The Annals of Statistics
2011, Vol. 39, No. 2, 1282–1309
DOI: 10.1214/11-AOS876

OPTIMAL SELECTION OF REDUCED RANK ESTIMATORS OF HIGH-DIMENSIONAL MATRICES

BY FLORENTINA BUNEA[1], YIYUAN SHE AND MARTEN H. WEGKAMP[1]

Florida State University

We introduce a new criterion, the Rank Selection Criterion (RSC), for selecting the optimal reduced rank estimator of the coefficient matrix in multivariate response regression models. The corresponding RSC estimator minimizes the Frobenius norm of the fit plus a regularization term proportional to the number of parameters in the reduced rank model.

The rank of the RSC estimator provides a consistent estimator of the rank of the coefficient matrix; in general, the rank of our estimator is a consistent estimate of the *effective rank*, which we define to be the number of singular values of the target matrix that are appropriately large. The consistency results are valid not only in the classic asymptotic regime, when n, the number of responses, and p, the number of predictors, stay bounded, and m, the number of observations, grows, but also when either, or both, n and p grow, possibly much faster than m.

We establish minimax optimal bounds on the mean squared errors of our estimators. Our finite sample performance bounds for the RSC estimator show that it achieves the optimal balance between the approximation error and the penalty term.

Furthermore, our procedure has very low computational complexity, linear in the number of candidate models, making it particularly appealing for large scale problems. We contrast our estimator with the nuclear norm penalized least squares (NNP) estimator, which has an inherently higher computational complexity than RSC, for multivariate regression models. We show that NNP has estimation properties similar to those of RSC, albeit under stronger conditions. However, it is not as parsimonious as RSC. We offer a simple correction of the NNP estimator which leads to consistent rank estimation.

We verify and illustrate our theoretical findings via an extensive simulation study.

1. Introduction.

In this paper, we propose and analyze dimension reduction-type estimators for multivariate response regression models. Given m observations of the responses $Y_i \in \mathbb{R}^n$ and predictors $X_i \in \mathbb{R}^p$, we assume that the matrices $Y = [Y_1, \ldots, Y_m]'$ and $X = [X_1, \ldots, X_m]'$ are related via an unknown $p \times n$ matrix

Received October 2010; revised January 2011.

[1]Supported in part by NSF Grant DMS-10-07444.

MSC2010 subject classifications. 62H15, 62J07.

Key words and phrases. Multivariate response regression, reduced rank estimators, dimension reduction, rank selection, adaptive estimation, oracle inequalities, nuclear norm, low rank matrix approximation.

of coefficients A, and write this as

$$Y = XA + E, \tag{1}$$

where E is a random $m \times n$ matrix, with independent entries with mean zero and variance σ^2.

Standard least squares estimation in (1), under no constraints, is equivalent to regressing each response on the predictors separately. It completely ignores the multivariate nature of the possibly correlated responses, see, for instance, Izenman (2008) for a discussion of this phenomenon. Estimators restricted to have rank equal to a fixed number $k \leq n \wedge p$ were introduced to remedy this drawback. The history of such estimators dates back to the 1950s, and was initiated by Anderson (1951). Izenman (1975) introduced the term *reduced-rank regression* for this class of models and provided further study of the estimates. A number of important works followed, including Robinson (1973, 1974) and Rao (1980). The monograph on reduced rank regression by Reinsel and Velu (1998) has an excellent, comprehensive account of more recent developments and extensions of the model. All theoretical results to date for estimators of A constrained to have rank equal to a given value k are of asymptotic nature and are obtained for fixed p, independent of the number of observations m. Most of them are obtained in a likelihood framework, for Gaussian errors E_{ij}. Anderson (1999) relaxed this assumption and derived the asymptotic distribution of the estimate, when p is fixed, the errors have two finite moments, and the rank of A is known. Anderson (2002) continued this work by constructing asymptotic tests for rank selection, valid only for small and fixed values of p.

The aim of our work is to develop a non-asymptotic class of methods that yield reduced rank estimators of A that are easy to compute, have rank determined adaptively from the data, and are valid for any values of m, n and p, especially when the number of predictors p is large. The resulting estimators can then be used to construct a possibly much smaller number of new transformed predictors or can be used to construct the most important canonical variables based on the original X and Y. We refer to Chapter 6 in Izenman (2008) for a historical account of the latter.

We propose to estimate A by minimizing the sum of squares $\|Y - XB\|_F^2 = \sum_i \sum_j \{Y_{ij} - (XB)_{ij}\}^2$ plus a penalty $\mu r(B)$, proportional to the rank $r(B)$, over all matrices B. It is immediate to see, using Pythagoras' theorem, that this is equivalent with computing $\min_B \{\|PY - XB\|_F^2 + \mu r(B)\}$ or $\min_k \{\min_{B\,:\,r(B)=k} \|PY - XB\|_F^2 + \mu k\}$, with P being the projection matrix onto the column space of X. In Section 2.1, we show that the minimizer $\widehat{k}$ of the above expression is the number of singular values $d_k(PY)$ of PY that exceed $\mu^{1/2}$. This observation reveals the prominent role of the tuning parameter μ in constructing $\widehat{k}$. The final estimator $\widehat{A}$ of the target matrix A is the minimizer of $\|PY - XB\|_F^2$ over matrices B of rank $\widehat{k}$, and can be computed efficiently even for large p, using the procedure that we describe in detail in Section 2.1 below.

The theoretical analysis of our proposed estimator $\widehat{A}$ is presented in Sections 2.2–2.4. The rank of A may not be the most appropriate measure of sparsity in multivariate regression models. For instance, suppose that the rank of A is 100, but only three of its singular values are large and the remaining 97 are nearly zero. This is an extreme example, and in general one needs an objective method for declaring singular values as "large" or "small." We introduce in Section 2.1 a slightly different notion of sparsity, that of *effective rank*. The effective rank counts the number of singular values of the signal XA that are above a certain noise level. The relevant notion of noise level turns out to be the largest singular value of PE. This is central to our results, and influences the choice of the tuning sequence μ. In Appendix C, we prove that the expected value of the largest singular value of PE is bounded by $(q+n)^{1/2}$, where $q \le m \wedge p$ is the rank of X. The effective noise level is at most $(m+n)^{1/2}$, for instance in the model $Y = A + E$, but it can be substantially lower, of order $(q+n)^{1/2}$, in model (1).

In Section 2.2, we give tight conditions under which $\widehat{k}$, the rank of our proposed estimator $\widehat{A}$, coincides with the effective rank. As an immediate corollary, we show when $\widehat{k}$ equals the rank of A. We give finite sample performance bounds for $\|X\widehat{A} - XA\|_F^2$ in Section 2.3. These results show that $\widehat{A}$ mimics the behavior of reduced rank estimates based on the ideal effective rank, had this been known prior to estimation. If X has a restricted isometry property, our estimate is minimax adaptive. In the asymptotic setting, for $n + (m \wedge p) \ge n + q \to \infty$, all our results hold with probability close to one, for tuning parameter chosen proportionally to the square of the noise level.

We often particularize our main findings to the setting of Gaussian $N(0, \sigma^2)$ errors E_{ij} in order to obtain sharp, explicit numerical constants for the penalty term. To avoid technicalities, we assume that σ^2 is known in most cases, and we treat the case of unknown σ^2 in Section 2.4.

We contrast our estimator with the penalized least squares estimator $\widetilde{A}$ corresponding to a penalty term $\tau\|B\|_1$ proportional to the nuclear norm $\|B\|_1 = \sum_j d_j(B)$, the sum of the singular values of B. This estimator has been studied by, among others, Yuan et al. (2007) and Lu, Monteiro and Yuan (2010), for model (1). Nuclear norm penalized estimators in general models $y = \mathcal{X}(A) + \varepsilon$ involving linear maps $\mathcal{X}$ have been studied by Candès and Plan (2010) and Negahban and Wainwright (2009). A special case of this model is the challenging matrix completion problem, first investigated theoretically, in the noiseless case, by Candès and Tao (2010). Rohde and Tsybakov (2010) studied a larger class of penalized estimators, that includes the nuclear norm estimator, in the general model $y = \mathcal{X}(A) + \varepsilon$.

In Section 3, we give bounds on $\|X\widetilde{A} - XA\|_F^2$ that are similar in spirit to those from Section 2. While the error bounds of the two estimators are comparable, albeit with cleaner results and milder conditions for our proposed estimator, there is one aspect in which the estimates differ in important ways. The nuclear norm penalized estimator is far less parsimonious than the estimate obtained via our rank selection

criterion. In Section 3, we offer a correction of the former estimate that yields a correct rank estimate.

Section 4 complements our theoretical results by an extensive simulation study that supports our theoretical findings and suggests strongly that the proposed estimator behaves very well in practice, in most situations is preferable to the nuclear norm penalized estimator and it is always much faster to compute.

Technical results and some intermediate proofs are presented in Appendices A–D.

2. The rank selection criterion.

2.1. *Methodology.* We propose to estimate A by the penalized least squares estimator

$$\widehat{A} = \arg\min_B \{\|Y - XB\|_F^2 + \mu r(B)\}. \tag{2}$$

We denote its rank by $\widehat{k}$. The minimization is taken over all $p \times n$ matrices B. Here and in what follows $r(B)$ is the rank of B and $\|C\|_F = (\sum_i \sum_j C_{ij}^2)^{1/2}$ denotes the Frobenius norm for any generic matrix C. The choice of the tuning parameter $\mu > 0$ is discussed in Section 2.2. Since

$$\min_B \{\|Y - XB\|_F^2 + \mu r(B)\} = \min_k \Big\{ \min_{B, r(B)=k} \{\|Y - XB\|_F^2 + \mu k\} \Big\}, \tag{3}$$

one needs to compute the restricted rank estimators $\widehat{B}_k$ that minimize $\|Y - XB\|_F^2$ over all matrices B of rank k. The following computationally efficient procedure for calculating each $\widehat{B}_k$ has been suggested by Reinsel and Velu (1998). Let $M = X'X$ be the Gram matrix, M^- be its Moore–Penrose inverse and let $P = XM^-X'$ be the projection matrix onto the column space of X.

1. Compute the eigenvectors $V = [v_1, v_2, \ldots, v_n]$, corresponding to the ordered eigenvalues arranged from largest to smallest, of the symmetric matrix $Y'PY$.
2. Compute the least squares estimator $\widehat{B} = M^-X'Y$.
   Construct $W = \widehat{B}V$ and $G = V'$.
   Form $W_k = W[\cdot, 1:k]$ and $G_k = G[1:k, \cdot]$.
3. Compute the final estimator $\widehat{B}_k = W_k G_k$.

In step 2 above, W_k denotes the matrix obtained from W by retaining all its rows and only its first k columns, and G_k is obtained from G by retaining its first k rows and all its columns.

Our first result, Proposition 1 below, characterizes the minimizer $\widehat{k} = r(\widehat{A})$ of (3) as the number of eigenvalues of the square matrix $Y'PY$ that exceed μ or, equivalently, as the number of singular values of the matrix PY that exceed $\mu^{1/2}$. The final estimator of A is then $\widehat{A} = \widehat{B}_{\widehat{k}}$.

