, \quad a.s.$$

by the law of large number. Thus, we can invoke Theorem 5.41 in Vershynin (2012) for F^{n-k+1} and F^k to obtain

$$\begin{aligned}
\mathbb{P}\left\{d_1(F^{n-k+1}) \ge \sqrt{m} + t\sqrt{n-k+1}\right\} &\le (n-k+1)\exp(-c_0 t^2), \quad t \ge 0,\\
\mathbb{P}\left\{d_1(F^{k}) \le \sqrt{m} - t\sqrt{k}\right\} &\le k\exp(-c_1 t^2), \quad t \ge 0,
\end{aligned}$$

for some constant $c_0, c_1 > 0$. For any $\varepsilon \in (0,1)$, choose $t = \varepsilon\sqrt{m}/\sqrt{n-k+1}$ in the first display and $t = \varepsilon\sqrt{m}/\sqrt{k}$ in the second display, and get

$$\begin{aligned}
&\mathbb{P}\left\{(1-\varepsilon)\sqrt{m} \le d_k(E) \le (1+\varepsilon)\sqrt{m}\right\}\\
&\quad \ge 1 - \exp\left(-\frac{c_0\varepsilon^2 m}{n-k+1} + \log(n-k+1)\right) - \exp\left(-\frac{c_1\varepsilon^2 m}{k} + \log k\right)\\
&\quad \ge 1 - 2\exp\left(-c_2\varepsilon^2 m^{1-\alpha}\right)
\end{aligned}$$

for some constant $c_2 > 0$. We use $n = O(m^\alpha)$ in the last inequality. Taking the union bound over $1 \le k \le n$ concludes the proof of (C.1).

If $m = O(n^\alpha)$, write $F = E^T$. Using Lemma 3 again gives $d_k(\overline{F}_k) \le d_k(F) \le d_1(\overline{F}_{n-k+1})$ for any $1 \le k \le m$. We observe that $(\overline{F}_j)^T \in \mathbb{R}^{n\times j}$ has independent entries, identity covariance matrix and the ℓ_2 norm of each row equal to $\sqrt{j}$ almost surely. Repeating the previous arguments will prove (C.1).

We proceed to show (C.2) for $n = O(m^\alpha)$ only since the case $m = O(n^\alpha)$ can be easily extended. For any $1 \le k \le n$, notice that $\|E - (E)_k\|^2 = \sum_{j=k+1}^{n} d_j^2(E)$. (C.2) follows immediately from (C.1) and Lemma 6. This completes the proof. □

C.1. Proof of Theorem 17. Without loss of generality, we assume $\mathbb{E}[E_{ij}^2] = 1$. We start by defining the following event

$$\mathcal{E}'' := \{|\|E\|^2 - nm| \le \varepsilon nm\} \cap \{d_1(E) = \sqrt{m} + \sqrt{n}\}$$

for any $\varepsilon \in (0,1)$. By Lemma 6 and Bai and Yin (1993), we have $\mathbb{P}(\mathcal{E}'') \to 1$ as $n, m \to \infty$. To show $\widehat{k} \le r$, from Theorem 7, it suffices to show (3.7). This is indeed the case since

$$\text{(C.3)} \qquad \frac{2d_1^2(PE)}{\widehat{\sigma}^2} \le \frac{2d_1^2(E)}{\widehat{\sigma}^2} \overset{\mathcal{E}''}{\le} \frac{2(\sqrt{m}+\sqrt{n})^2}{1-\varepsilon} = \frac{2(\sqrt{m}+\sqrt{q})^2}{1-\varepsilon}(1+o(1))$$

by using $d_1(PE) \le d_1(P)d_1(E) = d_1(E)$ in the first inequality and choosing $\lambda = 2(\sqrt{m}+\sqrt{q})^2/(1-\varepsilon)$. On the other hand, to show $\widehat{k} \ge s$, from Theorem 9, we need to verify (3.9). By using (3.8), (3.10) and (C.3), it follows that

$$d_1(PE) + \sqrt{\lambda}\widehat{\sigma}_r \le \sqrt{\lambda}\left[\frac{1}{\sqrt{2}} + \sqrt{1+\delta}\right]\widehat{\sigma} \overset{\mathcal{E}}{\le} C\sigma(\sqrt{m}+\sqrt{q}) \le d_s(XA)$$

by choosing $C = \sqrt{(1+\varepsilon)/(1-\varepsilon)}(1+\sqrt{2(1+\delta)})$. This completes the proof. □

C.2. Proof of Theorem 18. Without loss of generality, we assume E_{ij} has unit variance. As in the proof of Theorem 14, we first need to check the following two properties for $\widehat{\lambda}_t$ and $\widehat{k}_t$ obtained in (5.2) and (5.3): (1) $\widehat{\lambda}_t$ is decreasing; (2) $\widehat{k}_t \le r$. Obviously, (1) is straightforward from (5.2) and (5.3). To show (2), we only prove the case $n = O(m^\alpha)$ since the case $m = O(n^\alpha)$ is similar to derive.

