Estimation of matrices with row sparsity

Olga Klopp, Alexandre B. Tsybakov

Estimation of matrices with row sparsity

O. Klopp, A. B. Tsybakov

September 1, 2015

Abstract

An increasing number of applications is concerned with recovering a sparse matrix from noisy observations. In this paper, we consider the setting where each row of the unknown matrix is sparse. We establish minimax optimal rates of convergence for estimating matrices with row sparsity. A major focus in the present paper is on the derivation of lower bounds.

1 Introduction

In recent years, there has been a great interest for the theory of estimation in high-dimensional statistical models under different sparsity scenarii. The main motivation behind sparse estimation is based on the observation that, in several practical applications, the number of variables is much larger than the number of observations, but the degree of freedom of the underlying model is relatively small. One example of such sparse estimation is the problem of estimating of a sparse regression vector from a set of linear measurements (see, e.g., [2], [5], [16], [23]). Another example is the problem of matrix recovery under the assumption that the unknown matrix has low rank (see, e.g., [8, 20, 14, 15]).

In some recent papers dealing with covariance matrix estimation, a different notion of sparsity was considered (see, for example, [7], [19]). This notion is based on sparsity assumptions on the rows (or columns) $M_{i\cdot}$ of matrix M. One can consider the hard sparsity assumption meaning that each row $M_{i\cdot}$ of M contains at most s non-zero elements, or soft sparsity assumption, based on imposing a certain decay rate on ordered entries of $M_{i\cdot}$. These notions of sparsity can be defined in terms of l_q-balls for $q \in [0,2)$, defined as

$$\mathbb{B}_q(s) = \left\{ v = (v_i) \in \mathbb{R}^{n_2} \, : \, \sum_{i=1}^{n_2} |v_i|^q \le s \right\} \tag{1}$$

where $s < \infty$ is a given constant. The case $q = 0$

$$\mathbb{B}_0(s) = \left\{ v = (v_i) \in \mathbb{R}^{n_2} \, : \, \sum_{i=1}^{n_2} \mathbb{I}(v_i \neq 0) \le s \right\} \tag{2}$$

corresponds to the set of vectors v with at most s non-zero elements. Here $\mathbb{I}(\cdot)$ denotes the indicator function and $s \geq 1$ is an integer.

In the present note, we consider this row sparsity setting in the matrix signal plus noise model. Suppose we have noisy observations $Y = (y_{ij})$ of an $n_1 \times n_2$ matrix $M = (m_{ij})$ where

$$y_{ij} = m_{ij} + \xi_{ij}, \quad i = 1, \dots, n_1, \quad j = 1, \dots, n_2, \tag{3}$$

here, ξ_{ij} are i.i.d Gaussian $\mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$, or sub-Gaussian random variables. We denote by $E = (\xi_{ij})$ the corresponding matrix of noise. We study the minimax optimal rates of convergence for the estimation of M assuming that there exist $q \in [0, 2)$ and s such that $M_{i\cdot} \in \mathbb{B}_q(s)$ for any $i = 1, \dots, n_1$.

The minimax rate of convergence characterizes the fundamental limitation of the estimation accuracy. It also captures the interdependence between the different parameters in the model. There is an rich line of work on such fundamental limits (see, for example, [13, 21, 11]). The minimax risk depends crucially on the choice of the norm in the loss function. In the present paper, we measure the estimation error in $\|\cdot\|_{2,p}$-(quasi)norm for $0 < p < \infty$ (for the definition see (4)).

For $n_1 = 1$, we obtain the problem of estimating of a vector belonging to a $\mathbb{B}_q(s)$ ball in $\mathbb{R}^{n_2}$. This problem was considered in a number of papers, see, for example, [9], [3], [1], [17]. Let η_{vect} denote the minimax rate of convergence with respect to the squared Euclidean norm in the vector case. It is interesting to note that the results of the present paper show that, for the case $p = 2$, the minimax rate of convergence for estimation of matrices under the row sparsity assumption is $n_1\eta_{vect}$. Thus, in this case, the problem reduces to estimation of each row separately. The additional matrix structure does not lead to improvement or deterioration of the rate of convergence. We show that it is also true for general p.

