Statistica Sinica **27** (2017), 115-145
doi:http://dx.doi.org/10.5705/ss.2013.167

HIGH DIMENSIONAL MATRIX ESTIMATION WITH UNKNOWN VARIANCE OF THE NOISE

Stéphane Gaïffas and Olga Klopp

CMAP, École Polytechnique and Modal'X,
University Paris Ouest Nanterre la Défense

Abstract: Assume that we observe a small set of entries or linear combinations of entries of an unknown matrix A_0 corrupted by noise. We propose a new method for estimating A_0 that does not rely on the knowledge or on an estimation of the standard deviation of the noise σ. Our estimator achieves, up to a logarithmic factor, optimal rates of convergence under Frobenius risk and, thus, has the same prediction performance as previously proposed estimators that rely on the knowledge of σ. Some numerical experiments show the benefits of this approach.

Key words and phrases: Low rank matrix estimation, matrix completion, matrix regression, unknown variance of the noise.

1. Introduction

In this paper we focus on the problem of high-dimensional matrix estimation from noisy observations with *unknown* variance of the noise. Our main interest is the high dimensional setting, that is, when the dimension of the unknown matrix is much larger than the sample size. Such problems arise in a variety of applications. In order to obtain a consistent procedure in this setting we need some additional constraints. In sparse matrix recovery a standard assumption is that the unknown matrix is exactly or near low-rank. Low-rank conditions are appropriate for such applications as recommendation systems, system identification, global positioning and remote sensing (for more details see Candes and Plan (2010)).

We propose a new method for approximate low-rank matrix recovery that does not rely on the knowledge or on an estimation of the standard deviation of the noise. Two particular settings are analysed in more details: matrix completion and multivariate linear regression.

In the matrix completion problem we observe a small set of entries of an unknown matrix. Moreover, the entries that we observe may be perturbed by some noise. Based on these observations we want to predict or reconstruct exactly the missing entries. A well-known example of matrix completion is the Netflix recommendation system. Suppose we observe a few movie ratings from a large data

matrix in which rows are users and columns are movies. Each user only watches a few movies compared to the total database of movies available on Netflix. The goal is to predict the missing ratings in order to be able to recommend movies to a person that he/she has not yet seen.

In the noiseless setting, if the unknown matrix has low rank and is "incoherent", it can be reconstructed exactly with high probability from a small set of entries. This result was first proved in Candès and Recht (2009) using nuclear norm minimization. A tighter analysis of the same convex relaxation was carried out in Candès and Tao (2010). For a simpler approach see Recht (2009) and Gross (2011). An alternative line of work was developed in Keshavan, Montanari and Oh (2010).

In a more realistic setting the observed entries are corrupted by noise. This question has been addressed by several authors, see, e.g., Candes and Plan (2010); Keshavan, Montanari and Oh (2010); Rohde and Tsybakov (2011); Negahban and Wainwright (2011, 2010); Koltchinskii (2011); Koltchinskii, Lounici and Tsybakov (2011); Gaiffas and Lecué (2011); Klopp (2011)). These results require knowledge of the noise variance, however, in practice such an assumption can be difficult to meet, and the estimation of σ is non-trivial in large scale problems. Thus, there is a gap between theory and practice.

The multivariate linear regression model is given by

$$U_i = V_i A_0 + E_i \qquad i = 1, \dots, l, \tag{1.1}$$

where U_i are $1 \times m_2$ vectors of response variables, V_i are $1 \times m_1$ vectors of predictors, A_0 is an unknown $m_1 \times m_2$ matrix of regression coefficients and E_i are random $1 \times m_2$ vectors of noise with independent entries and mean zero. This model arises in such applications as the analysis of gene array data, medical imaging, astronomical data analysis, psychometrics and many more.

Multivariate linear regression with unknown noise variance has been considered in Bunea, She and Wegkamp (2011); Giraud (2011). These papers study rank-penalized estimators. Bunea, She and Wegkamp (2011), who first introduced such estimators, proposed an unbiased estimator of σ that required an assumption on the dimensions of the problem. This assumption excludes an interesting case, the case when the sample size is smaller than the number of covariates. The method proposed in Giraud (2011) can be applied to this last case under a condition on the rank of the unknown matrix A_0. Our method, unlike the method of Bunea, She and Wegkamp (2011), can be applied to the case when the sample size is smaller than the number of covariates and our condition is weaker than the conditions in Giraud (2011). For more details see Section 3.

Usually, the variance of the noise is involved in the choice of the regularization parameter. Our main idea is to use the Frobenius norm instead of the squared

Frobenius norm as a goodness-of-fit criterion, penalized by the nuclear norm, which is now a well-established proxy for rank penalization in the compressed sensing literature Candès and Tao (2010); Gross (2011). Roughly, the idea is that in the KKT condition, the gradient of this "square-rooted" criterion is the regression score, which is pivotal with respect to the noise level, so that the theoretically optimal smoothing parameter does not depend on the noise level anymore.

This cute idea for dealing with an unknown noise level was first introduced for the square-root lasso by Belloni, Chernozhukov and Wang (2011) in the vector regression model setting. The estimators proposed in the present paper require a different analysis, with proofs that differ from the vector case. Other methods dealing with the unknown noise level in high-dimensional sparse regression include e.g., the scaled Lasso Sun and Zhang (2012) and the penalized Gaussian log-likelihood Städler, Bühlmann and van de Geer (2010). For a complete and comprehensive survey see Giraud, Huet and Verzelen (2012). It is an interesting open question whether these other methods could be adapted in the matrix setting.

1.1. Layout of the paper

This paper is organized as follows. In Section 1.2 we set notations. In Section 2 we consider the matrix completion problem under uniform sampling at random (USR). We propose a new square-root type estimator for which the choice of the regularization parameter λ is independent of σ. The main result, Theorem 2, shows that, in the case of USR matrix completion and under some mild conditions that link the rank and the "spikiness" of A_0, the prediction risk of our estimator measured in Frobenius norm is comparable to the sharpest bounds obtained until now.

In Section 3, we apply our ideas to the problem of matrix regression. We introduce a new square-root type estimator. For this construction, as in the case of matrix completion, we do not need to know or estimate the noise level. The main result for matrix regression gives, up to a logarithmic factor, minimax optimal bound on the prediction error $\|V(\hat{A} - A_0)\|_2^2$.

In Section 4 we give empirical results that confirms our theoretical findings.

1.2. Notation

For any matrices $A, B \in \mathbb{R}^{m_1 \times m_2}$, we define the scalar product $\langle A, B \rangle = \mathrm{tr}(A^T B)$, where $\mathrm{tr}(A)$ denotes the trace of the matrix A.

For $0 < q \leq \infty$ the Schatten-q (quasi-)norm of the matrix A is defined by

$$\|A\|_q = \left(\sum_{j=1}^{\min(m_1, m_2)} \sigma_j(A)^q \right)^{1/q} \text{for } 0 < q < \infty \text{ and } \|A\|_\infty = \sigma_1(A),$$

where $(\sigma_j(A))_j$ are the singular values of A ordered decreasingly.

We summarize the notations which we use throughout this paper: ∂G is the subdifferential of G; $S^\perp$ is the orthogonal complement of S; $\mathcal{P}_S$ is the orthogonal projector on the linear vector subspace S and $\mathcal{P}_S^\perp = 1 - \mathcal{P}_S$; $\|A\|_{\sup} = \max_{i,j} \mid a_{ij} \mid$ where $A = (a_{ij})$. In what follows we denote by c a numerical constant whose value can vary from one expression to the other and is independent from n, m_1, m_2. Set $m = m_1 + m_2$, $m_1 \wedge m_2 = \min(m_1, m_2)$ and $m_1 \vee m_2 = \max(m_1, m_2)$. The symbol $\lesssim$ means that the inequality holds up to multiplicative numerical constants.

2. Matrix Completion

In this section we construct a square-root estimator for the matrix completion problem under uniform sampling at random. Let $A_0 \in \mathbb{R}^{m_1 \times m_2}$ be an unknown matrix, and consider the observations (X_i, Y_i) satisfying the trace regression model

$$Y_i = \operatorname{tr}(X_i^T A_0) + \sigma \xi_i,\ i = 1, \dots, n. \tag{2.1}$$

Here, Y_i are real random variables; X_i are random matrices with dimension $m_1 \times m_2$. The noise variables ξ_i are independent, identically distributed and having distribution Φ such that

$$\mathbb{E}_\Phi(\xi_i) = 0,\ \mathbb{E}_\Phi(\xi_i^2) = 1, \tag{2.2}$$

and $\sigma > 0$ is the *unknown* standard deviation of the noise.

We assume that the design matrices X_i are i.i.d. uniformly distributed on the set

$$\mathcal{X} = \left\{ e_j(m_1) e_k^T(m_2), 1 \le j \le m_1, 1 \le k \le m_2 \right\}, \tag{2.3}$$

where $e_l(m)$ are the canonical basis vectors in $\mathbb{R}^m$. Note that when $X_i = e_j(m_1) e_k^T(m_2)$ we observe the (j,k)th entry of A_0 perturbed by some noise. When the number of observations, n, is much smaller then the total number of coefficients, $m_1\, m_2$, we consider the problem of estimating of A_0, i.e., the problem of reconstruction of many missing entries of A_0 from n observed coefficients.

In Koltchinskii, Lounici and Tsybakov (2011), the authors introduce the following estimator of A_0

$$\hat{A} = \underset{A \in \mathbb{R}^{m_1 \times m_2}}{\arg\min} \left\{ \| A - \mathbf{X} \|_2^2 + \lambda \|A\|_1 \right\}, \tag{2.4}$$

where

$$\mathbf{X} = \frac{m_1\, m_2}{n} \sum_{i=1}^n Y_i X_i. \tag{2.5}$$

For this estimator, the variance of the noise is involved in the choice of the regularisation parameter λ. We propose a new square-root type estimator

$$\hat{A}_{\lambda,\mu} = \underset{A \in \mathbb{R}^{m_1 \times m_2}}{\arg\min} \left\{ \|A - \mathbf{X}\|_2 + \lambda \|A\|_1 \right\}. \tag{2.6}$$

The first part of our estimator coincides with the square root of the data-depending term in (2.4). This is similar to the principle used to define the square-root lasso for the usual vector regression model, see Belloni, Chernozhukov and Wang (2011). Despite taking the square-root of the least squares criterion function, the problem (2.6) retains global convexity and can be formulated as a solution to a conic programming problem. For more details see Section 4.

We will consider the case of sub-Gaussian noise and matrices with uniformly bounded entries. Let a denote a constant such that

$$\|A_0\|_{\sup} \leq a. \tag{2.7}$$

We suppose that the noise variables ξ_i are such that

$$\mathbb{E}(\xi_i) = 0,\ \mathbb{E}(\xi_i^2) = 1 \tag{2.8}$$

and that there exists a constant K such that

$$\mathbb{E}\left[\exp(t\xi_i)\right] \leq \exp\left(\frac{t^2}{2K}\right) \tag{2.9}$$

for all $t > 0$. Normal $N(0,1)$ random variables are sub-Gaussian with $K = 1$ and (2.9) implies that ξ_i has Gaussian type tails:

$$\mathbb{P}\left\{|\xi_i| > t\right\} \leq 2\exp\left\{\frac{-t^2}{2K}\right\}.$$

Condition $\mathbb{E}\xi_i^2 = 1$ implies that $K \leq 1$. We set

$$\mathbf{M} = \frac{1}{m_1\, m_2}\left(\mathbf{X} - A_0\right), \tag{2.10}$$

and note that $\mathbf{M}$ is centred. Its operator and Frobenius norms play an important role in the choice of the regularisation parameter λ. We set

$$\Delta = \frac{\|\mathbf{M}\|_\infty}{\|\mathbf{M}\|_2}. \tag{2.11}$$

We provide a general oracle inequality for the prediction error of our estimator,with proof given in Appendix A.1.

Theorem 1. *Suppose that* $\rho/\sqrt{2\mathrm{rank}(A_0)} \geq \lambda \geq 3\Delta$ *for some* $\rho < 1$*, then*

$$\|\hat{A} - A_0\|_2^2 \leq \inf_{\sqrt{2\mathrm{rank}(A)} \leq \rho/\lambda}\Big\{\left(1-\rho\right)^{-1}\|A - A_0\|_2^2 + \Big(\frac{2\lambda\, m_1\, m_2}{1-\rho}\Big)^2\|\mathbf{M}\|_2^2\ \mathrm{rank}A\Big\},$$

where Δ *and* M *are defined in* (2.11) *and* (2.10)*.*

In order to specify the value of the regularization parameter λ, we need to estimate Δ with high probability. Therefore we use the following two lemmas.