Lemma 14 in Appendix B shows that the fitted matrix $X\widehat{A}$ is equal to $\sum_{j \le \widehat{k}} d_j u_j v_j'$ based on the singular value decomposition $UDV = \sum_j d_j u_j v_j'$ of the projection PY.

PROPOSITION 1. *Let $\lambda_1(Y'PY) \geq \lambda_2(Y'PY) \geq \cdots$ be the ordered eigenvalues of $Y'PY$. We have $\widehat{A} = \widehat{B}_{\widehat{k}}$ with*

$$\widehat{k} = \max\{k : \lambda_k(Y'PY) \geq \mu\}. \tag{4}$$

PROOF. For $\widehat{B}_k$ given above, and by the Pythagorean theorem, we have

$$\|Y - X\widehat{B}_k\|_F^2 = \|Y - PY\|_F^2 + \|PY - X\widehat{B}_k\|_F^2$$

and we observe that $X\widehat{B} = PY$. By Lemma 14 in Appendix B, we have

$$\|X\widehat{B} - X\widehat{B}_k\|_F^2 = \sum_{j>k} d_j^2(X\widehat{B}) = \sum_{j>k} d_j^2(PY) = \sum_{j>k} \lambda_j(Y'PY),$$

where $d_j(C)$ denotes the jth largest singular value of a matrix C. Then the penalized least squares criterion reduces to

$$\|Y - PY\|_F^2 + \left\{\sum_{j>k} \lambda_j(Y'PY) + \mu k\right\}$$

and we find that $\min_B\{\|Y - XB\|_F^2 + \mu r(B)\}$ equals

$$\|Y - PY\|_F^2 + \mu n + \min_k \sum_{j>k} \{\lambda_j(Y'PY) - \mu\}.$$

It is easy to see that $\sum_{j>k}\{\lambda_j(Y'PY) - \mu\}$ is minimized by taking k as the largest index j for which $\lambda_j(Y'PY) - \mu \geq 0$, since then the sum only consists of negative terms. This concludes our proof. □

REMARK. The two matrices $W_{\widehat{k}}$ and $G_{\widehat{k}}$, that yield the final solution $\widehat{A} = W_{\widehat{k}}G_{\widehat{k}}$, have the following properties: (i) $G_{\widehat{k}}G'_{\widehat{k}}$ is the identity matrix; and (ii) $W'_{\widehat{k}}MW_{\widehat{k}}$ is a diagonal matrix. Moreover, the decomposition of $\widehat{A}$ as a product of two matrices with properties (i) and (ii) is unique, see, for instance, Theorem 2.2 in Reinsel and Vélu (1998). As an immediate consequence, one can construct new orthogonal predictors as the columns of $Z = XW_{\widehat{k}}$. If $\widehat{k}$ is much smaller than p, this can result in a significant dimension reduction of the predictors' space.

2.2. *Consistent effective rank estimation.* In this section, we study the properties of $\widehat{k} = r(\widehat{A})$. We will state simple conditions that guarantee that $\widehat{k}$ equals $r = r(A)$ with high probability. First, we describe in Theorem 2 what $\widehat{k}$ estimates and what quantities need to be controlled for consistent estimation. It turns out that $\widehat{k}$ estimates the number of the singular values of the signal XA above the threshold $\mu^{1/2}$, for any value of the tuning parameter μ. The quality of estimation is controlled by the probability that this threshold level exceeds the largest singular value $d_1(PE)$ of the *projected* noise matrix PE. We denote the jth singular value of a generic matrix C by $d_j(C)$ and we use the convention that the singular values are indexed in decreasing order.

THEOREM 2. *Suppose that there exists an index $s \le r$ such that*

$$d_s(XA) > (1+\delta)\sqrt{\mu} \quad \text{and} \quad d_{s+1}(XA) < (1-\delta)\sqrt{\mu}$$

for some $\delta \in (0,1]$. Then we have

$$\mathbb{P}\{\widehat{k} = s\} \ge 1 - \mathbb{P}\{d_1(PE) \ge \delta\sqrt{\mu}\}.$$

PROOF. Using the characterization of $\widehat{k}$ given in Proposition 1, we have

$$\widehat{k} > s \iff \sqrt{\mu} \le d_{s+1}(PY),$$
$$\widehat{k} < s \iff \sqrt{\mu} \ge d_s(PY).$$

Therefore, $\mathbb{P}\{\widehat{k} \ne s\} = \mathbb{P}\{\sqrt{\mu} \le d_{s+1}(PY) \text{ or } \sqrt{\mu} \ge d_s(PY)\}$. Next, observe that $PY = XA + PE$ and $d_k(XA) < d_k(PY) + d_1(PE)$ for any k. Hence, $d_s(PY) \le \mu^{1/2}$ implies $d_1(PE) \ge d_s(XA) - \mu^{1/2}$, whereas $d_{s+1}(PY) \ge \mu^{1/2}$ implies that $d_1(PE) \ge \mu^{1/2} - d_{s+1}(XA)$. Consequently, we have

$$\mathbb{P}\{\widehat{k} \ne s\} \le \mathbb{P}\{d_1(PE) \ge \min(\sqrt{\mu} - d_{s+1}(XA), d_s(XA) - \sqrt{\mu})\}.$$

Invoke the conditions on $d_{s+1}(XA)$ and $d_s(XA)$ to complete the proof. □

Theorem 2 indicates that we can consistently estimate the index s provided we use a large enough value for our tuning parameter μ to guarantee that the probability of the event $\{d_1(PE) \le \delta\mu^{1/2}\}$ approaches one. We call s the *effective rank* of A relative to μ, and denote it by $r_e = r_e(\mu)$.

This is the appropriate notion of sparsity in the multivariate regression problem: we can only hope to recover those singular values of the signal XA that are above the noise level $\mathbb{E}[d_1(PE)]$. Their number, r_e, will be the target rank of the approximation of the mean response, and can be much smaller than $r = r(A)$. We regard the largest singular value $d_1(PE)$ as the relevant indicator of the strength of the noise. Standard results on the largest singular value of Gaussian matrices show that $\mathbb{E}[d_1(E)] \le \sigma(m^{1/2} + n^{1/2})$ and similar bounds are available for sub-Gaussian matrices, see, for instance, Rudelson and Vershynin (2010). Interestingly, the expected value of the largest singular value $d_1(PE)$ of the projected noise matrix is smaller: it is of order $(q+n)^{1/2}$ with $q = r(X)$. If E has independent $N(0,\sigma^2)$ entries the following simple argument shows why this is the case.

LEMMA 3. *Let $q = r(X)$ and assume that E_{ij} are independent $N(0,\sigma^2)$ random variables. Then*

$$\mathbb{E}[d_1(PE)] \le \sigma(\sqrt{n} + \sqrt{q})$$

and

$$\mathbb{P}\{d_1(PE) \ge \mathbb{E}[d_1(PE)] + \sigma t\} \le \exp(-t^2/2)$$

for all $t > 0$.

PROOF. Let $U\Lambda U'$ be the eigen-decomposition of P. Since P is the projection matrix on the column space of X, only the first q entries of Λ on the diagonal equal to one, and all the remaining entries equal to zero. Then, $d_1^2(PE) = \lambda_1(E'PE) = d_1^2(\Lambda U'E)$. Since E has independent $N(0,\sigma^2)$ entries, the rotation $U'E$ has the same distribution as E. Hence, $\Lambda U'E$ can be written as a $q \times n$ matrix with Gaussian entries on top of a $(m-q)\times n$ matrix of zeroes. Standard random matrix theory now states that $\mathbb{E}[d_1(\Lambda U'E)] \le \sigma(q^{1/2}+n^{1/2})$. The second claim of the lemma is a direct consequence of Borell's inequality, see, for instance, Van der Vaart and Wellner (1996), after recognizing that $d_1(\Lambda U'E)$ is the supremum of a Gaussian process. $\square$

In view of this result, we take $\mu^{1/2} > \sigma(n^{1/2}+q^{1/2})$ as our measure of the noise level. The following corollary summarizes the discussion above and lists the main results of this section: the proposed estimator based on the rank selection criterion (RSC) recovers consistently the effective rank r_e and, in particular, the rank of A.

COROLLARY 4. *Assume that E has independent $N(0,\sigma^2)$ entries. For any $\theta > 0$, set*

$$\mu = (1+\theta)^2\sigma^2\bigl(\sqrt{n}+\sqrt{q}\bigr)^2/\delta^2$$

with δ as in Theorem 2. Then we have, for any $\theta > 0$,

$$\mathbb{P}\{\widehat{k} \ne r_e(\mu)\} \le \exp\bigl(-\tfrac{1}{2}\theta^2(n+q)\bigr) \to 0 \qquad \text{as } q+n \to \infty.$$

In particular, if $d_r(XA) > 2\mu^{1/2}$ and $\mu^{1/2} = (1+\theta)\sigma(\sqrt{n}+\sqrt{q})$, then

$$\mathbb{P}\{\widehat{k} \ne r\} \le \exp\bigl(-\tfrac{1}{2}\theta^2(n+q)\bigr) \to 0 \qquad \text{as } q+n \to \infty.$$

REMARK. Corollary 4 holds when $q+n \to \infty$. If $q+n$ stays bounded, but $m \to \infty$, the consistency results continue to hold when q is replaced by $q\ln(m)$ in the expression of the tuning parameter μ given above. Lemma 3 justifies this choice. The same remark applies to all theoretical results in this paper.

REMARK. A more involved argument is needed in order to establish the conclusion of Lemma 3 when E has independent sub-Gaussian entries. We give this argument in Proposition 15 presented in Appendix C. Proposition 15 shows, in particular, that when $\mathbb{E}[\exp(tE_{ij})] \le \exp(t^2/\Gamma_E)$ for all $t > 0$, and for some $\Gamma_E < \infty$, we have

$$\mathbb{P}\bigl\{d_1^2(PE) \ge 32\Gamma_E(q+n)\bigl(\ln(5)+x\bigr)\bigr\} \le 2\exp\{-x(q+n)\}$$

for all $x > 0$. The conclusion of Corollary 4 then holds for $\mu = C_0\Gamma_E(n+q)$ with C_0 large enough. Moreover, all oracle inequalities presented in the next sections remain valid for this choice of the tuning parameter, if E has independent sub-Gaussian entries.

2.3. *Errors bounds for the RSC estimator.* In this section, we study the performance of $\widehat{A}$ by obtaining bounds for $\|X\widehat{A} - XA\|_F^2$. First, we derive a bound for the fit $\|X\widehat{B}_k - XA\|_F^2$, based on the restricted rank estimator $\widehat{B}_k$, for each value of k.

THEOREM 5. *Set $c(\theta) = 1 + 2/\theta$. For any $\theta > 0$, we have*

$$\|X\widehat{B}_k - XA\|_F^2 \leq \min_{1\leq k\leq \min(n,p)} \left\{ c^2(\theta) \sum_{j>k} d_j^2(XA) + 2(1+\theta)c(\theta)kd_1^2(PE) \right\}$$

with probability one.