A major focus in the present paper is on derivation of lower bounds, which is a key step in establishing minimax optimal rates of convergence. Our analysis is based on a new selection lemma (Lemma 1). The rest of the paper is organized as follows. In Section 1.1, we introduce the notation and some basic tools used throughout the paper. Section 2 establishes the minimax lower bounds for estimation of matrices with row sparsity in $\|\cdot\|_{2,p}$-norm, see Theorems 1 and 2. In Section 3, we derive the upper bounds on the risks using a reduction to the vector case. Most of the proofs are given in the appendix.

1.1 Definitions and notation

Let A be a matrix or a vector. For $0 < q < \infty$ and $A \in \mathbb{R}^{n_1 \times n_2} = (a_{ij})$, we denote by $\|A\|_q = \left(\sum_{i,j} |a_{ij}|^q\right)^{1/q}$ the elementwise l_q-(quasi-)norm of A, and by $\|A\|_0$ the number of non-zero coefficients of A:

$$\|A\|_0 = \sum_{i,j} \mathbb{I}(a_{ij} \neq 0)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. For any $A = (A_{1\cdot}, \dots, A_{n_1\cdot})^T \in \mathbb{R}^{n_1 \times n_2}$ and $p > 0$ define

$$\|A\|_{2,p} = \left(\sum_{i=1}^{n_1} \|A_{i\cdot}\|_2^p \right)^{1/p}. \tag{4}$$

For $p = 2$, $\|A\|_{2,2}$ is the elementwise l_2-norm of A and we will use the notation $\|\cdot\|_{2,2} = \|\cdot\|_2$. For $0 < p < 1$, we have the following inequality

$$\|A + A'\|_{2,p}^p \leq \|A\|_{2,p}^p + \|A'\|_{2,p}^p.$$

For $q \in [0, 2)$ and $s > 0$ we define the following class of matrices

$$\mathcal{A}(q, s) = \{A \in \mathbb{R}^{n_1 \times n_2} : A_{i\cdot} \in \mathbb{B}_q(s) \text{ for any } i = 1, \dots, n_1\}. \tag{5}$$

In the limiting case $q = 0$, we will also write

$$\mathcal{A}(s) = \{A \in \mathbb{R}^{n_1 \times n_2} : A_{i\cdot} \in \mathbb{B}_0(s) \text{ for any } i = 1, \dots, n_1\}. \tag{6}$$

We set $\mathbb{N}_{n_1 \times n_2} = \{(i, j) : 1 \leq i \leq n_1, 1 \leq j \leq n_2\}$. For two real numbers a and b we use the notation $a \wedge b := \min(a, b)$, $a \vee b := \max(a, b)$; we denote by $\lfloor x \rfloor$ the integer part of x; we use the symbol C for a generic positive constant, which is independent of n_1, n_2, s and σ and may take different values at different appearances.

2 Lower bounds

We start by establishing the minimax lower bounds for estimation of matrices over the classes $\mathcal{A}(s)$ (Theorem 1) and $\mathcal{A}(q, s)$ (Theorem 2). We denote by $\inf_{\hat{A}}$ the infimum over all estimators $\hat{A}$ with values in $\mathbb{R}^{n_1 \times n_2}$. Consider first the case $q = 0$.

Theorem 1. *Let $n_1, n_2 \geq 2$ and $p > 0$. Fix an integer $1 \leq s \leq n_2/2$. Assume that for $(i, j) \in \mathbb{N}_{n_1 \times n_2}$ the noise variables ξ_{ij} are i.i.d Gaussian $\mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$. Then,*

(i)

$$\inf_{\hat{A}} \sup_{A \in \mathcal{A}(s)} \mathbb{P}\left\{ \|\hat{A} - A\|_{2,p}^2 \geq C\, \sigma^2\, (n_1)^{2/p}\, s \log\left(\frac{e\, n_2}{s}\right) \right\} \geq \beta;$$

(ii)

$$\inf_{\hat{A}} \sup_{A \in \mathcal{A}(s)} \mathbb{E}\|\hat{A} - A\|_{2,p}^2 \geq \tilde{C}\, \sigma^2\, (n_1)^{2/p}\, s \log\left(\frac{e\, n_2}{s}\right).$$

where $0 < \beta < 1$, $C > 0$, and $\tilde{C} > 0$ are absolute constants.

Proof. It is enough to prove (i) since (ii) follows from (i) and Markov inequality. For a $A \in \mathbb{R}^{n_1 \times n_2}$, we denote by $\mathbb{P}_A$ the probability distribution of $\mathcal{N}(A, \sigma^2 I)$ Gaussian random vector where I denotes $(n_1 n_2) \times (n_1 n_2)$ identity matrix. We denote by $\mathrm{KL}(P, Q)$ the Kullback-Leibler divergence between the probability measures P and Q.

