Biometrika (2017), **xx**, x, *pp*. 1–19

Printed in Great Britain

Supplementary material for robust reduced-rank regression

BY Y. SHE

Department of Statistics, Florida State University,
117 N. Woodward Ave, Tallahassee, Florida 32306, U.S.A
yshe@stat.fsu.edu

K. CHEN

Department of Statistics, University of Connecticut,
215 Glenbrook Road U-4120, Storrs, Connecticut 06269, U.S.A.
kun.chen@uconn.edu

1. PROOFS

1·1. *Notation and definitions*

Given $\mathcal{I} \subset [n], \mathcal{J} \subset [p]$, $X(\mathcal{I}, \mathcal{J})$ denotes a submatrix of X by extracting the rows and columns indexed by $\mathcal{I}$ and $\mathcal{J}$, respectively. We use c, L to denote constants. They are not necessarily the same at each occurrence. Denote by $CS(A)$ the column space of A. Given $\mathcal{P}_A$, denote by $\mathcal{P}_A^{\perp}$ the projection onto its orthogonal complement. In addition to the definitions of thresholding function Θ and the multivariate thresholding function $\vec{\Theta}$, we will use a matrix threshold function.

DEFINITION 1 (MATRIX THRESHOLD FUNCTION). *Given any threshold function $\Theta(\cdot;\lambda)$, its matrix version Θ^{σ} is defined for $B \in \mathbb{R}^{n\times m}$ as follows*

$$\Theta^{\sigma}(B;\lambda) = U diag\{\Theta(\sigma_i^B;\lambda)\}\, V^{\mathrm{T}}, \tag{1}$$

where U, V, and σ_i^B are obtained from the SVD of B: $B = U diag(\sigma_i^B) V^{\mathrm{T}}$.

Finally, we describe a quantile thresholding $\Theta^{\#}(\cdot;\varrho,\eta)$ which is convenient in analyzing the constraint-type problems. It can be seen as a vector variant of the hard-ridge thresholding $\Theta_{HR}(t;\lambda,\eta) = t/(1+\eta)1_{|t|>\lambda}$ (She, 2009). Given $1 \le \varrho \le n$ and $\eta \ge 0$, $\Theta^{\#}(a;\varrho,\lambda) : \mathbb{R}^n \to \mathbb{R}^n$ is defined for any $a \in \mathbb{R}^n$ such that the ϱ largest components of a, in absolute value, are shrunk by a factor of $(1+\lambda)$ and the remaining components are all set to be zero. In the case of ties, a random tie breaking rule is used. We abbreviate $\Theta^{\#}(a;\varrho,0)$ to $\Theta^{\#}(a;\varrho)$.

1·2. *Proof of Theorem 1*

We show the proof detail for the penalized estimators. First, the loss term in the objective can be decomposed into

$$\begin{aligned}\mathrm{tr}\{(Y-XB)\Gamma(Y-XB)^{\mathrm{T}}\} &= \|Y\Gamma^{1/2} - XB\Gamma^{1/2}\|_{\mathrm{F}}^2 \\ &= \|\mathcal{P}_X Y\Gamma^{1/2} - XB\Gamma^{1/2}\|_{\mathrm{F}}^2 + \|\mathcal{P}_X^{\perp} Y\Gamma^{1/2}\|_{\mathrm{F}}^2.\end{aligned}$$

Let $Z = \mathcal{P}_X Y \Gamma^{1/2}$. Clearly, $\mathcal{P}_Z \subset \mathcal{P}_X$. Consider the following optimization problem

$$\min_A \frac{1}{2}\|Z - A\|_{\mathrm{F}}^2 + \sum_{s=1}^{p\wedge m} P(\sigma_s^A; \lambda). \tag{2}$$

From the proof of Proposition 2.1 in She (2013), the following results can be obtained: (i) any optimal solution $\hat{A}$ to (2) must satisfy $\hat{A} \in \mathcal{P}_Z$; (ii) $A_o = \Theta^\sigma(Z; \lambda)$ gives a particular minimizer of (2), and $\|\hat{A} - A_o\|_* \le C(\lambda)$ holds for any $\hat{A}$, where $\|\cdot\|_*$ represents the nuclear norm and $C(\lambda)$ is a function dependent on the regularization parameter only. From (i), $X\hat{B}\Gamma^{1/2}$ is always a solution to (2). It suffices to study the breakdown point of A_o.

Because $X \ne 0$, there must exist $i \in [n]$ such that the ith column of $\mathcal{P}_X$ is not 0. Let $\tilde{Y} = Y + Me_i e_1^{\mathrm{T}}$. where e_i is the unit vector with the ith entry being 1. Due to the construction of $\tilde{Y}$ and the positive-definiteness of Γ,

$$\|\mathcal{P}_X\tilde{Y}\Gamma^{1/2}\|_{\mathrm{F}}^2 = M^2\|\mathcal{P}_X e_i e_1^{\mathrm{T}}\Gamma^{1/2}\|_{\mathrm{F}}^2 + 2M\langle \mathcal{P}_X Y, e_i e_1^{\mathrm{T}}\Gamma\rangle + \|\mathcal{P}_X Y\Gamma^{1/2}\|_{\mathrm{F}}^2 \to +\infty$$

as $M \to \infty$. That is, given λ, $\Theta^\sigma(\mathcal{P}_X\tilde{Y}\Gamma^{1/2}; \lambda)$ thresholds the singular values of $\mathcal{P}_X\tilde{Y}\Gamma^{1/2}$ the sum of which can be made arbitrarily large as M increases. It follows from the definition of Θ that $\sup_M \|\Theta^\sigma(\mathcal{P}_X\tilde{Y}\Gamma^{1/2}; \lambda)\|_{\mathrm{F}} = \infty$.

The proof for the reduced-rank regression estimator follows similar lines and is omitted.

1·3. *Proof of Theorem 2*

Part (i): The proof of this part is based on the following two lemmas.

LEMMA 1. *Given an arbitrary thresholding rule Θ satisfying Definition 1 in the paper, let P be any function associated with Θ through*

$$P(t;\lambda) - P(0;\lambda) = P_\Theta(t;\lambda) + q(t;\lambda), \quad P_\Theta(t;\lambda) = \int_0^{|t|} [\sup\{s : \Theta(s;\lambda) \le u\} - u]\,\mathrm{d}u,$$

for some nonnegative $q(\theta;\lambda)$ satisfying $q\{\Theta(t;\lambda)\} = 0$ for all t. Then, $\hat{\beta} = \vec{\Theta}(y;\lambda)$ gives a globally optimal solution to

$$\min_{\beta\in\mathbb{R}^n} \frac{1}{2}\|y - \beta\|_2^2 + P(\|\beta\|_2; \lambda).$$

This result is implied by Lemma 1 of She (2012). It is worth mentioning that $\vec{\Theta}(y;\lambda)$ is not necessarily unique when Θ has discontinuities. Next we prove an identity.

LEMMA 2. *Given any thresholding rule $\Theta(t;\lambda)$, define $P_\Theta(t;\lambda) = \int_0^{|t|}\{\Theta^{-1}(u;\lambda) - u\}\,\mathrm{d}u$ where $\Theta^{-1}(u;\lambda) = \sup\{t : \Theta(t;\lambda) \le u\}$. Then the following identity holds for any $r \in \mathbb{R}$*

$$\frac{1}{2}\{r - \Theta(r;\lambda)\}^2 + P_\Theta\{\Theta(r;\lambda);\lambda\} = \int_0^{|r|} \psi(t;\lambda)\,\mathrm{d}t, \tag{3}$$

where $\psi(t;\lambda) = t - \Theta(t;\lambda)$.

Proof. Without loss of generality, assume $r \ge 0$. By definition, $\int_0^r \psi(t;\lambda)\,\mathrm{d}t = r^2/2 - \int_0^r \Theta(t;\lambda)\,\mathrm{d}t$ and $P_\Theta\{\Theta(r;\lambda);\lambda\} = \int_0^{\Theta(r;\lambda)} \Theta^{-1}(t;\lambda)\,\mathrm{d}t - r^2/2$. It suffices to show that

$$\int_0^{\Theta(r;\lambda)} \Theta^{-1}(t;\lambda)\,\mathrm{d}t + \int_0^r \Theta(t;\lambda)\,\mathrm{d}t = r\Theta(r;\lambda).$$

In fact, changing the order of integration, and using the monotone property of Θ, we get

$$\begin{aligned}\int_0^r \Theta(t;\lambda)\,\mathrm{d}t - r\Theta(r;\lambda) &= \int_0^r \mathrm{d}t \int_0^{\Theta(t;\lambda)} \mathrm{d}s - \int_0^{\Theta(r;\lambda)} r\,\mathrm{d}t \\ &= \int_0^{\Theta(r;\lambda)} \mathrm{d}s \int_{\Theta^{-1}(s;\lambda)}^{r} \mathrm{d}t - \int_0^{\Theta(r;\lambda)} r\,\mathrm{d}t \\ &= -\int_0^{\Theta(r;\lambda)} \Theta^{-1}(t;\lambda)\,\mathrm{d}t.\end{aligned}$$

The conclusion thus follows. □

We have the pieces in place to prove part (i) of the theorem. Without loss of generality, assume $\Gamma = I$. Let $f(B,C) = \mathrm{tr}\{(Y - XB - C)(Y - XB - C)^{\mathrm{T}}\}/2 + \sum_{i=1}^n P(\|\Gamma^{1/2}c_i\|_2;\lambda)$, and $g(B) = \sum_{i=1}^n \rho(\|(y_i - B^{\mathrm{T}}x_i)\|_2;\lambda)$. By Lemma 1, fixing B, $\hat{C} = (c_1\ \ldots\ c_n)^{\mathrm{T}}$ with $\hat{c}_i = \vec{\Theta}(y_i - B^{\mathrm{T}}x_i;\lambda)$ gives an optimal solution to $\min_C f(B,C)$. For this $\hat{C}$, $f(B,\hat{C}) = g(B)$ holds by Lemma 2.

Part (ii): The proof follows similar lines of that of Part (i), based on the quantile thresholding and Lemma C.1 in She et al. (2013). The details are omitted.