PROOF. By the definition of $\widehat{B}_k$,

$$\|Y - X\widehat{B}_k\|_F^2 \leq \|Y - XB\|_F^2$$

for all $p \times n$ matrices B of rank k. Working out the squares, we obtain

$$\begin{aligned} \|X\widehat{B}_k - XA\|_F^2 &\leq \|XB - XA\|_F^2 + 2\langle E, X\widehat{A} - XB\rangle_F \\ &= \|XB - XA\|_F^2 + 2\langle PE, X\widehat{A} - XB\rangle_F \end{aligned}$$

with

$$\langle C, D\rangle_F = \operatorname{tr}(C'D) = \operatorname{tr}(D'C) = \sum_i \sum_j C_{ij}D_{ij}$$

for generic $m \times n$ matrices C and D. The inner product $\langle C, D\rangle_F$, operator norm $\|C\|_2 = d_1(C)$ and nuclear norm $\|D\|_1 = \sum_j d_j(D)$ are related via the inequality $\langle C, D\rangle_F \leq \|C\|_2\|D\|_1$. As a consequence we find

$$\begin{aligned} \langle PE, X\widehat{B}_k - XB\rangle_F &\leq d_1(PE)\|X\widehat{B}_k - XB\|_1 \\ &\leq d_1(PE)\sqrt{2k}\|X\widehat{B}_k - XB\|_F \\ &\leq d_1(PE)\sqrt{2k}\{\|X\widehat{B}_k - XA\|_F + \|XB - XA\|_F\}. \end{aligned}$$

Using the inequality $2xy \leq x^2/a + ay^2$ with $a > 0$ twice, we obtain that $\|X\widehat{B}_k - XA\|_F^2$ is bounded above by

$$\frac{1+b}{b}\|XB - XA\|_F^2 + \frac{1}{a}\|X\widehat{B}_k - XA\|_F^2 + (a+b)(2k)d_1^2(PE).$$

Hence we obtain, for any $a, b > 0$, the inequality

$$\|X\widehat{B}_k - XA\|_F^2 \leq \frac{a}{a-1}\left\{ \frac{1+b}{b}\|XB - XA\|_F^2 + 2(a+b)kd_1^2(PE) \right\}.$$

Lemma 14 in the Appendix B states that the minimum of $\|XA - XB\|_F^2$ over all matrices B of rank k is achieved for the GSVD of A and the minimum equals $\sum_{j>k} d_j^2(XA)$. The claim follows after choosing $a = (2+\theta)/2$ and $b = \theta/2$. □

COROLLARY 6. *Assume that E has independent $N(0,\sigma^2)$ entries. Set $c(\theta) = 1 + 2/\theta$. Then, for any $\theta, \xi > 0$, the inequality*

$$\|X\widehat{B}_k - XA\|_F^2 \leq \min_{1\leq k\leq \min(n,p)} \left\{ c^2(\theta) \sum_{j>k} d_j^2(XA) + 2c(\theta)(1+\theta)(1+\xi)^2\sigma^2 k(n+q) \right\}$$

holds with probability $1 - \exp(-\xi^2(n+q)/2)$. In addition,

$$\mathbb{E}[\|X\widehat{B}_k - XA\|_F^2] \lesssim \sum_{j>k} d_j^2(XA) + \sigma^2 k(n+q).$$

The symbol $\lesssim$ means that the inequality holds up to multiplicative numerical constants.

PROOF. Set $t = (1+\xi)^2\sigma^2(\sqrt{n}+\sqrt{q})^2$ for some $\xi > 0$. From Lemma 3, it follows that

$$\mathbb{P}\{d_1^2(PE) \geq t\} = \mathbb{P}\{d_1(PE) \geq (1+\xi)\sigma(\sqrt{n}+\sqrt{q})\} \leq \exp(-\xi^2(n+q)/2).$$

The first claim follows now from this bound and Theorem 5. From Lemma 16, it follows that $\mathbb{E}[d_1^2(PE)] \leq \nu^2 + \nu\sqrt{2\pi} + 2$ for $\nu = \mathbb{E}[d_1(PE)] \leq \sigma(\sqrt{n}+\sqrt{q})$. This proves the second claim. □

Theorem 5 bounds the error $\|X\widehat{B}_k - XA\|_F^2$ by an approximation error, $\sum_{j>k} d_j^2(XA)$, and a stochastic term, $kd_1^2(PE)$, with probability one. The approximation error is decreasing in k and vanishes for $k > r(XA)$.

The stochastic term increases in k and can be bounded by a constant times $k(n+q)$ with overwhelming probability and in expectation, for Gaussian errors, by Corollary 6 above. More generally, the same bound (up to constants) can be proved for sub-Gaussian errors. Indeed, for C_0 large enough, Proposition 15 in Appendix C, states that $\mathbb{P}\{d_1^2(PE) \leq C_0(n+q)\} \leq 2\exp\{-(n+q)\}$.

We observe that $k(n+q)$ is essentially the number of free parameters of the restricted rank problem. Indeed, our parameter space consists of all $p \times n$ matrices B of rank k and each matrix has $k(n+q-k)$ free parameters. Hence, we can interpret the bound in Corollary 6 above as the squared bias plus the dimension of the parameter space.

Remark (ii), following Corollary 8 below, shows that $k(n+q)$ is also the minimax lower bound for $\|X\widehat{B}_k - XA\|_F^2$, if the smallest eigenvalue of $X'X$ is larger than a strictly positive constant. This means that $X\widehat{B}_k$ is a minimax estimator under this assumption.

We now turn to the penalized estimator $\widehat{A}$ and show that it achieves the best (squared) bias-variance trade-off among all rank restricted estimators $\widehat{B}_k$ for the appropriate choice of the tuning parameter μ in the penalty $\text{pen}(B) = \mu r(B)$.

THEOREM 7. *We have, for any* $\theta > 0$, *on the event* $(1+\theta)d_1^2(PE) \le \mu$,

$$\|X\widehat{A} - XA\|_F^2 \le c^2(\theta)\|XB - XA\|_F^2 + 2c(\theta)\mu k \tag{5}$$

for any $p \times n$ *matrix* B. *In particular, we have, for* $\mu \ge (1+\theta)d_1^2(PE)$

$$\|X\widehat{A} - XA\|_F^2 \le \min_k \left\{ c^2(\theta) \sum_{j>k} d_j^2(XA) + 2c(\theta)\mu k \right\} \tag{6}$$

and

$$\|X\widehat{A} - XA\|_F^2 \le 2c(\theta)\mu r. \tag{7}$$

PROOF. By the definition of $\widehat{A}$,

$$\|Y - X\widehat{A}\|_F^2 + \mu r(\widehat{A}) \le \|Y - XB\|_F^2 + \mu r(B)$$

for all $p \times n$ matrices B. Working out the squares, we obtain

$$\begin{aligned}
&\|X\widehat{A} - XA\|_F^2 \\
&\quad \le \|XB - XA\|_F^2 + 2\mu r(B) + 2\langle E, X\widehat{A} - XB\rangle_F - \mu r(\widehat{A}) - \mu r(B) \\
&\quad = \|XB - XA\|_F^2 + 2\mu r(B) + 2\langle PE, X\widehat{A} - XB\rangle_F - \mu r(\widehat{A}) - \mu r(B).
\end{aligned}$$

Next, we observe that

$$\begin{aligned}
&\langle PE, X\widehat{A} - XB\rangle_F \\
&\quad \le d_1(PE)\|X\widehat{A} - XB\|_1 \\
&\quad \le d_1(PE)\{r(X\widehat{A}) + r(XB)\}^{1/2}\|X\widehat{A} - XB\|_F \\
&\quad \le d_1(PE)\{r(\widehat{A}) + r(B)\}^{1/2}\{\|X\widehat{A} - XA\|_F + \|XB - XA\|_F\}.
\end{aligned}$$

Consequently, using the inequality $2xy \le x^2/a + ay^2$ twice, we obtain, for any $a > 0$ and $b > 0$,

$$\begin{aligned}
\|X\widehat{A} - XA\|_F^2 \le{}& \|XB - XA\|_F^2 + \frac{1}{a}\|X\widehat{A} - XA\|_F^2 + \frac{1}{b}\|XB - XA\|_F^2 \\
&+ 2\mu r(B) + (a+b)\{r(\widehat{A}) + r(B)\}d_1^2(PE) - \mu\{r(\widehat{A}) + r(B)\}.
\end{aligned}$$

Hence, if $(a+b)d_1^2(PE) - \mu \le 0$, we obtain

$$\|X\widehat{A} - XA\|_F^2 \le \frac{a}{a-1}\left\{ \frac{1+b}{b}\|XB - XA\|_F^2 + 2\mu r(B) \right\}$$

for any $a > 1$ and $b > 0$. Lemma 14 in Appendix B evaluates the minimum of $\|XA - XB\|_F^2$ over all matrices B of rank k and shows that it equals $\sum_{j>k} d_j^2(XA)$. We conclude our proof by choosing $a = 1 + \theta/2$ and $b = \theta/2$. □

REMARK. The first two parts of the theorem show that $\widehat{A}$ achieves the best (squared) bias-variance trade-off among all reduced rank estimators $\widehat{B}_k$ if $\mu > d_1^2(PE)$. Moreover, the index k which minimizes $\sum_{j>k}\{d_j^2(XA) + \mu k\}$ essentially coincides with the effective rank $r_e = r_e(\mu)$ defined in the previous section. Therefore, the fit of the selected estimator $X\widehat{A}$ is comparable with that of the estimator $X\widehat{B}_k$ with rank $k = r_e$. Since the ideal r_e depends on the unknown matrix A, this ideal estimator cannot be computed. Although our estimator $\widehat{A}$ is constructed independently of r_e, it mimics the behavior of the ideal estimator $\widehat{B}_{r_e}$ and we say that the bound on $\|X\widehat{A} - XA\|_F^2$ adapts to $r_e \le r$.

The last part of our result is a particular case of the second part, but it is perhaps easier to interpret. Taking the index k equal to the rank r, the bias term disappears and the bound reduces to $rd_1^2(PE)$ up to constants. This shows clearly the important role played by r in the estimation accuracy: the smaller the rank of A, the smaller the estimation error.

For Gaussian errors, we have the following precise bounds.