We define the following event which is analogous to (B.7) in the proof of Proposition 12.

$$\mathcal{E}''' := \bigcap_{k=1}^{q}\left\{\frac{\|E-(E)_{(2k)\wedge n}\|^2}{d_1^2(E)\vee\left[d_{2k+1}^2(E)+d_{2k+2}^2(E)\right]} \ge (1-\varepsilon)[n/2-k]_+\right\} \bigcap \left\{2d_1^2(E) \le \widehat{\lambda}_0\widehat{\sigma}^2\right\}$$

On the event $\mathcal{E}'''$, $\widehat{k}_t \le r$ follows by the same arguments used in the proof of Proposition 12 except PE and PY are now replaced by E and Y, respectively. Thus, it suffices to show $\mathcal{E}'''$ holds with high probability. First, observe that, for any $\varepsilon_1, \varepsilon_2 \in (0,1)$, and some constant c, c',

$$\begin{aligned}\mathbb{P}\left\{2d_1^2(E) \le \widehat{\lambda}_0\widehat{\sigma}^2\right\} &\ge 1 - \mathbb{P}\{d_1^2(E) \ge (1+\varepsilon_1)^2 m\} - \mathbb{P}\{\widehat{\sigma}^2 \le (1-\varepsilon_2)nm\} \\ &\ge 1 - \exp(-c\varepsilon_1^2 m) - c'(\varepsilon_2^2 nm)^{-1}.\end{aligned}$$

We use Lemmas 6 and 7 in the second inequality. Choosing ε in $\widehat{\lambda}_0$ such that $1+\varepsilon = (1+\varepsilon_1)^2/(1-\varepsilon_2)$ proves the last event in $\mathcal{E}'''$. For the other intersect

event, the event holds trivially if $2k \geq q$. If $2k < q$, Lemma 7 guarantees $\mathcal{E}'''$ holds with probability tending 1.

Finally, the results of Theorem 17 follow from the same arguments in the proof of Theorem 14 except PE and PY are now replaced by E and Y, respectively. □

C.3. Proofs of Theorems 19 and 20. Recall that we use $d_1(PE) \leq d_1(E)$ in the proof of Theorems 17 and 18. Thus, the proofs for Theorems 19 and 20 remain the same as those for Theorems 17 and 18. □

APPENDIX D: ORACLE INEQUALITY

While our main interest is the study of the rank estimator $\widehat{k}$, we briefly mention an oracle inequality for our estimators $X\widehat{A} := (PY)_{\widehat{k}}$ based on GRS and STRS of the mean XA.

THEOREM 8. *Let $C > 2$. On the event $\lambda \geq C d_1^2(PE)/\widehat{\sigma}^2$, we have*

$$\|X\widehat{A} - XA\|^2 \leq \frac{C+2}{C-2} \min_{0 \leq k \leq K_\lambda} \left\{ \left(\frac{C+2}{C-2} + 8(\rho - 1) \right) \sum_{j>k} d_j^2(XA) + 3\rho\lambda\widehat{\sigma}^2 k \right\},$$

with $\rho := (nm)/(nm - \lambda K_\lambda)$.

PROOF. First we notice that

$$\frac{\|Y - (PY)_{\widehat{k}}\|^2}{nm - \lambda\widehat{k}} \leq \frac{\|Y - (PY)_k\|^2}{nm - \lambda k} \leq \frac{\|Y - (XA)_k\|^2}{nm - \lambda k}.$$

The first inequality follows from the optimality of $\widehat{k}$; the second inequality is a consequence of Pythagoras' identity and the Eckart and Young (1936) inequality. Rewriting the above display yields

$$\|Y - (PY)_{\widehat{k}}\|^2 \leq \|Y - (XA)_k\|^2 \left\{ 1 + \frac{2\lambda k}{nm - \lambda k} - \frac{\lambda(k + \widehat{k})}{nm - \lambda k} \right\},$$

and after working out the squares, we obtain

$$\begin{aligned} \|(PY)_{\widehat{k}} - XA\|^2 \leq{}& \|XA - (XA)_k\|^2 + \frac{2\lambda k}{nm - \lambda k}\|Y - (XA)_k\|^2 \\ &- \frac{\lambda(k + \widehat{k})}{nm - \lambda k}\|Y - (XA)_k\|^2 + 2\langle E, (PY)_{\widehat{k}} - (XA)_k \rangle. \end{aligned} \tag{D.1}$$