1·4. *Proofs of Theorem 3 & Theorem 6*

Recall that $P_1(t;\lambda) = \lambda|t|$, $P_0(t;\lambda) = (\lambda^2/2)1_{t\neq 0}$, $P_H(t;\lambda) = (-t^2/2 + \lambda|t|)1_{|t|<\lambda} + (\lambda^2/2)1_{|t|\geq\lambda}$. For convenience, $P_{2,1}(C;\lambda)$ is used to denote $\lambda\|C\|_{2,1}$, and $P_{2,0}$ and $P_{2,H}$ are used similarly.

By definition, $(\hat{B},\hat{C})$ satisfies the following inequality for any (B,C) with $r(B) \leq r$,

$$\frac{1}{2}M(\hat{B} - B^*, \hat{C} - C^*) \leq \frac{1}{2}M(B - B^*, C - C^*) + P(C;\lambda) - P(\hat{C};\lambda) + \langle \mathcal{E}, X\Delta^B + \Delta^C\rangle. \tag{4}$$

Here, $\Delta^B = \hat{B} - B$, $\Delta^C = \hat{C} - C$ and so $r(\Delta^B) \leq 2r$.

LEMMA 3. *For any given $1 \leq J \leq n, 1 \leq r \leq m \wedge p$, define $\Gamma_{r,J} = \{(B,C) \in \mathbb{R}^{p\times m} \times \mathbb{R}^{n\times m} : r(B) \leq r, J(C) = J\}$. Then there exist universal constants $A_0, C, c > 0$ such that for any $a \geq 2b > 0$, the following event*

$$\sup_{(B,C)\in\Gamma_{r,J}} \Big\{2\langle\mathcal{E}, XB + C\rangle - \frac{1}{a}\|XB + C\|_F^2 - \frac{1}{b}P_{2,H}(C;\lambda) - aA_0\sigma^2 r(m+q)\Big\} \geq a\sigma^2 t \tag{5}$$

occurs with probability at most $c'\exp(-ct)$, where $\lambda = A\lambda^o$, $\lambda^o = \sigma(m + \log n)^{1/2}$, $A = (abA_1)^{1/2}$, $A_1 \geq A_0$, and $t \geq 0$.

Let $l_H(B,C,r) = 2\langle\mathcal{E}, XB + C\rangle - \|XB + C\|_{\mathrm{F}}^2/a - P_{2,H}(C;\lambda)/b - aA_0\sigma^2 r(m+q)$. Define

$$R = \sup_{1\leq J\leq n, 1\leq r\leq m\wedge p}\ \sup_{(B,C)\in\Gamma_{r,J}} l_H(B,C,r).$$

From Lemma 3, it is easy to see $ER \le ac\sigma^2$. Substituting the bound below into (4),

$$\begin{aligned}
2\langle \mathcal{E}, X\Delta^B + \Delta^C\rangle \le& \frac{1}{a}\|X\Delta^B + \Delta^C\|_{\mathrm{F}}^2 + \frac{1}{b}P_{2,H}(\Delta^C;\lambda) + 2aA_0\sigma^2 r(m+q) + R\\
\le& \frac{2}{a}M(B-B^*, C-C^*) + \frac{2}{a}M(\hat{B}-B^*, \hat{C}-C^*)\\
&+ 2aA_0\sigma^2 r(m+q) + R + \frac{1}{b}P_{2,H}(\Delta^C;\lambda),
\end{aligned}$$

we have

$$\begin{aligned}
(1-\frac{2}{a})M(\hat{B}-B^*, \hat{C}-C^*) \le& (1+\frac{2}{a})M(B-B^*, C-C^*) + 2aA_0\sigma^2 r(m+q) + R\\
&+ 2P(C;\lambda) - 2P(\hat{C};\lambda) + \frac{1}{b}P_{2,H}(\Delta^C;\lambda).
\end{aligned}$$

It remains to deal with $2P(C;\lambda) - 2P(\hat{C};\lambda) + P_{2,H}(\Delta^C;\lambda)/b$ which is denoted by I below.

(i) Due to the sub-additivity of the function P_H that is concave on $[0,\infty)$,

$$\begin{aligned}
I &\le 2P(C;\lambda) - 2P_{2,H}(\hat{C};\lambda) + \frac{1}{b}P_{2,H}(\Delta^C;\lambda)\\
&\le 2P(C;\lambda) + \frac{1}{b}P_{2,H}(C;\lambda) + \frac{1}{b}P_{2,H}(\hat{C};\lambda) - 2P_{2,H}(\hat{C};\lambda)\\
&\le (2+\frac{1}{b})P(C;\lambda),
\end{aligned}$$

if $b \ge 1/2$. Theorem 3 can be obtained by choosing $a=4$, $b=1/2$, and $\lambda = A\lambda^o$ with $A \ge (2A_0)^{1/2}$.

(ii) When P is the group ℓ_1 penalty as in Theorem 6, by the sub-additivity of P, we have

$$\begin{aligned}
I &\le 2P_{2,1}(C;\lambda) - 2P_{2,1}(\hat{C};\lambda) + \frac{1}{b}P_{2,1}(\Delta^C;\lambda)\\
&\le 2A\lambda^o\{(1+\theta)\|\Delta^C_{\mathcal{J}}\|_{2,1} - (1-\theta)\|\Delta^C_{\mathcal{J}^c}\|_{2,1}\}\\
&\le 2A(1-\theta)\lambda^o\{(1+\vartheta)\|\Delta^C_{\mathcal{J}}\|_{2,1} - \|\Delta^C_{\mathcal{J}^c}\|_{2,1}\},
\end{aligned}$$

where $\mathcal{J}(C)$ and $J(C)$ are abbreviated to $\mathcal{J}$, J, respectively, and we set $b = 1/(2\theta)$, $\theta = \vartheta/(2+\vartheta)$. From the regularity condition, $(1+\vartheta)\|\Delta^C_{\mathcal{J}}\|_{2,1} - \|\Delta^C_{\mathcal{J}^c}\|_{2,1} \le KJ^{1/2}\|(I - \mathcal{P}_{X\Delta^B})\Delta^C\|_{\mathrm{F}} \le KJ^{1/2}\|X\Delta^B + \Delta^C\|_{\mathrm{F}}$, and so

$$\begin{aligned}
I &\le 2A(1-\theta)\lambda^o KJ^{1/2}\|X\Delta^B + \Delta^C\|_{\mathrm{F}}\\
&\le \frac{2}{a}M(B-B^*, C-C^*) + \frac{2}{a}M(\hat{B}-B^*, \hat{C}-C^*) + aA^2(1-\theta)^2K^2(\lambda^o)^2 J.
\end{aligned}$$

Taking $a = 4 + 1/\theta$, $b = 1/(2\theta)$, and $A \ge (abA_0)^{1/2}$ gives the conclusion in Theorem 6.

Proof of Lemma 3

Proof. Define

$$l_H(B,C,r) = 2\langle \mathcal{E}, XB + C\rangle - \frac{1}{a}\|XB+C\|_{\mathrm{F}}^2 - \frac{1}{b}P_{2,H}(C;\lambda) - aA_0\sigma^2 r(m+q).$$

Similarly, define $l_0(B,C,r)$ with $P_{2,0}$ in place of $P_{2,H}$ in the above. Let $\mathcal{A}_H = \{\sup_{(B,C)\in\Gamma_{r,J}} l_H(B,C,r) \ge at\sigma^2\}$, and $\mathcal{A}_0 = \{\sup_{(B,C)\in\Gamma_{r,J}} l_0(B,C,r) \ge at\sigma^2\}$.

Since $\mathcal{A}_H \subset \{\sup_{(B,C):r(B)\le r} l_H(B,C,r) \ge at\sigma^2\}$, the occurrence of $\mathcal{A}_H$ implies that

$$l_H(B^o, C^o, r) \ge at\sigma^2, \tag{6}$$

for any (B^o, C^o) that solves

$$\min_{B:r(B)\le r, C} \frac{1}{a}\|XB + C\|_{\mathrm{F}}^2 - 2\langle \mathcal{E}, XB + C\rangle + \frac{1}{b}P_{2,H}(C;\lambda). \tag{7}$$

LEMMA 4. *Given any $\theta \ge 1$, there exists a globally optimal solution C^o to $\min_C \|Y - C\|_F^2/2 + \theta P_{2,H}(C;\lambda)$ such that for any $i : 1 \le i \le n$, either $c_i^o = 0$ or $\|c_i^o\|_2 \ge \lambda\theta^{1/2} \ge \lambda$.*

See She (2012) for its proof. From Lemma 4 and $a \ge 2b$, (6) further indicates that there exists an optimal solution (B^o, C^o) such that $l_0(B^o, C^o, r) \ge at\sigma^2$. Hence $\mathcal{A}_H \subset \mathcal{A}_0$ and it suffices to show $\mathrm{pr}(\mathcal{A}_0) \le C\exp(-ct)$.

Let $\mathcal{J} = \mathcal{J}(C)$ for short. Denote by $I_{\mathcal{J}}$ the submatrix of $I_{n\times n}$ formed by the columns indexed by $\mathcal{J}$. We write the stochastic term into

$$\begin{aligned} 2\langle \mathcal{E}, XB + C\rangle =& 2\langle \mathcal{E}, \mathcal{P}_{I_{\mathcal{J}}}^{\perp} XB\rangle + 2\langle \mathcal{E}, \mathcal{P}_{I_{\mathcal{J}}}(XB + C)\rangle \\ \equiv& 2\langle \mathcal{E}, A_1\rangle + 2\langle \mathcal{E}, A_2\rangle, \end{aligned} \tag{8}$$

and $\|A_1\|_{\mathrm{F}}^2 + \|A_2\|_{\mathrm{F}}^2 = \|XB + C\|_{\mathrm{F}}^2$.

LEMMA 5. *Given $X \in \mathbb{R}^{n\times p}$, $1 \le J \le n$, $1 \le r \le m \wedge p$, define $\Gamma_{r,J}^1 = \{A \in \mathbb{R}^{n\times m} : \|A\|_F \le 1, r(A) \le r, CS(A) \subset CS\{X(\mathcal{J}^c, :)\}$ for some $\mathcal{J} : |\mathcal{J}| = J\}$. Let*

$$P_o^1(J, r) = \sigma^2\left[\{q \wedge (n - J)\}r + (m - r)r + \log\binom{n}{J}\right].$$

Then for any $t \ge 0$,

$$\mathrm{pr}\Big[\sup_{A\in\Gamma_{r,J}^1} \langle \mathcal{E}, A\rangle \ge t\sigma + \{LP_o^1(J, r)\}^{1/2}\Big] \le c'\exp(-ct^2), \tag{9}$$

where $L, c, c' > 0$ are universal constants.

