



HAL
open science

Low rank Multivariate regression

Christophe Giraud

► **To cite this version:**

| Christophe Giraud. Low rank Multivariate regression. 2010. hal-00521219v1

HAL Id: hal-00521219

<https://hal.science/hal-00521219v1>

Preprint submitted on 26 Sep 2010 (v1), last revised 21 Jun 2011 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

LOW RANK MULTIVARIATE REGRESSION

CHRISTOPHE GIRAUD

ABSTRACT. We consider in this paper the multivariate regression problem, when the target regression matrix A is close to a low rank matrix. Our primary interest is on the practical case where the variance of the noise is unknown. Our main contribution is to propose in this setting a criterion to select among a family of low rank estimators and prove a non-asymptotic oracle inequality for the resulting estimator. We also investigate the easier case where the variance of the noise is known and outline that the penalties appearing in our criteria are minimal (in some sense). These penalties involve the expected value of the Ky-Fan quasi-norm of some random matrices. These quantities can be evaluated easily in practice and upper-bounds can be derived from recent results in random matrix theory.

1. INTRODUCTION

We build on ideas introduced in a paper of Bunea, She and Wegkamp [7] for the multivariate regression problem

$$(1) \quad Y = XA + \sigma E$$

where Y is a $m \times n$ matrix of response variables, X is a $m \times p$ matrix of predictors, A is a $p \times n$ matrix of regression coefficients and E is a $m \times n$ random matrix with i.i.d. entries. We assume for simplicity that the entries $E_{i,j}$ are standard Gaussian, yet all the results can be extended to the case where the entries are sub-Gaussian.

An important issue in multivariate regression is to estimate A or XA when the matrix A has a low rank or can be well approximated by a low rank matrix, see Izenman [12]. In this case, a small number of linear combinations of the predictors catch most of the non-random variation of the response Y . This framework arises in many applications, among which analysis of fMRI image data [9], analysis of EEG data decoding [2], neural response modeling [6] or genomic data analysis [7].

When the variance σ^2 is known, the strategy developed by Bunea *et al.* [7] for estimating A or XA is the following. Writing $\|\cdot\|$ for the Frobenius norm and \hat{A}_r for the minimizer of $\|Y - X\hat{A}\|$ over the matrices \hat{A} of rank at most r , the matrix XA is estimated by $X\hat{A}_{\hat{r}}$, where \hat{r} minimizes the criterion

$$(2) \quad \text{Crit}_{\sigma^2}(r) = \|Y - X\hat{A}_r\|^2 + \text{pen}(r)\sigma^2.$$

Date: 23 septembre 2010.

Bunea *et al.* [7] consider a penalty $\text{pen}(r)$ linear in r et provide clean non-asymptotic bounds on $\|X\hat{A}_{\hat{r}} - XA\|^2$, on $\|\hat{A}_{\hat{r}} - A\|^2$ and on the probability that the estimated rank \hat{r} coincides with the rank of A .

Our main contributions are first to exhibit a minimal sublinear penalty for the Criterion (2) and second to propose and analyze a criterion to handle the case where σ^2 is unknown. Let us denote by q the rank of X and by $E_{q \times n}$ a $q \times n$ random matrix with i.i.d. standard Gaussian entries. The penalties we introduce involve the expected value of the Ky-Fan quasi-norm of the random matrix $E_{q \times n}$, namely

$$S_r(E_{q \times n}) = \sqrt{\sum_{k=1}^r \sigma_k^2(E_{q \times n})}$$

where $\sigma_k(E_{q \times n})$ stands for the k -th largest singular value of $E_{q \times n}$. More precisely, in the case where σ^2 is known, we show that the penalty $\text{pen}(r) = \mathbb{E}[S_r(E_{q \times n})]^2$ is minimal (in some sense) for the Criterion (2). For the case of unknown variance, we prove a non-asymptotic oracle-like inequality for the criterion

$$(3) \quad \text{Crit}(r) = \log(\|Y - X\hat{A}_r\|^2) + \text{pen}(r).$$

when

$$\text{pen}(r) \geq -\log\left(1 - K \frac{\mathbb{E}(S_r(E_{q \times n}))^2}{nm - 1}\right), \quad \text{with } K > 1.$$

The study of multivariate regression with rank constraints dates back to Anderson [1] and Izenman [11]. The question of rank selection has only been recently addressed by Anderson [1] in an asymptotic setting (with p fixed) and by Bunea *et al.* [7] in a non-asymptotic framework. We refer to the latter article for additional references. In parallel, a series of recent papers study the estimator $\hat{A}_\lambda^{\ell^1}$ obtained by minimizing

$$\|Y - X\hat{A}\|^2 + \lambda \sum_k \sigma_k(\hat{A})$$

see among others Yuan *et al.* [18], Bach [3], Neghaban and Wainwright [15], Lu *et al.* [14] and Rohde and Tsybakov [16]. Due to the " ℓ^1 " penalty $\sum_k \sigma_k(\hat{A})$, the estimator $\hat{A}_\lambda^{\ell^1}$ has a small rank for λ large enough and it is proven to have good statistical properties under some hypotheses on the design matrix X . We refer to Bunea *et al.* [7] for a detailed analysis of the similarities and the differences between $\hat{A}_\lambda^{\ell^1}$ and their estimator.

Our paper is organized as follows. In the next section, we give a few results on $\mathbb{E}(S_r(E_{q \times n}))^2$ and on the estimator $X\hat{A}_r$. In Section 3, we analyze the simple case where the variance σ^2 is known, which gives us some insights for the Section 4 where the case of unknown variance is tackled. In Section 5, we comment on the extension of the results to the case of sub-Gaussian errors and we outline that our theory provides a theoretically grounded criterion (in a non-asymptotic framework) to select the number r of components to be kept in a principal component analysis. Finally, we carry out an empirical study in Section 6.

2. A FEW FACTS ON $\mathbb{E}(S_r(E_{q \times n}))^2$ AND $X\hat{A}_r$

2.1. **Bounds on $\mathbb{E}(S_r(E_{q \times n}))^2$.** The expectation $\mathbb{E}(S_r(E_{q \times n}))$ can be evaluated numerically with a few lines of R-code. From a more theoretical point of view, we have the following bounds.

Lemma 1. *Assume that $q \leq n$. Then for any $r \leq q$, we have $\mathbb{E}(S_r(E_{q \times n}))^2 \geq r(n - 1/q)$ and*

$$\mathbb{E}(S_r(E_{q \times n}))^2 \leq \min \left\{ r(\sqrt{n} + \sqrt{q})^2, nq - \sum_{k=r+1}^q (\sqrt{n} - \sqrt{k})^2, r + \sum_{k=1}^r (\sqrt{n} + \sqrt{q - k + 1})^2 \right\}.$$

When $q > n$ the same result holds with q and n switched. In particular, for $r = \min(n, q)$, we have

$$qn - 1 \leq \mathbb{E}(S_{\min(n,q)}(E_{q \times n}))^2 = \mathbb{E}(\|E_{q \times n}\|)^2 \leq qn.$$

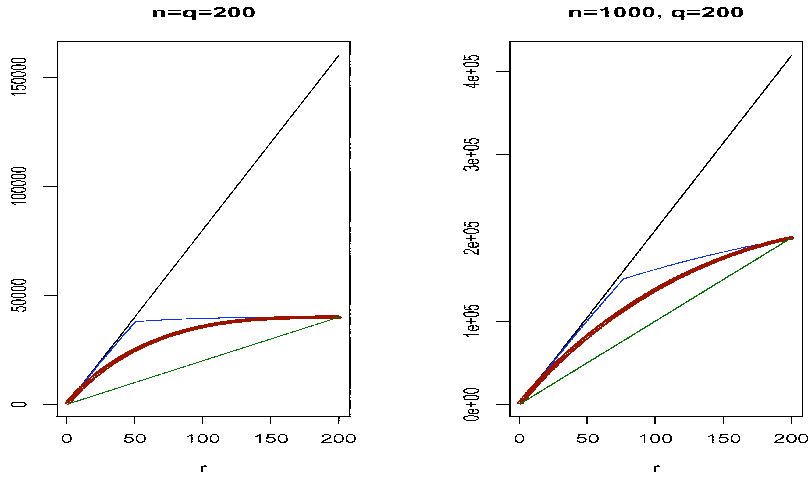


FIGURE 1. In red $r \rightarrow \mathbb{E}(S_r(E_{q \times n}))^2$, in black $r \rightarrow r(\sqrt{n} + \sqrt{q})^2$, in blue the upper-bound of Lemma 1, in green the lower bound. Left: $q = n = 200$. Right: $q = 200$ and $n = 1000$.