COROLLARY 8. *Assume that E has independent $N(0, \sigma^2)$ entries. Set*

$$\operatorname{pen}(B) = (1+\theta)(1+\xi)^2\big(\sqrt{n} + \sqrt{q}\big)^2\sigma^2 r(B)$$

with $\theta, \xi > 0$ arbitrary. Let $c(\theta) = 1 + 2/\theta$. Then, we have

$$\mathbb{P}\left[\|X\widehat{A} - XA\|_F^2 \le \min_{1\le k\le \min(n,p)}\left\{c^2(\theta)\sum_{j>k} d_j^2(XA) + 2c(\theta)\mu k\right\}\right] \ge 1 - \exp\left\{-\frac{\xi^2(n+q)}{2}\right\}$$

and

$$\begin{aligned}\mathbb{E}[\|X\widehat{A} - XA\|_F^2] &\le \min_{1\le k\le \min(n,p)}\left[c^2(\theta)\sum_{j>k} d_j^2(XA) + 2(1-\theta)c(\theta)(1+\xi)^2\sigma^2\big(\sqrt{n}+\sqrt{q}\big)^2 k\right] \\ &\quad + 4(1+\theta)c(\theta)\min(n,p)\sigma^2(1+\xi^{-1})\exp\big(-\xi^2(n+q)\big).\end{aligned}$$

PROOF. Recall from the proof of Theorem 7 that

$$\|X\widehat{A} - XA\|_F^2 \le \frac{2+\theta}{\theta}\left\{\frac{2+\theta}{\theta}\|XB - XA\|_F^2 + 2\operatorname{pen}(B) + R\right\}$$

with R defined by

$$\begin{aligned}R &= (1+\theta)\{r(\widehat{A}) + r(B)\}d_1^2(PE) - \operatorname{pen}(\widehat{A}) - \operatorname{pen}(B) \\ &\le 2(1+\theta)\max_{1\le k\le \min(n,p)} k\big\{d_1^2(PE) - (1+\xi)^2\big(\sqrt{n}+\sqrt{q}\big)^2\sigma^2\big\}.\end{aligned}$$

For $\tilde{E} = E/\sigma$, a matrix of independent $N(0, 1)$ entries, we have

$$R \le 2(1+\theta)\sigma^2 \max_{1\le k\le \min(n,p)} k\{d_1^2(P\tilde{E}) - (1+\xi)^2(\sqrt{n}+\sqrt{q})^2\}$$

$$\le 2\min(n, p)(1+\theta)\sigma^2\big(d_1^2(P\tilde{E}) - (1+\xi)^2(\sqrt{n}+\sqrt{q})^2\big)_+.$$

Apply Lemma 16 in Appendix D to deduce that

$$\mathbb{E}[R] \le 4\min(n, p)\frac{1+\xi}{\xi}(1+\theta)\sigma^2\exp\big(-\xi^2(n+q)/2\big).$$

The conclusion follows immediately. □

REMARKS. (i) We note that for $n+q$ large,

$$\mathbb{E}[\|X\widehat{A} - XA\|_F^2] \lesssim \min_{1\le k\le \min(n,p)}\Big[\sum_{j>k} d_j^2(XA) + \sigma^2(\sqrt{n}+\sqrt{q})^2 k\Big]$$

as the remainder term in the bound of $\mathbb{E}[\|X\widehat{A} - XA\|_F^2]$ in Corollary 8 converges exponentially fast in $n+q$, to zero.

(ii) Assuming that E has independent $N(0, \sigma^2)$ entries, the RSC estimator corresponding to the penalty $\mathrm{pen}(B) = C\sigma^2(n^{1/2} + q^{1/2})^2 r(B)$, for any $C > 1$, is minimax adaptive, for matrices X having a restricted isometry property (RIP), of the type introduced and discussed in Candès and Plan (2010) and Rohde and Tsybakov (2010). The RIP implies that $\|XA\|_F^2 \ge \rho\|A\|_F^2$, for all matrices A of rank at most r and for some constant $0 < \rho < 1$. For fixed design matrices X, this is equivalent with assuming that the smallest eigenvalue $\lambda_p(M)$ of the $p \times p$ Gram matrix $M = X'X$ is larger than ρ. To establish the minimax lower bound for the mean squared error $\|X\widehat{A} - XA\|_F^2$, notice first that our model (1) can be rewritten as $y_i = \mathrm{trace}(Z_i'A) + \varepsilon_i$, with $1 \le i \le mn$, via the mapping $(a, b) \to i = a + (b-1)n$, where $1 \le a \le m$, $1 \le b \le n$, $y_i =: Y_{ab} \in \mathbb{R}$ and $Z_i =: X_a'e_b \in M_{p\times n}$. Here $X_a \in \mathbb{R}^p$ denotes the ath row of X, e_b is the row vector in $\mathbb{R}^n$ having the bth component equal to 1 and the rest equal to zero, and $M_{p\times n}$ is the space of all $p \times n$ matrices. Then, under RIP, the lower bound follows directly from Theorem 5 in Rohde and Tsybakov (2010); see also Theorem 2.5 in Candès and Plan (2010) for minimax lower bounds on $\|\widehat{A} - A\|_F^2$.

(iii) The same type of upper bound as the one of Corollary 8 can be proved if the entries of E are sub-Gaussian: take $\mathrm{pen}(B) = C(n+q)r(B)$ for some C large enough, and invoke Proposition 15 in Appendix C.

(iv) Although the error bounds of $\|X\widehat{A} - XA\|_F$ are guaranteed for all X and A, the analysis of the estimation performance of $\widehat{A}$ depends on X. If $\lambda_p(M) \ge \rho > 0$, for some constant ρ, then, provided $\mu > (1+\theta)d_1^2(PE)$ with $\theta > 0$ arbitrary,

$$\|\widehat{A} - A\|_F^2 \le \frac{c(\theta)}{\lambda_p(M)}\min_{k\le r}\Big[c(\theta)\sum_{j>k} d_j^2(XA) + 2\mu k\Big]$$

follows from Theorem 7.

(v) Our results are slightly more general than stated. In fact, our analysis does not require that the postulated multivariate linear model $Y = XA + E$ holds exactly. We denote the expected value of Y by Θ and write $Y = \Theta + E$. We denote the projection of Θ onto the column space of X by XA, that is, $P\Theta = XA$. Because minimizing $\|Y - XB\|_F^2 + \mu r(B)$ is equivalent with minimizing $\|PY - XB\|_F^2 + \mu r(B)$ by Pythagoras' theorem, our least squares procedure estimates XA, the mean of PY. The statements of Theorems 2 and 7 remain unchanged, except that XA is the mean of the projection PY of Y, not the mean of Y itself.

2.4. *A data adaptive penalty term.* In this section, we construct a data adaptive penalty term that employs the unbiased estimator

$$S^2 = \|Y - PY\|_F^2/(mn - qn)$$

of σ^2. Set, for any $\theta > 0$, $\xi > 0$ and $0 < \delta < 1$,

$$\operatorname{pen}(B) = \frac{1+\theta}{1-\delta}(1+\xi)^2\big(\sqrt{n} + \sqrt{q}\big)^2 S^2 r(B).$$

Notice that the estimator S^2 requires that $n(m-q)$ be large, which holds whenever $m \gg q$ or $m - q \geq 1$ and n is large. The challenging case $m = q \ll p$ is left for future research.

THEOREM 9. *Assume that E is an $m \times n$ matrix with independent $N(0, \sigma^2)$ entries. Using the penalty given above we have, for $c(\theta) = 1 + 2/\theta$,*

$$\begin{aligned}
&\mathbb{E}[\|X\widehat{A} - XA\|_F^2] \\
&\quad \leq \min_{1 \leq k \leq \min(n,p)} \left[c^2(\theta) \sum_{j>k} d_j^2(XA) + 2(1+\theta)c(\theta)(1+\xi)^2\sigma^2\big(\sqrt{n} + \sqrt{q}\big)^2 k \right] \\
&\qquad + 4(1+\theta)c(\theta)\min(n, p)\sigma^2(1+\xi^{-1})\exp\left(-\frac{\xi^2(n+q)}{2}\right) \\
&\qquad + 4(1+\theta)c(\theta)\min(n, p)\sigma^2\big(2 + \big(\sqrt{n} + \sqrt{q}\big)^2 + \big(\sqrt{n} + \sqrt{q}\big)\sqrt{2\pi}\big) \\
&\qquad\quad \times \exp\left\{-\frac{\delta^2 n(m-q)}{4(1+\delta)}\right\}.
\end{aligned}$$

PROOF. Set $\tilde{E} = \sigma^{-1}E$. We have, for any $p \times n$ matrix B,

$$\begin{aligned}
\|X\widehat{A} - XA\|_F^2 &\leq \frac{2+\theta}{\theta}\left[\frac{2+\theta}{\theta}\|XB - XA\|_F^2 + 2\operatorname{pen}(B)\right] \\
&\quad + 2\frac{2+\theta}{\theta}(1+\theta)\sigma^2 \\
&\qquad \times \max_{1 \leq k \leq \min(n,p)} k\left\{d_1^2(P\tilde{E}) - \frac{(1+\xi)^2(\sqrt{n} + \sqrt{q})^2 S^2}{(1-\delta)\sigma^2}\right\}.
\end{aligned}$$

It remains to bound the expected value of

$$\max_{k\le\min(n,p)} k\left\{d_1^2(P\tilde{E}) - (1+\xi)^2\big(\sqrt{n}+\sqrt{q}\big)^2\frac{S^2}{(1-\delta)\sigma^2}\right\}$$
$$\le \min(n,p)\left(d_1^2(P\tilde{E}) - (1+\xi)^2\big(\sqrt{n}+\sqrt{q}\big)^2\frac{S^2}{(1-\delta)\sigma^2}\right)_+.$$

We split the expectation into two parts: $S^2 \ge (1-\delta)\sigma^2$ and its complement. We observe first that

$$\mathbb{E}\left[\left(d_1^2(P\tilde{E}) - (1+\xi)^2\big(\sqrt{n}+\sqrt{q}\big)^2\frac{S^2}{(1-\delta)\sigma^2}\right)_+ 1_{\{S^2\ge(1-\delta)\sigma^2\}}\right]$$
$$\le \mathbb{E}\big[\big(d_1^2(P\tilde{E}) - (1+\xi)^2\big(\sqrt{n}+\sqrt{q}\big)^2\big)_+\big]$$
$$\le 2(1+\xi^{-1})\min(n,p)\exp\big(-\xi\big(\sqrt{n}+\sqrt{q}\big)/2\big),$$

using Lemma 16 for the last inequality. Next, we observe that

$$\mathbb{E}\left[\left(d_1^2(P\tilde{E}) - (1+\xi)^2\big(\sqrt{n}+\sqrt{q}\big)^2\frac{S^2}{(1-\delta)\sigma^2}\right)_+ 1_{\{S^2\le(1-\delta)\sigma^2\}}\right]$$
$$\le \mathbb{E}\big[d_1^2(P\tilde{E})1_{\{S^2\le(1-\delta)\sigma^2\}}\big] = \mathbb{E}\big[d_1^2(P\tilde{E})1_{\{\|(I-P)\tilde{E}\|_F^2\le(1-\delta)(nm-nq)\}}\big].$$

Since $P\tilde{E}$ and $(I-P)\tilde{E}$ are independent, and $\|(I-P)\tilde{E}\|_F^2$ has a χ^2_{nm-nq} distribution, we find

$$\mathbb{E}\left[\left(d_1^2(P\tilde{E}) - (1+\xi)^2\big(\sqrt{n}+\sqrt{q}\big)^2\frac{S^2}{(1-\delta)\sigma^2}\right)_+ 1_{\{S^2\le(1-\delta)\sigma^2\}}\right]$$
$$\le \mathbb{E}[d_1^2(P\tilde{E})]\mathbb{P}\{\|(I-P)\tilde{E}\|_F^2 \le (1-\delta)(nm-nq)\}$$
$$\le \big(\big(\sqrt{n}+\sqrt{q}\big)^2 + \sqrt{2\pi}\big(\sqrt{n}+\sqrt{q}\big)+2\big)\exp\left\{-\frac{\delta^2}{4(1+\delta)}n(m-q)\right\},$$

using Lemmas 16 and 17 in Appendix D for the last inequality. This proves the result. □

REMARK. We see that for large values of $n+q$ and $n(m-q)$,

$$\mathbb{E}[\|X\widehat{A}-XA\|_F^2] \lesssim \min_{1\le k\le\min(n,p)}\left[\sum_{j>k} d_j^2(XA) + \sigma^2\big(\sqrt{n}+\sqrt{q}\big)^2 k\right]$$

as the additional terms in the theorem above decrease exponentially fast in $n+q$ and $n(m-q)$. This bound is similar to the one in Corollary 8, obtained for the RSC estimator corresponding to the penalty term that employs the theoretical value of σ^2.