The proof follows similar lines of the proof of Lemma 4 in She (2017) and is omitted. Now, we can bound the the first term on the right hand side of (8) as follows

$$\begin{aligned} &2\langle \mathcal{E}, A_1\rangle - \frac{1}{a}\|A_1\|_{\mathrm{F}}^2 - 2aLP_o^1(J, r) \\ \le& 2\langle \mathcal{E}, A_1/\|A_1\|_{\mathrm{F}}\rangle\|A_1\|_{\mathrm{F}} - 2\|A_1\|_{\mathrm{F}}\{LP_o^1(J, r)\}^{1/2} - \frac{1}{2a}\|A_1\|_{\mathrm{F}}^2 \\ \le& 2a\left[\langle \mathcal{E}, A_1/\|A_1\|_{\mathrm{F}}\rangle - \{LP_o^1(J, r)\}^{1/2}\right]_+^2 + \frac{1}{2a}\|A_1\|_{\mathrm{F}}^2 - \frac{1}{2a}\|A_1\|_{\mathrm{F}}^2 \\ =& 2a\left[\langle \mathcal{E}, A_1/\|A_1\|_{\mathrm{F}}\rangle - \{LP_o^1(J, r)\}^{1/2}\right]_+^2. \end{aligned}$$

By Lemma 5, for L large enough,

$$\mathrm{pr}\{2\langle \mathcal{E}, A_1\rangle - \frac{1}{a}\|A_1\|_{\mathrm{F}}^2 - 2aLP_o^1(J, r) > \frac{1}{2}at\sigma^2\} \le c'\exp(-ct).$$

Similarly, for the second term on the right hand side of (8),

$$\mathrm{pr}\{2\langle \mathcal{E}, A_2\rangle - \frac{1}{a}\|A_2\|_{\mathrm{F}}^2 - 2aLP_o^2(J,r) > \frac{1}{2}at\sigma^2\} \ \le c' \exp(-ct),$$

where

$$P_o^2(J,r) = \sigma^2 \left\{ Jm + \log \binom{n}{J} \right\},$$

and L is a large constant. Applying the union bound gives

$$\mathrm{pr}[2\langle \mathcal{E}, XB + C\rangle - \frac{1}{a}\|XB + C\|_{\mathrm{F}}^2 - 2aL\sigma^2\{(q+m-r)r + Jm + J\log(en/J)\} > at\sigma^2]$$
$$\le c' \exp(-ct)\cdot \quad (10)$$

The conclusion follows. □

1·5. *Proof of Theorem 4*

Similar to Section 1·4, we have

$$\frac{1}{2}M(\hat{B} - B^*, \hat{C} - C^*) \le \frac{1}{2}M(B - B^*, \hat{C} - C^*) + \langle \mathcal{E}, X\Delta^B + \Delta^C\rangle,$$

where $\Delta^B = \hat{B} - B$, $\Delta^C = \hat{C} - C$. Let $\tilde{r} = r(\Delta^B)$ and $\tilde{J} = J(\Delta^C)$. Then from (10) in the proof of Lemma 3,

$$2\langle \mathcal{E}, X\Delta^B + \Delta^C\rangle \le \frac{1}{a}\|X\Delta^B + \Delta^C\|_{\mathrm{F}}^2 - 2aL\sigma^2\{(q+m)\tilde{r} + \tilde{J}m + \tilde{J}\log(en/\tilde{J})\} \ + R,$$

where $ER \le ac\sigma^2$. The oracle inequality can be shown following the lines of Section 1·4, noticing that $\tilde{r} \le 2r$, $\tilde{J} \le 2\varrho$ and $\tilde{J}\log(2en/\tilde{J}) \le 2\varrho\log(en/\varrho)$.

1·6. *Proof of Theorem 5*

The proof is based on the general reduction scheme in Chapter 2 of Tsybakov (2009). We consider two cases.

Case (i) $(q+m)r \ge Jm + J\log(en/J)$. Suppose the SVD of X is $X = UDV^{\mathrm{T}}$ with D of size $q \times q$. Given an arbitrary estimator $(\hat{B}, \hat{C})$, let $\hat{A} = V^{\mathrm{T}}\hat{B}$ and $\tilde{\mathcal{S}}(r,J) = \{(A,C) \in \mathbb{R}^{q\times m} \times \mathbb{R}^{n\times m} : r(A) \le r, J(C) \le J\}$. Then

$$\sup_{(B^*,C^*)\in\mathcal{S}(r,J)} \mathrm{pr}\{\|XB^* - X\hat{B} + C^* - \hat{C}\|_{\mathrm{F}}^2 \ge cP_o(J,r)\}$$
$$\ge \sup_{(A^*,C^*)\in\tilde{\mathcal{S}}(r,J)} \mathrm{pr}\{\|UDA^* - UD\hat{A} + C^* - \hat{C}\|_{\mathrm{F}}^2 \ge cP_o(J,r)\},$$

because for any $A : r(A) \le r$, $B = VA$ satisfies $r(B) \le r$. The new design matrix UD has q columns, and it is easy to see that for any $A \in \mathbb{R}^{q\times m}$,

$$\underline{\kappa}\|A\|_{\mathrm{F}}^2 \le \|UDA\|_{\mathrm{F}}^2 \le \overline{\kappa}\|A\|_{\mathrm{F}}^2, \quad (11)$$

where $\underline{\kappa} = \sigma_{\min}^2(X)$ and $\overline{\kappa} = \sigma_{\max}^2(X)$ as defined in the theorem. Therefore, without any loss of generality we assume $X \in \mathbb{R}^{n\times q}$ and and $B \in \mathbb{R}^{q\times m}$ in the rest of the proof.

Consider a signal subclass

$$\mathcal{B}^1(r) = \{B = (b_{jk}), C = 0 : b_{jk} \in \{0, \gamma R\} \text{ if } (j,k) \in [q] \times [r/2] \cup [r/2] \times [m]$$
$$b_{jk} = 0 \text{ otherwise}\}\cdot$$

where $R = \sigma/(\overline{\kappa}^{1/2})$, and $\gamma > 0$ is a small constant to be chosen later. Clearly, $|\mathcal{B}^1(r)| = 2^{(q+m-r/2)r/2}$, $\mathcal{B}^1(r) \subset \mathcal{S}(r, J)$, and $r(B_1 - B_2) \le r$, for any $B_1, B_2 \in \mathcal{B}^1(r)$. Also, since $r \le q \wedge m$, $(q + m - r/2)r/2 \ge c(q + m)r$ for some constant c.

Let $\rho(B_1, B_2) = \| \operatorname{vec}(B_1) - \operatorname{vec}(B_2)\|_0$, the Hamming distance between $\operatorname{vec}(B_1)$ and $\operatorname{vec}(B_2)$. By the Varshamov-Gilbert bound, cf. Lemma 2.9 in Tsybakov (2009), there exists a subset $\mathcal{B}^{10}(r) \subset \mathcal{B}^1(r)$ such that

$$\log|\mathcal{B}^{10}(r)| \ge c_1 r(q + m), \quad \rho(B_1, B_2) \ge c_2 r(q + m), B_1, B_2 \in \mathcal{B}^{10}, B_1 \ne B_2$$

for some universal constants $c_1, c_2 > 0$. Then $\|B_1 - B_2\|_{\mathrm{F}}^2 = \gamma^2 R^2 \rho(B_1, B_2) \ge c_2 \gamma^2 R^2 (q + m)r$. It follows from (11) that

$$\|XB_1 - XB_2\|_{\mathrm{F}}^2 \ge c_2 \underline{\kappa} \gamma^2 R^2 (q + m)r \tag{12}$$

for any $B_1, B_2 \in \mathcal{B}^{10}$, $B_1 \ne B_2$, where $\underline{\kappa}/\overline{\kappa}$ is a positive constant.

For Gaussian models, the Kullback-Leibler divergence of $\mathcal{MN}(XB_2, \sigma^2 I \otimes I)$, denoted by P_{B_2}, from $\mathcal{MN}(XB_1), \sigma^2 I \otimes I)$, denoted by P_{B_1}, is

$$\mathcal{K}(\mathcal{P}_{B_1}, \mathcal{P}_{B_2}) = \frac{1}{2\sigma^2}\|XB_1 - XB_2\|_{\mathrm{F}}^2.$$

Let P_0 be $\mathcal{MN}(0, \sigma^2 I \otimes I)$. By (11) again, for any $B : r(B) \le r$, we have

$$\mathcal{K}(P_0, P_B) \le \frac{1}{2\sigma^2}\overline{\kappa}\gamma^2 R^2 \rho(0, B) \le \frac{\gamma^2}{\sigma^2}\overline{\kappa} R^2 (q + m)r,$$

where we used $\rho(B_1, B_2) \le r(q + m)$. Therefore,

$$\frac{1}{|\mathcal{B}^{10}|}\sum_{B \in \mathcal{B}^{10}} \mathcal{K}(P_0, P_B) \le \gamma^2 r(q + m). \tag{13}$$

Combining (12) and (13) and choosing a sufficiently small value for γ, we can apply Theorem 2.7 of Tsybakov (2009) to get the desired lower bound.

Case (ii) $(q + m)r < Jm + J\log(en/J)$. Define a signal subclass

$$\begin{aligned}\mathcal{B}^2(J) = &\{B, C = (c_1, \dots, c_n)^{\mathrm{T}} : B = 0, c_i = 0 \text{ or } \gamma R(1^{\mathrm{T}}, b^{\mathrm{T}})^{\mathrm{T}} \\ &\text{with } 1 = (1 \ \dots \ 1)^{\mathrm{T}} \in \mathbb{R}^{m - \lceil m/2 \rceil}, b \in \{0, 1\}^{\lceil m/2 \rceil}, J(C) \le J\}.\end{aligned}$$

where

$$R = \frac{\sigma}{\overline{\kappa}^{1/2}}\left\{1 + \frac{\log(en/J)}{m}\right\}^{1/2},$$

and $\gamma > 0$ is a small constant. Clearly, $\mathcal{B}^2(J) \subset \mathcal{S}(r, J)$. By Stirling's approximation,

$$\log|\mathcal{B}^2(J)| \ge \log\binom{n}{J} + \log 2^{Jm/2} \ge J\log(n/J) + Jm(\log 2)/2 \ge c\{J\log(en/J) + Jm\}$$

for some universal constant c. Applying Lemma 8.3 in Rigollet & Tsybakov (2011) and the Varshamov-Gilbert bound, there exists a subset $\mathcal{B}^{20}(J) \subset \mathcal{B}^2(J)$ such that

$$\log|\mathcal{B}^{20}(J)| \ge c_1\{J\log(en/J) + Jm\} \text{ and } \rho(B_1, B_2) \ge c_2 Jm, \forall B_1, B_2 \in \mathcal{B}^{20}, B_1 \ne B_2$$

for some universal constants $c_1, c_2 > 0$. The afterward treatment follows the same lines as in (i) and the details are omitted.