Proof. For notational simplicity we write $E' = E_{q \times n}$. The case $r = 1$ follows from Slepian's Lemma, see Davidson and Szarek [8] Chapter 8. For $r > 1$, we note that

$$\mathbb{E}(S_r(E'))^2 \leq \min \left\{ r \mathbb{E}(S_1(E'))^2, \sum_{k=1}^r \mathbb{E}(\sigma_k^2(E')) \right\}.$$

The first upper bound $\mathbb{E}(S_r(E_{q \times n}))^2 \leq r(\sqrt{n} + \sqrt{q})^2$ follows. For the second upper bound, we note that

$$\sum_{k=1}^r \mathbb{E}(\sigma_k^2(E')) \leq \mathbb{E}(\|E'\|^2) - \sum_{k=r+1}^q \mathbb{E}(\sigma_k(E'))^2.$$

The interlacing inequalities [10] ensure that $\sigma_k(E'_k) \leq \sigma_k(E')$ where E'_k is the matrix made of the k first rows of E' . The bound then follows from $\mathbb{E}(\sigma_k(E'_k)) \geq \sqrt{n} - \sqrt{k}$, see [8].

Let us turn to the third bound. The map $E' \rightarrow \sigma_k(E')$ is 1-Lipschitz so, writing M_k for the median of σ_k , the concentration inequality for Gaussian random variables ensures that $(M_k - \sigma_k(E'))_+ \leq \xi_+$ and $(\sigma_k(E') - M_k)_+ \leq \xi'_+$ where ξ_+ and ξ'_+ are the positive part of two standard Gaussian random variables. As a consequence we have

$$\begin{aligned} \mathbb{E}(\sigma_k^2(E')) - \mathbb{E}(\sigma_k(E'))^2 + (M_k - \mathbb{E}(\sigma_k(E')))^2 &= \mathbb{E}((\sigma_k(E') - M_k)_+^2) + \mathbb{E}((M_k - \sigma_k(E'))_+^2) \\ &\leq \mathbb{E}(\xi_+^2) + \mathbb{E}(\xi'_+^2) = 1, \end{aligned}$$

and thus $\mathbb{E}(\sigma_k^2(E')) \leq \mathbb{E}(\sigma_k(E'))^2 + 1$.

Furthermore, the interlacing inequalities [10] ensure that $\sigma_k(E') \leq \sigma_1(E'_{q-k+1})$. We can then bound $\mathbb{E}(\sigma_k(E'))$ by

$$\mathbb{E}(\sigma_k(E')) \leq \sqrt{n} + \sqrt{q - k + 1}$$

which leads to the last upper bound.

For the lower bound, we start from $S_k(E')^2 \geq \|E'\|^2 k/q$ (sum of a decreasing sequence) and use again the Gaussian concentration inequality to get

$$\mathbb{E}(\|E'\|^2) - 1 = nq - 1 \leq \mathbb{E}(\|E'\|)^2$$

and concludes that $k(nq - 1)/q \leq \mathbb{E}(S_k(E'))^2$. \square

Finally, we mention that for large values of q and n asymptotics are provided by the Marcenko-Pastur law. Actually, assume that q, n and r go to infinity with $q/n \rightarrow \beta < 1$ and $r/q \rightarrow \alpha < 1$. Then, we have

$$\mathbb{E}[S_r(E_{q \times n})]^2 \sim nq \int_{x_\alpha}^{(1+\sqrt{\beta})^2} x f_\beta(x) dx$$

where

$$f_\beta(x) = (2\pi\beta x)^{-1} \sqrt{(x - (1 - \sqrt{\beta})^2)((1 + \sqrt{\beta})^2 - x)} \mathbf{1}_{[(1-\sqrt{\beta})^2, (1+\sqrt{\beta})^2]}$$

and x_α fulfills

$$\int_{x_\alpha}^{(1+\sqrt{\beta})^2} x f_\beta(x) dx = \alpha.$$

2.2. Computation of $X\hat{A}_r$. Next lemma provides a useful formula for $X\hat{A}_r$.

Lemma 2. Write P for the projection matrix $P = X(X^*X)^+X^*$, with $(X^*X)^+$ the Moore-Penrose pseudo-inverse of X^*X . Then, for any $r \leq q$ we have $X\hat{A}_r = (PY)_r$ where $(PY)_r$ minimizes $\|PY - B\|^2$ over the matrices B of rank at most r .

As a consequence, writing $PY = U\Sigma V^*$ for the singular value decomposition of PY , the matrix $X\hat{A}_r$ is given by $X\hat{A}_r = U\Sigma_r V^*$, where Σ_r is obtained from Σ by setting $(\Sigma_r)_{i,i} = 0$ for $i \geq r + 1$.

Proof of Lemma 2. We note that that $\|PY - P(PY)_r\|^2 \leq \|PY - (PY)_r\|^2$ and $\text{rank}(P(PY)_r) \leq r$, so $P(PY)_r = (PY)_r$. In particular, we have $(PY)_r = X\tilde{A}_r$, with $\tilde{A}_r = (X^*X)^+X^*(PY)_r$. Since the rank of $X\hat{A}_r$ is also at most r , we have

$$\begin{aligned} \|Y - X\tilde{A}_r\|^2 &= \|Y - PY\|^2 + \|PY - (PY)_r\|^2 \\ &\leq \|Y - PY\|^2 + \|PY - X\hat{A}_r\|^2 = \|Y - X\hat{A}_r\|^2. \end{aligned}$$

Since the rank of \tilde{A}_r is not larger than r , we then have $\tilde{A}_r = \hat{A}_r$.

3. THE CASE OF KNOWN VARIANCE

In this section we revisit the results of Bunea *et al.* [7] for the case where σ^2 is known. This analysis will give us some insights for the case of unknown variance. Next theorem states an oracle inequality for the selection Criterion (2) with penalty fulfilling $\text{pen}(r) \geq K\mathbb{E}(S_r(E_{q \times n}))^2$ for $K > 1$. Later on, we will prove that the penalty $\text{pen}(r) = \mathbb{E}(S_r(E_{q \times n}))^2$ is minimal in some sense.

Theorem 1. Assume that for some $K > 1$ we have

$$(4) \quad \text{pen}(r) \geq K\mathbb{E}(S_r(E_{q \times n}))^2 \quad \text{for all } r \leq \min(n, q).$$

Then, when \hat{r} is selected by minimizing (2) the estimator $\hat{A} = \hat{A}_{\hat{r}}$ satisfies

$$(5) \quad \mathbb{E}(\|X\hat{A} - XA\|^2) \leq c(K) \min_r \left\{ \mathbb{E}(\|XA - X\hat{A}_r\|^2) + \text{pen}(r)\sigma^2 + \sigma^2 \right\}$$

for some positive constant $c(K)$ depending on K only.