To prove (i) we use Theorem 2.5 in [21]. It is enough to check that there exists a finite subset Ω' of $\mathcal{A}(s)$ such that for any two distinct B, B' in Ω' we have

(a) $\|B - B'\|_{2,p}^2 \geq C\, \sigma^2\, (n_1)^{p/2}\, s\, \log\left(\frac{e\, n_2}{s}\right)$,

(b) $\mathrm{KL}(\mathbb{P}_B, \mathbb{P}_{B'}) \leq \alpha \log\left(\mathrm{card}\, \Omega'\right)$

for some constants $C > 0$ and $0 < \alpha < 1/8$.

Denote by $\{0,1\}^s_{n_1 \times n_2}$ the set of all matrices $A = (a_{ij}) \in \mathbb{R}^{n_1 \times n_2}$ such that $a_{ij} \in \{0, 1\}$ and each row of A contains exactly s ones. For any two matrices $A = (a_{ij})$ and $A' = (a'_{ij})$ in $\{0,1\}^s_{n_1 \times n_2}$ define the Hamming distance

$$\mathrm{d}_H(A, A') = \sum_{(i,j) \in \mathbb{N}_{n_1 \times n_2}} \mathbb{I}_{\{a_{ij} \neq a'_{ij}\}}.$$

We use of the following selection lemma proved in Appendix A.

Lemma 1. *Let $n_1, n_2 \geq 2$ and $1 \leq s \leq n_2/2$. Then, there exists a subset Ω of $\{0,1\}^s_{n_1 \times n_2}$ such that for some numerical constant $C \geq 10^{-5}$*

$$\log(|\Omega|) \geq C\, n_1\, s\, \log\left(\frac{e\, n_2}{s}\right) \tag{7}$$

and, for any two distinct A, A' in Ω, the Hamming distance satisfies

$$\mathrm{d}_H(A, A') \geq \frac{n_1\, (s+1)}{16}. \tag{8}$$

Fix $0 < \gamma < 1$ and define

$$\Omega' = \left\{ \sigma\, \gamma \sqrt{\log\left(\frac{e\, n_2}{s}\right)}\, A \quad : \quad A \in \Omega \right\}$$

where Ω is a set satisfying the conditions of Lemma 1. For $p = 2$ using (8) we obtain that for any two distinct B, B' in Ω'

$$\|B - B'\|_2^2 \geq \frac{\gamma^2\, \sigma^2\, n_1\, s}{16} \log\left(\frac{e\, n_2}{s}\right).$$

This implies (a) for $p = 2$. For $p \neq 2$ we will use the following elementary lemma, cf. Appendix B.

Lemma 2. *If $A = (a_{ij})$ and $A' = (a'_{ij})$ are two elements of $\{0,1\}^s_{n_1 \times n_2}$ such that $\mathrm{d}_H(A, A') \geq \dfrac{n_1\, (s+1)}{16}$, then the cardinality of the set $J(A, A') = \left\{ 1 \leq i \leq n_1 \,:\, \sum_{j=1}^{n_2} \mathbb{I}_{\{a_{ij} \neq a'_{ij}\}} > \dfrac{s}{32} \right\}$ is greater than or equal to $\dfrac{n_1}{64}$.*

Lemma 2 implies that for any two distinct B, B' in Ω'

$$\begin{aligned} \|B - B'\|_{2,p}^2 &\geq \gamma^2\, \sigma^2\, \log\left(\frac{e\, n_2}{s}\right) \left(\left(\frac{s}{32}\right)^{p/2} \frac{n_1}{64}\right)^{2/p} \\ &\geq \frac{\gamma^2\, \sigma^2}{64^{1+2/p}} n_1^{2/p} s \log\left(\frac{e\, n_2}{s}\right), \end{aligned} \tag{9}$$

which yields (a) for $p \neq 2$.