3. Comparison with nuclear norm penalized estimators. In this section, we compare our RSC estimator $\widehat{A}$ with the alternative estimator $\widetilde{A}$ that minimizes

$$\|Y - XB\|_F^2 + 2\tau\|B\|_1$$

over all $p \times n$ matrices B.

THEOREM 10. *On the event $d_1(X'E) \le \tau$, we have, for any B,*

$$\|X\widetilde{A} - XA\|_F^2 \le \|XB - XA\|_F^2 + 4\tau\|B\|_1.$$

PROOF. By the definition of $\widetilde{A}$,

$$\|Y - X\widetilde{A}\|_F^2 + 2\tau\|\widetilde{A}\|_1 \le \|Y - XB\|_F^2 + 2\tau\|B\|_1$$

for all $m \times n$ matrices B. Working out the squares, we obtain

$$\|\widetilde{X}A - XA\|_F^2 \le \|XB - XA\|_F^2 + 2\tau\|B\|_1 + 2\langle X'E, \widetilde{A} - B\rangle_F - 2\tau\|\widetilde{A}\|_1.$$

Since

$$\langle X'E, \widetilde{A} - B\rangle_F \le \|X'E\|_2\|\widetilde{A} - B\|_1 \le \tau\|\widetilde{A} - B\|_1$$

on the event $d_1(X'E) \le \tau$, we obtain the claim using the triangle inequality. □

We see that $\widetilde{A}$ balances the bias term $\|XA - XB\|_F^2$ with the penalty term $\tau\|B\|_1$, provided $\tau > d_1(X'E)$. Since $X'E = X'PE + X'(I - P)E = X'PE$, we have $d_1(X'E) \le d_1(X)d_1(PE)$. We immediately obtain the following corollary using the results for $d_1(PE)$ of Lemma 3.

COROLLARY 11. *Assume that E has independent $N(0, \sigma^2)$ entries. For*

$$\tau = (1 + \theta)d_1(X)\sigma\big(\sqrt{n} + \sqrt{q}\big)$$

with $\theta > 0$ arbitrary, we have

$$\mathbb{P}\{\|X\widetilde{A} - XA\|_F^2 \le \|XB - XA\|_F^2 + 4\tau\|B\|_1\} \ge 1 - \exp\{-\tfrac{1}{2}\theta^2(n + q)\}.$$

The same result, up to constants, can be obtained if the errors E_{ij} are sub-Gaussian, if we replace σ in the choice of τ above by a suitably large constant C. The proof of this generalization uses Proposition 15 in Appendix C in lieu of Lemma 3. The same remark applies for all the results in this section.

The next result obtains an oracle inequality for $\widetilde{A}$ that resembles the oracle inequality for the RSC estimator $\widehat{A}$ in Theorem 7. We stress the fact that Theorem 12 below requires that $\lambda_p(X'X) > 0$; this was not required for the derivation of the oracle bound on $\|X\widehat{A} - XA\|_F^2$ in Theorem 7, which holds for all X. We denote the condition number of $M = X'X$ by $c_0(M) = \lambda_1(M)/\lambda_p(M)$.

THEOREM 12. *Assume that E has independent $N(0,\sigma^2)$ entries. For*

$$\tau = (1+\theta)d_1(X)\sigma\big(\sqrt{n}+\sqrt{q}\big)$$

with $\theta > 0$ arbitrary, we have

$$\|X\widetilde{A} - XA\|_F^2 \lesssim \min_{k\le r}\left(\sum_{j=k+1}^{r} d_j^2(XA) + c_0(M)k\sigma^2(n+q)\right).$$

Furthermore,

$$\|\widetilde{A} - A\|_F^2 \lesssim c_0(M)\sum_{j=k+1}^{r} d_j^2(A) + \frac{c_0(M)}{\lambda_p(M)}k\sigma^2(n+q).$$

Both inequalities hold with probability at least $1-\exp(-\theta^2(n+q)/2)$. The symbol $\lesssim$ means that the inequality holds up to multiplicative numerical constants (depending on θ).

To keep the paper self contained, we give a simple proof of this result in Appendix A. Similar results for the NNP estimator of A in the general model $y = \mathcal{X}(A)+\varepsilon$, where $\mathcal{X}$ is a random linear map, have been obtained by Negahban and Wainwright (2009) and Candès and Plan (2010), each under different sets of assumptions on $\mathcal{X}$. We refer to Rohde and Tsybakov (2010) for more general results on Schatten norm penalized estimators of A in the model $y = \mathcal{X}(A)+\varepsilon$, and a very thorough discussion on the assumptions on $\mathcal{X}$ under which these results hold.

Theorem 10 shows that the error bounds of the nuclear norm penalized (NNP) estimator $\widetilde{A}$ and the RSC estimator $\widehat{A}$ are comparable, although it is worth pointing out that our bounds for $\widehat{A}$ are much cleaner and obtained under fewer restrictions on the design matrix. However, there is one aspect in which the two estimators differ radically: correct rank recovery. We showed in Section 2.2 that the RSC estimator corresponding to the effective value of the tuning sequence μ_e has the correct rank and achieves the optimal bias-variance trade-off. This is also visible in the left panel of Figure 1 which shows the plots of the MSE and rank of the RSC estimate as we varied the tuning parameter of the procedure over a large grid. The numbers on the vertical axis correspond to the range of values of the rank of the estimator considered in this experiment, 1 to 25. The rank of A is 10. We notice that for the same range of values of the tuning parameter, RSC has both the smallest MSE value and the correct rank. We repeated this experiment for the NNP estimator. The right panel shows that the smallest MSE and the correct rank are *not* obtained for the same value of the tuning parameter. Therefore, a different strategy for correct rank estimation via NNP is in order. Rather than taking the rank of $\widetilde{A}$ as the estimator of the rank of A, we consider instead, for $M = X'X$,

$$\tilde{k} = \max\{k : d_k(M\widetilde{A}) > 2\tau\}. \tag{8}$$

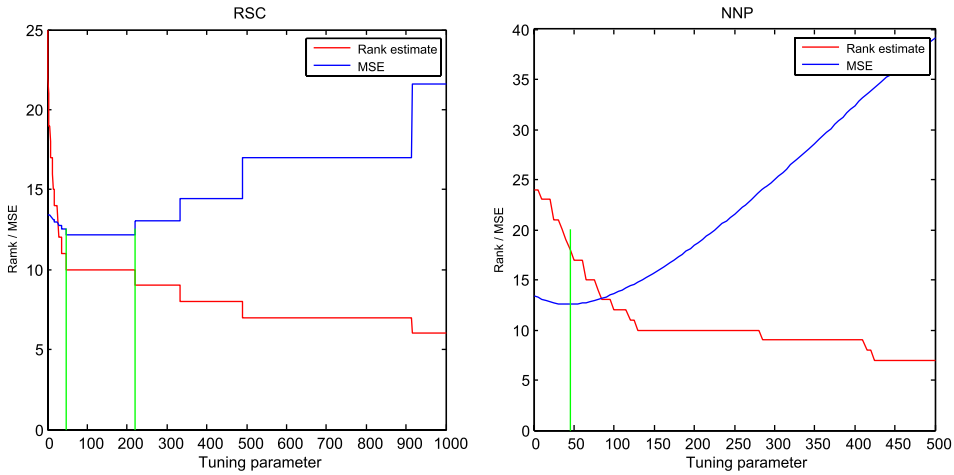

FIG. 1. *The MSE and rank of the estimators RSC* (left) *and NNP* (right) *as a function of the tuning parameter. The rank estimate and MSE curves are plotted together for a better view of the effect of tuning on different estimation aspects.*

THEOREM 13. *Let* $r = r(A)$ *and assume that* $d_r(MA) > 4\tau$. *Then*

$$\mathbb{P}\{\tilde{k} \neq r\} \leq \mathbb{P}\{d_1(X'E) > \tau\}.$$

If E *has independent* $N(0, \sigma^2)$ *entries and* $\tau = (1+\theta)\sigma d_1(X)(\sqrt{n} + \sqrt{q})$, *the above probability is bounded by* $\exp(-\theta^2(n+q)/2)$.

PROOF. After computing the sub-gradient of $f(B) = \|Y - XB\|_F^2 + 2\tau\|B\|_1$, we find that $\tilde{A}$ is a minimizer of $f(B)$ if and only if there exists a matrix J with $d_1(J) \leq 1$ such that $X'X(\tilde{A} - A) = X'E + \tau UJV'$, where $\tilde{A} = UDV'$ is the full SVD and U and V are orthonormal matrices. The matrix J is obtained from D by setting $J_{ii} = 0$ if $D_{ii} = 0$ and $J_{ii} \leq 1$ if $D_{ii} > 0$. Therefore,

$$d_1(M\tilde{A} - MA) \leq d_1(X'E) + \tau.$$

From Horn and Johnson [(1985), page 419],

$$|d_k(M\tilde{A}) - d_k(MA)| \leq d_1(M\tilde{A} - MA) \leq 2\tau$$

for all k, on the event $d_1(X'E) \leq \tau$. This means that $d_k(M\tilde{A}) > 2\tau$ for all $k \leq r$ and $d_k(M\tilde{A}) < 2\tau$ for all $k > r$, since $d_r(MA) > 4\tau$ and $d_{r+1}(MA) = 0$. The result now follows. □

4. Empirical studies.

4.1. *RSC vs. NNP.* We performed an extensive simulation study to evaluate the performance of the proposed method, RSC, and compare it with the NNP method.

The RSC estimator $\widehat{A}$ was computed via the procedure outlined in Section 2.1. This method is computationally efficient in large dimensions. Its computational complexity is the same as that of PCA. Our choice for the tuning parameter μ was based on our theoretical findings in Section 2. In particular, Corollaries 4 and 8 guarantee good rank selection and prediction performance of RSC provided that μ is just a little bit larger than $\sigma^2(\sqrt{n}+\sqrt{q})^2$. Under the assumption that $q < m$, we can estimate σ^2 by S^2; see Section 2.4 for details. In our simulations, we used the adaptive tuning parameter $\mu_{\mathrm{adap}} = 2S^2(n+q)$. We experimented with other constants and found that the constant equal to 2 was optimal; constants slightly larger than 2 gave very similar results.