Lemma 1. *For $n > 8(m_1 \wedge m_2)\log^2 m$, with probability at least $1 - 3/m$, one has*

$$\|\mathbf{M}\|_\infty \leq (c_*\sigma + 2a)\sqrt{\frac{2\log(m)}{(m_1 \wedge m_2)n}}, \tag{2.12}$$

where c_* is a numerical constant that depends only on K. If ξ_i are $N(0,1)$, then we can take $c_* = 6.5$.

Proof. The bound (2.12) is stated in Lemmas 2 and 3 in Koltchinskii, Lounici and Tsybakov (2011). A closer inspection of the proof of Proposition 2 in Koltchinskii (2011) gives an estimation on c_* in the case of Gaussian noise. For more details see Appendix A.4.

The next result is proven in Appendix A.5.

Lemma 2. *Suppose that $4\,n \leq m_1 m_2$. Then, for $\mathbf{M}$ defined in (2.10), with probability at least $1 - 2/m_1 m_2 - c_1 \exp\{-c_2 n\}$, one has*

(i)
$$2\Big(\frac{\|A_0\|_2^2}{n m_1 m_2} + \frac{\sigma^2}{n}\Big) \geq \|\mathbf{M}\|_2^2 \geq \frac{\sigma^2}{2\,n};$$

(ii)
$$\Big\|\frac{1}{n}\sum_{i=1}^n Y_i X_i\Big\|_2^2 \geq \frac{\|A_0\|_2^2}{n\,m_1 m_2} \geq \frac{4\,\|A_0\|_2^2}{(m_1 m_2)^2};$$

(iii)
$$\|\mathbf{M}\|_2 \geq \frac{1}{2}\Big\|\frac{1}{n}\sum_{i=1}^n Y_i X_i\Big\|_2,$$

where (c_1, c_2) are numerical constants that depend only on K, a and σ.

The condition on λ in Theorem 1 is that $\lambda \geq 3\Delta$. Using Lemmas 1 and 2, we can choose

$$\lambda = 2c_*\sqrt{\frac{\log m}{m_1 \wedge m_2}} + 4a\sqrt{\frac{2\,n\log m}{m_1 \wedge m_2}}\frac{1}{\|\sum_{i=1}^n Y_i X_i\|_2}. \tag{2.13}$$

In (2.13) λ is data-driven and independent of σ. With this choice of λ, the assumption of Theorem 1, $\rho/\sqrt{\mathrm{rank}(A_0)} \geq \lambda$, takes the form

$$\frac{\rho}{\sqrt{\mathrm{rank}(A_0)}} \geq 2c_*\sqrt{\frac{\log m}{m_1 \wedge m_2}} + 4a\sqrt{\frac{2\,n\log m}{m_1 \wedge m_2}}\frac{1}{\|\sum_{i=1}^n Y_i X_i\|_2}. \tag{2.14}$$

Using (ii) of Lemma 2 we get that (2.14) is satisfied with a high probability if

$$\frac{\rho}{\sqrt{\operatorname{rank}(A_0)}} \geq 2\, c_* \sqrt{\frac{\log m}{m_1 \wedge m_2}} + \frac{4\, a \sqrt{m_1 m_2}}{\|A_0\|_2} \sqrt{\frac{2 \log m}{m_1 \wedge m_2}}. \tag{2.15}$$

As m_1 and m_2 are large, the first term on the rhs of (2.15) is small. Thus (2.15) is essentially equivalent to

$$\rho \geq 4 \sqrt{\frac{2 \log m}{(m_1 \wedge m_2)}} \sqrt{\operatorname{rank}(A_0)}\, \alpha_{sp}, \tag{2.16}$$

where $\alpha_{sp} = \sqrt{m_1 m_2}\, \|A_0\|_{\sup} / \|A_0\|_2$ is the *spikiness ratio* of A_0. The notion of "spikiness" was introduced in Negahban and Wainwright (2010). We have that $1 \leq \alpha_{sp} \leq \sqrt{m_1 m_2}$ and is large for matrices where some "large" coefficients emerge as spikes among very "small" coefficients. For instance, $\alpha_{sp} = 1$ if all the entries of A_0 are equal to some constant and $\alpha_{sp} = \sqrt{m_1 m_2}$ if A_0 has only one non-zero entry.

Condition (2.16) is a kind of trade-off between "spikiness" and rank. If α_{sp} is bounded by a constant then, up to a logarithmic factor, $\operatorname{rank}(A_0)$ can be of the order $m_1 \wedge m_2$, which is its maximal possible value. If our matrix is "spiky", then we need low rank. To gain some intuition consider the case of square matrices. Typically, matrices with both high spikiness ratio and high rank look almost diagonal. Under uniform sampling if $n \ll m_1 m_2$, with high probability we do not observe diagonal (i.e. non-zero) elements.

Theorem 2. *Suppose* (2.8)−(2.7) *are satisfied and* λ *is as in* (2.13). *Assume that* $8(m_1 \wedge m_2) \log^2 m < n \leq m_1 m_2/4$ *and that* (2.15) *holds for some* $\rho < 1$. *Then, with probability at least* $1 - 4/m - c_1 \exp\{-c_2 n\}$,

$$\frac{1}{m_1 m_2} \|\hat{A} - A_0\|_2^2 \leq C_* \frac{(m_1 \vee m_2)}{n} \operatorname{rank}(A_0) \log m. \tag{2.17}$$

Here $C_* = 16\left(2c_*\sigma^2 + (18 + 2c_*)a^2\right)/(1-\rho)^2$, c_* *is an absolute constant that depends only on* K *and* (c_1, c_2) *are numerical constants that depend only on* K, a *and* σ.

Proof. This is a consequence of Theorem 1 for $A = A_0$. From (2.13) we get

$$\|\hat{A} - A_0\|_2^2 \leq \frac{8(m_1 m_2)^2}{(1-\rho)^2} \left(c_* \sqrt{\frac{4 \log m}{m_1 \wedge m_2}} + 2a \sqrt{\frac{2\, n \log m}{m_1 \wedge m_2}} \frac{1}{\left\| \sum_{i=1}^n Y_i X_i \right\|_2} \right)^2 \times \|\mathbf{M}\|_2^2 \operatorname{rank}(A_0). \tag{2.18}$$

Using the triangle inequality and (ii) of Lemma 2, we compute

$$\|\mathbf{M}\|_2 \le \Big\|\frac{1}{n}\sum_{i=1}^n Y_i X_i\Big\|_2 + \frac{1}{m_1 m_2}\|A_0\|_2$$
$$\le \frac{3}{2}\Big\|\frac{1}{n}\sum_{i=1}^n Y_i X_i\Big\|_2. \tag{2.19}$$

Using (i) of Lemma 2 and (2.19), from (2.18) we get

$$\|\hat{A} - A_0\|_2^2 \le \frac{16\log(m)(m_1 m_2)^2}{(1-\rho)^2(m_1 \wedge m_2)}\Big(2\,c_*\Big(\frac{\|A_0\|_2^2}{n m_1 m_2} + \frac{\sigma^2}{n}\Big) + \frac{18a^2}{n}\Big)\mathrm{rank}(A_0).$$

Then, we use $\|A_0\|_2^2 \le a^2 m_1 m_2$ to obtain

$$\frac{\|\hat{A} - A_0\|_2^2}{m_1 m_2} \le \frac{16\log(m)(m_1 \vee m_2)}{(1-\rho)^2 n}\left(2c_*\sigma^2 + (18 + 2c_*)a^2\right)\mathrm{rank}(A_0).$$

This completes the proof of Theorem 2.

Theorem 2 guarantees that the normalized Frobenius error $\| \hat{A} - A_0 \|_2 / \sqrt{m_1 m_2}$ of the estimator $\hat{A}$ is small whenever $n > C(m_1 \vee m_2)\log(m)\mathrm{rank}(A_0)$ with a constant C large enough. This quantifies the sample size, n, necessary for successful matrix completion from noisy data with unknown variance of the noise. Remarkably, this sampling size is the same as in the case of known variance of the noise. In Theorem 2 we have an additional restriction $4n \le m_1 m_2$. In the matrix completion setting the number of observed entries n is always smaller then the total number of entries $m_1 m_2$ and this condition can be replaced by $n \le \alpha m_1 m_2$ for some $\alpha < 1$.

Theorem 2 leads to the same rate of convergence as previous results on matrix completion which treat σ as known. In order to compare our bounds to those obtained in past works on noisy matrix completion, we start by describing the result of Keshavan, Montanari and Oh (2010). Under sampling without replacement sampling scheme and sub-Gaussian errors, the estimator proposed in Keshavan, Montanari and Oh (2010) satisfies, with high probability,

$$\frac{1}{m_1 m_2}\|\hat{A} - A_0\|_2^2 \lesssim k^4\sqrt{\alpha}\frac{(m_1 \vee m_2)}{n}\mathrm{rank}(A_0)\log n. \tag{2.20}$$

Here $k = \sigma_{\max}(A_0)/\sigma_{\min}(A_0)$ is the condition number and $\alpha = (m_1 \vee m_2)/(m_1 \wedge m_2)$ is the aspect ratio. Comparing (2.20) and (2.17), we see that our bound is better: it does not involve the multiplicative coefficient $k^4\sqrt{\alpha}$ which can be large.

Negahban and Wainwright (2010) propose an estimator which, in the case of USR matrix completion and sub-exponential noise, satisfies

$$\frac{1}{m_1 m_2}\|\hat{A} - A_0\|_2^2 \lesssim \alpha_{sp}\frac{m}{n}\mathrm{rank}(A_0)\log m. \tag{2.21}$$

Here α_{sp} is the spikiness ratio of A_0. For α_{sp} bounded by a constant, (2.21) gives the same bound as Theorem 2. The construction of $\hat{A}$ in Negahban and Wainwright (2010) requires a priori information on the spikiness ratio of A_0 and on σ. This is not the case for our estimator.

The estimator proposed in Koltchinskii, Lounici and Tsybakov (2011) achieves the same bound as ours. In addition to prior information on $\|A_0\|_{\sup}$, their method also requires prior information on σ. In the case of Gaussian errors, this rate of convergence is optimal up to a logarithmic factor (cf., Theorem 6 of Koltchinskii, Lounici and Tsybakov (2011)) for the class of matrices $\mathcal{A}(r,a)$ defined as follows: for given r and a, $A_0 \in \mathcal{A}(r,a)$ if and only if the rank of A_0 is bounded by r and all the entries of A_0 are bounded in absolute value by a.

An important difference with previous works on matrix completion is that Theorem 2 requires an additional growth restriction on λ, $\rho/\sqrt{2\text{rank}(A_0)} \geq \lambda$. The consequence of this restriction is that our method can not be applied to matrices which have both large spikiness ratio and large rank. Note that the square-root lasso estimator also requires an additional growth restriction on λ (see Theorem 1 in Belloni, Chernozhukov and Wang (2011)). We may think that these restrictions is the price of not knowing σ in our framework.

3. Matrix Regression

In this section we apply our method to matrix regression. The matrix regression model is given by

$$U_i = V_i A_0 + E_i \qquad i = 1, \ldots, n, \tag{3.1}$$

where U_i are $1 \times m_2$ vectors of response variables; V_i are $1 \times m_1$ vectors of predictors; A_0 is an unknown $m_1 \times m_2$ matrix of regression coefficients; E_i are random $1 \times m_2$ noise vectors with independent entries E_{ij}. We suppose that E_{ij} has mean zero and *unknown* standard deviation σ. Set $V = \left(V_1^T, \ldots, V_n^T\right)^T$, $U = \left(U_1^T, \ldots, U_n^T\right)^T$ and $E = \left(E_1^T, \ldots, E_n^T\right)^T$.

We propose new estimator of A_0 using again the idea of the square-root estimators:

$$\hat{A} = \underset{A \in \mathbb{R}^{m_1 \times m_2}}{\arg\min} \left\{ \|U - V A\|_2 + \lambda \|V A\|_1 \right\},$$

where $\lambda > 0$ is a regularization parameter. This estimator can be formulated as a solution to a conic programming problem. For more details see Section 4.