1·7. *Proof of Theorem 7*

The first conclusion follows from the block coordinate descent design and the optimality of the multivariate thresholding for solving the C-optimization problem (She, 2012).

When the continuity condition holds, $\vec{\Theta}(Y - XB; \lambda)$ is the unique minimizer of $\min_C F(B, C)$; see Lemma 1 of She (2012). But in general, the problem of $\min_B F(B, C)$ subject to $r(B) \leq r$ may not have a unique solution. The accumulation point result is an application of Zangwill's Global Convergence Theorem (Luenberger & Ye, 2008), and the proof proceeds along similar lines of the proof of Theorem 7 of Bunea et al. (2012). The details are omitted.

To get the stationarity guarantee when $q(\cdot; \lambda) \equiv 0$, we can write the problem as $\min \|Y - XSV^{\mathrm{T}} - C\|_{\mathrm{F}}^2/2 + \sum_{i=1}^n P_\Theta(\|c_i\|_2; \lambda)$ subject to $(S, V, C) \in \mathbb{R}^{p\times r} \times \mathbb{O}^{m\times r} \times \mathbb{R}^{n\times m}$, where $\mathbb{O}^{m\times r} = \{V \in \mathbb{R}^{m\times r} : V^{\mathrm{T}}V = I\}$. Then one can view the problem as an unconstrained one on the manifold $\mathbb{R}^{p\times r} \times \mathbb{O}^{m\times r} \times \mathbb{R}^{n\times m}$, and define the Remannian gradient with respect to V; see Theorem 6 of Bunea et al. (2012) for more detail.

1·8. *Proof of Theorem 8*

First, by a bit of algebra we have the following result.

LEMMA 6. *For any $(\hat{B}, \hat{C})$ defined in the theorem, we have*

$$(\hat{B}, \hat{C}) \in \arg\min_{(B,C)} g(B, C; B^-, C^-)|_{B^-=\hat{B}, C^-=\hat{C}} \text{ s.t. } r(B) \leq r,$$

where g is constructed by $g(B, C; B^-, C^-) = l(B^-, C^-) + P_{2,\Theta}(C; \lambda) + \langle XB^- + C^- - Y, XB - XB^- + C - C^-\rangle + \|XB - XB^-\|_F^2/2 + \|C - C^-\|_F^2/2$, with $l(B, C) = \|XB + C - Y\|_F^2/2$ and $P_{2,\Theta}(C; \lambda) = \sum_{i=1}^n P_\Theta(\|c_i\|_2; \lambda)$.

The following result can be obtained from Lemma 2 in She (2012).

LEMMA 7. *Let $Q(C) = \|C - Y\|_F^2/2 + P_{2,\Theta}(C; \lambda)$ and $C^o = \vec{\Theta}(Y; \lambda)$. Assume that $\vec{\Theta}$ is continous at Y. Then for any C, $Q(C) - Q(C^o) \geq (1 - \mathcal{L}_\Theta)\|C - C^o\|_F^2/2$.*

LEMMA 8. *Let $Q(B) = \|XB - Y\|_F^2/2$ and $B^o = \mathcal{R}(X, Y, r)$ which is of rank r. Then for any $B : r(B) \leq r/(1 + \alpha)$ with $\alpha \geq 0$, $Q(B) - Q(B^o) \geq \{1 - (1 + \alpha)^{-1/2}\}\|XB - XB^o\|_F^2/2$.*

The lemma follows from Proposition 2.2 of She (2013) and Lemma 9 below.

LEMMA 9. *The optimization problem $\min_{\beta\in\mathbb{R}^p} l(\beta) = \|y - \beta\|_2^2/2$ s.t. $\|\beta\|_0 \leq q$ has $\hat{\beta} = \Theta^\#(y; q)$ as a globally optimal solution. Assume that $J(\hat{\beta}) = q$, where $J(\cdot) = \|\cdot\|_0$. Then for any β with $J(\beta) \leq s = q/\theta$ and $\theta \geq 1$, we have $l(\beta) - l(\hat{\beta}) \geq \{1 - \mathcal{L}(\mathcal{J}, \hat{\mathcal{J}})\}\|\hat{\beta} - \beta\|_2^2/2$ where $\mathcal{L}(\mathcal{J}, \hat{\mathcal{J}}) = (|\mathcal{J} \setminus \hat{\mathcal{J}}|/|\hat{\mathcal{J}} \setminus \mathcal{J}|)^{1/2} \leq (s/q)^{1/2} = \theta^{-1/2}$, $\mathcal{J} = \mathcal{J}(\beta)$ and $\hat{\mathcal{J}} = \mathcal{J}(\hat{\beta})$.*

With Lemmas 6, 7, and 8 available, the conclusion results from Theorem 2 of She (2016).

Proof of Lemma 9

Proof. Let $\mathcal{J}_1 = \mathcal{J} \cap \hat{\mathcal{J}}$, $\mathcal{J}_2 = \hat{\mathcal{J}} \setminus \mathcal{J}$ and $\mathcal{J}_3 = \mathcal{J} \setminus \hat{\mathcal{J}}$. Then $\beta = \beta_{\mathcal{J}_1} + \beta_{\mathcal{J}_3}$ and $\hat{\beta} = \beta_{\mathcal{J}_1} + \beta_{\mathcal{J}_2}$. By writing $\beta_{\mathcal{J}_1} = y_{\mathcal{J}_1} + \delta_{\mathcal{J}_1}$ and $\beta_{\mathcal{J}_3} = y_{\mathcal{J}_3} + \delta_{\mathcal{J}_3}$, we have

$$l(\beta) - l(\hat{\beta}) = \frac{1}{2}\|\delta_{\mathcal{J}_1}\|_2^2 + \frac{1}{2}\|y_{\mathcal{J}_2}\|_2^2 + \frac{1}{2}\|\delta_{\mathcal{J}_3}\|_2^2 - \frac{1}{2}\|y_{\mathcal{J}_3}\|_2^2$$

$$\frac{1}{2}\|\hat{\beta} - \beta\|_2^2 = \frac{1}{2}\|\delta_{\mathcal{J}_1}\|_2^2 + \frac{1}{2}\|y_{\mathcal{J}_2}\|_2^2 + \frac{1}{2}\|y_{\mathcal{J}_3} + \delta_{\mathcal{J}_3}\|_2^2.$$

The key lies in the comparison between $\|y_{\mathcal{J}_2}\|_2^2 + \|\delta_{\mathcal{J}_3}\|_2^2 - \|y_{\mathcal{J}_3}\|_2^2$ and $\|y_{\mathcal{J}_2}\|_2^2 + \|y_{\mathcal{J}_3} + \delta_{\mathcal{J}_3}\|_2^2$. Let $K \leq 1$ satisfy

$$\frac{1}{2}\|y_{\mathcal{J}_2}\|_2^2 + \frac{1}{2}\|\delta_{\mathcal{J}_3}\|_2^2 - \frac{1}{2}\|y_{\mathcal{J}_3}\|_2^2 \geq \frac{K}{2}\|y_{\mathcal{J}_2}\|_2^2 + \frac{K}{2}\|y_{\mathcal{J}_3} + \delta_{\mathcal{J}_3}\|_2^2,$$

which is equivalent to

$$(1-K)\|y_{\mathcal{J}_2}\|_2^2 + \|\delta_{\mathcal{J}_3}\|_2^2 \geq K\|y_{\mathcal{J}_3} + \delta_{\mathcal{J}_3}\|_2^2 + \|y_{\mathcal{J}_3}\|_2^2. \tag{14}$$

By construction, $|y_i| \geq |y_j|$ for any $i \in \mathcal{J}_2$ and $j \in \mathcal{J}_3$. Thus $\|y_{\mathcal{J}_2}\|_2^2/J_2 \geq \|y_{\mathcal{J}_3}\|_2^2/J_3$, from which it follows that (14) is implied by

$$(1-K)\frac{J_2}{J_3}\|y_{\mathcal{J}_3}\|_2^2 + \|\delta_{\mathcal{J}_3}\|_2^2 \geq (1+K)\|y_{\mathcal{J}_3}\|_2^2 + K\|\delta_{\mathcal{J}_3}\|_2^2 + 2K\langle y_{\mathcal{J}_3}, \delta_{\mathcal{J}_3}\rangle,$$

or

$$\frac{(1-K)(J_2/J_3) - (1+K)}{K}\|y_{\mathcal{J}_3}\|_2^2 + \frac{1-K}{K}\|\delta_{\mathcal{J}_3}\|_2^2 \geq 2\langle y_{\mathcal{J}_3}, \delta_{\mathcal{J}_3}\rangle.$$

Therefore, the largest possible K satisfies

$$\frac{(1-K)(J_2/J_3) - (1+K)}{K} \times \frac{1-K}{K} = 1$$

or $(1-K)^2 = J_3/J_2$. This gives

$$\mathcal{L} = 1 - K = (J_3/J_2)^{1/2} \leq \{(J_3 + J_1)/(J_2 + J_1)\}^{1/2} = (J/\hat{J})^{1/2} \leq \theta^{-1/2}.$$