We compared the RSC estimator with the NNP estimator $\widetilde{A}$ and with the proposed trimmed or calibrated NNP estimator, denoted in what follows by $\mathrm{NNP}^{(c)}$. The NNP estimator is the minimizer of the convex criterion $\|Y - XB\|_F^2 + 2\tau\|B\|_1$. By the equivalent SDP characterization of the NNP-norm given in Fazel (2002), the original minimization problem is equivalent to the convex optimization problem

$$\begin{aligned} &\min_{B\in\mathbb{R}^{p\times n}, W_1\in S^{n-1}, W_2\in S^{p-1}} \|Y - XB\|_F^2 + \tau\big(\mathrm{Tr}(W_1) + \mathrm{Tr}(W_2)\big) \\ &\text{s.t.} \quad \begin{bmatrix} W_1 & B^T \\ B & W_2 \end{bmatrix} \succeq 0. \end{aligned} \tag{9}$$

Therefore, the NNP estimator can be computed by adapting the general convex optimization algorithm `SDPT3` [Toh, Todd and Tütüncü (1999)] to (9). Alternatively, Bregman iterative algorithms can be developed; see Ma, Goldfarb and Chen (2009) for a detailed description of the main idea. Their code, however, is specifically designed for matrix completion and does not cover the multivariate regression problem. We implemented this algorithm for the simulation study presented below. The $\mathrm{NNP}^{(c)}$ is our calibration of the NNP estimator, based on Theorem 13. For a given value of the tuning parameter τ we calculate the NNP estimator $\widetilde{A}$ and obtain the rank estimate $\widetilde{r}$ from (8). We then calculate the calibrated $\mathrm{NNP}^{(c)}$ estimator as the reduced rank estimator $\widehat{B}_{\widetilde{r}}$, with rank equal to $\widetilde{r}$, following the procedure outlined in Section 2.1.

In our simulation study, we compared the rank selection and the estimation performances of the RSC estimator $\mathrm{RSC}|_{\mathrm{adap}}$, corresponding to μ_{adap}, with the optimally tuned RSC estimator, and the optimally tuned NNP and $\mathrm{NNP}^{(c)}$ estimators. The last three estimators are called $\mathrm{RSC}|_{\mathrm{val}}$, $\mathrm{NNP}|_{\mathrm{val}}$ and $\mathrm{NNP}^{(c)}|_{\mathrm{val}}$. They correspond to those tuning parameters μ_{val}, τ_{val} and τ_{val}, respectively, that gave the best prediction accuracy, when prediction was evaluated on a very large independent validation set. This comparison helps us understand the true potential of each method in an ideal situation, and allows us to draw a stable performance comparison between the proposed adaptive RSC estimator and the best possible versions of RSC and NNP.

We considered the following *large sample-size* set up and *large dimensionality* set up.

EXPERIMENT 1 ($m > p$). We constructed the matrix of dependent variables $X = [x_1, x_2, \ldots, x_m]'$ by generating its rows x_i as i.i.d. realizations from a multivariate normal distribution MVN($\mathbf{0}$, Σ), with $\Sigma_{jk} = \rho^{|j-k|}$, $\rho > 0$, $1 \le j, k \le p$. The coefficient matrix $A = bB_0B_1$, with $b > 0$, B_0 is a $p \times r$ matrix and B_1 is a $r \times n$ matrix. All entries in B_0 and B_1 are i.i.d. $N(0, 1)$. Each row in $Y = [y_1, \ldots, y_m]'$ is then generated as $y_i = x_i'A + E_i$, $1 \le i \le m$, with E_i denoting the ith row of the noise matrix E which has $m \times n$ independent $N(0, 1)$ entries E_{ij}.

EXPERIMENT 2 [$p > m(> q)$]. The sample size in this experiment is relatively small. X is generated as $X_0\Sigma^{1/2}$, where $\Sigma_{jk} = \rho^{|j-k|} \in \mathbb{R}^{p\times p}$, $X_0 = X_1X_2$, $X_1 \in \mathbb{R}^{m\times q}$, $X_2 \in \mathbb{R}^{q\times p}$ and all entries of X_1, X_2 are i.i.d. $N(0, 1)$. The coefficient matrix and the noise matrix are generated in the same way as in Experiment 1. Since $p > m$, this is a much more challenging setup than the one considered in Experiment 1. Note however that q, the rank of X, is required to be strictly less than m.

Each simulated model is characterized by the following control parameters: m (sample size), p (number of independent variables), n (number of response variables), r (rank of A), ρ (design correlation), q (rank of the design) and b (signal strength). In Experiment 1, we set $m = 100$, $p = 25$, $n = 25$, $r = 10$, and varied the correlation coefficient $\rho = 0.1, 0.5, 0.9$ and signal strength $b = 0.1, 0.2, 0.3, 0.4$. All combinations of correlation and signal strength are covered in the simulations. The results are summarized in Table 1. In Experiment 2, we set $m = 20$, $p = 100$, $n = 25$, $q = 10$, $r = 5$, and varied the correlation $\rho = 0.1, 0.5, 0.9$ and signal strength $b = 0.1, 0.2, 0.3$. The corresponding results are reported in Table 2. In both tables, MSE(A) and MSE(XA) denote the 40% trimmed-means of $100 \cdot \|A - \hat{B}\|_F^2/(pn)$ and $100 \cdot \|XA - X\hat{B}\|_F^2/(mn)$, respectively. We also report the median rank estimates (RE) and the successful rank recovery percentages (RRP).

Summary of simulation results.

(i) We found that the RSC estimator corresponding to the adaptive choice of the tuning parameter $\mu_{\text{adap}} = 2S^2(n + q)$ has excellent performance. It behaves as well as the RSC estimator that uses the parameter μ tuned on the large validation set or the RSC estimator corresponding to the theoretical $\mu = 2\sigma^2(n + q)$.

(ii) When the signal-to-noise ratio SNR $:= d_r(XA)/(\sqrt{q} + \sqrt{n})$ is moderate or high, with values approximately 1, 1.5 and 2, corresponding to $b = 0.2, 0.3, 0.4$, and for low to moderate correlation between the predictors ($\rho = 0.1, 0.5$), RSC has

TABLE 1

Performance comparisons of Experiment 1, *in terms of mean squared errors* [MSE(XA), MSE(A)], *median rank estimate* (*RE*), *and rank recovery percentage* (*RRP*)

		$\mathbf{RSC}\vert_{\mathbf{adap}}$	$\mathbf{RSC}\vert_{\mathbf{val}}$	$\mathbf{NNP}\vert_{\mathbf{val}}$	$\mathbf{NNP}^{(c)}\vert_{\mathbf{val}}$
			$b = 0.1$		
$\rho = 0.9$	MSE(XA), MSE(A)	16.6, 5.3	16.3, 5.2	11.5, 3.0	16.5, 5.3
	RE, RRP	6, 0%	6, 0%	12, 0%	6, 0%
$\rho = 0.5$	MSE(XA), MSE(A)	18.7, 1.4	18.1, 1.4	16.2, 1.1	18.1, 1.4
	RE, RRP	8, 0%	9, 40%	16.5, 0%	9, 35%
$\rho = 0.1$	MSE(XA), MSE(A)	19.3, 1.0	18.0, 0.9	16.9, 0.8	18.0, 0.9
	RE, RRP	9, 0%	10, 75%	17, 0%	10, 65%
			$b = 0.2$		
$\rho = 0.9$	MSE(XA), MSE(A)	18.4, 7.0	17.9, 7.1	15.9, 5.4	17.9, 7.1
	RE, RRP	8, 0%	9, 20%	16, 0%	9, 15%
$\rho = 0.5$	MSE(XA), MSE(A)	16.7, 1.3	16.7, 1.3	18.9, 1.5	16.7, 1.3
	RE, RRP	10, 100%	10, 100%	19, 0%	10, 100%
$\rho = 0.1$	MSE(XA), MSE(A)	16.5, 0.9	16.5, 0.9	19.2, 1.0	16.5, 0.9
	RE, RRP	10, 100%	10, 100%	18, 0%	10, 100%
			$b = 0.3$		
$\rho = 0.9$	MSE(XA), MSE(A)	17.4, 7.0	17.3, 6.9	17.7, 6.7	17.3, 7.0
	RE, RRP	10, 65%	10, 95%	18, 0%	10, 80%
$\rho = 0.5$	MSE(XA), MSE(A)	16.4, 1.3	16.4, 1.3	19.8, 1.6	16.4, 1.3
	RE, RRP	10, 100%	10, 100%	19, 0%	10, 100%
$\rho = 0.1$	MSE(XA), MSE(A)	16.4, 0.9	16.4, 0.9	19.9, 1.1	16.4, 0.9
	RE, RRP	10, 100%	10, 100%	19, 0%	10, 100%
			$b = 0.4$		
$\rho = 0.9$	MSE(XA), MSE(A)	16.8, 6.6	16.8, 6.7	18.7, 7.4	16.8, 6.8
	RE, RRP	10, 100%	10, 100%	18, 0%	10, 85%
$\rho = 0.5$	MSE(XA), MSE(A)	16.3, 1.3	16.3, 1.3	20.3, 1.7	16.3, 1.3
	RE, RRP	10, 100%	10, 100%	20, 0%	10, 100%
$\rho = 0.1$	MSE(XA), MSE(A)	16.3, 0.9	16.3, 0.9	20.3, 1.1	16.3, 0.9
	RE, RRP	10, 100%	10, 100%	20, 0%	10, 100%

excellent behavior in terms of rank selection and means squared errors. Interestingly, NNP does not have optimal behavior in this set-up: its mean squared errors are slightly higher than those of the RSC estimator. When the noise is very large relative to the signal strength, corresponding to $b = 0.1$ in Table 1, or when the correlation between some covariates is very high, $\rho = 0.9$ in Table 1, NNP may be slightly more accurate than the RSC.

(iii) The NNP does not recover the correct rank, when its regularization parameter is tuned by validation. Both Tables 1 and 2 show that the correct rank r

TABLE 2
Performance comparisons of Experiment 2, in terms of mean squared errors [MSE(XA), MSE(A)], *median rank estimate* (*RE*), *and rank recovery percentage* (*RRP*)

		$\mathbf{RSC}\vert_{\mathbf{adap}}$	$\mathbf{RSC}\vert_{\mathbf{val}}$	$\mathbf{NNP}\vert_{\mathbf{val}}$	$\mathbf{NNP}^{(c)}\vert_{\mathbf{val}}$
			$b = 0.1$		
$\rho = 0.9$	MSE(XA), MSE(A)	29.4, 3.9	29.4, 3.9	36.4, 3.9	29.4, 3.9
	RE, RRP	5, 100%	5, 100%	10, 0%	5, 100%
$\rho = 0.5$	MSE(XA), MSE(A)	29.1, 3.9	29.1, 3.9	37.2, 3.9	29.1, 3.9
	RE, RRP	5, 100%	5, 100%	10, 0%	5, 100%
$\rho = 0.1$	MSE(XA), MSE(A)	29.0, 3.9	29.0, 3.9	37.2, 4.0	29.0, 3.9
	RE, RRP	5, 100%	5, 100%	10, 0%	5, 100%
			$b = 0.2$		
$\rho = 0.9$	MSE(XA), MSE(A)	28.9, 15.7	28.9, 15.7	38.7, 15.7	28.9, 15.7
	RE, RRP	5, 100%	5, 100%	10, 0%	5, 100%
$\rho = 0.5$	MSE(XA), MSE(A)	28.6, 15.7	28.6, 15.7	39.0, 15.7	28.6, 15.7
	RE, RRP	5, 100%	5, 100%	10, 0%	5, 100%
$\rho = 0.1$	MSE(XA), MSE(A)	28.7, 15.8	28.7, 15.8	38.7, 15.8	28.7, 15.8
	RE, RRP	5, 100%	5, 100%	10, 0%	5, 100%
			$b = 0.3$		
$\rho = 0.9$	MSE(XA), MSE(A)	28.8, 35.3	28.8, 35.3	39.2, 35.3	28.8, 35.3
	RE, RRP	5, 100%	5, 100%	10, 0%	5, 100%
$\rho = 0.5$	MSE(XA), MSE(A)	28.5, 35.4	28.5, 35.4	39.5, 35.4	28.5, 35.4
	RE, RRP	5, 100%	5, 100%	10, 0%	5, 100%
$\rho = 0.1$	MSE(XA), MSE(A)	28.6, 35.5	28.6, 35.5	39.3, 35.5	28.6, 35.5
	RE, RRP	5, 100%	5, 100%	10, 0%	5, 100%

($r = 10$ in Experiment 1 and $r = 5$ in Experiment 2) is overestimated by NNP. Our trimmed estimator, $\text{NNP}^{(c)}$, provides a successful improvement over NNP in this respect. This supports Theorem 13.