With $\mathcal{P}_V$ denote the orthogonal projector on the linear span of the columns of matrix V, set

$$\Delta' = \frac{\|\mathcal{P}_V(E)\|_\infty}{\|E\|_2}.$$

Minor modifications in the proof of Theorem 1 yield the following result.

Theorem 3. *If* $\rho/\sqrt{2\mathrm{rank}(VA_0)} \geq \lambda \geq 3\Delta'$ *for some* $\rho < 1$*, then*

$$\left\|V\left(\hat{A}-A_0\right)\right\|_2^2 \leq \inf_{\sqrt{2\mathrm{rank}(VA)}\leq\rho/\lambda}\left\{\frac{\|V\,(A-A_0)\|_2^2}{1-\rho}+\left(\frac{2\lambda}{1-\rho}\right)^2\|E\|_2^2\,\mathrm{rank}(VA)\right\}.$$

Proof. The proof follows the lines of the proof of Theorem 1, it is given in Appendix A.7.

To get the oracle inequality in a closed form it remains to specify the value of regularization parameter λ such that $\lambda \geq 3\Delta'$. This requires some assumptions on the distribution of the noise $(E_{ij})_{i,j}$. We consider the case of Gaussian errors. Suppose that $E_{ij} = \sigma\xi_{ij}$ where ξ_{ij} are normal $N(0,1)$ random variables. In order to estimate $\|\mathcal{P}_V E\|_\infty$ we use the following.

Lemma 3 (Bunea, She and Wegkamp (2011), Lemma 3)**.** *Let* $r = \mathrm{rank}(V)$ *and assume that* E_{ij} *are independent* $N(0,\sigma^2)$ *random variables. Then*

$$\mathbb{E}(\|\mathcal{P}_V E\|_\infty) \leq \sigma(\sqrt{m_2}+\sqrt{r})$$

and

$$\mathbb{P}\left\{\|\mathcal{P}_V E\|_\infty \geq \mathbb{E}(\|\mathcal{P}_V E\|_\infty) + \sigma t\right\} \leq \exp\left\{\frac{-t^2}{2}\right\}.$$

We use Bernstein's inequality to get a bound on $\|E\|_2$. Let $\alpha < 1$. With probability at least $1 - 2\exp\left\{-c\,\alpha^2\, n\, m_2\right\}$, one has

$$(1+\alpha)\sigma\sqrt{n\,m_2} \geq \|E\|_2 \geq (1-\alpha)\sigma\sqrt{n\,m_2}. \tag{3.2}$$

Let $\beta > 0$ and take $t = \beta\left(\sqrt{m_2}+\sqrt{r}\right)$ in Lemma 3. Then, using (3.2), we can take

$$\lambda = \frac{(1+\beta)\left(\sqrt{m_2}+\sqrt{r}\right)}{(1-\alpha)\sqrt{n\,m_2}}. \tag{3.3}$$

Put $\gamma = [(1+\beta)/(1-\alpha)] > 1$. Thus, condition $\rho/\sqrt{2\,\mathrm{rank}(VA_0)} \geq \lambda$ gives

$$\mathrm{rank}(VA_0) \leq \frac{\rho^2 n\, m_2}{2\gamma^2\left(\sqrt{m_2}+\sqrt{r}\right)^2} \tag{3.4}$$

and we get the following result.

Theorem 4. *Assume that* ξ_{ij} *are independent* $N(0,1)$*. Pick* λ *as in* (3.3)*. Assume* (3.4) *is satisfied for some* $\rho < 1$*,* $\alpha < 1$ *and* $\beta > 0$*. Then, with probability at least* $1 - 2\exp\{-c(m_2+r)\}$*, we have that*

$$\left\|V\left(\hat{A}-A_0\right)\right\|_2^2 \lesssim \sigma^2(m_2+r)\,\mathrm{rank}(VA_0).$$

Proof. This is a consequence of Theorem 3.

We now compare condition (3.4) with the conditions obtained in Bunea, She and Wegkamp (2011); Giraud (2011). In Bunea, She and Wegkamp (2011), the authors introduce a new rank-penalised estimator and consider both cases when the variance of the noise is known or not. In the case of known variance of the noise, minimax optimal bounds on the mean squared errors are established (this does not need growth restriction on λ and, thus, applies to all $\mathrm{rank}(VA_0)$). When the variance of the noise is unknown, an unbiased estimator of σ^2 is proposed:

$$S^2 = \frac{\|U - PU\|_2^2}{nm_2 - qm_2},$$

where P is the projection matrix on the column space of V and q is the rank of V. This estimator requires an assumption on the dimensions of the problem. In particular it requires that $m_2(n-r)$ be large, which holds whenever $n \gg r$ or $n-r \geq 1$ and m_2 is large. This condition excludes an interesting case $n = r \ll m_2$. On the other hand (3.4) is satisfied for $n = r \ll m_2$ if $\mathrm{rank}(A_0) \lesssim n$ where we used $\mathrm{rank}(VA_0) \leq r \wedge \mathrm{rank}(A_0)$.

The method of Giraud (2011) requires the following condition to be satisfied

$$\mathrm{rank}(A_0) \leq \frac{C_1(n\,m_2 - 1)}{C_2\left(\sqrt{m_2} + \sqrt{r}\right)^2} \tag{3.5}$$

with some constants $C_1 < 1$ and $C_2 > 1$. This is quite similar to (3.4). As $\mathrm{rank}(VA_0) \leq \mathrm{rank}(A_0)$, (3.4) is weaker than (3.5). To the opposite of Giraud (2011), our results are valid for all A_0 provided that

$$r \leq \frac{\rho^2 n\, m_2}{2\gamma^2\left(\sqrt{m_2} + \sqrt{r}\right)^2}.$$

For large $m_2 \gg n$, this condition roughly means that $n > cr$ for some constant c.

4. Simulations

In this section, we give empirical results that confirms our theoretical findings. We illustrate the fact that using the Frobenius norm instead of the square Frobenius norm as a goodness-of-fit criterion makes the optimal smoothing parameter λ independent of the noise level, allowing for a better stability of the procedure with respect to the noise level, as compared to other state-of-the-art procedures. We focus on the matrix regression problem only, since our conclusions are the same for matrix completion. We compare in particular the following procedures:

$$\mathrm{argmin}_A\Big\{\frac{1}{2}\|U - VA\|_2^2 + \lambda\|A\|_1\Big\}, \tag{4.1}$$

which is based on the classical least-squares penalized by the trace norm,

$$\operatorname{argmin}_A \Big\{ \|U - VA\|_2 + \lambda \|A\|_1 \} \tag{4.2}$$

which uses trace norm penalization with square-root least squares, and

$$\operatorname{argmin}_A \Big\{ \|U - VA\|_2 + \lambda \|VA\|_1 \Big\} \tag{4.3}$$

which is the procedure introduced in this paper. We illustrate in particular the fact that (4.2) and (4.3), which are based on a goodness-of-fit using the Frobenius norm instead of the squared Frobenius norm, provide a choice of λ which is independent of the noise level σ.

4.1. Optimization algorithms

In this section, we describe the convex optimization algorithms used for solving problems (4.1), (4.2) and (4.3). For this we need to introduce the proximal operator Bauschke and Combettes (2011) $\mathrm{prox_g}$ of a convex, proper, low-semicontinuous function g, given by

$$\mathrm{prox_g(W)} = \arg\min_{\mathrm{Y}} \Big\{ \frac{1}{2} \|\mathrm{W} - \mathrm{Y}\|_2^2 + \mathrm{g(Y)} \Big\}.$$

In the algorithms described below, we need to compute such proximal operator for specific functions. The proximal operator of the trace norm is given by spectral soft-thresholding,

$$\mathrm{prox_{tg}(W)} = \mathcal{S}_{\mathrm{t}}(\mathrm{W}) \quad \text{for} \quad \mathrm{g(W)} = \|\mathrm{W}\|_1$$

for any $t > 0$, where

$$\mathcal{S}_t(W) = U_W \operatorname{diag}[(\sigma_1(W) - t)_+ \ \cdots \ (\sigma_{\operatorname{rank}(W)}(W) - t)_+] V_W^\top,$$

with $U_W \operatorname{diag}[\sigma_1(W) \ \cdots \ \sigma_{\operatorname{rank}(W)}(W)] V_W^\top$ the singular value decomposition of W, with the columns of U_W and V_W being the left and right singular vectors of W, and $\sigma_1(W) \geq \cdots \geq \sigma_{\operatorname{rank}(W)}(W)$ its singular values.

Problem (4.1) is solved using accelerated proximal gradient, also known as Fista Beck and Teboulle (2009), since the loss is gradient-Lipschitz. Fista allows to minimize an objective of the form

$$F(A) = f(A) + g(A),$$

where f is smooth (gradient-Lipshitz) with Lipschitz constant $L = \|V\|_\infty$ (the operator norm of V) and g is prox-capable. In our setting we consider $f(A) = (1/2)\|U - VA\|_2^2$ and $g(A) = \lambda \|A\|_1$, so that $\nabla f(A) = V^\top (VA - U)$ and

Algorithm 1 Fista

Require: Starting points $B^1 = A^0$, Lipschitz constant $L > 0$ for ∇f, $t_1 = 1$

for $k = 0, 1, 2, \ldots$ **do**
  $A^k \leftarrow \mathrm{prox}_{\mathrm{L}^{-1}\mathrm{g}}(\mathrm{B}^{\mathrm{k}} - \frac{1}{\mathrm{L}}\nabla \mathrm{f}(\mathrm{B}^{\mathrm{k}}))$
  $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$
  $B^{k+1} = A^k + \frac{t_k - 1}{t_{k+1}}(A^k - A^{k-1})$
end for
return A^k

$\mathrm{prox}_{\mathrm{tg}}(\mathrm{A}) = \mathcal{S}_{\mathrm{t}}(\mathrm{A})$. The Fista algorithm is described in Algorithm 1 below. In our experiments we used backtracking linesearch, instead of fixing the step-size constant and equal to $1/L$.

Problem (4.2) is solved using a primal-dual algorithm Chambolle and Pock (2011), see Algorithm 2. It allows to minimize an objective of the form

$$F(A) = f(KA) + g(A), \tag{4.4}$$

where both f and g are prox-capable (with f non-smooth) and K a linear operator. In our setting we choose this time $K = V$, $f(A) = \|A - U\|_2$ and $g(A) = \lambda\|A\|_1$. It is easily proved that

$$\mathrm{prox}_{\mathrm{tf}}(\mathrm{A}) = \begin{cases} U & \text{if } \|A - U\|_2 \leq t, \\ A - t\frac{A-U}{\|A-U\|_2} & \text{if } \|A - U\|_2 > t, \end{cases}$$

which allows to instantiate Algorithm 2 for problem (4.2), using also the Moreau's identity $\mathrm{prox}_{\mathrm{f}^*}(\mathrm{A}) - \mathrm{A} - \mathrm{prox}_{\mathrm{f}}(\mathrm{A})$, see Bauschke and Combettes (2011), where f^* is the Fenchel conjugate of f. In Algorithm 2 we use the heuristics described in Chambolle and Pock (2011) to choose the step-sizes η and τ.