Since both $(XA)_k$ and $(PY)_{\widehat{k}}$ are in the column space of P, see, for instance, (Giraud, 2015, page 124), we have $2\langle E, (PY)_{\widehat{k}} - (XA)_k\rangle = 2\langle PE, (PY)_{\widehat{k}} - (XA)_k\rangle$, and the norm duality and the elementary inequality $2xy \le ax^2 + y^2/a$ for $a > 0$, yield

$$\begin{aligned} 2\langle E, (PY)_{\widehat{k}} - (XA)_k\rangle \le\ & 2d_1(PE)\cdot\sqrt{k+\widehat{k}}\cdot\|(PY)_{\widehat{k}} - (XA)_k\| \\ \le\ & (a+b)(k+\widehat{k})d_1^2(PE) + \frac{1}{a}\|(PY)_{\widehat{k}} - XA\|^2 + \frac{1}{b}\|XA - (XA)_k\|^2 \end{aligned} \tag{D.2}$$

for any $a, b > 0$. Moreover, we find

$$\begin{aligned} \|Y - (XA)_k\|^2 &= \|XA - (XA)_k\|^2 + \|E\|^2 + 2\langle E, XA - (XA)_k\rangle \\ &\le 3\|XA - (XA)_k\|^2 + \frac{3}{2}\|E\|^2 \end{aligned} \tag{D.3}$$

and

$$\|Y - (XA)_k\|^2 \ge \frac{1}{2}\|E\|^2 - \|XA - (XA)_k\|^2 \tag{D.4}$$

Finally, combining (D.1) with (D.2), (D.3) and (D.4) gives

$$\begin{aligned} \frac{a-1}{a}\|X\widehat{A} - XA\|^2 \le & \left[\frac{b+1}{b} + \frac{6\lambda k}{(nm-\lambda k)} + \frac{\lambda(k+\widehat{k})}{(nm-\lambda k)}\right]\|XA - (XA)_k\|^2 \\ & + \frac{3\lambda k}{nm-\lambda k}\|E\|^2 + (k+\widehat{k})\left[(a+b)d_1^2(PE) - \frac{\lambda/2}{nm-\lambda k}\|E\|^2\right]. \end{aligned}$$

By using $(nm)/(nm-\lambda\ell) \le (nm)/(nm-\lambda K_\lambda)$ for all $\ell \le K_\lambda$ and setting $a = 1 + b$ and $b = C/2$, the claim follows after a little algebra. □

Contrary to the rank consistency results in Sections 3 and 4, no lower bound on the non-zero singular values $d_j(XA)$ of the signal XA, nor any assumption on X is required. It is clear that our selected estimator achieves the optimal bias-variance tradeoff, as discussed in Bunea, She and Wegkamp (2011); Giraud (2011, 2015). In the above oracle inequality, the quantity

$$\rho := \frac{nm}{nm - \lambda K_\lambda}$$

could be large under some circumstances. Nevertheless, the mild rank constraint (3.10) with $\delta > 0$ guarantees the upper bound $\rho \le 1 + \delta$. Theorem 8 and the exponential inequalities in (3.2) – (3.4) immediately yield the following corollary.

COROLLARY 9. *Suppose E has i.i.d. $N(0,\sigma^2)$ entries. For $\lambda > 2(\sqrt{m}+\sqrt{q})^2$, the event*

$$\|X\widehat{A} - XA\|^2 \lesssim \rho \min_{0\le k\le K_\lambda} \left\{ \sum_{j>k} d_j^2(XA) + \lambda\sigma^2 k \right\} \tag{D.5}$$

holds with probability tending to 1, as $nm \to \infty$ and $q+m \to \infty$.

We use the notation $\lesssim$ for inequalities that hold up to multiplicative constants. The probability of event (D.5) converges (to 1) exponentially fast in mn and $m+q$.

PROOF. We can write $\lambda := C'(\sqrt{m}+\sqrt{q})^2$ for some $C' = C(1+C_0)^2/(1-C_1) > 2$ with $C_0 > 0$, $0 < C_1 < 1$ and C equal to the one in Theorem 8. From inequalities (3.2) – (3.4), we have

$$\begin{aligned}
\mathbb{P}\left\{ C d_1^2(PE) \ge \lambda \frac{\|E\|^2}{nm} \right\} \le\ & \mathbb{P}\left\{ d_1^2(PE) \ge (1+C_0)^2(\sqrt{m}+\sqrt{q})^2\sigma^2 \right\} \\
& + \mathbb{P}\left\{ \|E\|^2 \le (1-C_1)nm\sigma^2 \right\} \\
\le\ & \exp\left\{ -C_0^2(\sqrt{m}+\sqrt{q})^2/2 \right\} + \exp\left\{ -C_1^2 nm/4 \right\}.
\end{aligned}$$

This proves the claim. □

For GRS and STRS, $\widehat{k} = r$ holds with overwhelming probability. On this event, Theorem 10 provides a cleaner and tighter bound for the fit $\|X\widehat{A} - XA\|$.