In additional simulations, we found that, especially for low or moderate SNRs, the NNP parameter tuning problem is much more challenging than the RSC parameter tuning. NNP cannot accurately estimate A and consistently select the rank at the same time, for the same value of the tuning parameter. This echoes the findings presented in Figure 1, and is to be expected: in NNP regularization, the threshold value τ also controls the amount of shrinkage, which should be mild for large samples with relatively low contamination. This is the case for moderate SNR and moderate correlation between predictors: the tuned τ tends to be too small, so it cannot introduce enough sparsity. The same continues to be true for slightly larger values of τ that compensate for high noise level and very high correlation between predictors. In summary, one may not be able to build an accurate *and* parsimonious model via the NNP method, without further adjustments.

Overall, RSC is recommended over the NNP estimators, especially when we suspect that the SNR is not very low. With large validation tuning, NNP$^{(c)}$ has the same properties as RSC – they coincide when both methods select the same rank. But in general, the rank estimation via NNP$^{(c)}$ is much more difficult to tune and much more computationally involved than RSC.

For data with low SNR, an immediate extension of the RSC estimator that involves a second penalty term, of ridge-type, may induce the right amount of shrinkage needed to offset the noise in the data. This conjecture will be investigated carefully in future research.

4.2. *Tightness of the rank consistency results*. It can be shown, using arguments similar to those used in the proof of Theorem 2, that

$$\mathbb{P}\{\widehat{k} \neq r\} \geq P_1 \equiv \mathbb{P}\big\{\sqrt{\mu} \leq d_{2r+1}(PE) \text{ or } d_1(PE) < \sqrt{\mu} - d_r(XA)\big\}.$$

On the other hand, the proof of Theorem 2 reveals that

$$\mathbb{P}\{\widehat{k} \neq r\} \leq P_2 \equiv \mathbb{P}\big\{d_1(PE) \geq \min\big(\sqrt{\mu}, d_r(XA) - \sqrt{\mu}\big)\big\}.$$

Suppose now that $2\mu^{1/2} < d_r(XA)$ and that r is small. Then P_1 equals $\mathbb{P}\{d_{2r+1}(PE) \geq \sqrt{\mu}\}$ and is close to $P_2 = \mathbb{P}\{d_1(PE) \geq \sqrt{\mu}\}$ for a sparse model. Of course, if μ is much larger than $d_r^2(XA)$, then P_2 cannot be small. We use this observation to argue that, if the goal is consistent rank estimation, then we can deviate only very little from the requirement $2\mu^{1/2} < d_r(XA)$. This strongly suggests that the sufficient condition given in Corollary 4 for consistent rank selection is tight. We empirically verified this conjecture for signal-to-noise ratios larger than 1 by comparing $\mu_1 = d_r^2(XA)$ with μ_u, the ideal upper bound of that interval of values of μ that give the correct rank. The value of μ_u was obtained in the simulation experiments by searching along solution paths obtained as follows. We constructed 100 different pairs (X, A) following the simulation design outlined in the subsection above. Each pair was obtained by varying the signal strength b, correlation ρ, the rank of A and m, n, p. For each run we computed the solution path, as in Figure 1 of the previous section. From the solution path we recorded the upper value of the μ interval for which the correct rank was recovered. We plotted the resulting (μ_1, μ_u) pairs in Figure 2 and we conclude that the theoretical bound on μ in Corollary 4 is tight.

APPENDIX A: PROOF OF THEOREM 12

The starting point is the inequality

$$\|X\widetilde{A} - XA\|_F^2 \leq \|XB - XA\|_F^2 + 2\tau\{\|\widetilde{A} - B\|_1 + \|B\|_1 - \|\widetilde{A}\|_1\}$$

that holds on the event $d_1(X'E) \leq \tau$. The inequality can be deduced from the proof of Theorem 10. Then, by Lemmas 3.4 and 2.3 in Recht, Fazel and Parrilo (2010) there exist two matrices $\widetilde{A}_1$ and $\widetilde{A}_2$ such that:

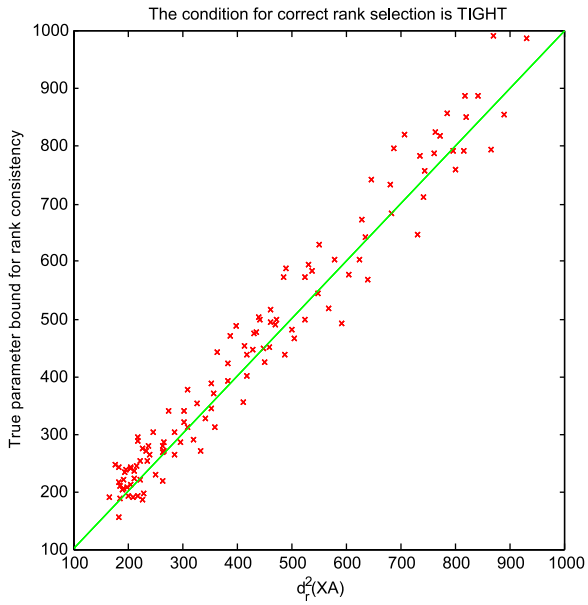

FIG. 2. *Tightness of the consistency condition.*

(i) $\widetilde{A} = \widetilde{A}_1 + \widetilde{A}_2$,
(ii) $r(\widetilde{A}_1) \le 2r(B)$,
(iii) $\|\widetilde{A} - B\|_1 = \|\widetilde{A}_1 - B\|_1 + \|\widetilde{A}_2\|_1$,
(iv) $\|\widetilde{A} - B\|_F^2 = \|\widetilde{A}_1 - B\|_F^2 + \|\widetilde{A}_2\|_F^2 \ge \|\widetilde{A}_1 - B\|_F^2$,
(v) $\|\widetilde{A}\|_1 = \|\widetilde{A}_1\|_1 + \|\widetilde{A}_2\|_1$.

Using the display above, we find

$$\begin{aligned}
&\|X\widetilde{A} - XA\|_F^2 \\
&\quad \le \|XB - XA\|_F^2 + 2\tau\{\|\widetilde{A}_1 - B\|_1 + \|\widetilde{A}_2\|_1 + \|B\|_1 - \|\widetilde{A}_1\|_1 - \|\widetilde{A}_2\|_1\} \\
&\qquad\qquad \text{by (i), (iii) and (v)} \\
&\quad \le \|XB - XA\|_F^2 + 4\tau\|\widetilde{A}_1 - B\|_1 \\
&\quad \le \|XB - XA\|_F^2 + 4\tau\sqrt{r(\widetilde{A}_1 - B)}\|\widetilde{A}_1 - B\|_F \qquad \text{by Cauchy–Schwarz} \\
&\quad \le \|XB - XA\|_F^2 + 4\tau\sqrt{3r(B)}\|\widetilde{A} - B\|_F \qquad \text{by (ii) and (iv).}
\end{aligned}$$

Using $\lambda_p(M)\|\widetilde{A} - B\|_F^2 \le \|X\widetilde{A} - XB\|_F^2$ and $2xy \le x^2/2 + 2y^2$, we obtain

$$\tfrac{1}{2}\|X\widetilde{A} - XA\|_F^2 \le \tfrac{3}{2}\|XB - XA\|_F^2 + 24\tau^2 r(B)/\lambda_p(M).$$

The proof is complete by choosing the truncated GSVD B' under metric M, see Lemma 14 below.

APPENDIX B: GENERALIZED SINGULAR VALUE DECOMPOSITION

We consider the functional

$$G(B) = \|XB_0 - XB\|_F^2 = \operatorname{tr}\big((B - B_0)'M(B - B_0)\big)$$

with $M = X'X = NN$ and B_0 is a fixed $p \times n$ matrix of rank r. By the Eckhart–Young theorem, we have the lower bound

$$G(B) \geq \sum_{j>k} d_j^2(XB_0)$$

for all $p \times n$ matrices B of rank k. We now show that this infimum is achieved by the generalized singular value decomposition (GSVD) under metric M, limited to its k largest generalized singular values. Following Takane and Hunter [(2001), pages 399–400], the GSVD of B_0 under metric M is UDV' where U is an $p \times r$ matrix, $U'MU = I_r$, V is an $n \times r$ matrix, $V'V = I_r$ and D is a diagonal $r \times r$ matrix, and $NB_0 = NUDV'$. It can be computed via the (regular) SVD $\bar{U}\bar{D}\bar{V}'$ of NB_0. It is easy to verify that the choices $U = N^-\bar{U}$, where N^- is the Moore–Penrose inverse of N, $D = \bar{D}$ and $V = \bar{V}$ satisfy the above conditions. This means in particular that the generalized singular values d_j are the regular singular values of NB_0. Let $B_k = U_k D_k V_k'$ by retaining as usual the first k columns of U and V.

LEMMA 14. *Let B_k be the GSVD of B_0 under metric M, restricted to the k largest generalized singular values. We have*

$$\|XB_0 - XB_k\|_F^2 = \sum_{j>k} d_j^2(XB_0).$$

PROOF. Since $NB_0 = NUDV'$ and $NB_k = NU_k D_k V_k'$, we obtain

$$\Delta = NB_0 - NB_k = N \sum_{j>k} u_j v_j' d_j = NU_{(k)} D_{(k)} V_{(k)}'$$

using the notation $U_{(k)}$ for the $p \times (r-k)$ matrix consisting of the last $r-k$ column vectors of U, $D_{(k)}$ is the diagonal $(r-k) \times (r-k)$ matrix based on the last $r-k$ singular values, and $V_{(k)}$ for the $n \times (r-k)$ matrix consisting of the last $r-k$ column vectors of V. Finally,

$$\begin{aligned}\|XB_0 - XB_k\|_F^2 &= \|\Delta\|_F^2 = \big\|NU_{(k)}D_{(k)}V_{(k)}'\big\|_F^2 \\ &= \operatorname{tr}\big(V_{(k)}D_{(k)}U_{(k)}'MU_{(k)}D_{(k)}V_{(k)}'\big) \\ &= \operatorname{tr}\big(V_{(k)}D_{(k)}I_{(k)}D_{(k)}V_{(k)}'\big) = \operatorname{tr}\big(D_{(k)}^2\big) = \sum_{j>k} d_j^2.\end{aligned}$$

Recall that in the construction of the GSVD, the generalized singular values d_j are the singular values of NB_0. Since

$$d_j^2(NB_0) = \lambda_j(B_0'MB_0) = \lambda_j(B_0'X'XB_0) = d_j^2(XB_0),$$

the claim follows. □

REMARK. The rank restricted estimator $\widehat{B}_k$ given in Section 2.1 is the GSVD of the least squares estimator $\widehat{B}$ under the metric $M = X'X$, see Takane and Hwang (2007).