Algorithm 2 Primal-dual algorithm

Require: Starting points $A^0, \bar{A}^0, Z^0$, step-sizes $\eta, \tau > 0$ such that

for $k = 0, 1, 2, \ldots$ **do**
  $Z^{k+1} \leftarrow \mathrm{prox}_{\eta \mathrm{f}^*}(\mathrm{Z}^{\mathrm{k}} + \eta \mathrm{V}\bar{\mathrm{A}}^{\mathrm{k}})$
  $A^{k+1} \leftarrow \mathrm{prox}_{\tau \mathrm{g}}(\mathrm{A}^{\mathrm{k}} - \tau \mathrm{V}^{\top}\mathrm{Z}^{\mathrm{k}+1})$
  $\bar{A}^{k+1} \leftarrow A^{k+1} + \theta(A^{k+1} - A^k)$
end for
return A^k

Problem (4.3) is solved using parallel splitting Bauschke and Combettes (2011). First, we need to reformulate the problem. If $\hat{A}$ is a solution to (4.3), then any $\hat{A} + B$ with $B \in \ker(V)$, where $\ker(V) = \{A \in \mathbb{R}^{m_1 \times m_2} : VA = 0\}$, is also a solution. Thus, we solve the problem on a splitted variable $W = VA$. We define

the linear space $\mathrm{col(V)} = \{\mathrm{W} \in \mathbb{R}^{\mathrm{n}\times \mathrm{m}_2} : \exists \mathrm{A} \in \mathbb{R}^{\mathrm{m}_1 \times \mathrm{m}_2}, \mathrm{VA} = \mathrm{W}\}$. Then, we have

$$V\hat{A} = \operatorname*{arg\,min}_{W \in \mathrm{col(V)}} \|U - W\|_2 + \lambda \|W\|_1,$$

so that we end up with the problem

$$\text{minimize} \quad \|U - W\|_2 + \lambda\|W\|_1 + \delta_{\mathrm{col(V)}}(W), \tag{4.5}$$

where $\delta_C(X)$ stands for the indicator function of a convex set C, defined by $\delta_C(X) = 0$ when $X \in C$ and $\delta_C(X) = +\infty$ when $X \notin C$. Then, we solve (4.5) using parallel splitting Bauschke and Combettes (2011). Each function in (4.5) is prox-capable. Take

$$f_1(W) = \|U - W\|_2, \quad f_2(W) = \lambda\|W\|_1, \quad f_3(W) = \delta_{\mathrm{col(V)}}(W)$$

with $\mathrm{prox_{f_1}}$ and $\mathrm{prox_{f_2}}$ as above. We have that

$$\mathrm{prox_{f_3}(W)} = \mathcal{P}_{\mathrm{col(V)}}(\mathrm{W}) = \mathrm{V}(\mathrm{V}^\top \mathrm{V})^\dagger \mathrm{V}^\top \mathrm{W},$$

where $\mathcal{P}_{\mathrm{col(V)}}$ is the projection operator onto the set col(V), and where $Z^\dagger$ stands for the pseudo-inverse of Z. The parallel splitting algorithm is described in Algorithm 3.

Algorithm 3 Parallel splitting

Require: Step-sizes $\gamma > 0$, $\tau_k \in [0, 2]$, initial values W_1^0, W_2^0, W_3^0

1: **for** $k = 0, 1, 2, \ldots$ **do**
2: $\quad P^k \leftarrow \frac{1}{3}(W_1^k + W_2^k + W_3^k)$
3: $\quad Z_i^k \leftarrow \mathrm{prox}_{\gamma \mathrm{f_i}}(\mathrm{W_i^k})$ for $i = 1, 2, 3$
4: $\quad Q^k \leftarrow \frac{1}{3}(Z_1^k + Z_2^k + Z_3^k)$
5: $\quad W_i^{k+1} \leftarrow W_i^k + \tau_k(2Q^k - P^k - Z_i^k)$ for $i = 1, 2, 3$
6: **end for**
7: **return** P^k

Convergence is guaranteed for $\tau_k \in [0, 2]$ such that $\sum_{k\geq 0} \tau_k(2 - \tau_k) = +\infty$, see Bauschke and Combettes (2011), we simply choose $\tau_k = 1.9$ in our experiments. An alternative (but somewhat less direct) method for solving (4.5) is to write an equivalent conic formulation, and smooth the primal objective by adding a strongly convex term. Then, the corresponding dual problem can be solved using first order techniques. This method, called `TFOCS`, is the one described in Becker, Candès and Grant (2011) for solving general convex cone problems.

4.2. Numerical illustration

We give several numerical illustrations. First, we show that the optimal choice of λ is almost independent of the noise level for the procedures (4.2) and

(4.3), while it needs to be increased with σ for procedure (4.1). This fact is illustrated in Figures 1 and 2. Then, we compare the best prediction errors (among prediction errors obtained for several λ) of solutions of problems (4.1), (4.2) and (4.3). This is illustrated in Tables 1 and 2.

We simulate data as follows. We pick at random A_1 and A_2 as, respectively, $m_1 \times r$ and $m_2 \times r$ matrices with $N(0,1)$ i.i.d. entries, and we fix $A_0 = A_1 A_2^\top$, which is a $m_1 \times m_2$ matrix with rank r a.s. We pick at random a $n \times m_1$ matrix V, with lines $V_i \in \mathbb{R}^{m_1}$, $i = 1, \dots, n$, distributed as a centered Gaussian vectors with covariance equal to the Toeplitz matrix $\Sigma = (\rho^{-|i-j|})_{1 \le i,j \le m_1}$. We finally compute $U = VA_0 + \sigma E$, where the noise matrix E contains $N(0,1)$ i.i.d. entries and $\sigma > 0$ is the standard deviation.

We consider the setting $n = 1{,}000$, $m_1 = 200$, $m_2 = 100$, $r = 10$ and $\rho = 0.5$, called "experiment 1" in Figures and Tables, while we choose $n = 200$, $m_1 = 100$, $m_2 = 400$ and other parameters unchanged for "experiment 2".

In Figures 1 and 2, Tables 1 and 2 we consider values of σ in $\{0.1, 0.5, 1.0, 5.0\}$, and for each value of σ we plot the prediction error $\|V(\hat{A}_\lambda - A_0)\|_2$ for a parameter λ in a grid. We repeat this 10 times, and plot each time the prediction error in Figure 1 and print the average best prediction errors (and standard deviation) in Table 1.

The conclusion of this experiment is the following. The minimum of the prediction error is achieved for a parameter λ that increases with σ for procedure (4.1), while it is almost constant for procedures (4.2) and (4.3). This confirms numerically the fact that when using square-root least-squares instead of least-squares, the optimal choice of λ can be done independently of the noise level. Also, the minimum prediction errors of each procedure are of the same order for experiment 1, with a slight advantage for procedure (4.3) for each considered value of σ, while there is a strong advantage for procedure (4.3) for experiment 2, which corresponds to the case where the number of tasks m_2 is larger than the sample size n.

Appendix. Proofs

A.1. Proof of Theorem 1

The proof of Theorem 1 is based on the ideas of the proof of Theorem 1 in Koltchinskii, Lounici and Tsybakov (2011). However, as the statistical structure of our estimator is different from theirs, the proof requires several modifications and additional information on the behaviour of the estimator. This information is given in Lemmas A.1 and A.2. In particular, Lemma A.1 provides a bound on the rank of our estimator. Its proof is given in Appendix A.2.

Lemma A.1. $\operatorname{rank}(\hat{A}) \le 1/\lambda^2$.

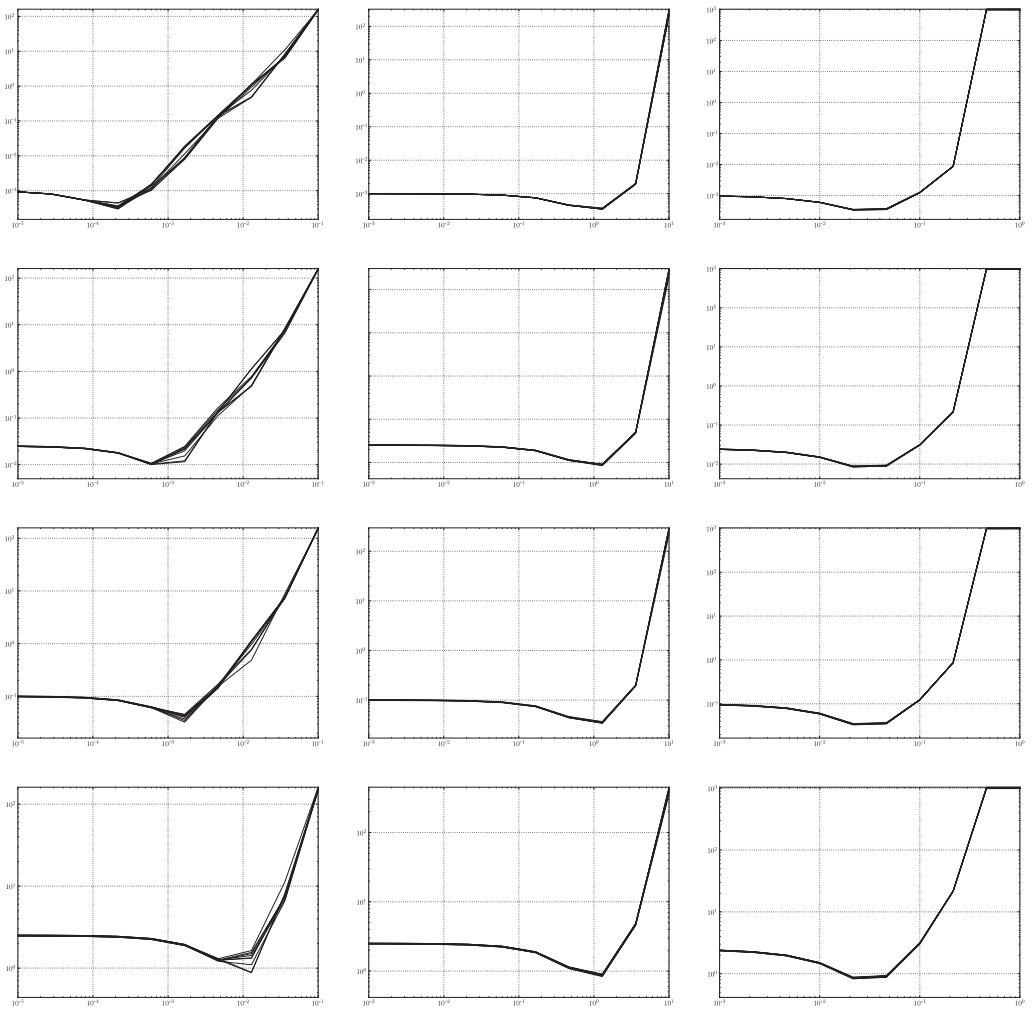

Figure 1. Prediction errors (y-axis) for experiment 1 (see text) for a varying λ (x-axis) for procedure (4.1) (first column), procedure (4.2) (second column) and procedure (4.3) (third column). We plot the estimation errors over 10 simulated datasets (corresponding to a line in each figure), for an increasing noise level $\sigma = 0.1$ (first line), $\sigma = 0.5$ (second line), $\sigma = 1.0$ (third line), $\sigma = 5.0$ (fourth line). We can observe that the optimum λ for (4.1) increases with σ (see the position of minimum along the first column), while it can be kept almost constant for procedures (4.2) and (4.3).

Lemma A.2. Suppose that $\rho/\sqrt{\mathrm{rank}(A_0)} \geq \lambda \geq 3\Delta$ for some $\rho < 1$, then

$$\|\hat{A} - \mathbf{X}\|_2 \ \geq \left(\frac{3 - \sqrt{1+\rho^2}}{3 + \sqrt{1+\rho^2}}\right) \|A_0 - \mathbf{X}\|_2. \tag{A.1}$$

Table 1. Average best prediction error (and standard deviation) for experiment 1 of the considered procedures for several values σ. Procedure (4.3) introduced in this paper always leads to a slight improvement.

Noise level σ	0.1	0.5	1.0	5.0
Procedure (4.1)	3.56e-04	1.03e-02	4.01e-02	1.17e+00
	(4.90e-05)	(2.23e-04)	(4.02e-03)	(1.52e-01)
Procedure (4.2)	3.54e-04	8.87e-03	3.54e-02	8.72e-01
	(8.66e-06)	(2.01e-04)	(8.34e-04)	(2.17e-02)
Procedure (4.3)	**3.47e-04**	**8.65e-03**	**3.43e-02**	**8.54e-01**
	(5.16e-06)	(1.44e-04)	(6.73e-04)	(1.56e-02)

Table 2. Average best prediction error (and standard deviation) for experiment 2 of the considered procedures for several values σ. Procedure (4.3) introduced in this paper leads to a strong improvement in this case.

Noise level σ	0.1	0.5	1.0	5.0
Procedure (4.1)	1.50e-02	6.37e-02	2.24e-01	6.87e+00
	(7.82e-03)	(5.59e-03)	(1.42e-02)	(1.17e-01)
Procedure (4.2)	2.05e-03	5.01e-02	2.01e-01	4.95e+00
	(5.37e-05)	(4.93e-04)	(1.79e-03)	(5.63e-02)
Procedure (4.3)	**1.64e-03**	**4.10e-02**	**1.64e-01**	**3.93e+00**
	(2.61e-05)	(4.40e-04)	(2.78e-03)	(5.87e-02)

If $\hat{A} = \mathbf{X}$, then (A.1) implies that $A_0 = \mathbf{X}$ and we get $\|\hat{A} - A_0\|_2 = 0$.