The risk bound (5) ensures that the risk of the estimator \hat{A} is not larger (up to a constant) than the minimum over r of the sum of the risk of the estimator \hat{A}_r plus the penalty term $\text{pen}(r)\sigma^2$. This risk bound bears some similitude with the bound (4) of Theorem 6 in Bunea *et al.* [7]. Let us comment on the two main differences between the above theorem and Theorem 6 in [7]. First, in Theorem 6 of [7], the condition on the penalty is $\text{pen}(r) \geq 32(n+q)r$ for the Gaussian case. According to Lemma 1, our Condition (4) is weaker and more importantly Proposition 1 below proves that it is minimal in some sense. The second difference lies in the nature of the bound (5). Theorem 6 in [7] provides a clean bound which holds with high probability, but the size of the latter probability is

directly tuned by parameter μ of the penalty $\text{pen}(r) = \mu(n+q)r$. As a consequence, the bound in Theorem 6 of [7] cannot be integrated to obtain a risk bound as (5). Theorem 1 is then an improvement of Theorem 6 of Bunea *et al* [7], and its proof is interesting since it paves the way for the unknown variance case.

Proof of Theorem 1. The inequality $\text{Crit}_{\sigma^2}(\hat{r}) \leq \text{Crit}_{\sigma^2}(r)$ gives

$$(6) \quad \|X\hat{A} - XA\|^2 \leq \|X\hat{A}_r - XA\|^2 + \text{pen}(r)\sigma^2 + 2\sigma \langle E, X\hat{A} - X\hat{A}_r \rangle - \text{pen}(\hat{r})\sigma^2.$$

Next lemma provides an upper bound for the scalar product.

Lemma 3. *Fix $r \leq \min(n, q)$ and $\eta > 0$. Then, writing A_r for the best approximation of A with rank at most r , we have for any $k \leq \min(n, q)$*

$$(7) \quad 2\sigma |\langle E, X\hat{A}_k - XA_r \rangle| \leq \frac{1}{1+\eta} \|X\hat{A}_k - XA\|^2 + \frac{1+1/\eta}{(1+\eta)^2} \|XA - XA_r\|^2 \\ + (1+\eta)^2 (1+1/\eta)\sigma^2 U_r + (1+\eta)^3 \sigma^2 S_k(PE)^2$$

where $P = X(X^*X)^+X^*$ as in Lemma 1 and U_r is a random variable such that $\mathbb{E}(U_r) \leq r \min(n, q)$.

Iterating twice the inequality $ab \leq a^2/c + cb^2$ gives

$$2\sigma |\langle E, X\hat{A}_k - XA_r \rangle| \\ \leq \frac{1}{1+\eta} \|X\hat{A}_k - XA\|^2 + \frac{1+1/\eta}{(1+\eta)^2} \|XA - XA_r\|^2 + (1+\eta)^2 \sigma^2 \frac{\langle E, X\hat{A}_k - XA_r \rangle^2}{\|X\hat{A}_k - XA_r\|^2}.$$

We write $XA_r = U\Gamma_r V^*$ for the singular value decomposition of XA_r , with the convention that the diagonal entries of Γ_r are decreasing. Since the rank of XA_r is upper bounded by the rank of A_r , the $m \times n$ diagonal matrix Γ_r has at most r non zeros elements. Assume first that $n \leq q$. Denoting by I_r the $m \times m$ diagonal matrix with $(I_r)_{i,i} = 1$ if $i \leq r$ and $(I_r)_{i,i} = 0$ if $i > r$ and writing $I_{-r} = I - I_r$ and $\hat{B}_k = U^*X\hat{A}_kV$, we have

$$\frac{\langle E, X\hat{A}_k - XA_r \rangle^2}{\|X\hat{A}_k - XA_r\|^2} = \frac{\langle U^*PEV, \hat{B}_k - \Gamma_r \rangle^2}{\|\hat{B}_k - \Gamma_r\|^2} \\ = \frac{\left(\langle U^*PEV, I_r(\hat{B}_k - \Gamma_r) \rangle + \langle U^*PEV, I_{-r}\hat{B}_k \rangle \right)^2}{\|I_r(\hat{B}_k - \Gamma_r)\|^2 + \|I_{-r}\hat{B}_k\|^2} \\ \leq (1+\eta^{-1}) \frac{\langle U^*PEV, I_r(\hat{B}_k - \Gamma_r) \rangle^2}{\|I_r(\hat{B}_k - \Gamma_r)\|^2} + (1+\eta) \frac{\langle U^*PEV, I_{-r}\hat{B}_k \rangle^2}{\|I_{-r}\hat{B}_k\|^2}.$$

The first term is upper bounded by

$$\frac{\langle U^*PEV, I_r(\hat{B}_k - \Gamma_r) \rangle^2}{\|I_r(\hat{B}_k - \Gamma_r)\|^2} \leq \|I_r U^*PEV\|^2 = U_r$$

and the expected value of the right-hand side fulfills

$$\mathbb{E}(U_r) = n\|I_r U^* P U\|^2 = n\|U^* P U I_r\|^2 \leq nr.$$

Since the rank of $I_{-r} \widehat{B}_k$ is at most k , the second term can be bounded by

$$\frac{\langle U^* P E V, I_{-r} \widehat{B}_k \rangle^2}{\|I_{-r} \widehat{B}_k\|^2} \leq \sup_{\text{rank}(B) \leq k} \frac{\langle U^* P E V, B \rangle^2}{\|B\|^2} = S_k(U^* P E V)^2 = S_k(P E)^2.$$

Putting pieces together gives (7) for $n \leq q$. The case $n > q$ can be treated in the same way, starting from

$$\widehat{B}_k - \Gamma_r = (\widehat{B}_k - \Gamma_r) I_r + \widehat{B}_k I_{-r}$$

with I_r and I_{-r} two $n \times n$ diagonal matrices defined as above. \square

Combining the inequalities (6) and (7) with $\eta = ((1 + K)/2)^{1/3} - 1 > 0$, we obtain

$$\begin{aligned} \frac{\eta}{1 + \eta} \|X \widehat{A} - X A\|^2 &\leq \|X \widehat{A}_r - X A\|^2 + 2 \frac{1 + 1/\eta}{(1 + \eta)^2} \|X A - X A_r\|^2 + 2 \text{pen}(r) \sigma^2 \\ &\quad + 2(1 + \eta)^2 (1 + \eta^{-1}) \sigma^2 U_r + \frac{K + 1}{2} \sigma^2 S_r(P E)^2 - \text{pen}(r) \sigma^2 \\ &\quad + \sigma^2 \sum_{k=1}^{\min(n, q)} \left(\frac{K + 1}{2} S_k(P E)^2 - \text{pen}(k) \right)_+. \end{aligned}$$

The map $E \rightarrow S_k(P E)$ is 1-Lipschitz and convex, so there exist a standard Gaussian random variable ξ such that $S_k(P E) \leq \mathbb{E}(S_k(P E)) + \xi_+$ and then

$$\begin{aligned} \mathbb{E} \left(\frac{K + 1}{2} S_k(P E)^2 - \text{pen}(k) \right)_+ &\leq \frac{1 + K}{2} \mathbb{E} \left(\xi_+^2 + 2\xi_+ \mathbb{E}(S_k(P E)) - \frac{K - 1}{K + 1} \mathbb{E}(S_k(P E))^2 \right)_+ \\ &\leq c_1(K) \exp(-c_2(K) \mathbb{E}(S_k(P E))^2). \end{aligned}$$

Since $S_k(P E)$ is distributed as $S_k(E_{q \times n})$, Lemma 1 gives that $\mathbb{E}(S_k(P E))^2 \geq k \max(n, q) - 1$ and the series

$$\sum_{k=1}^{\min(n, q)} \mathbb{E} \left(\frac{K + 1}{2} S_k(P E)^2 - \text{pen}(k) \right)_+$$

can be upper-bounded by $c_1(K) e^{c_2(K)} (1 - e^{-c_2(K)})^{-1} e^{-c_2(K) \max(n, q)}$. Finally, $\mathbb{E}(U_r) \leq r \min(n, q)$ is bounded by $1 + \text{pen}(r)$ and $\|X A - X A_r\|^2$ is smaller than $\mathbb{E}(\|X A - X \widehat{A}_r\|^2)$, so there exists some constant $c(K) > 0$ such that (11) holds. \square

Next proposition shows that choosing a penalty $\text{pen}(r) = K \mathbb{E}(S_r(E_{q \times n}))^2$ with $K < 1$ can lead to a strong overfitting.