The proof is complete. □

1·9. *Proof of Theorem 9*

Let $h(B, C; A) = 1/\{mn - AP(B, C)\}$. It follows from $1/(1-\delta) \geq \exp(\delta)$ for any $0 \leq \delta < 1$ and $\exp(\delta) \geq 1/(1-\delta/2)$ for any $0 \geq \delta < 2$ that

$$\begin{aligned} mn\|Y - X\hat{B} - \hat{C}\|_{\mathrm{F}}^2\, h(\hat{B}, \hat{C}; A/2) \leq& \|Y - X\hat{B} - \hat{C}\|_{\mathrm{F}}^2 \exp\{\delta(\hat{B}, \hat{C})\} \\ \leq& \|Y - XB^* - C^*\|_{\mathrm{F}}^2 \exp\{\delta(B^*, C^*)\} \\ \leq& \|Y - XB^* - C^*\|_{\mathrm{F}}^2\, h(B^*, C^*; A)mn. \end{aligned}$$

Since $h(\hat{B}, \hat{C}; A/2) > 0$, we have

$$\|Y - X\hat{B} - \hat{C}\|_{\mathrm{F}}^2 \leq \|Y - XB^* - C^*\|_{\mathrm{F}}^2\, h(B^*, C^*; A)/h(\hat{B}, \hat{C}; A/2).$$

With a bit of algebra, we get

$$\begin{aligned} M(\hat{B} - B^*, \hat{C} - C^*) \leq& \|\mathcal{E}\|_{\mathrm{F}}^2\{h(B^*, C^*; A)/h(\hat{B}, \hat{C}; 0{\cdot}5A) - 1\} \\ &+ 2\langle \mathcal{E}, X\hat{B} - XB^* + \hat{C} - C^*\rangle \\ \leq& \frac{A\|\mathcal{E}\|_{\mathrm{F}}^2}{mn\sigma^2 - A\sigma^2 P(B^*, C^*)}\sigma^2 P(B^*, C^*) - \frac{0{\cdot}5A\|\mathcal{E}\|_{\mathrm{F}}^2}{mn\sigma^2}\sigma^2 P(\hat{B}, \hat{C}) \\ &+ 2\langle \mathcal{E}, X\hat{B} - XB^* + \hat{C} - C^*\rangle. \end{aligned}$$

We give a finer treatment of the last stochastic term than that in the proof of Lemma 3, to show that $\langle \mathcal{E}, X\hat{B} - XB^* + \hat{C} - C^*\rangle$ can be bounded by $P(B^*, C^*) + P(\hat{B}, \hat{C})$ up to a multiplicative constant with high probability. Let $\Delta^B = \hat{B} - B^*$, $\Delta^C = \hat{C} - C^*$, $\hat{\mathcal{J}} = \mathcal{J}(\hat{C})$, $\mathcal{J}^* = \mathcal{J}(C^*)$, $\hat{r} = r(\hat{B})$, $r^* = r(C^*)$. In the following, given any index set $\mathcal{J} \subset [n]$, we denote by $I_{\mathcal{J}}$ the submatrix of $I_{n\times n}$ formed by the columns indexed by $\mathcal{J}$, and abbreviate $\mathcal{P}_{I_{\mathcal{J}}}$ to $\mathcal{P}_{\mathcal{J}}$. Let $\mathcal{P}_1 = \mathcal{P}_{\mathcal{J}^*}$, $\mathcal{P}_2 = \mathcal{P}_{(\mathcal{J}^*)^c\cap\hat{\mathcal{J}}}$, $\mathcal{P}_3 = \mathcal{P}_{(\mathcal{J}^*\cup\hat{\mathcal{J}})^c}$, and $\mathcal{P}_{rs}$ be the orthogonal projection onto the row space of XB^* which is of rank $\leq r^*$. Then

$$
\begin{aligned}
&X\Delta^B - \Delta^C\\
=&\,\mathcal{P}_1(X\Delta^B - \Delta^C) + \mathcal{P}_2(X\Delta^B - \Delta^C) + \mathcal{P}_3(X\Delta^B - \Delta^C)\mathcal{P}_{rs} + \mathcal{P}_3(X\Delta^B - \Delta^C)\mathcal{P}_{rs}^{\perp}\\
\equiv&\,\Delta_1 + \Delta_2 + \Delta_3 + \Delta_4,
\end{aligned}
$$

and $\sum_{i=1}^4 \|\Delta_i\|_{\mathrm{F}}^2 = \|X\Delta^B - \Delta^C\|_{\mathrm{F}}^2$. Then $CS(\Delta_1) \subset \mathcal{P}_{\mathcal{J}^*}$, $CS(\Delta_2) \subset \mathcal{P}_{\hat{\mathcal{J}}}$, $r(\Delta_3) \leq r^*$, and $r(\Delta_4) = r(\mathcal{P}_3 X\Delta^B\mathcal{P}_{rs}^{\perp}) = r(\mathcal{P}_3 X\hat{B}\mathcal{P}_{rs}^{\perp}) \leq \hat{r}$. The stochastic term can then be handled in a way similar to that in Lemma 3. For example, we can use the following result to handle $\langle \mathcal{E}, \Delta_4\rangle$.

LEMMA 10. *Given* $X \in \mathbb{R}^{n\times p}$, $1 \leq J_1, J_2 \leq n$, $1 \leq r \leq m \wedge p$, *define* $\Gamma_{r,J_1,J_2} = \{A \in \mathbb{R}^{n\times m} : \|A\|_F \leq 1, r(A) \leq r, CS(A) \subset CS[X\{(\mathcal{J}_1 \cup \mathcal{J}_2)^c, :\}]$ *for some* $\mathcal{J}_1, \mathcal{J}_2 : |\mathcal{J}_1| = J_1, |\mathcal{J}_2| = J_2\}$. *Let*

$$
P_o(J_1, J_2, r) = \sigma^2\left\{qr + (m-r)r + \log\binom{n}{J_1} + \log\binom{n}{J_2}\right\}.
$$

Then for any $t \geq 0$,

$$
\mathrm{pr}\Big[\sup_{A\in\Gamma_{r,J_1,J_2}} \langle \mathcal{E}, A\rangle \geq t\sigma + \{LP_o(J_1, J_2, r)\}^{1/2}\Big] \leq c' \exp(-ct^2), \tag{15}
$$

where $L, c, c' > 0$ *are universal constants.*

Following the lines of the proof of Theorem 2 in She (2017), we can show that for any constants $a, b, a' > 0$ satisfying $4b > a$, the following event

$$
2\langle \mathcal{E}, X\Delta^B - \Delta^C\rangle \leq 2(1/a + 1/a')M(\hat{B} - B^*, \hat{C} - C^*) + 8bL\sigma^2\{P(\hat{B}, \hat{C}) + P(B^*, C^*)\}
$$

occurs with probability at least $1 - c_1' n^{-c_1}$ for some $c_1, c_1' > 0$, where L is a sufficiently large constant.

Let γ and γ' be constants satisfying $0 < \gamma < 1, \gamma' > 0$. On $\mathcal{A} = \{(1-\gamma)mn\sigma^2 \leq \|\mathcal{E}\|_{\mathrm{F}}^2 \leq (1+\gamma')mn\sigma^2\}$, we have

$$
\begin{aligned}
&\frac{A\|\mathcal{E}\|_{\mathrm{F}}^2}{mn\sigma^2 - A\sigma^2 P(B^*, C^*)}\sigma^2 P(B^*, C^*) - \frac{0{\cdot}5A\|\mathcal{E}\|_{\mathrm{F}}^2}{mn\sigma^2}\sigma^2 P(\hat{B}, \hat{C})\\
\leq&\,\frac{(1+\gamma')AA_0}{A_0 - A}\sigma^2 P(B^*, C^*) - 0{\cdot}5(1-\gamma)A\sigma^2 P(\hat{B}, \hat{C}).
\end{aligned}
$$

From Laurent & Massart (2000), the complement of $\mathcal{A}$ occurs with probability at most $c_2' \exp(-c_2 mn)$, where c_2, c_2' are dependent on constants γ, γ'. With A_0 large enough, we can choose a, a', b, A such that $(1/a + 1/a') < 1/2$, $4b > a$, and $16bL \leq (1-\gamma)A$. The conclusion results.

1·10. *Theorem 10*

THEOREM 10. *Let* $(\hat{B}, \hat{C}) = \arg\min_{(B,C)} \|Y - XB - C\|_F^2/2 + \lambda\|C\|_{2,1}$ *subject to* $r(B) \leq r$, $\lambda = A\sigma(m + \log n)^{1/2}$ *where* $r \geq r^* \geq 1$ *and* A *is a large enough constant. Assume that* X *satisfies* $(1+\vartheta)\lambda\|C'_{\mathcal{J}^*}\|_{2,1} + n\|B'\|_F^2 \leq \lambda\|C'_{\mathcal{J}^{*c}}\|_{2,1} + \sigma\zeta\{(m+q)r\}^{1/2}\|XB' + C'\|_F$

for all B' and C' with $r(B') \le 2r$, where $\vartheta > 0$ is a constant and $\zeta \ge 0$. Then, we have

$$E(\|\hat{B} - B^*\|_F^2) \lesssim \sigma^2(1+\zeta^2)\frac{(m+q)r}{n}.$$

Proof. A careful examination of the proof of Theorem 3 shows that for any $a \ge 2b > 0$,

$$(1-\frac{1}{a})M(\hat{B}-B^*, \hat{C}-C^*) \le 2aA_0\sigma^2 r(m+q) + R + 2P(C^*;\lambda) - 2P(\hat{C};\lambda) + \frac{1}{b}P_{2,H}(\hat{C}-C^*;\lambda),$$

where $\lambda = A\lambda^o$, $\lambda^o = \sigma(m+\log n)^{1/2}$, $A = (abA_1)^{1/2}$, $A_1 \ge A_0$ with A_0 a universal constant, and $ER \le ac\sigma^2$.