When $\hat{A} \neq \mathbf{X}$, we use the fact that the subdifferential of the convex function $A \to \|A\|_1$ is the following set of matrices (cf., Watson (1992))

$$\partial\|A\|_1 = \left\{ \sum_{j=1}^{\operatorname{rank}(A)} u_j(A) v_j^T(A) + \mathcal{P}_{S_1^\perp(A)} W \mathcal{P}_{S_2^\perp(A)} \ : \ \|W\|_\infty \leq 1 \right\}. \tag{A.2}$$

Here $u_j(A)$ and $v_j(A)$ are respectively the left and right orthonormal singular vectors of A, $S_1(A)$ is the linear span of $\{u_j(A)\}$, $S_2(A)$ is the linear span of $\{v_j(A)\}$. For simplicity we write u_j and v_j instead of $u_j(A)$ and $v_j(A)$. A necessary condition for an extremum in (2.6) implies that there exists $\hat{V} \in \partial\|\hat{A}\|_1$ such that, for any $A \in \mathbb{R}^{m_1 \times m_2}$,

$$\frac{2\langle \hat{A} - \mathbf{X}, \hat{A} - A \rangle}{2\|\hat{A} - \mathbf{X}\|_2} + \lambda \langle \hat{V}, \hat{A} - A \rangle \leq 0. \tag{A.3}$$

By the monotonicity of subdifferentials of convex functions we have that $\langle \hat{V} - V, \hat{A} - A \rangle \geq 0$, where $V \in \partial\|A\|_1$. Then (A.3) and $2\langle \hat{A} - A_0, \hat{A} - A \rangle =$

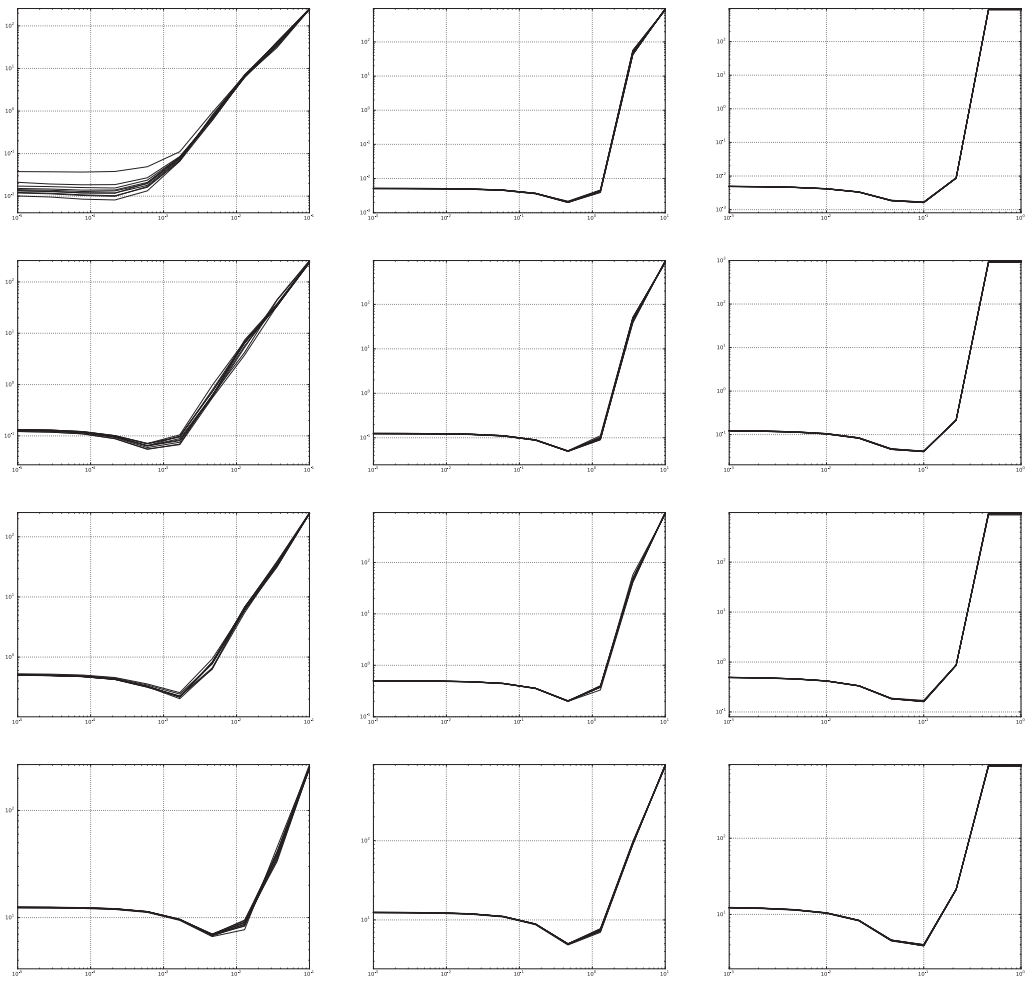

Figure 2. Prediction errors (y-axis) for experiment 2 (see text) for a varying λ (x-axis) for procedure (4.1) (first column), procedure (4.2) (second column) and procedure (4.3) (third column). We plot the estimation errors over 10 simulated datasets (corresponding to a line in each figure), for an increasing noise level $\sigma = 0.1$ (first line), $\sigma = 0.5$ (second line), $\sigma = 1.0$ (third line), $\sigma = 5.0$ (fourth line). We can observe that the optimum λ for (4.1) increases with σ (see the position of minimum along the first column), while it can be kept almost constant for procedures (4.2) and (4.3).

$\|\hat{A} - A_0\|_2^2 + \|\hat{A} - A\|_2^2 - \|A - A_0\|_2^2$ imply that

$$\|\hat{A} - A_0\|_2^2 + \|\hat{A} - A\|_2^2 + 2\lambda\|\hat{A} - \mathbf{X}\|_2 \left\langle \mathcal{P}_{S_1^\perp(A)} W \mathcal{P}_{S_2^\perp(A)}, \hat{A} - A \right\rangle$$
$$\leq \|A - A_0\|_2^2 + 2\langle \mathbf{X} - A_0, \hat{A} - A\rangle - 2\lambda\|\hat{A} - \mathbf{X}\|_2 \left\langle \sum_{j=1}^{r} u_j v_j^T, \hat{A} - A \right\rangle . \quad \text{(A.4)}$$

For B, a $m_1 \times m_2$ matrix, let $\mathbf{Pr}_A(B) = B - \mathcal{P}_{S_1^\perp(A)} B \mathcal{P}_{S_2^\perp(A)}$. Since

$$\mathbf{Pr}_A(B) = \mathcal{P}_{S_1^\perp(A)} B \mathcal{P}_{S_2(A)} + \mathcal{P}_{S_1(A)} B$$

and $\text{rank}(\mathcal{P}_{S_i(A)} B) \leq \text{rank}(A)$ we have that $\text{rank}(\mathbf{Pr}_A(B)) \leq 2\text{rank}(A)$.

Consider each term in (A.4) separately. First, using the trace duality and the triangle inequality, we get

$$\begin{aligned}\langle \mathbf{X} - A_0, \hat{A} - A \rangle &\leq \|\mathbf{X} - A_0\|_\infty \|\hat{A} - A\|_1 \\ &\leq \|\mathbf{X} - A_0\|_\infty \left\| \mathbf{Pr}_A \left(\hat{A} - A \right) \right\|_1 \\ &\quad + \|\mathbf{X} - A_0\|_\infty \left\| \mathcal{P}_{S_1^\perp(A)} \left(\hat{A} - A \right) \mathcal{P}_{S_2^\perp(A)} \right\|_1 . \end{aligned} \tag{A.5}$$

Note that $\left\| \sum_{j=1}^r u_j v_j^T \right\|_\infty = 1$. Then, the trace duality implies that

$$\left\langle \sum_{j=1}^r u_j v_j^T, \hat{A} - A \right\rangle = \left\langle \sum_{j=1}^r u_j v_j^T, \mathbf{Pr}_A \left(\hat{A} - A \right) \right\rangle \leq \left\| \mathbf{Pr}_A \left(\hat{A} - A \right) \right\|_1 . \tag{A.6}$$

From the trace duality, we get that there exists W with $\|W\|_\infty \leq 1$ such that

$$\begin{aligned}\left\langle \mathcal{P}_{S_1^\perp(A)} W \mathcal{P}_{S_2^\perp(A)}, \hat{A} - A \right\rangle &= \left\langle W, \mathcal{P}_{S_1^\perp(A)} \left(\hat{A} - A \right) \mathcal{P}_{S_2^\perp(A)} \right\rangle \\ &= \left\| \mathcal{P}_{S_1^\perp(A)} \left(\hat{A} - A \right) \mathcal{P}_{S_2^\perp(A)} \right\|_1 . \end{aligned} \tag{A.7}$$

Using (A.1) and the definition of λ we derive

$$\begin{aligned}&\lambda \|\hat{A} - \mathbf{X}\|_2 \left\| \mathcal{P}_{S_1^\perp(A)} \, \hat{A} \, \mathcal{P}_{S_2^\perp(A)} \right\|_1 \\ &\geq \lambda \frac{3 - \sqrt{1+\rho^2}}{3 + \sqrt{1+\rho^2}} \|A_0 - \mathbf{X}\|_2 \left\| \mathcal{P}_{S_1^\perp(A)} \, \hat{A} \, \mathcal{P}_{S_2^\perp(A)} \right\|_1 \\ &\geq 3 \frac{3 - \sqrt{1+\rho^2}}{3 + \sqrt{1+\rho^2}} \|A_0 - \mathbf{X}\|_\infty \|\mathcal{P}_{S_1^\perp(A)} \, \hat{A} \, \mathcal{P}_{S_2^\perp(A)}\|_1 . \end{aligned} \tag{A.8}$$

Since $6[3 - \sqrt{1+\rho^2}]/[3 + \sqrt{1+\rho^2}] \geq 2$ for any $\rho < 1$, putting (A.5), (A.6) and (A.8) into (A.4) yields

$$\begin{aligned}\|\hat{A} - A_0\|_2^2 + \|\hat{A} - A\|_2^2 \leq \|A - A_0\|_2^2 &+ 2\|\mathbf{X} - A_0\|_\infty \|\mathbf{Pr}_A \left(\hat{A} - A \right)\|_1 \\ &+ 2\lambda \|\hat{A} - \mathbf{X}\|_2 \|\mathbf{Pr}_A \left(\hat{A} - A \right)\|_1 . \end{aligned} \tag{A.9}$$

Using the triangle inequality and the fact that

$$\left\| \mathbf{Pr}_A \left(\hat{A} - A \right) \right\|_1 \leq \sqrt{2\text{rank}(A)} \|\hat{A} - A\|_2$$

we get

$$
\begin{aligned}
&2\left\|\mathbf{X}-A_0\right\|_\infty \left\|\mathbf{Pr}_A\left(\hat{A}-A\right)\right\|_1 + 2\lambda\left\|\hat{A}-\mathbf{X}\right\|_2 \left\|\mathbf{Pr}_A\left(\hat{A}-A\right)\right\|_1 \\
&\leq 2\Big(\left\|\mathbf{X}-A_0\right\|_\infty + \lambda\left\|\mathbf{X}-A_0\right\|_2\Big)\sqrt{2\mathrm{rank}(A)}\left\|\hat{A}-A\right\|_2 \\
&\quad +2\lambda\|\hat{A}-A_0\|_2\sqrt{2\mathrm{rank}(A)}\|\hat{A}-A\|_2. \qquad \text{(A.10)}
\end{aligned}
$$

From the definition of λ we get that $\left\|\mathbf{X}-A_0\right\|_\infty \leq \lambda\left\|\mathbf{X}-A_0\right\|_2/3$. For A such that $\lambda\sqrt{2\mathrm{rank}(A)} \leq \rho$, (A.10) implies

$$
\begin{aligned}
\|\hat{A}-A_0\|_2^2 + \|\hat{A}-A\|_2^2 \leq \|A-A_0\|_2^2 + \frac{8}{3}\lambda\|\mathbf{X}-A_0\|_2\sqrt{2\mathrm{rank}(A)}\|\hat{A}-A\|_2 \\
+2\rho\|\hat{A}-A_0\|_2\|\hat{A}-A\|_2.
\end{aligned}
$$

Using $2ab \leq a^2+b^2$ twice we compute

$$
(1-\rho)\|\hat{A}-A_0\|_2^2 \leq \|A-A_0\|_2^2 + \frac{4}{1-\rho}\lambda^2\|\mathbf{X}-A_0\|_2^2\mathrm{rank}(A)
$$

which implies the statement of Theorem 1.