Proposition 1. *Assume that $A = 0$ and that \hat{r} is any minimizer of the Criterion (2) with $\text{pen}(r) = K\mathbb{E}(S_r(E_{q \times n}))^2$ for some $K < 1$. Then, setting $\alpha = 1 - \sqrt{(1+K)/2} > 0$ we have*

$$\mathbb{P}\left(\hat{r} \geq \frac{1-K}{4} \times \frac{nq-1}{(\sqrt{n}+\sqrt{q})^2}\right) \geq 1 - e^{\alpha^2/2} \frac{e^{-\alpha^2 \max(n,q)/2}}{1 - e^{-\alpha^2 \max(n,q)/2}}.$$

A direct consequence of Proposition 1 is that the risk bound (5) cannot hold when Condition (4) is replaced by $\text{pen}(r) = K\mathbb{E}(S_r(E_{q \times n}))^2$ with $K < 1$. In this sense, Condition (4) is minimal.

Proof Proposition 1. For simplicity we consider first the case where $m = q$. We set

$$\Omega_0 = \{\|E\| \geq (1-\alpha)\mathbb{E}(\|E\|)\} \bigcap_{r=1}^{\min(n,m)} \{S_r(E) \leq (1+\alpha)\mathbb{E}(S_r(E))\}.$$

According to the Gaussian concentration inequality we have

$$\begin{aligned} \mathbb{P}(\Omega_0) &\geq 1 - \sum_{r=1}^{\min(n,m)} e^{-\alpha^2 \mathbb{E}(S_r(E))^2/2} \\ &\geq 1 - e^{\alpha^2/2} \sum_{r=1}^{\min(n,m)} e^{-\alpha^2 r \max(n,m)/2} \end{aligned}$$

where the last bound follows from Lemma 1. Furthermore, Lemma 2 gives that $X\hat{A}_r = Y_r (= E_r)$, where Y_r is the matrix M minimizing $\|Y - M\|^2$ over the matrices of rank at most r . As a consequence, writing $m^* = \min(n, m)$, we have on Ω_0

$$\begin{aligned} \text{Crit}_{\sigma^2}(m^*) - \text{Crit}_{\sigma^2}(r) &= K\mathbb{E}(S_{m^*}(E))^2 - (S_{m^*}(E)^2 - S_r(E)^2) - K\mathbb{E}(S_r(E))^2 \\ &\leq ((1+\alpha)^2 - K)\mathbb{E}(S_r(E))^2 - ((1-\alpha)^2 - K)\mathbb{E}(S_{m^*}(E))^2 \\ &\leq 2\mathbb{E}(S_r(E))^2 - \frac{1-K}{2}\mathbb{E}(S_{m^*}(E))^2 \\ &< 2r(\sqrt{n} + \sqrt{m})^2 - \frac{1-K}{2}(nm - 1). \end{aligned}$$

We then conclude that on Ω_0 we have $\hat{r} \geq \frac{1-K}{4} \times \frac{nm-1}{(\sqrt{n}+\sqrt{m})^2}$.

When $q < m$, we start from $\|Y - X\hat{A}_r\|^2 = \|Y - PY\|^2 + \|PY - X\hat{A}_r\|^2$ with $P = X(X^*X)^+X^*$ and follow the same lines, replacing everywhere E by PE and m by q . \square

4. THE CASE OF UNKNOWN VARIANCE

4.1. Main results. We consider now the case where the variance σ^2 is unknown. For a given $r_{\max} \leq \min(n, q)$, we propose to select $\hat{r} \in \{1, \dots, r_{\max}\}$ by minimizing over

$\{1, \dots, r_{\max}\}$ the Criterion (3), namely

$$\text{Crit}(r) = \log(\|Y - X\hat{A}_r\|^2) + \text{pen}(r).$$

We note that the Criterion (3) is equivalent to the criterion

$$(8) \quad \text{Crit}'(r) = \|Y - X\hat{A}_r\|^2 \left(1 + \frac{\text{pen}'(r)}{nm}\right),$$

with $\text{pen}'(r) = nm(e^{\text{pen}(r)} - 1)$. This last criterion bears some similitude with the Criterion (2). Indeed, the Criterion (8) can be written as

$$\|Y - X\hat{A}_r\|^2 + \text{pen}'(r)\hat{\sigma}_r^2,$$

with $\hat{\sigma}_r^2 = \|Y - X\hat{A}_r\|^2/(nm)$, which is the maximum likelihood estimator of σ^2 associated to \hat{A}_r . To facilitate comparisons with the case of known variance, we will work henceforth with the Criterion (8). Next theorem provides an upper bound for the risk of the estimator $X\hat{A}_{\hat{r}}$.

Theorem 2. *Assume that for some $K > 1$ we have both*

$$(9) \quad K\mathbb{E}(S_{r_{\max}}(E_{q \times n}))^2 + 1 < nm$$

$$(10) \quad \text{and } \text{pen}'(r) \geq \frac{K\mathbb{E}(S_r(E_{q \times n}))^2}{1 - \frac{1}{nm}(1 + K\mathbb{E}(S_r(E_{q \times n}))^2)}, \text{ for } r \leq r_{\max}.$$

Then, when \hat{r} is selected by minimizing (8) over $\{1, \dots, r_{\max}\}$, the estimator $\hat{A} = \hat{A}_{\hat{r}}$ satisfies

$$(11) \quad \mathbb{E}(\|X\hat{A} - XA\|^2) \leq c(K) \min_{r \leq r_{\max}} \left\{ \mathbb{E}(\|X\hat{A}_r - XA\|^2) \left(1 + \frac{\text{pen}'(r)}{nm}\right) + (\text{pen}'(r) + 1)\sigma^2 \right\}.$$

for some constant $c(K) > 0$ depending only on K .

Theorem 2 is proved in the next subsection. Let us compare Theorem 2 with Theorem 1. The two main differences lie in Condition (10) and in the form of the risk bound (11). Condition (10) is more stringent than Condition (4). More precisely, when r is small compared to q and n , both conditions are close, but when r is of a size comparable to q or n , Condition (10) is much stronger than (4). In the case where $m = q$, it even enforces a blow up of the penalty $\text{pen}'(r)$ when r tends to $\min(n, m)$. This blow up is actually necessary to avoid overfitting, since in this case the residual sum of squares $\|Y - X\hat{A}_r\|^2$ tends to 0 when r increases. The second major difference between Theorem 2 and Theorem 1 lies in the multiplicative factor $(1 + \text{pen}'(r)/nm)$ in the right-hand side of the risk bound (11). We mention that such a term also appears in the univariate regression framework with unknown variance, see Theorem 2 in Baraud *et al.* [4]. Due to this term, the bound (11) is not (strictly speaking) an oracle bound. To obtain an oracle

bound, we have to add the condition that there exists some positive constant C such that $\text{pen}'(r) \leq Cnm$ for all $r \leq r_{\max}$. With such a condition, the bound (11) becomes an oracle inequality which ensures that the risk of \hat{A} can be upper bounded (up to a constant) by the minimum over r of the sum of the risk of \hat{A}_r plus the penalty term $\text{pen}'(r)\sigma^2$.