THEOREM 10. *On the event $\widehat{k} = r$, we have*

$$\|X\widehat{A} - XA\|^2 \le 4r d_1^2(PE)$$

PROOF. From the optimality of $X\widehat{A}$, we have

$$\|Y - X\widehat{A}\|^2 \le \|Y - (XA)_r\|^2 = \|E\|^2$$

on the event $\{\widehat{r} = r\}$. This implies

$$\|X\widehat{A} - XA\|^2 \le 2|\langle E, X\widehat{A} - (XA)_r\rangle| \le 2\|X\widehat{A} - XA\|(\sqrt{r} d_1(PE))$$

and the result follows. □

Specialized to our STRS procedure, we have the following corollary.

COROLLARY 11. *Assume E has i.i.d. $N(0,\sigma^2)$ entries. On the event (4.7) in Theorem 14, if we choose and update λ according to (4.2) and (4.4), then we have $\widehat{k} = r$ with overwhelming probability. Hence, for some constant $C > 4$, we have*

$$\mathbb{P}\left\{\|X\widehat{A} - XA\|^2 \le Cr(\sqrt{q}+\sqrt{m})^2\sigma^2\right\} \to 1$$

as $n \to \infty$ and $(q \vee m) \to \infty$.

The convergence rate in Corollary 11 is exponentially fast. The proof follows immediately from Theorem 14, Theorem 10, (3.2) and $\mathbb{E}[d_1(PE)] \le \sigma(\sqrt{m}+\sqrt{q})$.

APPENDIX E: STRS WITH DETERMINISTIC BOUNDS

When E has i.i.d. $N(0,\sigma^2)$ entries, a deterministic bound for updating λ in Section 4 can be derived as follows. We define $M := m \vee q$ and recall that $N = q \wedge m$. For any $0 < \varepsilon < 1$, we choose $\widetilde{\lambda}_0 = 2(1+\varepsilon)(\sqrt{m}+\sqrt{q})^2$ and let $\widetilde{k}_0$ be selected from (1.3) by using $\widetilde{\lambda}_0$. For given $\widetilde{k}_t \ge 1$ with $t \ge 0$, we set

(E.1)
$$\widetilde{\lambda}_{t+1} = \begin{cases} nm \Big/ \left[(1-\varepsilon)\widetilde{R}_t/\widetilde{U}_t + \widetilde{k}_t\right], & \text{if } 2\widetilde{k}_t \le N; \\ nm \Big/ \left[(1-\varepsilon)(n-q)m/((\sqrt{m}+\sqrt{q})^2+1) + \widetilde{k}_t\right], & \text{if } 2\widetilde{k}_t \ge N \end{cases}$$

with

$$\widetilde{R}_t := \max\Bigg\{nm - \sum_{j=1}^{2\widetilde{k}_t}\left(\sqrt{M}+\sqrt{N-j+1}\right)^2 - 2\widetilde{k}_t, \\ (n-q)m + \sum_{j=2\widetilde{k}_t+1}^{N}\left(\sqrt{M}-\sqrt{j}\right)^2\Bigg\} \tag{E.2}$$

and

$$\widetilde{U}_t := \max\left\{(\sqrt{m}+\sqrt{q})^2+1, \sum_{j=2\widetilde{k}_t+1}^{2\widetilde{k}_t+2}\left(\sqrt{M}+\sqrt{N-j+1}\right)^2 + 2\right\}. \tag{E.3}$$

After updating $\widetilde{\lambda}_t$, we select $\widetilde{k}_t$ as

$$\widetilde{k}_t := \operatorname*{arg\,min}_{\widetilde{k}_{t-1} \le k \le \widetilde{K}_t} \frac{\|Y - (PY)_k\|^2}{nm - \widetilde{\lambda}_t k}$$

where $\widetilde{K}_t := \lfloor nm/\widetilde{\lambda}_t \rfloor \wedge q \wedge m$. The procedure stops when $\widetilde{k}_t = \widetilde{k}_{t+1}$. For this deterministic self-tuning procedure, we show that results analogous to Proposition 12 and Theorem 14 are still guaranteed.