A.2. Proof of Lemma A.1.

That $\hat{A}$ is the minimum of (2.6) implies that $0 \in \partial F(\hat{A})$. For $\hat{A} \neq \mathbf{X}$, (A.2) implies that there exists a matrix W such that $\|W\|_\infty \leq 1$ and

$$
\frac{\hat{A}-\mathbf{X}}{\|\hat{A}-\mathbf{X}\|_2} = -\lambda\sum_{j=1}^{\mathrm{rank}(\hat{A})} u_j(\hat{A})v_j^T(\hat{A}) - \lambda\mathcal{P}_{S_1^\perp(\hat{A})}W\mathcal{P}_{S_2^\perp(\hat{A})}. \qquad \text{(A.11)}
$$

Calculating the $\|\cdot\|_2^2$ norm of both sides of (A.11) we get that $1 \geq \lambda^2\mathrm{rank}(\hat{A})$.

When $\hat{A} = \mathbf{X}$, instead of the differential of $\|\hat{A}-\mathbf{X}\|_2$ we use its subdifferential: in (A.11) the term $(\hat{A}-\mathbf{X})/\|\hat{A}-\mathbf{X}\|_2$ is replaced by a matrix $\tilde{W}$ such that $\|\tilde{W}\|_2 \leq 1$ and we get, again, $1 \geq \lambda^2\mathrm{rank}(\hat{A})$.

A.3. Proof of Lemma A.2.

If $A_0 = \mathbf{X}$, then, trivially $\|\hat{A}-\mathbf{X}\|_2 \geq 0$. If $A_0 \neq \mathbf{X}$, by the convexity of the function $A \to \|A-\mathbf{X}\|_2$, we have

$$
\begin{aligned}
\|\hat{A}-\mathbf{X}\|_2 - \|A_0-\mathbf{X}\|_2 &\geq \frac{\langle A_0-\mathbf{X}, \hat{A}-A_0\rangle}{\|A_0-\mathbf{X}\|_2} \\
&\geq -\frac{\|A_0-\mathbf{X}\|_\infty}{\|A_0-\mathbf{X}\|_2}\|\hat{A}-A_0\|_1 \\
&\geq -\frac{\|A_0-\mathbf{X}\|_\infty}{\|A_0-\mathbf{X}\|_2}\sqrt{\mathrm{rank}(\hat{A})+\mathrm{rank}(A_0)}\|\hat{A}-A_0\|_2. \qquad \text{(A.12)}
\end{aligned}
$$

Using Lemma A.1, the bound $\rho/\sqrt{\text{rank}(A_0)} \geq \lambda$, and the triangle inequality, from (A.12) we get

$$\|\hat{A} - \mathbf{X}\|_2 - \|A_0 - \mathbf{X}\|_2 \geq -\frac{\sqrt{1+\rho^2}}{\lambda} \frac{\|A_0 - \mathbf{X}\|_\infty}{\|A_0 - \mathbf{X}\|_2} \left(\|\hat{A} - \mathbf{X}\|_2 + \|A_0 - \mathbf{X}\|_2 \right). \tag{A.13}$$

Note that $(\|A_0 - \mathbf{X}\|_\infty)/(\lambda \|A_0 - \mathbf{X}\|_2) \leq 1/3$, which finally leads to

$$\Big(1 + \frac{\sqrt{1+\rho^2}}{3}\Big) \|\hat{A} - \mathbf{X}\|_2 \ \geq \Big(1 - \frac{\sqrt{1+\rho^2}}{3}\Big) \|A_0 - \mathbf{X}\|_2,$$

and completes the proof of Lemma A.2.

A.4. Proof of Lemma 1.

Our goal is to get a numerical estimate of c_* in the case of Gaussian noise. Let $Z_i = \xi_i \, (X_i - \mathbb{E}X_i)$ and

$$\sigma_Z = \max \Big\{ \Big\| \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left(Z_i Z_i^T \right) \Big\|_\infty^{1/2}, \Big\| \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left(Z_i^T \, Z_i \right) \Big\|_\infty^{1/2} \Big\} = \frac{1}{m_1 \wedge m_2}.$$

The constant c_* comes up in the proof of Lemma 2 in Koltchinskii, Lounici and Tsybakov (2011) in the estimation of

$$\Delta_1 = \Big\| \frac{1}{n} \sum_{i=1}^n \xi_i X_i \Big\|_\infty \leq \Big\| \frac{1}{n} \sum_{i=1}^n \xi_i \Big(X_i - \mathbb{E}X_i \Big) \Big\|_\infty + \frac{1}{\sqrt{m_1 m_2}} \Big| \frac{1}{n} \sum_{i=1}^n \xi_i \Big|.$$

A standard application of Markov's inequality gives that, with probability at least $1 - 1/m$

$$\frac{1}{\sqrt{m_1 m_2}} \Big| \frac{1}{n} \sum_{i=1}^n \xi_i \Big| \leq 2 \sqrt{\frac{\log m}{n m_1 m_2}}. \tag{A.14}$$

In Koltchinskii, Lounici and Tsybakov (2011), the authors estimate $\|(1/n) \sum_{i=1}^n \xi_i (X_i - \mathbb{E}X_i)\|_\infty$ using Koltchinskii (2011, Proposition 2). To get a numerical estimate of c_* we follow the lines of this proof. In order to simplify notation, we write $\| \ \|_\infty = \| \ \|$ and consider the case of Hermitian matrices of size m'. The extension to rectangular matrices is straightforward via self-adjoint dilation, cf., for example, 2.6 in Tropp (2011).

Let $Y_n = \sum_{i=1}^n Z_i$. In the proof of Koltchinskii (2011, Proposition 2), after following the standard derivation of the classical Bernstein inequality and using the Golden-Thompson inequality, the author finds that

$$\mathbb{P} \left(\|Y_n\| \geq t \right) \leq 2m' e^{-\lambda t} \|\mathbb{E} e^{\lambda Z_1}\|^n \tag{A.15}$$

and

$$\|\mathbb{E} e^{\lambda Z_1}\| \leq 1 + \lambda^2 \Big\| \mathbb{E} Z_1^2 \Big[\frac{e^{\lambda \|Z_1\|} - 1 - \lambda \|Z_1\|}{\lambda^2 \|Z_1\|^2} \Big] \Big\|. \tag{A.16}$$

Using that $\|Z_1\| \leq 2|\xi_i|$, from (A.16), we compute

$$\begin{aligned}\left\|\mathbb{E}e^{\lambda Z_1}\right\| &\leq 1 + \lambda^2 \left\|\mathbb{E}[(X_i - \mathbb{E}X_i)^2]\mathbb{E}\Big(\xi_i^2\Big[\frac{e^{2\lambda|\xi_i|} - 1 - 2\lambda|\xi_i|}{4\lambda^2\xi_i^2}\Big]\Big)\right\| \\ &\leq 1 + \lambda^2\sigma_Z^2\mathbb{E}\left(\frac{(2|\xi_i|)^2}{2!} + \frac{\lambda(2|\xi_i|)^3}{3!} + \cdots\right). \qquad \text{(A.17)}\end{aligned}$$

If $\lambda < 1$, then (A.17) implies

$$\left\|\mathbb{E}e^{\lambda Z_1}\right\| \leq 1 + \lambda^2\sigma_Z^2\mathbb{E}e^{2|\xi_i|} \leq 1 + 2\lambda^2\sigma_Z^2e^2 \leq \exp\{2\lambda^2\sigma_Z^2e^2\}.$$

Using this bound, from (A.15) we get

$$\mathbb{P}\left(\|Y_n\| \geq t\right) \leq 2m' \exp\{-\lambda t + 2\lambda^2\sigma_Z^2e^2\}.$$

It remains now to minimize the last bound with respect to $\lambda \in (0,1)$ to obtain that

$$\mathbb{P}\left(\|Y_n\| \geq t\right) \leq 2m' \exp\left\{-\frac{t^2}{4e^2\sigma_Z^2 n}\right\},$$

where we have supposed that n is large enough.

Putting $2m' \exp\left\{-t^2/(4\sigma_Z^2e^2n)\right\} = 1/(2m')$, we get $t = 2e\sqrt{2\log(2m')n/(m_1 \wedge m_2)}$. Using (A.14) we compute that $c_* \leq 2e+1 \leq 6.5$. This completes the proof of Lemma 1.

A.5. Proof of Lemma 2.

Let $\epsilon_i = \sigma\xi_i$. To prove (i) we compute

$$\begin{aligned}\langle \mathbf{M}, \mathbf{M}\rangle = \frac{\|A_0\|_2^2}{(m_1m_2)^2} &+ \Big(1 - \frac{2n}{m_1m_2}\Big)\underbrace{\frac{1}{n^2}\sum_{i=1}^n \langle A_0, X_i\rangle^2}_{\mathbf{I}} + \underbrace{\frac{1}{n^2}\sum_{i=1}^n \epsilon_i^2}_{\mathbf{II}} \\ &+ \Big(1 - \frac{n}{m_1m_2}\Big)\underbrace{\frac{2}{n^2}\sum_{i=1}^n \langle A_0, X_i\rangle\,\epsilon_i}_{\mathbf{III}} + \underbrace{\frac{4}{n^2}\sum_{i<j} \epsilon_i \langle A_0, X_j\rangle \langle X_i, X_j\rangle}_{\mathbf{IV}} \\ &+ \underbrace{\frac{2}{n^2}\sum_{i<j} \epsilon_i\epsilon_j \langle X_i, X_j\rangle}_{\mathbf{V}} + \underbrace{\frac{1}{n^2}\sum_{i\neq j} \langle A_0, X_i\rangle \langle A_0, X_j\rangle \langle X_j, X_i\rangle}_{\mathbf{VI}}. \qquad \text{(A.18)}\end{aligned}$$

We estimate each term in (A.18) separately, with a good probability.

I : We have that $\mathbb{E}((1/n^2)\sum_{i=1}^n \langle A_0, X_i\rangle^2) = \|A_0\|_2^2/nm_1m_2$ and $|\langle A_0, X_i\rangle| \leq a$. Using Hoeffding's inequality, we get that, with probability at least $1 - 2\exp\{-2\sigma^4 n/(8a)^2\}$

$$\frac{\|A_0\|_2^2}{nm_1m_2} + \frac{\sigma^2}{8\,n} \geq \frac{1}{n^2}\sum_{i=1}^{n} \langle A_0, X_i\rangle^2 \geq \frac{\|A_0\|_2^2}{nm_1m_2} - \frac{\sigma^2}{8\,n}.$$

II: ϵ_i^2 are sub-exponential random variables and $\mathbb{E}((1/n^2)\sum_{i=1}^n \epsilon_i^2) = \sigma^2/n$. Using Bernstein inequality for sub-exponentials random variables Vershynin (2012, Proposition 16) we get that, with probability at least $1 - 2\exp\{-cn\min[\sigma^2 K/8^2, \sigma\sqrt{K}/8]\}$

$$\frac{\sigma^2}{n} + \frac{\sigma^2}{8n} \geq \frac{1}{n^2}\sum_{i=1}^{n} \epsilon_i^2 \geq \frac{\sigma^2}{n} - \frac{\sigma^2}{8n}.$$

III: We have that $\mathbb{E}((2/n^2)\sum_{i=1}^n \langle A_0, X_i\rangle \epsilon_i) = 0$, using Hoeffding's type inequality for sub-Gaussian random variables Vershynin (2012, Proposition 10) we get that, with probability at least $1 - e\exp\left\{-c\sigma^2 Kn/a^2\right\}$

$$\frac{\sigma^2}{8n} \geq \frac{2}{n^2}\sum_{i=1}^{n} \langle A_0, X_i\rangle \epsilon_i \geq -\frac{\sigma^2}{8n}.$$

IV: We compute $\mathbb{E}((4/n^2)\sum_{i<j} \epsilon_i \langle A_0, X_j\rangle \langle X_i, X_j\rangle) = 0$. The following lemma is proven in Appendix A.6.