Let us discuss in more details the Conditions (9) and (10) and the condition $\text{pen}'(r) \leq Cnm$. We have $\mathbb{E}(S_r(E_{q \times n}))^2 < r(\sqrt{n} + \sqrt{q})^2$ so the Conditions (9) and (10) are satisfied as soon as

$$r_{\max} \leq \frac{nm - 1}{K(\sqrt{n} + \sqrt{q})^2} \quad \text{and} \quad \text{pen}'(r) \geq \frac{Kr(\sqrt{n} + \sqrt{q})^2}{1 - \frac{1}{nm}(1 + Kr(\sqrt{n} + \sqrt{q})^2)}, \quad \text{for } r \leq r_{\max}.$$

Furthermore, if

$$r_{\max} \leq \alpha \frac{nm - 1}{K(\sqrt{q} + \sqrt{n})^2} \quad \text{for some } 0 < \alpha < 1,$$

then the penalty $\text{pen}'(r)$ chosen by taking equality in (10) fulfills $\text{pen}'(r) \leq C_\alpha nm$ with $C_\alpha = \alpha(1 - \alpha)^{-1}$. In terms of the Criterion (3), the Condition (10) reads

$$\text{pen}(r) \geq -\log(1 - K\mathbb{E}(S_r(E_{q \times n}))^2/(nm - 1)).$$

When $\text{pen}(r)$ is defined by taking equality in the above inequality, we have $\text{pen}(r) \approx Kr(\sqrt{n} + \sqrt{q})^2/(nm)$ for small values of r , see Figure 2. Finally, next proposition, shows

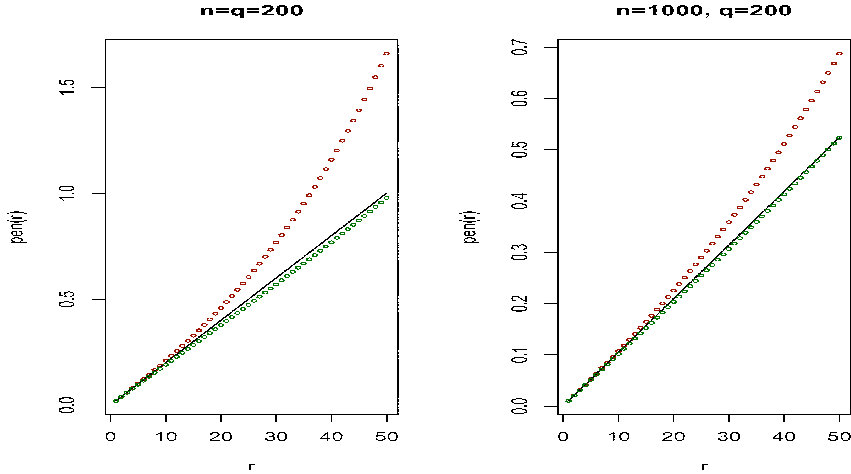


FIGURE 2. In green $\text{pen}(r) = -\log(1 - \mathbb{E}(S_r(E_{q \times n}))^2/(nq - 1))$, in black $\text{pen}(r) = r(\sqrt{n} + \sqrt{q})^2/(nq)$, in red $\text{pen}'(r)/(nq) = \mathbb{E}(S_r(E_{q \times n}))^2/(nq - 1 - \mathbb{E}(S_r(E_{q \times n}))^2)$. Left: $q = n = 200$. Right: $q = 200$ and $n = 1000$.

that the Condition (10) on $\text{pen}'(r)$ is necessary to avoid overfitting.

Proposition 2. Assume that $A = 0$ and that \hat{r} is any minimizer of Criterion (8) over $\{1, \dots, \min(n, q) - 1\}$ with

$$(12) \quad \text{pen}'(r) = \frac{K\mathbb{E}(S_r(E_{q \times n}))^2}{1 - \frac{K}{nm}\mathbb{E}(S_r(E_{q \times n}))^2} \quad \text{for some } K < 1.$$

Then, setting $\alpha = (1 - K)/4 > 0$ we have

$$\mathbb{P}\left(\hat{r} \geq \frac{1 - K}{8} \times \frac{nq - 1}{(\sqrt{n} + \sqrt{q})^2}\right) \geq 1 - 2e^{\alpha^2/2} \frac{e^{-\alpha^2 \max(n, q)/2}}{1 - e^{-\alpha^2 \max(n, q)/2}}.$$

The proof of Proposition 2 is postponed to the end of the section. A direct consequence of this proposition is that Theorem 2 cannot hold with Condition (10) replaced by (12). In this sense, Condition (10) is minimal.

4.2. Proof of Theorem 2. To simplify the formulae, we will note $\overline{\text{pen}}(r) = \text{pen}'(r)/(nm)$. The inequality $\text{Crit}'(\hat{r}) \leq \text{Crit}'(r)$ gives

$$\begin{aligned} & \|X\hat{A} - XA\|^2(1 + \overline{\text{pen}}(\hat{r})) \\ & \leq \|Y - X\hat{A}_r\|^2 - \sigma^2(1 + \overline{\text{pen}}(r))\|E\|^2 + \overline{\text{pen}}(r)\|Y - X\hat{A}_r\|^2 + \overline{\text{pen}}(r)\|E\|^2\sigma^2 \\ & \quad + 2(1 + \overline{\text{pen}}(\hat{r}))\sigma < E, X\hat{A} - XA > - \overline{\text{pen}}(\hat{r})\|E\|^2\sigma^2 \\ & \leq \left(2\sigma < E, XA_r - X\hat{A}_r > - \overline{\text{pen}}(r)\sigma^2\|E\|^2\right)_+ + (1 + 2\overline{\text{pen}}(r))\|XA - X\hat{A}_r\|^2 + 3\overline{\text{pen}}(r)\|E\|^2 \\ & \quad + (1 + \overline{\text{pen}}(\hat{r}))\left(2\sigma < E, X\hat{A} - XA_r > - \frac{\overline{\text{pen}}(\hat{r})}{1 + \overline{\text{pen}}(\hat{r})}\|E\|^2\sigma^2\right)_+ + 2\sigma\overline{\text{pen}}(\hat{r}) < E, XA_r - XA >. \end{aligned}$$

Dividing both side by $1 + \overline{\text{pen}}(\hat{r})$, we obtain

$$\|X\hat{A} - XA\|^2 \leq (1 + 2\overline{\text{pen}}(r))\|XA - X\hat{A}_r\|^2 + 3\overline{\text{pen}}(r)\|E\|^2 + 2\sigma | < E, XA - XA_r > | + |\Delta_r + \Delta_{\hat{r}}|$$

where

$$\Delta_k = \left(2\sigma | < E, X\hat{A}_k - XA_r > | - \frac{\overline{\text{pen}}(k)}{1 + \overline{\text{pen}}(k)}\|E\|^2\sigma^2\right)_+.$$

We first note that $\mathbb{E}(\|E\|^2) = nm$ and $2\sigma\mathbb{E}(| < E, XA - XA_r > |) \leq \sigma^2 + \|XA - XA_r\|^2$. Then, combining Lemma 3 with $\eta = (K^{1/6} - 1)$ and the following lemma with $\delta = \eta$ gives

$$\mathbb{E}(\|X\hat{A} - XA\|^2) \leq c(K) \left(\mathbb{E}(\|XA - X\hat{A}_r\|^2)(1 + \overline{\text{pen}}(r)) + (1 + nm\overline{\text{pen}}(r))\sigma^2\right),$$

for some $c(K) > 0$.

Lemma 4. Write P for the projection matrix $P = X(X^*X)^+X^*$, with $(X^*X)^+$ the Moore-Penrose pseudo-inverse of X^*X . For any $\delta > 0$ and $r \leq \min(n, q)$ such that $(1 + \delta)\mathbb{E}(S_r(PE)) \leq \sqrt{nm - 1}$, we have

$$(13) \quad \mathbb{E}((S_r(PE))^2 - (1 + \delta)^3\mathbb{E}(S_r(PE))^2\|E\|^2/(nm - 1))_+) \leq 4(1 + 1/\delta)e^{\delta^2/4}e^{-\delta^2 r \max(n, q)/4}.$$

As a consequence, we have

$$\begin{aligned} \mathbb{E} \left(\sup_{r \leq r_{\max}} (S_r(PE)^2 - (1 + \delta)^3 \mathbb{E}(S_r(E_{q \times n}))^2 \|E\|^2 / (nm - 1))_+ \right) \\ \leq 4(1 + 1/\delta) e^{\delta^2/4} \frac{e^{-\delta^2 \max(n,q)/4}}{1 - e^{-\delta^2 \max(n,q)/4}}. \end{aligned}$$

Proof of the Lemma.