PROPOSITION 12. *For all $t \geq 0$, we have $\widetilde{\lambda}_{t+1} \leq \widetilde{\lambda}_t$ and $\widetilde{k}_{t+1} \geq \widetilde{k}_t$. Moreover, $\widetilde{k}_t \leq r$ holds with probability converging to 1 as $(q \vee m) \to \infty$ and $n \to \infty$.*

PROOF OF PROPOSITION 12. We first show $\widetilde{\lambda}_t$ is decreasing in $\widetilde{k}_t$. For notational simplicity, we write k for $\widetilde{k}_t$ and assume $q \leq m$. The case $q > m$ can be obtained similarly. The decreasing property of (E.1) is easily seen if $2k \geq N$. Hence we focus on the case $2k < N$ and consider two sub-cases:

(1) Suppose $\widetilde{U}_t = \sum_{j=2k+1}^{2k+2} (\sqrt{m} + \sqrt{q-j+1})^2 + 2$. It suffices to show

$$A_k := \frac{(1-\varepsilon)\widetilde{R}_t}{\sum_{j=2k+1}^{2k+2} (\sqrt{m} + \sqrt{q-j+1})^2 + 2} + k$$

is increasing in k. It has the same form as (B.5) in the proof of Proposition 12. By repeating the arguments there, we can similarly show that A_k is increasing. Hence $\widetilde{\lambda}_t$ is decreasing.

(2) Suppose $\widetilde{U}_t = (\sqrt{m} + \sqrt{q})^2 + 1$. This is similar to (B.6). Therefore, the same arguments can be used to prove that $\widetilde{\lambda}_k$ is decreasing. By the same reasoning as the proof of Proposition 12, we have $\widetilde{k}_{t+1} \geq \widetilde{k}_t$. Finally, the proof for $\widetilde{k}_t \leq r$ follows immediately from Proposition 12 by observing that $\widetilde{\lambda}_t \geq \widehat{\lambda}_t$ using Lemma 4. □

THEOREM 13. *Assume E has i.i.d. $N(0, \sigma^2)$ entries. For any subsequence $k_0 < \cdots < k_T = r$ of $\{1, 2, \ldots, r\}$ with $T \leq r - 1$, we let $\lambda_0 = 2(1+\varepsilon)(\sqrt{m} + \sqrt{q})^2$ for any $\varepsilon \in (0, 1)$. Denote by λ_t the updated λ according to (E.1) by using k_t, for $t = 0, 1, \ldots, T-1$. On the event,*

$$\text{(E.4)} \quad d_{k_t}(XA) \geq C\sigma\sqrt{\lambda_t} \left[\frac{\sqrt{2}}{2} + \sqrt{\frac{nm}{nm - \lambda_t r}} \right], \quad t = 0, 1, \ldots, T.$$

for some $C > 1$, there exists $T' \leq T$ such that $\mathbb{P}\{\widetilde{k}_{T'} = r\} \to 1$, where $\widetilde{k}_0, \widetilde{k}_1, \ldots, \widetilde{k}_{T'}$ are the selected ranks from the procedure above.

PROOF OF THEOREM 13. The proof follows the same arguments as the proof of Theorem 14 by using Proposition 12. □

APPENDIX F: ADDITIONAL SIMULATIONS

F.1. Simulations of Monte Carlo simulations vs deterministic bounds. Depending on whether we update λ by Monte Carlo simulations (MC) or by the Deterministic Bounds (DB) in Section E, we denote by STRS-MC the procedure using MC and by STRS-DB the one using DB. We compare their performance for some non-Gaussian distributions. In particular, we generate E from either uniform$(-\sqrt{3}, \sqrt{3})$ or t_ν-distribution (d.f. $\nu = 6$). For each distribution, both low- and high-dimensional settings are considered. The low-dimensional case sets $\eta = 0.1$, $b_0 = 0.1$, $n = 300$, $m = p = q = 50$ and varies r between 0 and 15. For the high-dimensional case, we consider $\eta = 0.1$, $b_0 = 0.003$, $n = 200$, $m = 60$, $p = 300$, $q = 30$ and $0 \le r \le 15$. The mean selected ranks and rank recovery rate of the two methods for each setting are shown in Figure 1.