To check (b), note that $\mathrm{d}_H(A, A') \leq 2 n_1 s$ for all $A, A' \in \{0,1\}^s_{n_1 \times n_2}$. This implies

$$\mathrm{KL}(\mathbb{P}_B, \mathbb{P}_{B'}) = \frac{1}{2\,\sigma^2} \|B - B'\|_2^2 \leq \gamma^2\, n_1\, s\, \log\left(\frac{e\, n_2}{s}\right). \tag{10}$$

Since also $|\Omega| = |\Omega'|$, from (7) and (10) we deduce that (b) is satisfied with $\alpha < 1/8$ if $\gamma > 0$ is chosen sufficiently small. This completes the proof of Theorem 1. $\square$

Note that there are $\binom{n_2}{s}^{n_1}$ possible sparsity patterns which satisfy the hard sparsity condition on the rows. By standard bounds on binomial coefficients, we have $\log\left(\binom{n_2}{s}^{n_1}\right) \asymp n_1 s \log\left(\frac{n_2}{s}\right)$. Consequently, the rate $n_1 s \log\left(\frac{e n_2}{s}\right)$ corresponds to the logarithm of the number of models.

Let us turn out to the soft sparsity scenario. For any $0 < q < 2$ and $s > 0$ define the quantity

$$\eta(s) = \left(n_1\, s\, \left[\sigma^2\, \log\left(1 + \frac{\sigma^q\, n_2}{s}\right)\right]^{1-q/2}\right) \vee \left(n_1\, s^{2/q}\right) \vee \left(n_1\, n_2\, \sigma^2\right) \tag{11}$$

The minimax lower bound is given by the following theorem proved in Appendix C.

Theorem 2. *Let $n_1, n_2 \geq 2$. Fix $0 < q < 2$ and $s > 0$. Suppose that for $(i,j) \in \mathbb{N}_{n_1 \times n_2}$ the noise variables ξ_{ij} are i.i.d Gaussian $\mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$. Then, there exists a numerical constant c^* such that*

(i)

$$\inf_{\hat{A}} \sup_{A \in \mathcal{A}(q,s)} \mathbb{P}\left\{\|\hat{A} - A\|_2^2 \geq c^*\, \eta(s)\right\} \geq \beta,$$

where $0 < \beta < 1$ and

(ii)

$$\inf_{\hat{A}} \sup_{A \in \mathcal{A}(q,\delta)} \mathbb{E}\|\hat{A} - A\|_2^2 \geq c^*\, \eta(s).$$

3 Minimax rates of convergence

Consider the problem of estimating of a vector $v = (v_i) \in \mathbb{B}_q(s) \subset \mathbb{R}^{n_2}$ from noisy observations

$$y_i = v_i + \xi_i, \quad i = 1, \dots, n_2,$$

where ξ_{ij} are i.i.d. Gaussian $\mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$.

The non-asymptotic minimax optimal rate of convergence for estimation of v in the l_2-norm, obtained in [3], is given by

$$\eta_{vect}(s) = \sigma^2\, s\, \log\left(\frac{e\, n_2}{s}\right)$$

when $q = 0$ and by

$$\eta_{vect}(s) = \left(s\left[\sigma^2\, \log\left(1 + \frac{\sigma^q\, n_2}{s}\right)\right]^{1-q/2}\right) \vee \left(s^{2/q}\right) \vee \left(n_2\, \sigma^2\right)$$

when $0 < q < 2$.

We see that, for $p = 2$, the lower bounds given by Theorems 1 and 2 are $n_1\eta_{vect}(s)$ in the case of hard sparsity and $n_1\eta_{vect}(s)$ in the case of soft sparsity. We get the same rate as when estimating each row separately. This implies that, in this particular case, the additional matrix structure does not lead to improvement or to deterioration of the rate of convergence.

As shown below and in view of the lower bounds of Theorems 1 and 2, optimal rates for arbitrary p can be also obtained from vector estimation method. It suffices to apply to the rows of M a minimax optimal method for vector estimation on $\mathbb{B}_q(s)$ balls. One can take, for example, the following penalized least squares estimator $\hat{M}$ of M (cf. [3]):

$$\hat{M} = \underset{A \in \mathbb{R}^{n_1 \times n_2}}{\operatorname{argmin}} \left\{ \|Y - A\|_2^2 + \lambda \|A\|_0 \log\left(\frac{e\, n_1\, n_2}{\|A\|_0 \vee 1}\right)\right\} \tag{12}$$

where $\lambda > 0$ is a regularization parameter. The penalty in (12) is inspired by the hard thresholding penalty $\|A\|_0$, which leads to $\hat{m}_{ij}$ that are thresholded values of y_{ij} (see, for instance [12], page 138).