Writing $t = (1 + \delta)\mathbb{E}(S_r(PE))/\mathbb{E}(\|E\|) \leq 1$, the map $E \rightarrow S_r(PE) - t\|E\|$ is $\sqrt{2}$ -Lipschitz. Gaussian concentration inequality then ensures that

$$\begin{aligned} S_r(PE) &\leq t\|E\| + \mathbb{E}(S_r(PE) - t\|E\|) + 2\sqrt{\xi} \\ &\leq t\|E\| + \left(2\sqrt{\xi} - \delta \mathbb{E}(S_r(PE))\right)_+, \end{aligned}$$

with ξ a standard exponential random variable. We then get that

$$S_r(PE)^2 \leq (1 + \delta)t^2\|E\|^2 + 4(1 + 1/\delta) \left(\sqrt{\xi} - \delta \mathbb{E}(S_r(PE))/2\right)_+^2$$

and

$$\begin{aligned} \mathbb{E} \left((S_r(PE)^2 - (1 + \delta)t^2\|E\|^2)_+ \right) &\leq 4(1 + 1/\delta) \mathbb{E} \left(\left(\sqrt{\xi} - \delta \mathbb{E}(S_r(PE))/2\right)_+^2 \right) \\ &\leq 4(1 + 1/\delta) e^{-\delta^2 \mathbb{E}(S_r(PE))^2/4}. \end{aligned}$$

The bound (13) then follows from $\mathbb{E}(S_r(PE))^2 \geq r \max(n, q) - 1$ and $\mathbb{E}(\|E\|)^2 \geq nm - 1$.

4.3. Proof of Proposition 2. As in the proof of Proposition 1, we restrict for simplicity to the case where $q = m$, the general case being treated similarly. We write $\overline{\text{pen}}(r) = \text{pen}'(r)/(nm)$ and for any integer $r^* \in [\min(n, m)/2, \min(n, m) - 1]$, we set

$$\Omega_* = \{S_{r^*}(E) \geq (1 - \alpha)\mathbb{E}(S_{r^*}(E))\} \bigcap_{r=1}^{\min(n,m)} \{S_r(E) \leq (1 + \alpha)\mathbb{E}(S_r(E))\}.$$

According to the Gaussian concentration inequality we have

$$\begin{aligned} \mathbb{P}(\Omega_*) &\geq 1 - 2 \sum_{r=1}^{\min(n,m)} e^{-\alpha^2 \mathbb{E}(S_r(E))^2/2} \\ &\geq 1 - 2e^{\alpha^2/2} \sum_{r=1}^{\min(n,m)} e^{-\alpha^2 r \max(n,m)/2} \end{aligned}$$

where the last bound follows from Lemma 1. For any $r \leq r^*$, we have on Ω_*

$$\begin{aligned} \text{Crit}'(r^*) - \text{Crit}'(r) &= \|E\|^2(\overline{\text{pen}}(r^*) - \overline{\text{pen}}(r)) + S_r(E)^2(1 + \overline{\text{pen}}(r)) - S_{r^*}(E)^2(1 + \overline{\text{pen}}(r^*)) \\ &\leq (1 + \alpha)^2(\mathbb{E}(\|E\|)^2(\overline{\text{pen}}(r^*) - \overline{\text{pen}}(r)) + \mathbb{E}(S_r(E))^2(1 + \overline{\text{pen}}(r))(1 + \alpha)^2 \\ &\quad - \mathbb{E}(S_{r^*}(E))^2(1 + \overline{\text{pen}}(r^*))(1 - \alpha)^2). \end{aligned}$$

Since $(\mathbb{E}(\|E\|)^2 \leq nm = KE(S_r(E))^2(1 + \overline{\text{pen}}(r))/\overline{\text{pen}}(r)$, we have

$$\begin{aligned} \text{Crit}'(r^*) - \text{Crit}'(r) &\leq (1 + \alpha)^2(1 - K)(1 + \overline{\text{pen}}(r))E(S_r(E))^2 \\ &\quad - ((1 - \alpha)^2 - (1 + \alpha)^2K)(1 + \overline{\text{pen}}(r^*))\mathbb{E}(S_{r^*}(E))^2 \\ &\leq (1 + \alpha)^2(1 - K)(1 + \overline{\text{pen}}(r^*)) [\mathbb{E}(S_r(E))^2 - (1 - (1 + \alpha)^{-2})\mathbb{E}(S_{r^*}(E))^2]. \end{aligned}$$

To conclude, we note that $\mathbb{E}(S_r(E))^2 < r(\sqrt{n} + \sqrt{m})^2$, $\mathbb{E}(S_{r^*}(E))^2 \geq (nm - 1)/2$ and $1 - (1 + \alpha)^{-2} \geq \alpha$, so the term in the bracket is smaller than

$$r(\sqrt{n} + \sqrt{m})^2 - \frac{1 - K}{8}(nm - 1)$$

which is negative when $r \leq \frac{1 - K}{8} \times \frac{nm - 1}{(\sqrt{n} + \sqrt{m})^2}$.

5. COMMENTS AND EXTENSIONS

5.1. Link with PCA. In the case where X is the identity matrix, namely $Y = A + E$, Principal Component Analysis (PCA) is a popular technique to estimate A . The matrix A is estimated by projecting the data Y on the r first principal components, the number r of components being chosen according to empirical or asymptotical criterions.

It turns out that the projection of the data Y on the r first principal components coincides with the estimator \hat{A}_r . The criterions (2) and (8) then provide a theoretically grounded way to select the number r of components. Theorem 1 and Theorem 2 ensure that the risk of the final estimate \hat{A}_r nearly achieves the minimum over r of the risks $\mathbb{E}(\|\hat{A}_r - A\|^2)$.

5.2. Sub-Gaussian errors. We have considered for simplicity the case of Gaussian errors, but the results can be extended to the case where the entries $E_{i,j}$ are i.i.d sub-Gaussian. In this case, the matrix PE will play the role of the matrix $E_{q \times n}$ in the Gaussian case. More precisely, combining recent results of Rudelson and Vershynin [17] and Bunea *et al.* [7] on sub-Gaussian random matrices, with concentration inequality for sub-Gaussian random variables [13] enables to prove an analog of Lemma 1 for $\mathbb{E}(S_r(PE))^2$ (with different constants). Then, the proof of Theorem 1 and Theorem 2 can be easily adapted, replacing the Condition (4) by

$$\text{pen}(r) \geq K\mathbb{E}(S_r(PE))^2, \quad \text{for } r \leq \min(q, n),$$

and the Conditions (9) and (10) by $K\mathbb{E}(S_{r_{\max}}(PE))^2 < \mathbb{E}(\|E\|)^2$ and

$$\text{pen}'(r) \geq \frac{K\mathbb{E}(S_r(PE))^2}{1 - K\mathbb{E}(S_r(PE))^2/\mathbb{E}(\|E\|)^2}, \quad \text{for } r \leq r_{\max}.$$

Analog of Proposition 1 and 2 also hold with different constants.