Set $b = 1/(2\theta)$, $\theta = \vartheta/(2+\vartheta)$. Then

$$\begin{aligned}(1-\frac{1}{a})M(\hat{B}-B^*, \hat{C}-C^*) \le\,& 2(1-\theta)\lambda\{(1+\vartheta)\|(\hat{C}-C^*)_{\mathcal{J}^*}\|_{2,1} - \|(\hat{C}-C^*)_{\mathcal{J}^{*c}}\|_{2,1}\} \\ &+ 2aA_0\sigma^2 r(m+q) + R \\ \le\,& 2(1-\theta)\Big[\sigma\zeta\{(m+q)r\}^{1/2}\{M(\hat{B}-B^*, \hat{C}-C^*)\}^{1/2} \\ &- n\|\hat{B}-B^*\|_{\mathrm{F}}^2\Big] + 2aA_0\sigma^2 r(m+q) + R.\end{aligned}$$

The conclusion follows by applying Hölder's inequality and setting, say, $a = 2 + 1/\theta$, $b = 1/2\theta$ and $A \ge (abA_0)^{1/2}$. □

2. SIMULATIONS

2·1. *Simulation setups*

We consider three model setups. In Models I and II, we set $n = 100$, $p = 12$, $m = 8$, and $r^* = 3$. The design matrix X is generated by sampling its n rows from $N(0, \Delta_0)$, where Δ_0 is with diagonal elements 1 and off-diagonal elements 0.5. This brings in wide-range predictor correlation. The rows of the error matrix $\mathcal{E}$ are generated as independently and identically distributed samples from $N(0, \sigma^2\Sigma_0)$. Models I and II differ in their error structures. In Model I, we set $\Sigma_0 = I$, whereas in Model II, Σ_0 has the same compound symmetry structure as Δ_0. In each simulation, σ^2 is computed to control the signal to noise ratio, defined as the ratio between the r^*th singular value of XB^* and $\|\mathcal{E}\|_{\mathrm{F}}$.

Model III is a high-dimensional setup with $n = 100, p = 500, m = 50, r^* = 3$ and $q = 10$. As such, there are 25,000 unknown parameters in the coefficient matrix, posing a challenging high-dimensional problem. The design is generated as $X = X_1X_2\Delta_0^{1/2}$, where $X_1 \in \mathbb{R}^{n\times q}$, $X_2 \in \mathbb{R}^{q\times p}$, and all entries of X_1 and X_2 are independently and identically distributed samples from $N(0,1)$. The error structure is the same as in Model II.

In each of the three models, B^* is randomly generated as $B^* = B_1B_2^{\mathrm{T}}$ in each simulation, where $B_1 \in \mathbb{R}^{p\times r^*}$, $B_2 \in \mathbb{R}^{m\times r^*}$ and all entries in B_1 and B_2 are independently and identically distributed samples from $N(0,1)$. Outliers are then added by setting the first $n \times O\%$ rows of C^* to be nonzero, where $O\% \in \{5\%, 10\%, 15\%\}$. Concretely, the jth entry in any outlier row of C^* is α times the standard deviation of the jth column of XB^*, where $1 \le j \le m$ and $\alpha = 2, 4$. To make the problem even more challenging, we modify all entries of the first two rows of the design to 10. This yields some outliers with high leverage values. Finally, the response Y is generated as $Y = XB^* + C^* + \mathcal{E}$. Overall, the signal is contaminated by both random errors and gross outliers. Under each setting, the entire data generation process described above is replicated 200 times.

2·2. *Methods and evaluation metrics*

We compare the proposed robust reduced-rank regression with several robust regression approaches and rank reduction methods. There exist many robust multivariate regression methods in the traditional large-n setting. We mainly consider the MM-estimator by Tatsuoka & Tyler (2000), using its implementation provided by the R package `FRB` and the default settings therein. Other robust estimators including the S-estimator (Aelst & Willems, 2005) and the GS-estimator (Roelant et al., 2009) were also examined; we omit their results here, as they were similar to or slightly worse than those of the MM-estimator. None of these classical methods is applicable in high dimensions, and so they were only used on the datasets generated according to Models I and II.

For reduced-rank methods, we consider the plain reduced-rank regression (Bunea et al., 2011) and the reduced-rank ridge regression (Mukherjee & Zhu, 2011; She, 2013), both tuned by 10-fold cross validation. The latter method combines rank reduction and shrinkage estimation, which can potentially improve the predictive performance of the former when the predictors exhibit strong correlation.

We also consider a three-step fitting-detection-refitting procedure. Specifically, the first step is to fit a plain reduced-rank regression using all data; in the second step, the value of the residual sum of squares is computed for each of the n observation rows, and exactly $n \times O\%$ observations with the largest residual sum of squares are labeled as outliers and discarded; at the third step, the plain reduced-rank regression is refitted with the rest of the observations. This method can be regarded as a naive oracle procedure, as it relies on the knowledge of the true number of outliers.

As for the proposed robust reduced-rank regression, we used the ℓ_0 penalized form and the predictive information criterion for tuning. Our method allows the incorporation of the error structure through setting the weighting matrix Γ; see Equation (8) of the paper. To investigate the impact of weighting, we considered both $\Gamma = I$ and $\Gamma = \hat{\Sigma}^{-1}$ in the setting of Model II, where $\hat{\Sigma}$ is a robust estimate of $\Sigma = \sigma^2\Sigma_0$ from MM-estimation. Since it is in general difficult to estimate Σ in high dimensional settings, for the data generated in Model III we just set $\Gamma = I$. For each rank value $r = 1, \ldots, \min(n, q)$, we compute the solutions over a grid of 100 λ values equally spaced on the log scale, corresponding to a proper interval of the proportion of outliers given by $[v_L, v_U]$. We take $v_L = 0$ and $v_U \approx 0{\cdot}4$, as in practice the proportion of outliers is usually under 40%. All the methods are implemented in a user-friendly R package.

To characterize estimation accuracy robustly, we report the 10% trimmed mean of the mean squared error from all runs,

$$\mathrm{Err}(\hat{B}) = \|XB^* - X\hat{B}\|_{\mathrm{F}}^2/(mn).$$

In Model II, we additionally report the 10% trimmed mean of the weighted mean squared errors from all runs, defined as

$$\mathrm{Err}(\hat{B};\Sigma) = \mathrm{tr}\{(XB^* - X\hat{B})\Sigma^{-1}(XB^* - X\hat{B})^{\mathrm{T}}\}/(mn),$$

where $\Sigma = \sigma^2\Sigma_0$ is the true error covariance matrix. Similarly, the prediction error is defined as

$$\mathrm{Err}(\hat{B}, \hat{C}) = \|XB^* + C^* - X\hat{B} - \hat{C}\|_{\mathrm{F}}^2/(mn).$$

While the robust reduced-rank regression explicitly estimates C^*, this is not the case for the other approaches. In the plain reduced-rank regression and the reduced-rank ridge regression, $\hat{C}$ is set as a zero matrix, while in the MM estimation and the three-step procedure, the rows in $\hat{C}$

corresponding to the identified outliers are filled with model residuals in $Y - X\hat{B}$. The leverage points, if exists, are removed from X in the above calculations.

To evaluate the rank selection performance, we report the average of rank estimates from all runs. To examine the outlier detection performance, we report the average masking rate, i.e., the fraction of undetected outliers, the average swamping rate, i.e., the fraction of good points labeled as outliers, and the frequency of correct joint outlier detection, i.e., the fraction of simulations with no masking and no swamping.

2·3. *Simulation results*

Tables 1–3 summarize the simulation results of Models I–III, respectively, for $\alpha = 2$ and signal to noise ratio 0.75. We omit the results in other settings since they deliver similar messages.

In Models I and II, the MM-estimates achieved better predictive performance than both reduced-rank regression and reduced-rank ridge regression. This demonstrates that when severe outliers are present, it is pivotal to perform robust estimation. Even in these low-dimensional settings, the proposed robust reduced-rank regression outperforms all other methods, and perfectly detects all outliers jointly. MM-estimation can also achieve pretty low masking rates, but this comes at the cost of increasing false positives, which translates to efficiency loss. In particular, when the errors become correlated, our robust reduced-rank regression still showed impressive performance in both prediction and outlier detection. Additionally, the inverse covariance weighting did show some improvements over the identity weighting, but the gain was small.

Both reduced-rank regression and reduced-rank ridge regression tended to overestimated the rank in the presence of highly leveraged outliers. This complies with the theoretical results, cf. Remark 7 following Theorem 6. In contrast, robust reduced-rank regression achieved nearly perfect rank selection in all the experiments. The three-step procedure relies on the accuracy of the estimated model residuals, and often fails in the presence of leverage points. In practice, making a judgement of the number of outliers is critical. One merit of the proposed method is that the theoretically justified predictive information criterion can choose suitable parameters regardless of the size of n, m, or p, leading to an automatic identification of the right amount of outlyingness from a predictive learning perspective.

Similar conclusions can be drawn from the comparison in the high-dimensional model. Indeed, according to Table 3, the robust reduced-rank regression showed comparable or better performance than the other methods in almost all categories.

2·4. *Size of K*

We performed numerical experiments to study the size of K in the regularity condition of Theorem 6, which also plays a role in the final oracle inequality (26). It is easy to see that the condition is implied by the restricted eigenvalue condition $\|\Delta^C_{\mathcal{J}}\|_{\mathrm{F}}^2 \leq \{K^2/(1+\vartheta)^2\}\|X\Delta^B + \Delta^C\|_{\mathrm{F}}^2$, for all (Δ^B, Δ^C) in a cone defined by $r(\Delta^B) \leq 2r$, $\|\Delta^C_{\mathcal{J}^c}\|_{2,1} \leq (1+\vartheta)\|\Delta^C_{\mathcal{J}}\|_{2,1}$. Such a type of regularity conditions is commonly assumed in large-p analysis, and because of the restricted cone, K often does not grow as fast as p, m or n (van de Geer & Bühlmann, 2009; Bunea et al., 2011). We verified this by computer experiments using the Gaussian designs in the simulation models. See Table 4 for more detail.