The penalized least squares estimator defined in (12) can be computed efficiently. Let $y_{(j)}$ denote the jth largest in absolute value component of Y. The estimator $\hat{M}$ is obtained by thresholding the coefficients of Y: we keep $y_{(j)}$ such that

$$y_{(j)}^2 > \lambda\left(\log(e\, n_1\, n_2) + \sum_{i=2}^{j} (-1)^{i+j+1}\, i\, \log(i)\right)$$

and set all other coefficients equal to zero.

In what follows we assume that the noise variables ξ_{ij} are zero-mean and sub-Gaussian, which means that they satisfy the following assumption.

Assumption 1. *$\mathbb{E}(\xi_{ij}) = 0$ and there exists a constant $K > 0$ such that*

$$(\mathbb{E}|\xi_{ij}|^p)^{1/p} \leq K\sqrt{p} \quad \textit{for all} \quad p \geq 1$$

for any $1 \leq i \leq n_1$ and $1 \leq j \leq n_2$.

This assumption on the noise variables means that their distribution is dominated by the distribution of a centered Gaussian random variable. This class of distributions is rather wide. Examples of sub-Gaussian random variables are Gaussian or bounded random variables. In particular, Assumption 1 implies that $\mathbb{E}\left(\xi_{ij}^2\right) \leq 2\,K^2$.

The next theorem presents oracle inequalities for the penalized least squares estimator $\hat{M}$, both in probability and in expectation.

Theorem 3. *Let $\hat{M}$ be the penalized least squares estimator defined in (12), $a > 1$ and $\lambda = 2a\,K_0\,K^2$ where $K_0 > 0$ is large enough. Suppose that Assumption 1 holds. Then, for any $\Delta > 0$*

$$\|M-\hat{M}\|_2^2 \leq \inf_{A\in\mathbb{R}^{n_1\times n_2}} \left\{\frac{a+1}{a-1}\|M-A\|_2^2 + C\,K^2\,\|A\|_0 \log\left(\frac{e\,n_1\,n_2}{\|A\|_0 \vee 1}\right)\right\} + \frac{2\,a^2}{a-1}\Delta \tag{13}$$

with probability at least $1 - 2\exp\left\{-\frac{C_0\,\Delta}{K^2}\right\}$, and

$$\mathbb{E}\,\|M-\hat{M}\|_2^2 \leq \inf_{A\in\mathbb{R}^{n_1\times n_2}} \left\{\frac{a+1}{a-1}\|M-A\|_2^2 + C\,K^2\|A\|_0 \log\left(\frac{e\,n_1\,n_2}{\|A\|_0 \vee 1}\right)\right\} + \tilde{C}\,K^2 \tag{14}$$

where C, C_0 and $\tilde{C}$ are numerical constants.

For the particular case of Gaussian noise, the result (14) of Theorem 3 is proved in [3], and the result (13) in [4]. Theorem 3 extends the analysis to the case of sub-Gaussian noise. The prooof is given in Appendix D.

Now suppose that $M \in \mathcal{A}(s)$. Using Theorem 3 and the inequality

$$\|\hat{M} - M\|_{2,p} \leq n_1^{1/p-1/2}\|\hat{M} - M\|_2$$

that holds for any $0 < p \leq 2$ we obtain the following corollary.

Corollary 1. *Let $\hat{M}$ be the penalized least squares estimator defined in (12) with $\lambda = K_0\,K^2$ where $K_0 > 0$ is large enough. Suppose that Assumption 1 holds and that $M \in \mathcal{A}(s)$. Then, for all $0 < p \leq 2$ and for any $\Delta > 0$*

$$\|\hat{M} - M\|_{2,p}^2 \leq C\,K^2\,n_1^{2/p} s \log\left(\frac{e\,n_2}{s}\right) + \Delta \tag{15}$$

with probability at least $1 - 2\exp\left\{-\frac{C_2\,\Delta}{K^2}\right\}$, and

$$\mathbb{E}\|\hat{M} - M\|_{2,p}^2 \leq C\,K^2\,n_1^{2/p} s \log\left(\frac{e\,n_2}{s}\right). \tag{16}$$

These inequalities shows that, for $0 < p \leq 2$, the penalized least squares estimator (12) achieves the rate of convergence given by Theorem 1.This implies that this rate is minimax optimal.