5.3. Selecting among arbitrary estimators. Our theory provides a procedure to select among the family of estimators $\{\widehat{A}_r, r \leq r_{\max}\}$. It turns out that it can be extended to arbitrary (finite) families of estimators $\{A_\lambda, \lambda \in \Lambda\}$ such as the nuclear norm penalized estimator family $\{\widehat{A}_\lambda^{\ell_1}, \lambda \in \Lambda\}$. The most straightforward way is to replace everywhere \widehat{A}_r by \widehat{A}_λ and $\text{pen}(r)$ by $\underline{\text{pen}}(\lambda)$, with $\underline{\text{pen}}(\lambda) = \text{pen}(\text{rank}(\widehat{A}_\lambda))$. In the spirit of Baraud *et al.* [5], we may also consider more refined criterions such as

$$\text{Crit}_\alpha(\lambda) = \min_{r \leq r_{\max}} \left\{ (\|Y - X\widehat{A}_{\lambda,r}\|^2 + \|X\widehat{A}_\lambda - X\widehat{A}_{\lambda,r}\|^2) \left(1 + \frac{\text{pen}'(r)}{nm} \right) \right\},$$

where $\alpha > 0$ and $\widehat{A}_{\lambda,r}$ minimizes $\|B - \widehat{A}_\lambda\|$ over the matrices B of rank at most r . Analog of Theorem 2 can be derived for such criterions, but we will not pursue in that direction.

6. EMPIRICAL STUDY

6.1. Simulation setting. We perform the same simulation study as the one described in Section 4.1 of Bunea *et al.* [7]. Our goal is to compare the estimator introduced in Section 4 with the estimator \widehat{A}_{BSW} proposed by Bunea, She and Wegkamp [7]. The latter estimator is given by $\widehat{A}_{BSW} = \widehat{A}_{\hat{r}}$ where \hat{r} minimizes the criterion

$$\text{Crit}_\lambda(r) = \|Y - X\widehat{A}_r\|^2 + \lambda(n + \text{rank}(X))r$$

with λ chosen by V -fold cross validation on a fine grid (here we set $V = 5$).

We recall briefly the simulation setting. The rows of the matrix X are drawn independently according to a centered Gaussian distribution with covariance matrix $\Sigma_{i,j} = \rho^{|i-j|}$, $\rho > 0$. For a positive b , the matrix A is given by $A = bB_{p \times r}B_{r \times n}$, where the entries of the B matrices are i.i.d. standard Gaussian. For $r \leq \min(n, p)$, the rank of the matrix A is then r with probability one. In the first experiment, $m = 100$, $p = n = 25$, $r = 10$, ρ varies in $\{0.1, 0.5, 0.9\}$ and b varies in $\{0.1, 0.2, 0.3, 0.4\}$. In the second experiment, $m = 20$, $p = 100$, $n = 25$, $r = 10$, ρ varies in $\{0.1, 0.5, 0.9\}$ and b varies in $\{0.1, 0.2, 0.3, 0.4\}$.

6.2. Results. We consider henceforth the selection Criterion (8) with

$$\text{pen}'(r) = \frac{K\mathbb{E}(S_r(E_{q \times n}))^2}{1 - \frac{1}{nq}(1 + K\mathbb{E}(S_r(E_{q \times n}))^2)}.$$

We have no theoretical hint concerning the choice of $K > 1$ and there is actually no reason for the existence of a universal "optimal" constant K . We then choose K by 5-fold cross-validation among a small grid of values between 1 and 3. We write \widehat{A}_K for the resulting estimator $\widehat{A}_{\hat{r}}$.

The results of the first experiment are reported in Figure 3 and those of the second experiment in Figure 4. In the first column the estimated risk ratio

$$\mathcal{R}(\rho, b) = \mathbb{E}(\|XA - X\hat{A}_K\|^2) / \mathbb{E}(\|XA - X\hat{A}_{BSW}\|^2)$$

are plotted for each value of ρ and b . The boxplots of the second column compare the performances of estimators $X\hat{A}_K$ and $X\hat{A}_{BSW}$ to that of the estimator $X\hat{A}_{10}$ that we would use if we knew that the rank of A is 10. The boxplots give for each value of ρ the distribution of the ratios

$$(14) \quad \|XA - X\hat{A}_{BSW}\|^2 / \|XA - X\hat{A}_{10}\|^2 \quad \text{and} \quad \|XA - X\hat{A}_K\|^2 / \|XA - X\hat{A}_{10}\|^2.$$

Finally, we plot in the last column the mean estimated ranks $\mathbb{E}(\hat{r}_K)$ and $\mathbb{E}(\hat{r}_{BSW})$ for each value of b and ρ .

In Experiment 1 (large sample size), both estimators \hat{A}_K and \hat{A}_{BSW} performs very similarly. In Experiment 2 (small sample size), the estimator \hat{A}_K performs significantly better than \hat{A}_{BSW} . Most of the estimated risk ratios $\mathcal{R}(\rho, b)$ are smaller than 0.8 and the boxplots show that the estimator \hat{A}_K is more stable in this experiment.

REFERENCES

- [1] T.W. Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distribution. *Annals of Mathematical Statistics* 22 (1951), 327–351.
- [2] C. W. Anderson, E. A. Stolz, and S. Shamsunder. Multivariate autoregressive models for classification of spontaneous electroencephalogram during mental tasks. *IEEE Trans. on bio-medical engineering*, 45 no 3 (1998), 277–286.
- [3] F. Bach. Consistency of trace norm minimization, *Journal of Machine Learning Research*, 9 (2008), 1019-1048.
- [4] Y. Baraud, C. Giraud and S. Huet. Gaussian model selection with unknown variance. *Annals of Statistics*, Vol. 37 (2009), no2, 630–672.
- [5] Y. Baraud, C. Giraud and S. Huet. Estimator selection in the Gaussian setting. arXiv:1007.2096v1
- [6] E. N. Brown, R. E. Kass, and P. P. Mitra. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature Neuroscience*, 7 no 5 (2004), 456–461.
- [7] F. Bunea, Y. She and M. Wegkamp. Adaptive rank Penalized Estimators in Multivariate Regression. arXiv:1004.2995v1 (2010)
- [8] Davidson and Szarek. *Handbook of the Geometry of Banach Spaces*. North-Holland Publishing Co., Amsterdam, 2001.
- [9] L. Harrison, W. D. Penny, and K. Friston. Multivariate autoregressive modeling of fmri time series. *NeuroImage*, 19 (2004), 1477-1491
- [10] R.A. Horn and C.R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, Cambridge, 1994.
- [11] A.J. Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate analysis* 5 (1975), 248–262.
- [12] A.J. Izenman. *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer, New York, 2008.
- [13] M. Ledoux. *The concentration of measure phenomenon*. Mathematical Surveys and Monographs, 89. American Mathematical Society, Providence, 2001.

- [14] Z. Lu, R. Monteiro and M. Yuan. Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression. *Mathematical Programming* (to appear).
- [15] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. arXiv:0912.5100v1 (2009)
- [16] A. Rohde, A.B. Tsybakov. Estimation of High-Dimensional Low-Rank Matrices. arXiv:0912.5338v2 (2010)
- [17] M. Rudelson, R. Vershynin. Non-asymptotic theory of random matrices: extreme singular values. arXiv:1003.2990v2 (2010)
- [18] M. Yuan, A. Ekici, Z. Lu and R. Monteiro. Dimension Reduction and Coefficient Estimation in Multivariate Linear Regression. *Journal of the Royal Statistical Society, Series B*, 69 (2007), 329-346.