Lemma A.3. Suppose that $n \leq m_1m_2$. With probability at least $1 - 2/(m_1m_2)$,

$$\sum_{i<j} \langle X_i, X_j\rangle \leq n.$$

Lemma A.3 and a Hoeffding-type inequality imply that, with probability at least $1 - 2/m_1m_2 - e\exp\left\{-c\sigma^2 nK/a^2\right\}$,

$$\frac{\sigma^2}{8n} \geq \frac{4}{n^2}\sum_{i<j} \epsilon_i \langle A_0, X_j\rangle \langle X_i, X_j\rangle \geq -\frac{\sigma^2}{8n}.$$

V: We have that $\mathbb{E}\left(2/n^2 \sum_{i<j} \epsilon_i\epsilon_j \langle X_i, X_j\rangle\right) = 0$. Using the Bernstein inequality for sub-exponentials random variables Vershynin (2012, Proposition 16) and Lemma A.3 we get that, with probability at least $1 - 2\exp\{-cn\min[\sigma^2 K/8^2, \sigma\sqrt{K}/8]\}$

$$\frac{\sigma^2}{8n} \geq \frac{2}{n^2}\sum_{i<j} \epsilon_i\epsilon_j \langle X_i, X_j\rangle \geq -\frac{\sigma^2}{8n}.$$

VI: We compute that

$$\begin{aligned}\mathbb{E}\left(\frac{1}{n^2}\sum_{i\neq j} \langle A_0, X_i\rangle \langle A_0, X_j\rangle \langle X_j, X_i\rangle\right) &= \frac{1}{n^2}\sum_{i\neq j} \langle \mathbb{E}\left(\langle A_0, X_j\rangle X_j\right), \mathbb{E}\left(\langle A_0, X_i\rangle X_i\right)\rangle \\ &= \frac{1}{n^2}\sum_{i\neq j} \frac{\|A_0\|_2^2}{(m_1m_2)^2} \\ &\leq \frac{\|A_0\|_2^2}{(m_1m_2)^2}.\end{aligned}$$

Using Lemma A.3 and Hoeffding's type inequality for sub-Gaussian random variables (cf., Vershynin (2012, Proposition 10)), we get that, with probability at least $1 - 2/m_1m_2 - 2\exp\left\{-2\sigma^4 n/(8a)^2\right\}$

$$\frac{1}{n^2}\sum_{i\neq j}\langle A_0, X_i\rangle\langle A_0, X_j\rangle\langle X_j, X_i\rangle \leq \frac{\|A_0\|_2^2}{(m_1m_2)^2} + \frac{\sigma^2}{8n}.$$

To obtain the lower bound, note that, for $i \neq j$, $\langle X_i, X_j\rangle \neq 0$ iff $X_i = X_j$. This implies that $\sum_{i\neq j}\langle A_0, X_i\rangle\langle A_0, X_j\rangle\langle X_j, X_i\rangle \geq 0$. We use that $2n < m_1m_2$ to get

$$\frac{\|A_0\|_2^2}{(m_1m_2)^2} + \left(1 - \frac{2n}{m_1m_2}\right)\frac{1}{n^2}\sum_{i=1}^{n}\langle A_0, X_i\rangle^2 \geq 0.$$

Putting the lower bounds in $\mathbf{II} - \mathbf{V}$ together we compute from (A.18)

$$\|\mathbf{M}\|_2^2 \geq \frac{\sigma^2}{2\,n}.$$

To obtain the upper bound, we use the upper bounds in $\mathbf{I} - \mathbf{VI}$. From (A.18) we get

$$\|\mathbf{M}\|_2^2 \leq \frac{2\,\|A_0\|_2^2}{(m_1m_2)^2} + \frac{\|A_0\|_2^2}{nm_1m_2} + \frac{14\sigma^2}{8\,n} \leq 2\left(\frac{\|A_0\|_2^2}{nm_1m_2} + \frac{\sigma^2}{n}\right),$$

where we have used that $2n \leq m_1m_2$. This completes the proof of part (i) in Lemma 2.

To prove (ii) we use that $\langle X_i, X_i\rangle = 1$ and $\langle X_i, X_j\rangle \neq 0$ iff $X_i = X_j$. We compute

$$\begin{aligned}
\frac{1}{n^2}\left\langle\sum_{i=1}^{n}Y_iX_i, \sum_{i=1}^{n}Y_iX_i\right\rangle &= \frac{1}{n^2}\sum_{i=1}^{n}Y_i^2 + \frac{2}{n^2}\sum_{i<j}Y_iY_j\langle X_i, X_j\rangle \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left(\langle A_0, X_i\rangle^2 + \epsilon_i^2 + 2\langle A_0, X_i\rangle\epsilon_i\right) \\
&\quad + \frac{2}{n^2}\sum_{i<j}\langle A_0, X_i\rangle^2\langle X_i, X_j\rangle \\
&\quad + \frac{4}{n^2}\sum_{i<j}\epsilon_i\langle A_0, X_j\rangle\langle X_i, X_j\rangle + \frac{2}{n^2}\sum_{i<j}\epsilon_i\epsilon_j\langle X_i, X_j\rangle.
\end{aligned}$$

This implies that

$$\frac{1}{n^2}\left\langle \sum_{i=1}^n Y_iX_i, \sum_{i=1}^n Y_iX_i \right\rangle \geq \underbrace{\frac{1}{n^2}\sum_{i=1}^n \langle A_0, X_i\rangle^2}_{\mathbf{I}} + \underbrace{\frac{1}{n^2}\sum_{i=1}^n \epsilon_i^2}_{\mathbf{II}} + \underbrace{\frac{2}{n^2}\sum_{i=1}^n \langle A_0, X_i\rangle\, \epsilon_i}_{\mathbf{III}}$$
$$+ \underbrace{\frac{4}{n^2}\sum_{i<j} \epsilon_i \langle A_0, X_j\rangle \langle X_i, X_j\rangle}_{\mathbf{IV}} + \underbrace{\frac{2}{n^2}\sum_{i<j} \epsilon_i\epsilon_j \langle X_i, X_j\rangle}_{\mathbf{V}}. \tag{A.19}$$

Using the lower bounds for $\mathbf{I} - \mathbf{V}$ we get from (A.19) that

$$\frac{1}{n^2}\left\langle \sum_{i=1}^n Y_iX_i, \sum_{i=1}^n Y_iX_i \right\rangle \geq \frac{\|A_0\|_2^2}{nm_1m_2}$$

which proves the part (ii) of Lemma 2.

(iii) is a consequence of (ii). For $4n \leq m_1m_2$ (ii) implies

$$\frac{1}{4n^2}\left\langle \sum_{i=1}^n Y_iX_i, \sum_{i=1}^n Y_iX_i \right\rangle \geq \frac{\|A_0\|_2^2}{(m_1m_2)^2}.$$

Now we complete the proof of part (iii) of Lemma 2 using that

$$\|\mathbf{M}\|_2 \geq \Big\|\frac{1}{n}\sum_{i=1}^n Y_iX_i\Big\|_2 - \frac{\|A_0\|_2}{m_1m_2}.$$

A.6. Proof of Lemma A.3

For $i \neq j$, X_i and X_j are independent. We compute the expectation

$$\mathbb{E}\Big(\sum_{i<j}\langle X_i, X_j\rangle\Big) = \sum_{i<j}\langle \mathbb{E}X_i, \mathbb{E}X_j\rangle = \frac{n(n-1)}{2m_1m_2}$$

and the variance

$$\mathbb{E}\Big(\Big(\sum_{i<j}\langle X_i, X_j\rangle\Big)^2\Big) - \Big(\mathbb{E}\Big(\sum_{i<j}\langle X_i, X_j\rangle\Big)\Big)^2$$
$$= \mathbb{E}\Big(\sum_{\substack{i<j\\ i'<j'}}\langle X_i, X_j\rangle\langle X_{i'}, X_{j'}\rangle\Big) - \sum_{\substack{i<j\\ i'<j'}}\mathbb{E}\Big(\langle X_i, X_j\rangle\Big)\mathbb{E}\left(\langle X_{i'}, X_{j'}\rangle\right).$$

When i, j, i', j' are all distinct, $\mathbb{E}\left(\langle X_i, X_j\rangle\langle X_{i'}, X_{j'}\rangle\right)$ is cancelled by the corresponding term in $\sum_{\substack{i<j\\ i'<j'}}\mathbb{E}\left(\langle X_i, X_j\rangle\right)\mathbb{E}\left(\langle X_{i'}, X_{j'}\rangle\right)$.

It remains to consider the five cases: (1) $i = i'$ and $j = j'$; (2) $i = i'$ and $j \neq j'$; (3) $i \neq i'$ and $j = j'$; (4) $i = j'$ and $j \neq i'$; (5) $i' = j$ and $j' \neq i$.

Case (1): As $\langle X_i, X_j \rangle$ takes only two values 0 or 1,

$$\mathbb{E}\left(\langle X_i, X_j \rangle^2\right) = \mathbb{E}\left(\langle X_i, X_j \rangle\right) = \frac{1}{m_1 m_2}.$$

Cases (2)-(5): in these cases, we need to calculate $\mathbb{E}\left(\langle X_i, X_k \rangle \langle X_k, X_j \rangle\right)$ for $i \neq j$ and $k \notin \{i, j\}$. Note that $\mathcal{P}_{X_k} = \langle \cdot, X_k \rangle X_k$ is the orthogonal projector on the vector space spanned by X_k. We compute

$$\mathbb{E}\mathcal{P}_{X_k} = \frac{1}{m_1 m_2}\mathrm{Id},$$

where Id is the identity application on $\mathbb{R}^{m_1 \times m_2}$. Then, we get

$$\begin{aligned}
\mathbb{E}\left(\langle \langle X_i, X_k \rangle X_k, X_j \rangle\right) &= \mathbb{E}\left(\langle \mathcal{P}_{X_k}\left(X_i\right), X_j \rangle\right) \\
&= \langle \mathbb{E}\left(\mathcal{P}_{X_k}\right)\left(\mathbb{E}X_i\right), \mathbb{E}X_j \rangle \\
&= \frac{1}{m_1 m_2} \langle \mathbb{E}X_i, \mathbb{E}X_j \rangle \\
&= \frac{1}{\left(m_1 m_2\right)^2}.
\end{aligned}$$

These terms are cancelled by the corresponding terms in $\sum_{\substack{i<j \\ i'<j'}} \mathbb{E}(\langle X_i, X_j \rangle)\mathbb{E}(\langle X_{i'}, X_{j'} \rangle)$ as

$$\mathbb{E}\left(\langle X_i, X_k \rangle\right)\mathbb{E}\left(\langle X_k, X_j \rangle\right) = \frac{1}{\left(m_1 m_2\right)^2}.$$

Finally we get that

$$\mathbb{E}\Big(\Big(\sum_{i<j} \langle X_i, X_j \rangle\Big)^2\Big) - \Big(\mathbb{E}\Big(\sum_{i<j} \langle X_i, X_j \rangle\Big)\Big)^2 \leq \frac{n(n-1)}{2m_1 m_2}.$$

The Bienaymé-Tchebychev inequality implies that

$$\mathbb{P}\Big(\sum_{i<j} \langle X_i, X_j \rangle \geq n\Big) \leq \frac{n(n-1)}{2m_1 m_2 \left(n - n(n-1)/2m_1 m_2\right)^2} \leq \frac{2}{m_1 m_2}$$

when $m_1 m_2 \geq n$. This completes the proof of Lemma A.3.

A.7. Proof of Theorem 3.

The following lemma is the counterpart of Lemma A.1 in the present setting. It is proven in Appendix A.8.

Lemma A.4. $\operatorname{rank}(V\hat{A}) \leq 1/\lambda^2$.

We need an auxiliary result that corresponds to Lemma A.2; it is proven in Appendix A.9.