The next corollary shows that the estimator (12) also achieves the minimax rate of convergence in a more general setting when $M \in \mathcal{A}(q, s)$ for $0 < q < 2$. For any $0 < q < 2$ and $s > 0$ define the quantity

$$\psi(s) = \left(n_1\, s \left[K^2 \log\left(1 + \frac{K^q\, n_2}{s}\right)\right]^{1-q/2} \right) \vee \left(n_1\, s^{2/q} \right) \vee \left(n_1\, n_2\, K^2 \right). \quad (17)$$

Corollary 2. *Let $\hat{M}$ be the penalized least squares estimator defined in (12) with $\lambda = K_0\, K^2$ where $K_0 > 0$ is large enough. Suppose that Assumption 1 holds and $M \in \mathcal{A}(q, s)$. Then, there exists numerical constant C^* such that for any $\Delta > 0$*

$$\|\hat{M} - M\|_2^2 \leq C^*\, \psi(s) + \Delta$$

with probability at least $1 - 2\exp\left\{-\frac{C_2\,\Delta}{K^2}\right\}$, *and*

$$\mathbb{E}\|\tilde{M} - M\|_{2,p}^2 \leq C^*\, \psi(s).$$

We give the proof of Corollary 2 in Appendix F. If the noise variables ξ_{ij} are i.i.d Gaussian $\mathcal{N}(0, \sigma^2)$, we have $\psi(s) = \eta(s)$. Thus, the rate of convergence given by (11) is minimax optimal.

A Proof of Lemma 1

To prove Lemma 1 we use the Varshamov-Gilbert bound. The volume (cardinality) V_1 of $\{0,1\}^s_{n_1 \times n_2}$ is

$$V_1 = \binom{n_2}{s}^{n_1}.$$

Note that the volume of the Hamming ball of radius $n_1(s+1)/2$ in $\{0,1\}^s_{n_1\times n_2}$ is smaller than the volume V_2 of the Hamming ball of the same radius in a larger space of all matrices $A = (a_{ij}) \in \mathbb{R}^{n_1 \times n_2}$ such that $a_{ij} \in \{0,1\}$ and A contains at most $n_1 s$ ones. Let $K = \left\lfloor \dfrac{n_1(s+1)}{2} \right\rfloor$ where $\lfloor x \rfloor$ denotes the integer part of x. A standard bound implies

$$V_2 = \sum_{i=1}^{K} \binom{n_1 n_2}{i} \leq \left(\frac{e n_1 n_2}{K}\right)^K \leq \left(\frac{2 e n_2}{s+1}\right)^{n_1(s+1)/2}$$

where we use that $f(x) = x \log\left(\dfrac{e n_1 n_2}{x}\right)$ is growing for $x \leq n_1 n_2$.

In order to lower bound V_1 we use Stirling's formula (see, e.g., [10, p. 54]): for any $j \in \mathbb{N}$

$$j! = j^{j+1/2} e^{-j} \sqrt{2\pi}\, \psi(j) \quad \text{with} \\ e^{(12j+1)^{-1}} < \psi(j) < e^{(12j)^{-1}}. \tag{18}$$

Using (18) we get

$$\binom{n_2}{s} \geq \frac{e^{-1/6} \left(\frac{n_2}{s}\right)^{n_2+1/2}}{\sqrt{2\pi\, s} \left(\frac{n_2}{s} - 1\right)^{n_2-s+1/2}}. \tag{19}$$

Now, the Varshamov-Gilbert bound implies that there exists a subset Ω of $\{0,1\}^s_{n_1\times n_2}$ such that $\mathrm{d}_H(A, A') > \frac{n_1(s+1)}{2}$ for any $A, A' \in \Omega$, $A \neq A'$ and

$$|\Omega| \geq \frac{\binom{n_2}{s}^{n_1}}{\left(\frac{2en_2}{s+1}\right)^{n_1(s+1)/2}} \geq \left(\frac{e^{-1/6} \left(\frac{n_2}{s}\right)^{n_2+1/2} (s+1)^{\frac{s+1}{2}}}{\sqrt{2\pi\, s} \left(\frac{n_2}{s} - 1\right)^{n_2-s+1/2} (2en_2)^{\frac{s+1}{2}}} \right)^{n_1}$$

which implies

$$\begin{aligned} \log|\Omega| \geq n_1 \Bigg[& -\frac{1}{6} - \frac{1}{2}\log s - \log(\sqrt{2\pi}) + (n_2 + 1/2)\log\left(\frac{n_2}{s}\right) + \frac{s+1}{2}\log(s+1) \\ & -(n_2 - s + 1/2)\log\left(\frac{n_2}{s} - 1\right) - \frac{s+1}{2}\log(2en_2) \Bigg] \\ \geq n_1 & \left[-\frac{1}{6} - \frac{1}{2}\log s - \log(\sqrt{2\pi}) + s\log\left(\frac{n_2}{s} - 1\right) - \frac{s+1}{2}\log\left(\frac{2en_2}{s+1}\right) \right]. \end{aligned} \tag{20}$$