APPENDIX C: LARGEST SINGULAR VALUES OF TRANSFORMATIONS OF SUB-GAUSSIAN MATRICES

We call a random variable W sub-Gaussian with sub-Gaussian moment Γ_W, if

$$\mathbb{E}[\exp(tW)] \le \exp(t^2/\Gamma_W)$$

for all $t > 0$. Markov's inequality implies that W has Gaussian type tails:

$$\mathbb{P}\{|W| > t\} \le 2\exp\{-t^2/(2\Gamma_W)\}$$

holds for any $t > 0$. Normal $N(0, \sigma^2)$ random variables are sub-Gaussian with $\Gamma_W = \sigma^2$. General results on the largest singular values of matrices E with sub-Gaussian entries can be found in the survey paper by Rudelson and Vershynin (2010). The analysis of our estimators require bounds for the largest singular values of PE and $X'E$, for which the standard results on E do not apply directly.

PROPOSITION 15. *Let E be a $m \times n$ matrix with independent sub-Gaussian entries E_{ij} with sub-Gaussian moment Γ_E. Let X be an $m \times p$ matrix of rank q and let $P = X(X'X)^- X'$ be the projection matrix on $R[X]$. Then, for each $x > 0$,*

$$\mathbb{P}\big\{d_1^2(PE) \ge 32\Gamma_E\big((n+q)\ln(5) + x\big)\big\} \le 2\exp(-x).$$

In particular,

$$\mathbb{E}[d_1(PE)] \le 15\Gamma_E\sqrt{n+q}.$$

PROOF. Let S^{n-1} be the unit sphere in $\mathbb{R}^n$. First, we note that

$$\|PE\|_2 = \sup_{u\in S^{p-1}, v\in S^{n-1}} \langle Pu, Ev\rangle = \sup_{u\in U, v\in S^{n-1}} \langle u, Ev\rangle$$

with $U = PS^{p-1} = \{u = Ps : s \in S^{p-1}\}$. Let $\mathcal{M}$ be a δ-net of U and $\mathcal{N}$ be a δ-net for S^{n-1} with $\delta = 1/2$. Since the dimension of U is q and $\|u\| \le 1$ for each $u \in U$, we need at most 5^q elements in $\mathcal{M}$ to cover U and 5^n elements to cover S^{n-1}, see Kolmogorov and Tihomirov (1961). A standard discretization trick, see, for instance, Rudelson and Vershynin [(2010), proof of Proposition 2.4], gives

$$\|PE\|_2 \le 4 \max_{u\in\mathcal{M}, v\in\mathcal{N}} \langle u, Ev\rangle.$$

Next, we write $\langle u, Ev\rangle = \sum_{i=1}^m u_i\langle E_i, v\rangle$ and note that each $\langle E_i, v\rangle$ is sub-Gaussian with moment Γ_E, as

$$\mathbb{E}[\exp(t\langle E_i, v\rangle)] = \prod_{j=1}^n \mathbb{E}[\exp(tv_jE_{ij})] \le \exp\Big(t^2\sum_j v_j^2/\Gamma_E\Big) = \exp(t^2/\Gamma_E).$$

It follows that each term in $\sum_{i=1}^m u_i \langle E_i, v\rangle$ is sub-Gaussian, and $\langle u, Ev\rangle$ is sub-Gaussian with sub-Gaussian moment $\Gamma_E \sum_{i=1}^m u_i^2 = \Gamma_E$. This implies the tail bound

$$\mathbb{P}\{|\langle u, Ev\rangle| > t\} \le 2\exp\{-t^2/(2\Gamma_E)\}$$

for each fixed u and v and all $t > 0$. Combining the previous two steps, we obtain

$$\mathbb{P}\{\|PE\|_2 \ge 4t\} \le 5^{n+q} 2\exp\{-t^2/(2\Gamma_E)\}$$

for all $t > 0$. Taking $t^2 = 2\{\ln(5)(n+q) + x\}\Gamma_E$, we obtain the first claim. The second claim follows from this tail bound. □

APPENDIX D: AUXILIARY RESULTS

LEMMA 16. *Let X be a nonnegative random variable with $\mathbb{E}[X] = \mu$ and $\mathbb{P}\{X - \mu \ge t\} \le \exp(-t^2/2)$ for all $t \ge 0$. Then we have*

$$\mathbb{E}[X^2] \le \mu^2 + \mu\sqrt{2\pi} + 2.$$

Moreover, for any $\xi > 0$, we have

$$\mathbb{E}\big[(X^2 - (1+\xi)^2\mu^2)_+\big] \le 2(1+\xi^{-1})\exp(-\xi^2\mu^2/2).$$

PROOF. The following string of inequalities are self-evident:

$$\begin{aligned}\mathbb{E}[X^2] &= \int_0^\infty \mathbb{P}\{X^2 \ge x\}\,\mathrm{d}x \le \mu^2 + \int_\mu^\infty 2x\mathbb{P}\{X \ge x\}\,\mathrm{d}x \\ &\le \mu^2 + \int_0^\infty 2(x+\mu)\exp\left(-\frac{1}{2}x^2\right)\mathrm{d}x = \mu^2 + \mu\sqrt{2\pi} + 2.\end{aligned}$$

This proves our first claim. The second claim is easily deduced as follows:

$$\begin{aligned}\mathbb{E}\big[(X^2 - (1+\xi)^2\mu^2)_+\big] &\le \mathbb{E}\big[X^2 1_{\{X \ge (1+\xi)\mu\}}\big] = \int_{(1+\xi)\mu}^\infty 2t\mathbb{P}\{X \ge t\}\,\mathrm{d}t \\ &\le (1+\xi^{-1})\int_{\xi\mu}^\infty 2t\exp(-t^2/2)\,\mathrm{d}t \\ &= 2(1+\xi^{-1})\exp(-\xi^2\mu^2/2).\end{aligned}$$

The proof of the lemma is complete. □

LEMMA 17. *Let Z_d be a χ_d^2 random variable with d degrees of freedom. Then*

$$\mathbb{P}\{Z_d - d \le -x\sqrt{2d}\} \le \exp\left(-\frac{x^2}{2 + 2x\sqrt{2/d}}\right).$$

In particular, for any $0 < t < 1$,

$$\mathbb{P}\{Z_d \le (1-t)d\} \le \exp\{-t^2 d/4(1+t)\}.$$

PROOF. See Cavalier et al. [(2002), page 857] for the first claim. The second claim follows by taking $x = t(d/2)^{1/2}$. □

Acknowledgments. We would like to thank Emmanuel Candès, Angelika Rohde and Sasha Tsybakov for stimulating conversations in Oberwolfach, Tallahassee and Paris, respectively. We also thank the Associate Editor and the referees for their constructive remarks.

REFERENCES

ANDERSON, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Statist.* **22** 327–351. MR0042664

ANDERSON, T. W. (1999). Asymptotic distribution of the reduced rank regression estimator under general conditions. *Ann. Statist.* **27** 1141–1154. MR1740118

ANDERSON, T. W. (2002). Specification and misspecification in reduced rank regression. *Sankhyā Ser. A* **64** 193–205. MR1981753

CANDÈS, E. J. and PLAN, Y. (2010). Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. Available at arxiv:1001.0339 [cs.IT].

CANDÈS, E. J. and TAO, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory* **56** 2053–2080. MR2723472

CAVALIER, L., GOLUBEV, G. K., PICARD, D. and TSYBAKOV, A. B. (2002). Oracle inequalities for inverse problems. *Ann. Statist.* **30** 843–874. MR1922543

FAZEL, M. (2002). Matrix rank minimization with applications. Ph.D. thesis, Stanford University.

HORN, R. A. and JOHNSON, C. R. (1985). *Matrix Analysis*. Cambridge Univ. Press, Cambridge. MR0832183

IZENMAN, A. J. (1975). Reduced-rank regression for the multivariate linear model. *J. Multivariate Anal.* **5** 248–264. MR0373179

IZENMAN, A. J. (2008). *Modern Multivariate Statistical Techniques*: *Regression*, *Classification*, *and Manifold Learning*. Springer, New York. MR2445017

KOLMOGOROV, A. N. and TIHOMIROV, V. M. (1961). ε-entropy and ε-capacity of sets in functional spaces. *Amer. Math. Soc. Transl.* (2) **17** 277–364. MR0124720

LU, Z., MONTEIRO, R. and YUAN, M. (2010). Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression. *Math. Program.* To appear.

MA, S., GOLDFARB, D. and CHEN, L. (2009). Fixed point and Bregman iterative methods for matrix rank minimization. Available at arxiv:0905.1643 [math.OC].

NEGAHBAN, S. and WAINWRIGHT, M. J. (2009). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. Available at arxiv:0912.5100v1 [math.ST].

RAO, C. R. (1980). Matrix approximations and reduction of dimensionality in multivariate statistical analysis. In *Multivariate Analysis*, *V* (*Proc. Fifth Internat. Sympos., Univ. Pittsburgh*, *Pittsburgh*, *PA*, 1978) 3–22. North-Holland, Amsterdam. MR0566327

RECHT, B., FAZEL, M. and PARRILO, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.* **52** 471–501. MR2680543

REINSEL, G. C. and VELU, R. P. (1998). *Multivariate Reduced-Rank Regression*: *Theory and Applications*. *Lecture Notes in Statist.* **136**. Springer, New York. MR1719704

ROBINSON, P. M. (1973). Generalized canonical analysis for time series. *J. Multivariate Anal.* **3** 141–160. MR0326959

ROBINSON, P. M. (1974). Identification, estimation and large-sample theory for regressions containing unobservable variables. *Internat. Econom. Rev.* **15** 680–692. MR0356376

ROHDE, A. and TSYBAKOV, A. B. (2010). Estimation of high-dimensional low-rank matrices. Available at arxiv:0912.5338v2 [math.ST].

RUDELSON, M. and VERSHYNIN, R. (2010). Non-asymptotic theory of random matrices: Extreme singular values. In *Proceedings of the International Congress of Mathematicians*. Hyderabad, India. To appear.

TAKANE, Y. and HUNTER, M. A. (2001). Constrained principal component analysis: A comprehensive theory. *Appl. Algebra Engrg. Comm. Comput.* **12** 391–419. MR1864610

TAKANE, Y. and HWANG, H. (2007). Regularized linear and kernel redundancy analysis. *Comput. Statist. Data Anal.* **52** 394–405. MR2409991

TOH, K. C., TODD, M. J. and TÜTÜNCÜ, R. H. (1999). SDPT3—a MATLAB software package for semidefinite programming, version 1.3. *Optim. Methods Softw.* **11/12** 545–581. MR1778429

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York. MR1385671

YUAN, M., EKICI, A., LU, Z. and MONTEIRO, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 329–346. MR2323756

DEPARTMENT OF STATISTICS
FLORIDA STATE UNIVERSITY
TALLAHASSEE, FLORIDA 32306-4330
USA
E-MAIL: flori@stat.fsu.edu
yshe@stat.fsu.edu
wegkamp@stat.fsu.edu