CMAP, UMR CNRS 7641, ECOLE POLYTECHNIQUE, ROUTE DE SACLAY, 91128 PALAISEAU CEDEX, FRANCE

E-mail address: christophe.giraud@polytechnique.edu

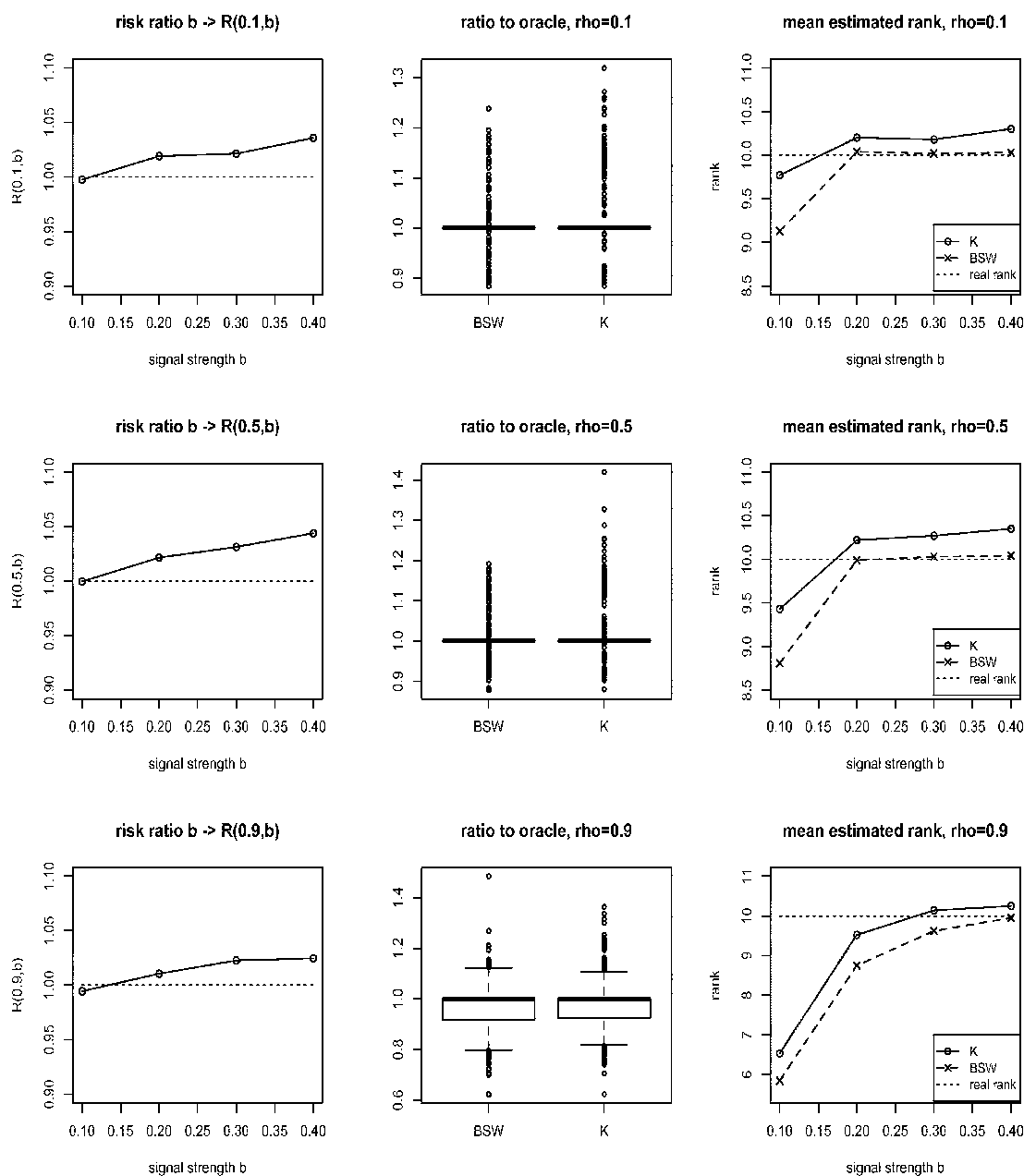


FIGURE 3. Experiment 1. Top to bottom $\rho = 0.1, 0.5, 0.9$. The index K refers to \hat{A}_K , the index BSW to \hat{A}_{BSW} . Left column: risk ratio $\mathcal{R}(\rho, b)$. Center: boxplots of (14). Right: mean estimated rank $\mathbb{E}(\hat{r}_K)$ and $\mathbb{E}(\hat{r}_{BSW})$.

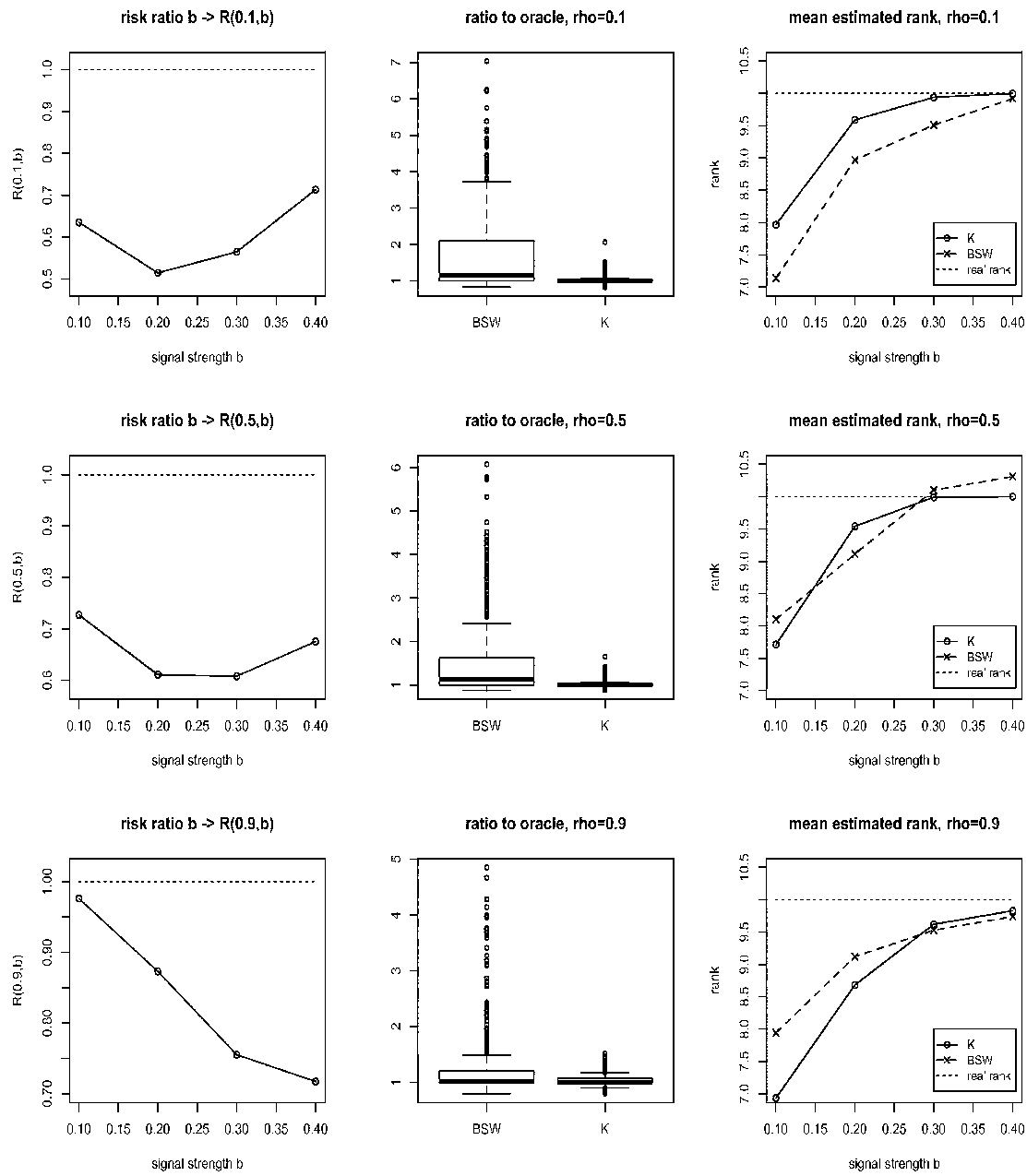


FIGURE 4. Experiment 2. Top to bottom $\rho = 0.1, 0.5, 0.9$. The index K refers to \hat{A}_K , the index BSW to \hat{A}_{BSW} . Left column: risk ratio $\mathcal{R}(\rho, b)$. Center: boxplots of (14). Right: mean estimated rank $\mathbb{E}(\hat{r}_K)$ and $\mathbb{E}(\hat{r}_{BSW})$.