1) We first consider the case $501 \leq s \leq n_2/8$. Using that $\frac{251s}{501} \geq \frac{s+1}{2}$ for $s \geq 501$, we get

$$\frac{s+1}{2}\log\left(\frac{2en_2}{s+1}\right) \leq \frac{251s}{501}\log\left(\frac{501en_2}{251s}\right) \leq \frac{98s}{100}\log\left(\frac{n_2}{s} - 1\right)$$

where the last inequality is valid for $n_2/s \geq 8$.

On the other hand, it is easy to see that for $501 \leq s \leq n_2/4$ we have

$$\frac{1}{2}\log s \leq 0,007s\log\left(\frac{n_2}{s} - 1\right) \quad \text{and} \quad \frac{1}{6} + \log(\sqrt{2\pi}) \leq 0,002s\log\left(\frac{n_2}{s} - 1\right).$$

Then, (20) implies

$$\log|\Omega| \geq 0.011 n_1 s \log\left(\frac{n_2}{s} - 1\right) \geq 0.01 n_1 s \log\left(\frac{en_2}{s}\right).$$

for $n_2/8 \geq s \geq 501$.

2) Consider next the case $s < 501$ and $s \leq n_2/8$. Now, instead of the set $\{0,1\}^s_{n_1\times n_2}$ we will deal with the set $\{0,1\}^1_{n_1\times l}$ where $l = \lfloor n_2/s \rfloor$. Using the same arguments as above, we will show that there exists a subset $\tilde{\Omega} \subset \{0,1\}^1_{n_1\times l}$ such that $\mathrm{d}_H(A,A') \geq n_1/2$ for any $A, A' \in \tilde{\Omega}$, $A \neq A'$ and $\log(\mathrm{card}\,\tilde{\Omega}) \geq C\, n_1 \log(e\, n_2)$. In this case, the previous values V_1 and V_2 are replaced by

$$V_1 = l^{n_1}, \qquad V_2 = \sum_{i=1}^{\lfloor n_1/2 \rfloor} \binom{n_1 l}{i} \leq (2el)^{n_1/2}$$

and

$$\log|\tilde{\Omega}| \geq \frac{n_1}{2}\left(2\log(l) - \log(2el)\right) \geq \frac{n_1 \log(l)}{10} \geq 10^{-4} n_1 s \log\left(\frac{en_2}{s}\right)$$

for $s < 501$ and $n_2/s \geq 8$. To embed $\tilde{\Omega}$ in $\{0,1\}^s_{n_1\times n_2}$ define

$$\Omega = \{A \in \{0,1\}^s_{n_1\times n_2} \,:\, A = (\underbrace{\tilde{A},\ldots,\tilde{A}}_{s\,\text{times}}, \mathbf{0})\,,\ \tilde{A} \in \tilde{\Omega}\,,\ \mathbf{0} \in \mathbb{R}^{n_1\times(n_2 - ls)}\}.$$

We have $\Omega \subset \{0,1\}^s_{n_1\times n_2}$, $\mathrm{card}\,\Omega = \mathrm{card}\,\tilde{\Omega}$ and $\mathrm{d}_H(A,A') \geq \dfrac{n_1(s+1)}{4}$ for any $A, A' \in \Omega$, $A \neq A'$.

3) In order to deal with the case $n_2/8 \leq s \leq n_2/4.5$ define $s' = \left\lfloor \dfrac{s}{2} \right\rfloor$ and $n_2' = n_2 - (s - s')$. Then, $n_2' \geq 8s'$ and we can apply the previous result. This implies that there exists a subset $\bar{\Omega}$ of $\{0,1\}^{s'}_{n_1\times n_2'}$ such that

$$\mathrm{d}_H(A,A') \geq \frac{n_1(s'+1)}{2} \geq \frac{n_1(s+1)}{4}$$

for any $A, A' \in \bar{\Omega}$, $A \neq A'$ and

$$\log(\mathrm{card}\,\bar{\Omega}) \geq 10^{-4} n_1\, s' \,\log\left(\frac{e\, n_2'}{s'}\right) \geq \frac{10^{-4}}{2} n_1\, s\, \log\left(\frac{e\, n_2}{s}\right)$$

where we used $n_2'/s' \geq n_2/s$.