Lemma A.5. Suppose that $\rho/\sqrt{\operatorname{rank}(VA_0)} \geq \lambda \geq 3\Delta'$ for some $\rho < 1$, then

$$\left\|V\hat{A} - U\right\|_2 \geq \left(\frac{3-\sqrt{1+\rho^2}}{3+\sqrt{1+\rho^2}}\right)\|E\|_2.$$

The proof of Theorem 3 is similar to the proof of the Theorem 1. We only sketch it. If $V\hat{A} \neq U$, a necessary condition of the extremum in (3.1) implies that there exists a $\hat{W} \in \partial\|V\hat{A}\|_1$ such that, for any $A \in \mathbb{R}^{m_1 \times m_2}$,

$$\frac{2\left\langle V\hat{A} - U, V\left(\hat{A} - A\right)\right\rangle}{2\left\|V\hat{A} - U\right\|_2} + \lambda\left\langle \hat{W}, V\left(\hat{A} - A\right)\right\rangle \leq 0 \tag{A.20}$$

and we get

$$\begin{aligned}
&\left\|V\left(\hat{A}-A_0\right)\right\|_2^2 + \left\|V\left(\hat{A}-A\right)\right\|_2^2 + 2\lambda\left\|V\hat{A}-U\right\|_2\left\langle \mathcal{P}_{S_1^\perp(VA)} W \mathcal{P}_{S_2^\perp(VA)}, V\left(\hat{A}-A\right)\right\rangle \\
&\quad \leq \left\|V\left(A-A_0\right)\right\|_2^2 + 2\left\langle E, V\left(\hat{A}-A\right)\right\rangle \\
&\qquad -2\lambda\left\|V\hat{A}-U\right\|_2\left\langle \sum_{j=1}^{\operatorname{rank}(VA)} u_j(VA)v_j(VA)^T, V\left(\hat{A}-A\right)\right\rangle.
\end{aligned} \tag{A.21}$$

Let $\mathbf{Pr}_{VA}(B) = B - \mathcal{P}_{S_1^\perp(VA)} B \mathcal{P}_{S_2^\perp(VA)}$. Then, the trace duality and the triangle inequality imply that

$$\begin{aligned}
\left\langle E, V\left(\hat{A}-A\right)\right\rangle &= \left\langle \mathcal{P}_V E, V\left(\hat{A}-A\right)\right\rangle \\
&\leq \|\mathcal{P}_V E\|_\infty \left\|V\left(\hat{A}-A\right)\right\|_1 \\
&\leq \|\mathcal{P}_V E\|_\infty \left\|\mathbf{Pr}_{VA}\left[V\left(\hat{A}-A\right)\right]\right\|_1 \\
&\quad + \|\mathcal{P}_V E\|_\infty \left\|\mathcal{P}_{S_1^\perp(VA)} V\left(\hat{A}-A\right)\mathcal{P}_{S_2^\perp(VA)}\right\|_1.
\end{aligned} \tag{A.22}$$

Using $6 \times [(3-\sqrt{1+\rho^2})/(3+\sqrt{1+\rho^2})] \geq 2$ for any $\rho < 1$ (A.21) implies that

$$\begin{aligned}
&\left\|V\left(\hat{A}-A_0\right)\right\|_2^2 + \left\|V\left(\hat{A}-A\right)\right\|_2^2 \\
&\quad \leq \left\|V\left(A-A_0\right)\right\|_2^2 + 2\|\mathcal{P}_V E\|_\infty \left\|\mathbf{Pr}_{VA}\left[V\left(\hat{A}-A\right)\right]\right\|_1 \\
&\qquad +2\lambda\left\|V\hat{A}-U\right\|_2 \left\|\mathbf{Pr}_{VA}\left[V\left(\hat{A}-A\right)\right]\right\|_1.
\end{aligned} \tag{A.23}$$

Now we use $\left\|\mathbf{Pr}_{VA}\left[V\left(\hat{A}-A\right)\right]\right\|_1 \leq \sqrt{2\mathrm{rank}(VA)}\left\|V\left(\hat{A}-A\right)\right\|_2$, $\|\mathcal{P}_V E\|_\infty \leq \lambda\|E\|_2/3$ and $\lambda\sqrt{2\,\mathrm{rank}(VA)} \leq \rho$ to conclude that

$$(1-\rho)\left\|V\left(\hat{A}-A_0\right)\right\|_2^2 \leq \left\|V\left(A-A_0\right)\right\|_2^2 + \frac{4\lambda^2}{1-\rho}\|E\|_2^2\,\mathrm{rank}(VA),$$

which implies the statement of Theorem 3.

A.8. Proof of Lemma A.4

That $\hat{A}$ is the minimum of (3.1) implies that $0 \in \partial G(\hat{A})$ where $G = \|U - VA\|_2 + \lambda\|VA\|_1$. Note that the subdifferential of the convex function $A \to \|VA\|_1$ is the following set of matrices

$$\partial\|VA\|_1 = V^T\left\{\sum_{j=1}^{\mathrm{rank}(VA)} u_j(VA)v_j^T(VA) + \mathcal{P}_{S_1^\perp(VA)}W\mathcal{P}_{S_2^\perp(VA)} \;:\; \|W\|_\infty \leq 1\right\},$$

where $S_1(VA)$ is the linear span of $\{u_j(VA)\}$ and $S_2(VA)$ is the linear span of $\{v_j(VA)\}$.

If $\hat{A}$ is such that $V\hat{A} \neq U$, we obtain that there exists a matrix W such that $\|W\|_\infty \leq 1$ and

$$V^T\frac{V\hat{A}-U}{\|V\hat{A}-U\|_2} = -\lambda V^T\left\{\sum_{j=1}^{\mathrm{rank}(VA)} u_j(VA)v_j^T(VA) + \mathcal{P}_{S_1^\perp(VA)}W\mathcal{P}_{S_2^\perp(VA)}\right\}.$$

This implies

$$V^T\mathcal{P}_V\frac{V\hat{A}-U}{\|V\hat{A}-U\|_2} = -\lambda V^T\mathcal{P}_V\left\{\sum_{j=1}^{\mathrm{rank}(VA)} u_j(VA)v_j^T(VA) + \mathcal{P}_{S_1^\perp(VA)}W\mathcal{P}_{S_2^\perp(VA)}\right\}. \tag{A.24}$$

Using $\mathcal{P}_V VA\left(v_j(VA)\right) = VA\left(v_j(VA)\right) = \sigma_j(VA)u_j(VA)$ and $\sigma_j \neq 0$ we get

$$\mathcal{P}_V u_j(VA) = u_j(VA). \tag{A.25}$$

For any w such that $\langle w, u_j(VA)\rangle = 0$ (A.25) implies that

$$\langle \mathcal{P}_V w, u_j(VA)\rangle = \langle w, u_j(VA)\rangle = 0. \tag{A.26}$$

By definition, $\mathcal{P}_{S_1^\perp(VA)}$ projects on the orthogonal complement of the linear span of $\{u_j(VA)\}$. Thus, (A.26) implies that $\mathcal{P}_V\mathcal{P}_{S_1^\perp(VA)}$ also projects on the subspace orthogonal to the linear span of $\{u_j(VA)\}$.

Note that $V^T\mathcal{P}_V B = 0$ implies $\mathcal{P}_V B = 0$ and we get from (A.24)

$$\mathcal{P}_V \frac{V\hat{A} - U}{\|V\hat{A} - U\|_2} = -\lambda\Big\{ \sum_{j=1}^{\mathrm{rank}(VA)} u_j(VA)v_j^T(VA) + \mathcal{P}_V\left[\mathcal{P}_{S_1^\perp(VA)} W \mathcal{P}_{S_2^\perp(VA)}\right]\Big\}. \tag{A.27}$$

Calculating the $\|\ \|_2^2$ norm of both sides on (A.27) we get that $1 \geq \lambda^2 \mathrm{rank}(V\hat{A})$. When $V\hat{A} = U$, instead of the differential of $\|U - VA\|_2$ we use its subdiffential.

A.9. Proof of Lemma A.4

If $VA_0 = U$, then we have, trivially, $\|V\hat{A} - U\|_2 \geq 0$. If $VA_0 \neq U$, by the convexity of function $A \to \|VA - U\|_2$, we have

$$\begin{aligned}
&\left\|V\hat{A} - U\right\|_2 - \|VA_0 - U\|_2 \\
&\quad \geq \frac{\left\langle VA_0 - U, V\left(\hat{A} - A_0\right)\right\rangle}{\|VA_0 - U\|_2} \\
&\quad = \frac{\left\langle \mathcal{P}_V(E), V\left(\hat{A} - A_0\right)\right\rangle}{\|VA_0 - U\|_2} \\
&\quad \geq -\frac{\|\mathcal{P}_V(E)\|_\infty}{\|E\|_2}\left\|V\left(\hat{A} - A_0\right)\right\|_1 \\
&\quad \geq -\frac{\|\mathcal{P}_V(E)\|_\infty}{\|E\|_2}\sqrt{\mathrm{rank}(VA_0) + \mathrm{rank}(V\hat{A})}\left\|V\left(\hat{A} - A_0\right)\right\|_2.
\end{aligned} \tag{A.28}$$

Using the bound $\rho/\sqrt{\mathrm{rank}(VA)} \geq \lambda$, Lemma A.4, and the triangle inequality from (A.28) we get

$$\begin{aligned}
&\left\|V\hat{A} - U\right\|_2 - \|VA_0 - U\|_2 \\
&\quad \geq -\frac{\sqrt{1+\rho^2}}{\lambda}\frac{\|\mathcal{P}_V(E)\|_\infty}{\|E\|_2}\left(\|V\hat{A} - U\|_2 + \|VA_0 - U\|_2\right).
\end{aligned}$$

By the definition of λ we have $\|\mathcal{P}_V(E)\|_\infty/\lambda\|E\|_2 \leq 1/3$. This leads to

$$\left(1 + \frac{\sqrt{1+\rho^2}}{3}\right)\|V\hat{A} - U\|_2 \geq \left(1 - \frac{\sqrt{1+\rho^2}}{3}\right)\|VA_0 - U\|_2,$$

and completes the proof of Lemma A.4.

Acknowledgements

It is a pleasure to thank A. Tsybakov for introducing us to this problem and for illuminating discussions.

References

Bauschke, H. H. and Combettes, P. L. (2011). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2**, 183-202.

Becker, S. R., Candès, E. J. and Grant, M. C. (2011). Templates for convex cone problems with applications to sparse signal recovery. *Math. Programm. Computat.* **3**, 165-218.

Belloni, A., Chernozhukov, V. and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika* 98, 791-806.

Bunea, F., She, Y. and Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.* **39**, 1282-1309.

Candes, E. J. and Plan, Y. (2010). Matrix completion with noise. *Proc. IEEE* **98**, 925-936.

Candès, E. J. and Recht, B. (2009). Exact matrix completion via convex optimization. *Found. Computat. Math.* **9**, 717–772.

Candès, E. J. and Tao, T. (2010). The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inform. Theory* **56**, 2053-2080.

Chambolle, A. and Pock, T. (2011). A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision* **40**, 120-145.

Gaiffas, S. and Lecué, G. (2011). Sharp oracle inequalities for high-dimensional matrix prediction. *IEEE Trans. Inform. Theory* **57**, 6942-6957.

Giraud, C. (2011). Low rank multivariate regression. *Electron. J. Statist.* **5**, 775-799.

Giraud, C., Huet, S. and Verzelen, N. (2012). High-dimensional regression with unknown variance. *Statist. Sci.* **27**, 500-518.

Gross, D. (2011). Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory* **57**, 1548-1566.

Keshavan, R. H., Montanari, A. and Oh, S. (2010). Matrix completion from noisy entries. *J. Mach. Learn. Res.* **11**, 2057-2078.

Keshavan, R.H., Montanari, A. and Oh, S. (2010). Matrix completion from a few entries. *IEEE Trans. Inform. Theory* **56**, 2980-2998.

Klopp, O. (2011). Rank penalized estimators for high-dimensional matrices. *Electron. J. Statist.* **5**, 1161-1183.

Koltchinskii, V. (2011). Von Neumann entropy penalization and low rank matrix estimation. *Ann. Statist.* **39**, 2936-2973.

Koltchinskii, V., Lounici, K. and Tsybakov, A. (2011). Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *Ann. Statist.* **39**, 2302-2329.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.* **39**, 1069-1097.

Negahban, S. and Wainwright, M. J. (2010). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *J. Mach. Learn. Res.* **13**, 1665-1697.

Recht, B. (2009). A simpler approach to matrix completion. *J. Mach. Learn. Res.* **12**, 3413-3430.

Rohde, A. and Tsybakov, A. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.* **39**, 887-930.

Städler, N., Bühlmann, P. and van de Geer, S. (2010). l_1-penalization for mixture regression models. *TEST* **19**, 209-256.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99**, 879-898.

Tropp, J.A. (2011). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **11**.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing, Theory and Applications* (Edited by Y. Eldar and G. Kutyniok). Cambridge University Press.

Watson, G. A. (1992). Characterization of the subdifferential of some matrix norms. *Linear Algebra Appl.* **170**, 33-45.

Centre de Mathématiques Appliquées, École Polytechnique, route de Saclay 91128 Palaiseau Cedex, France.

E-mail: stephane.gaiffas@cmap.polytechnique.fr

MODAL'X Université Paris Ouest Nanterre la Défense, 200 avenue de la République, 92001 Nanterre, France.

E-mail: kloppolga@math.cnrs.fr

(Received June 2013; accepted January 2016)