2·5. *Convex vs. nonconvex penalties*

We also experimented with using the convex group ℓ_1 penalty in the robust reduced-rank regression, which, according to Theorem 2, amounts to applying Huber's loss. Figures 1–3 show the boxplots of prediction errors for comparing various reduced-rank methods. Clearly, the group ℓ_1 penalization shows significant improvements over the ℓ_2-penalized or the ordinary reduced-

Table 1: Simulation results of Model I with $\alpha = 2$ and signal to noise ratio 0.75. The errors are reported with their standard errors in parentheses

	Err($\hat{B}$)	Err($\hat{B}$, $\hat{C}$)	Rank	Mask	Swamp	Detection
			5%			
MM	0·4 (0·2)	4·2 (1·7)	8·0	0%	3·7%	0%
RRR	2·9 (3·7)	6·1 (4·4)	3·6	100%	0%	0%
RRS	1·8 (0·8)	4·7 (1·7)	4·0	100%	0%	0%
RRO	0·3 (0·3)	1·2 (1)	3·1	18·1%	1%	28·5%
R^4	0·2 (0·1)	0·3 (0·1)	3·0	0%	0%	100%
			10%			
MM	0·4 (0·2)	12·3 (6)	8·0	0%	2·6%	1·5%
RRR	5·4 (5)	15·9 (8·5)	3·5	100%	0%	0%
RRS	3·5 (2·4)	14·3 (9·7)	4·1	100%	0%	0%
RRO	0·3 (0·2)	2 (1·3)	3·0	13·3%	1·5%	20·5%
R^4	0·2 (0·1)	0·4 (0·2)	3·0	0%	0%	100%
			15%			
MM	0·5 (0·4)	17·8 (6·6)	8·0	0·1%	1·4%	24%
RRR	4·4 (2·1)	17·9 (5·5)	3·8	100%	0%	0%
RRS	4 (2·5)	18·4 (6·1)	3·9	100%	0%	0%
RRO	0·5 (0·3)	2·3 (1·5)	3·0	8·9%	1·6%	27·5%
R^4	0·3 (0·2)	0·8 (0·5)	2·9	0%	0%	100%

MM, the robust MM-regression method; RRR, the reduced-rank regression; RRS, the reduced-rank ridge regression; RRO, the three-step procedure for reduced-rank estimation with outlier detection; R^4, the proposed robust reduced-rank regression with $\Gamma = I$; Rank, the average of rank estimates; Mask, the average masking rate; Swamp, the average swamping rate; Detection, the frequency of correct joint outlier detection.

rank regression when outliers occur, but its performance is still substantially worse and less stable than that of using the nonconvex group ℓ_0 penalization.

3. STOCK LOG-RETURN DATA

Consider the 52 weekly stock log-return data for nine of the ten largest American corporations in 2004 available from the R package MRCE (Rothman et al., 2010), with $y_t \in \mathbb{R}^9$ $(t = 1, \ldots, T)$ and $T = 52$. Chevron was excluded due to its drastic changes (Yuan et al., 2007). The nine time series are shown in Figure 4. For the purpose of constructing market factors that drive general stock movements, a reduced-rank vector autoregressive model can be used, i.e., $y_t = B^* y_{t-1} + e_t$, with B^* of low rank. By conditioning on the initial state y_0 and assuming the normality of e_t, the conditional likelihood leads to a least squares criterion, so the estimation of B^* can be formulated as a reduced-rank regression problem (Reinsel, 1997; Lütkepohl, 2007). However, as shown in the figure, several stock returns experienced short-term changes, and the autoregressive structure makes any outlier in the time series also a leverage point in the covariates.

Using the weekly log-returns in the first 26 weeks for training and those in the last 26 weeks for forecast, we analyzed the data with the reduced-rank regression and the proposed robust reduced-rank regression approach. While both methods resulted in unit-rank models, the robust reduced-rank regression automatically detected three outliers, i.e., the log-returns of Ford at weeks 5 and 17 and the log-return of General Motors at week 5.These correspond to two real major market

Table 2: Simulation results of Model II with $\alpha = 2$ and signal to noise ratio 0.75. The layout of the table is similar to that of Table 1

	Err($\hat{B}$)	Err($\hat{B}$; Σ)	Err($\hat{B}$, $\hat{C}$)	Rank	Mask	Swamp	Detection
			5%				
MM	0·4 (0·3)	0·4 (0·3)	6·9 (2·9)	8·0	0%	3·3%	0%
RRR	2·6 (2·4)	4·6 (4·3)	9·8 (6·2)	4·0	100%	0%	0%
RRS	1·9 (1·4)	3·3 (2·5)	8·5 (4·4)	4·3	100%	0%	0%
RRO	0·4 (0·3)	0·5 (0·3)	2·7 (1·8)	3·0	25·7%	1·4%	17%
R^4	0·2 (0·2)	0·2 (0·2)	0·3 (0·2)	3·0	0%	0·2%	84%
R^4_w	0·2 (0·1)	0·2 (0·2)	0·3 (0·2)	3·0	0%	0%	100%
			10%				
MM	0·5 (0·3)	0·5 (0·4)	21·2 (9·7)	8·0	0%	1·9%	12·5%
RRR	3·6 (1·1)	6·5 (2·3)	21·7 (9·1)	4·1	100%	0%	0%
RRS	4 (1·8)	7·4 (3·7)	24·6 (10·6)	4·0	100%	0%	0%
RRO	0·4 (0·2)	0·6 (0·3)	4·3 (2·1)	3·0	16·4%	1·8%	4·5%
R^4	0·3 (0·2)	0·4 (0·3)	0·7 (0·6)	3·0	0%	0%	99·5%
R^4_w	0·2 (0·1)	0·3 (0·2)	0·6 (0·4)	3·0	0%	0%	100%
			15%				
MM	0·4 (0·2)	0·4 (0·2)	31·3 (12·4)	8·0	0%	1·1%	46·5%
RRR	4·5 (2·7)	7·9 (5·2)	33·4 (13·4)	4·3	100%	0%	0%
RRS	4·8 (3·4)	8·7 (6·8)	36·5 (16·1)	4·0	100%	0%	0%
RRO	0·4 (0·2)	0·6 (0·2)	3·3 (1·4)	3·0	9·4%	1·7%	10%
R^4	0·2 (0·2)	0·3 (0·2)	0·6 (0·3)	3·0	0·3%	0%	95·5%
R^4_w	0·2 (0·1)	0·2 (0·1)	0·5 (0·2)	3·0	0%	0%	100%

R^4_w, the robust reduced-rank regression with $\Gamma = \hat{\Sigma}^{-1}$, where $\hat{\Sigma}$ is a robust estimate of $\Sigma = \sigma^2\Sigma_0$ obtained from MM-estimation. The other notations are the same as in Table 1.

disturbances attributed to the auto industry. Our robust method automatically took the outlying samples into account and led to a more reliable model. Table 5 displays the factor coefficients indicating how the stock returns are related to the estimated factors, and the p-values for testing the associations between the estimated factors and the individual stock return series using the data in the last 26 weeks. The stock factor estimated robustly has positive influence over all nine companies, and overall, it correlates with the series better according to the reported p-values. The out-of-sample prediction errors for least squares, reduced-rank regression and robust reduced-rank regression are $9{\cdot}97$, $8{\cdot}85$ and $6{\cdot}72$, respectively, when measured by mean square error, and are $5{\cdot}44$, $4{\cdot}52$ and $3{\cdot}58$, respectively, when measured by 40% trimmed mean square error. The robustification of rank reduction resulted in about 20% improvement in prediction.

Table 3: Simulation results of Model III with $\alpha = 2$ and signal to noise ratio 0.75. The values of actual Err($\hat{B}$) and Err($\hat{B}$, $\hat{C}$) are divided by 100 for better presentation. The layout of the table is similar to that of Table 1

	Err($\hat{B}$)	Err($\hat{B}$, $\hat{C}$)	Rank	Mask	Swamp	Detection
			5%			
RRR	2·5 (0·9)	15·5 (6·3)	4·0	100%	0%	0%
RRS	2·4 (0·9)	15·6 (6·3)	4·0	100%	0%	0%
RRO	1 (0·6)	3·9 (3·9)	3·0	11·3%	0·6%	67·5%
R^4	0·9 (0·5)	1·6 (0·9)	3·0	1·6%	0%	96%
			10%			
RRR	5·4 (2·3)	47·5 (18)	4·0	100%	0%	0%
RRS	5·1 (2·1)	47·8 (18)	4·0	100%	0%	0%
RRO	0·8 (0·4)	5·1 (4·6)	3·0	4·9%	0·5%	68·5%
R^4	0·7 (0·3)	2·2 (0·9)	3·0	0%	0%	100%
			15%			
RRR	8·7 (4·2)	77 (39·9)	4·0	100%	0%	0%
RRS	8 (3·6)	77·4 (40)	4·0	100%	0%	0%
RRO	1·4 (0·8)	11·9 (8·5)	3·0	9·7%	1·7%	24%
R^4	0·8 (0·3)	3·1 (1·1)	3·2	3·2%	0%	75·5%

Table 4: Magnitude of K in different cases of model dimensions

n	m	p	O%	K
60	60	200	10%	1.2
60	60	200	30%	1.6
60	120	2000	10%	1.6
60	120	2000	30%	2.2
120	60	2000	30%	1.7

n, the sample size; m, the number of responses; p, the number of predictors; O%, the proportion of outliers.

Table 5: Stock return example: the factor coefficients showing how the stock returns load on the estimated factors, and the p-values for testing the associations between the estimated factors and the stock returns using the data in the last 26 weeks

	Reduced-rank regression		Robust reduced-rank regression	
	coefficient	p-value	coefficient	p-value
Walmart	0·46	0·44	0·36	0·23
Exxon	−0·15	0·32	0·14	0·84
General Motors	0·96	0·42	0·90	0·02
Ford	1·20	0·64	0·59	0·18
General Electric	0·24	0·67	0·32	0·06
Conoco Phillips	−0·04	0·19	0·36	0·08
Citi Group	0·27	0·93	0·45	0·00
International Business Machines	0·36	0·42	0·57	0·13
American International Group	0·19	0·01	0·58	0·00

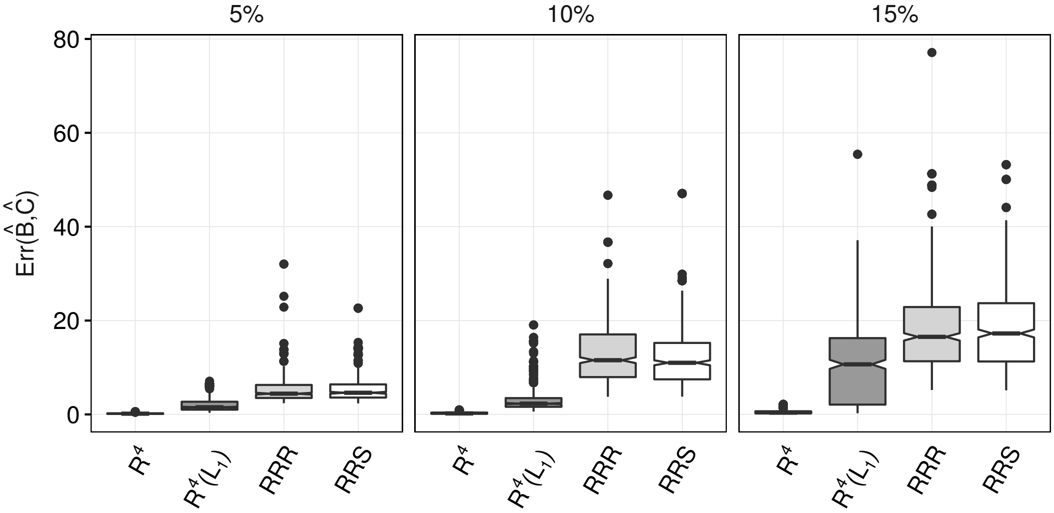

Fig. 1: Boxplots of prediction errors in Model I for comparing reduced-rank methods. RRR, the reduced-rank regression; RRS, the reduced-rank ridge regression; R^4, the proposed robust reduced-rank regression with the nonconvex group ℓ_0 penalty; $\mathrm{R}^4(L_1)$, the robust reduced-rank regression with the convex group ℓ_1 penalty.