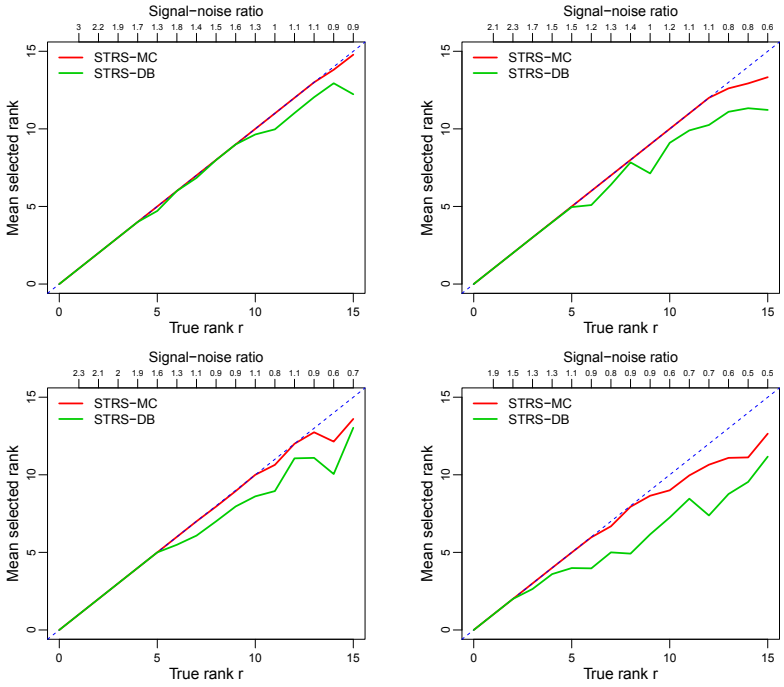

Fig 1: Comparison of STRS-MC and STRS-DB in the low-dimensional setting (top) and high-dimensional setting (bottom) for the uniform errors (left) and t_6 errors (right).

Both STRS-MC and STRS-DB work perfectly when the SNR is not small (say greater than 1.6). STRS-DB seems to require a slightly larger SNR, as expected, since the deterministic bounds are upper bounds of those in STRS-MC leading to a larger updated λ.

More importantly, we emphasize that STRS-MC is using Monte Carlo simulations based on $N(0,1)$. It also supports our conjecture that STRS-MC works for other distributions with heavier tails.

F.2. Simulation for KF methods. Giraud (2011) proposed to minimize

$$\frac{\|Y-(PY)_k\|^2}{nm-1-C\left(\mathbb{E}[\|G\|_{(2,k)}]\right)^2} \tag{F.1}$$

with $\|G\|_{2,k}^2=\sum_{i=1}^k d_i^2(G)$ and some tuning parameter $C>1$. It does not require to estimate σ^2 but still needs the selection of leading constant C. Giraud (2011) recommended to use $C=2$ based on Birgé and Massart (2007). We use KF to denote this method. In particular, we define KF-2 for choosing leading constant $C=2$ and KF-CV for choosing C via cross-validation. The simulation setting is slightly different from the low dimensional one in Experiment 1 and aims to give an example of an easy situation where KF fails to select the correct rank. Here we have $n=300$, $m=40$ and $p=q=35$. For illustration, we only vary the true rank from 10 to 35. We set $\eta=0.1$ and $b_0=20$. Note $b_0=20$ ensures a large SNR which should be the most ideal case for rank recovery. We compare the performance of KF-2, KF-CV and STRS. The plot of rank recovery and mean selected rank versus the true rank for different methods are shown in Figure 2. From the result, it is clear that neither KF-2 nor KF-CV consistently recovers the true rank while STRS does. But since KF was not developed for rank recovery, nor did it claim to have this property, this does not contradict the results in Giraud (2011).

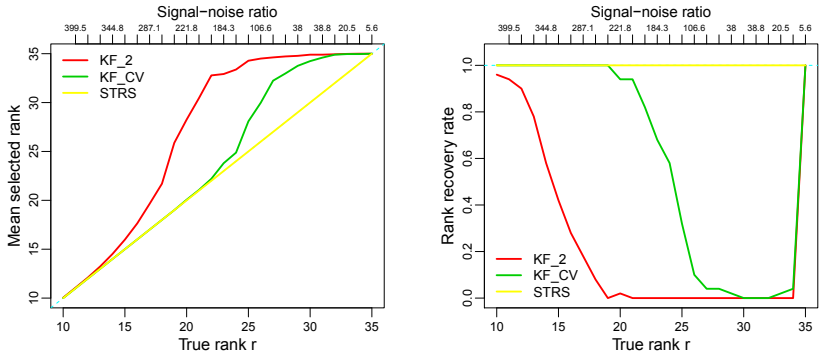

Fig 2: Plot of rank recovery rate and mean selected rank for KF and STRS.

F.3. Simulations to compare the error $\|X\widehat{A} - XA\|$. In this section, we compare three methods, STRS-MC, BSW-1.3 and KF-2, based on two criterions: (1) the fit $\|X\widehat{A} - XA\|/\sqrt{nm}$; (2) the selected rank.

In the low-dimensional setting, when X^TX is invertible, we also compute the prediction error $\|\widehat{A} - A\|/\sqrt{pm}$. We consider both situations when the model is correctly specified and when the model approximately holds.

F.3.1. *Exact low rank model.* We first consider the scenario when A has an exact low rank structure, in both low- and high-dimensional settings. In the low-dimensional setting, we choose $\rho = 0.1$, $n = 200$, $m = p = q = 50$, $r = 10$ and $b_0 \in \{0.02, 0.022, 0.024, \ldots, 0.046\}$. In the high-dimensional setting, we specify $\rho = 0.1$, $n = 150$, $p = 300$, $m = q = 50$, $r = 10$ and $b_0 \in \{0.0015, 0.0016, \ldots, 0.003\}$. Different grids of b_0 are chosen to maintain similar signal-to-noise ratio. The random additive errors in both settings are generated from $N(0, 1)$. Within each setting, we repeat 100 times. The averaged results are reported in Figure 3 demonstrating how criterions (1) and (2) of the three methods vary with b_0 in both low- and high-dimensional settings. STRS-MC dominates the other two methods as it produces a smaller error and it always selects a rank closer to the true rank than the other two methods.

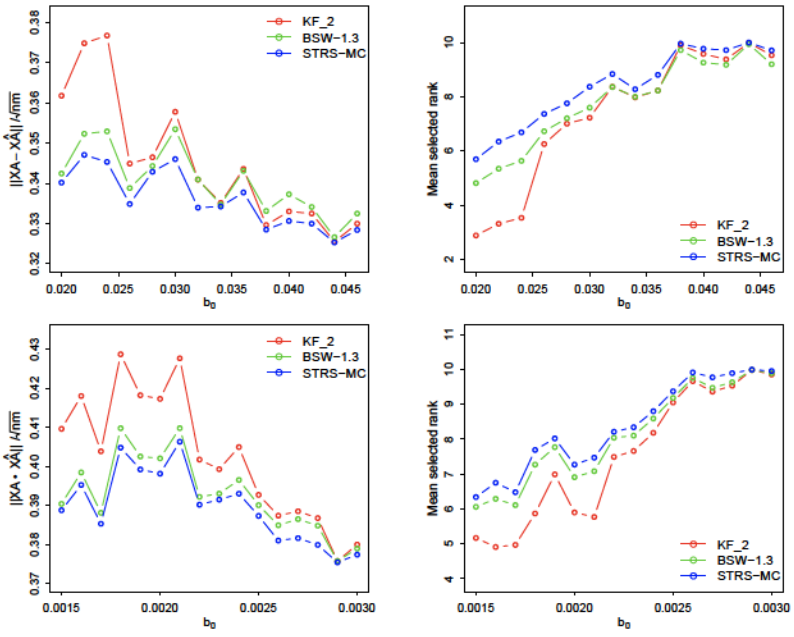

Fig 3: Criterions (1) and (2) of STRS-MC, KF-2 and BSW-1.3 as b_0 varies in the low-dimensional setting (first row) and high-dimensional setting (second row).

F.3.2. *Approximate low rank model.* We devote this part to testing the performance of different approaches when the model is mis-specified in that A doesn't have an exact low rank structure, but rather has a small effective rank with non-zero decaying singular values. To generate this type of A, we first generate an exact low rank matrix A^* with specified rank r based on our data generating mechanism described in Section 6.3. Next, we compute its singular value decomposition $A^* = UDV^T$ with $D = \text{diag}(d_1, \ldots, d_r, 0, \ldots, 0)$ and add polynomial decaying noise to the (zero valued) singular values d_j for $j \geq r$. Specifically, we take $d_j = d_r \cdot \gamma(j - r + 1)^{-\beta}$ for $j \geq r + 1$, with $\gamma \in (0, 1)$ and some positive integer β. The matrix $A = U\widetilde{D}V^T$ is our approximate low rank matrix with $\widetilde{D} = \text{diag}(d_1, \ldots, d_r, d_{r+1}, d_{r+2}, \ldots)$. Since the results are similar for different choices of γ and β, we only present the results of $\gamma = 0.8$ and $\beta = 1$. The rest of our simulation setup stays the same as in the exact low rank model case above. Criterions (1) and (2) for both low- and high-dimensional settings are shown in Figure 4. We find the same conclusion as before when the model is mis-specified: STRS-MC continues to outperform the other two methods by yielding the smallest $\|X\widehat{A} - XA\|$ and selecting a rank closer to the effective rank (of 10).