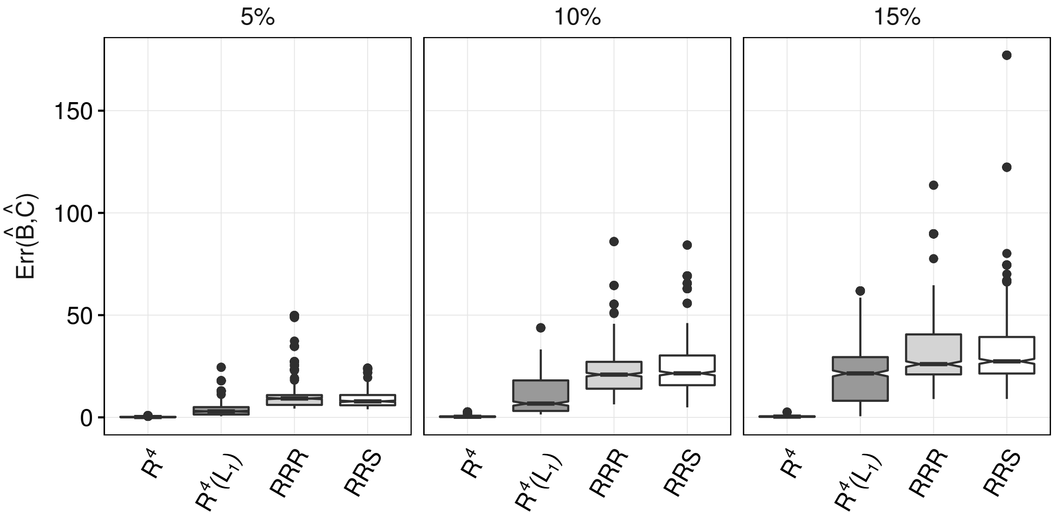

Fig. 2: Boxplots of prediction errors in Model II for comparing reduced-rank methods. The notations and layout are the same as in Figure 1.

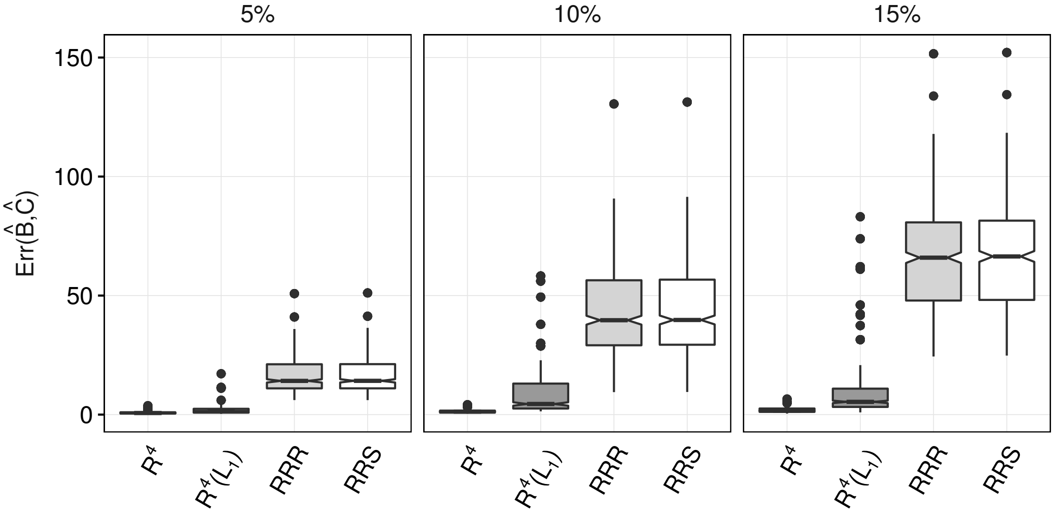

Fig. 3: Boxplots of prediction errors in Model III for comparing reduced-rank methods. The values of actual Err($\hat{B}$, $\hat{C}$) are divided by 100 to be consistent with Table 3. The notations and layout are the same as in Figure 1.

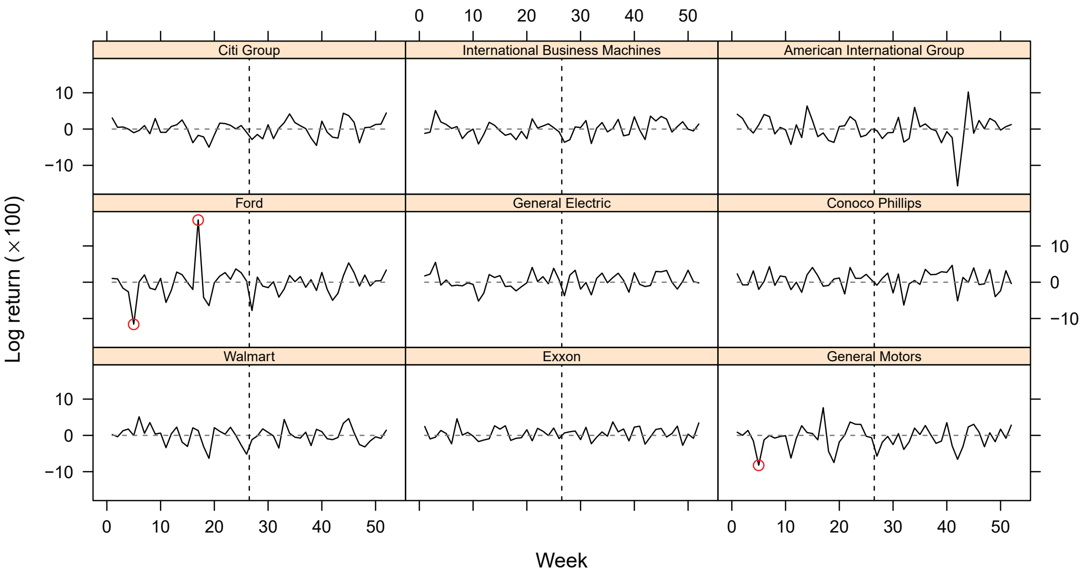

Fig. 4: Stock return example: scaled weekly log-returns of stocks in 2004. The log-returns of Ford at weeks 5 and 17 and the log-return of General Motors at week 5 are captured as outliers by fitting robust reduced-rank regression with data in the first 26 weeks; the corresponding points are indicated by the circles. The dashed line in each panel separates the series to two parts, i.e., the first 26 weeks for training and the last 26 weeks for testing. The horizontal line in each panel is drawn at zero height.

REFERENCES

AELST, S. V. & WILLEMS, G. (2005). Multivariate regression S-estimators for robust estimation and inference. *Statistica Sinica* **15**, 981–1001.

BUNEA, F., SHE, Y. & WEGKAMP, M. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Annals of Statistics* **39**, 1282–1309.

BUNEA, F., SHE, Y. & WEGKAMP, M. (2012). Joint variable and rank selection for parsimonious estimation of high dimensional matrices. *Annals of Statistics* **40**, 2359–2388.

LAURENT, B. & MASSART, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics* **28**, 1302–1338.

LUENBERGER, D. & YE, Y. (2008). *Linear and Nonlinear Programming*. New York: Springer-Verlag.

LÜTKEPOHL, H. (2007). *New Introduction to Multiple Time Series Analysis*. Springer-Verlag Berlin Heidelberg.

MUKHERJEE, A. & ZHU, J. (2011). Reduced rank ridge regression and its kernel extensions. *Statistical Analysis and Data Mining* **4**, 612–622.

REINSEL, G. C. (1997). *Elements of Multivariate Time Series Analysis*. New York: Springer-Verlag.

RIGOLLET, P. & TSYBAKOV, A. (2011). Exponential screening and optimal rates of sparse estimation. *Annals of Statistics* **39**, 731–771.

ROELANT, E., AELST, S. V. & CROUX, C. (2009). Multivariate generalized S-estimators. *Journal of Multivariate Analysis* **100**, 876–887.

ROTHMAN, A. J., LEVINA, E. & ZHU, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics* **19**, 947–962.

SHE, Y. (2009). Thresholding-based iterative selection procedures for model selection and shrinkage. *Electron. J. Statist.* **3**, 384–415.

SHE, Y. (2012). An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Computational Statistics & Data Analysis* **56**, 2976–2990.

SHE, Y. (2013). Reduced rank vector generalized linear models for feature extraction. *Statistics and Its Interface* **6**, 197–209.

SHE, Y. (2016). On the finite-sample analysis of θ-estimators. *Electron. J. Statist.* **10**, 1874–1895.

SHE, Y. (2017). Selective factor extraction in high dimensions. *Biometrika* **104**, 97–110.

SHE, Y., WANG, J., LI, H. & WU, D. (2013). Group iterative spectrum thresholding for super-resolution sparse spectral selection. *IEEE Transactions on Signal Processing* **61**, 6371–6386.

TATSUOKA, K. & TYLER, D. (2000). The uniqueness of S and M-functionals under nonelliptical distributions. *Annals of Statistics* **28**, 1219–1243.

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. New York: Springer-Verlag.

VAN DE GEER, S. A. & BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the lasso. *Electron. J. Statist.* **3**, 1360–1392.

YUAN, M., EKICI, A., LU, Z. & MONTEIRO, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *Journal of the Royal Statistical Society: Series B* **69**, 329–346.