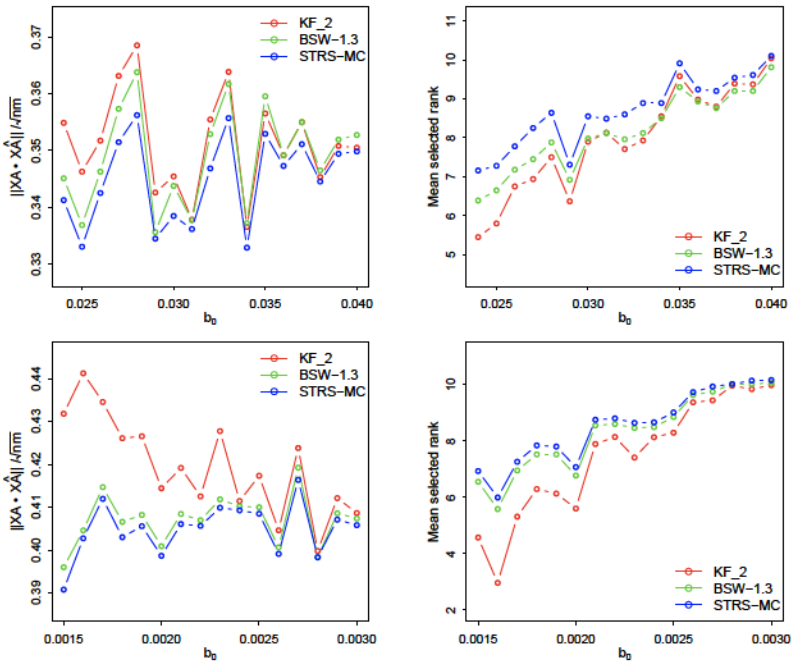

Fig 4: Criterions (1) and (2) of STRS-MC, KF-2 and BSW-1.3 as b_0 varies in the low-dimensional setting (top) and the high-dimensional setting (bottom).

F.3.3. *The prediction error* $\|\widehat{A} - A\|$. We compute the prediction error $\|\widehat{A} - A\|/\sqrt{pm}$ in the low-dimensional setting for both the exact low rank model and the approximate low rank model. Figure 5 shows that STRS-MC dominates the other two methods, demonstrating the importance of selecting a better rank.

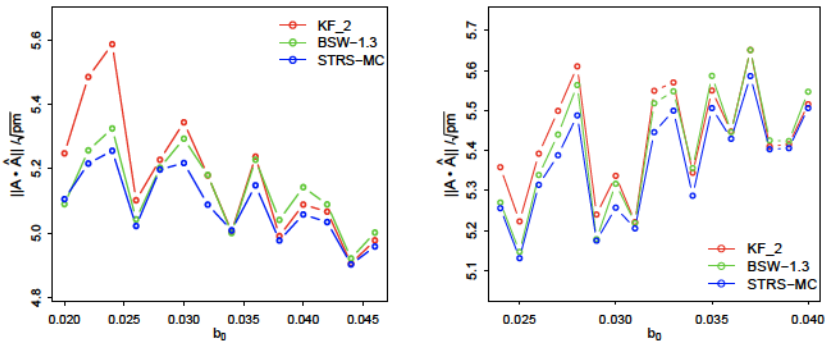

Fig 5: Plots of $\|\widehat{A} - A\|$ in the exact low-rank model (left) and in the approximate low-rank model (right).

REFERENCES

Bai, Z. D. and Yin, Y. Q. (1993). Limit of the Smallest Eigenvalue of a Large Dimensional Sample Covariance Matrix. *Ann. Probab.* **21** 1275–1294.

Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields* **138** 33-73.

Bunea, F., She, Y. and Wegkamp, M. H. (2011). Optimal Selection of Reduced Rank Estimators of High-Dimensional Matrices. *Annals of Statistics* **39** 1282-1309.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika* **1**.

Giraud, C. (2011). Low rank multivariate regression. *Electronic Journal of Statistics* **5** 775-799.

Giraud, C. (2015). *Introduction to high-dimensional statistics*. Monographs on Statistics and Applied Probability.

Horn, R. A. and Johnson, C. R. (2013). *Matrix analysis*. Cambridge University Press, Cambridge.

Vershynin, R. (2012). *Introduction to the non-asymptotic analysis of random matrices* 210-268. Cambridge University Press.

Xin Bing
Department of Statistical Science
Cornell University
101 Malott Hall Ithaca, New York 14853-3801
United States of America
E-mail: xb43@cornell.edu

Marten H. Wegkamp
Department of Mathematics &
Department of Statistical Science
Cornell University
432 Malott Hall
Ithaca, New York 14853-3801
United States of America
E-mail: marten.wegkamp@cornell.edu