To embed $\bar{\Omega}$ in $\{0,1\}^s_{n_1\times n_2}$ define

$$\Omega = \{A \in \{0,1\}^s_{n_1\times n_2} \,:\, A = (\bar{A}, \underbrace{\mathbf{1},\ldots,\mathbf{1}}_{s-s'\,\text{times}})\,,\ \bar{A} \in \bar{\Omega}\,,\ \mathbf{1} = (1,\ldots,1)^T \in \mathbb{R}^{n_1}\}.$$

We have $\Omega \subset \{0,1\}^s_{n_1\times n_2}$, $\mathrm{card}\,\Omega = \mathrm{card}\,\bar{\Omega}$ and $\mathrm{d}_H(A,A') \geq \dfrac{n_1(s+1)}{4}$ for any $A, A' \in \Omega$, $A \neq A'$.

Using exactly the same argument we can treat cases $n_2/4.5 \leq s \leq n_2/3$ and $n_2/3 \leq s \leq n_2/2$ to get the statement of Lemma 1.

B Proof of Lemma 2

Assume that $\operatorname{card}(J(A,A')) < \frac{n_1}{64}$. Then, denoting by $J^C(A,A')$ the complement of $J(A,A')$ and using that $\operatorname{card}\left(J^C(A,A')\right) \le n_1$, we get

$$\begin{aligned} \mathrm{d}_H(A,A') &\le 2s\,\operatorname{card}(J(A,A')) + \frac{s}{32}\operatorname{card}\left(J^C(A,A')\right) \\ &< 2s\,\frac{n_1}{64} + \frac{n_1 s}{32} = \frac{n_1 s}{16} \end{aligned}$$

which contradicts the premise of the lemma.

C Proof of Theorem 2.

It is enough to prove (i) since (ii) follows from (i) and the Markov inequality.

To prove (i) we use Theorem 2.5 in [21]. We define $k \ge 1$ be the largest integer satisfying

$$k \le s\,\sigma^{-q}\,\left(\log\left(1+\frac{n_2}{k}\right)\right)^{-q/2}. \tag{21}$$

If there is no $k \ge 1$ satisfying (21), take $k=0$. Set $\bar{k} = k \vee 1$ and $S = \bar{k} \wedge \frac{n_2}{2}$. Let $\Omega' \subset \{0,1\}^S_{n_1 \times n_2}$ be the set given by Lemma 1. We consider

$$\Omega = \left\{ \tau \left(\frac{\bar{\delta}}{S}\right)^{1/q} A \,:\, A \in \Omega' \right\}$$

where $0 < \tau < 1$ and $0 < \bar{\delta} \le s$ will be chosen later. It is easy to see that $\Omega \subset \mathcal{A}(q,s)$.

Since the noise variables ξ_{ij} are i.i.d Gaussian $\mathcal{N}(0,\sigma^2)$, for any two distinct B, B' in Ω, the Kullback-Leibler divergence $\mathrm{KL}(\mathbb{P}_B, \mathbb{P}_{B'})$ between $\mathbb{P}_B$ and $\mathbb{P}_{B'}$ is given by

$$\mathrm{KL}(\mathbb{P}_B, \mathbb{P}_{B'}) = \frac{\|B-B'\|_2^2}{2\,\sigma^2} \tag{22}$$

We consider now three cases, depending on the value of the integer k defined in (21).

Case (1): $k=0$. Since $k=0$, the inequality (21) is violated for $k=1$, so that

$$s \le \sigma^q\,\left(\log(1+n_2)\right)^{q/2}. \tag{23}$$

Here $S=1$ and we take $\bar{\delta} = s$. We have that for any two distinct B, B' in Ω,

$$\|B-B'\|_2^2 \ge \frac{n_1 \tau^2}{4.5}\,(s)^{2/q}. \tag{24}$$

On the other hand, by Lemma 1, we have that

$$\log|\Omega| \ge C\,n_1 \log(1+n_2)$$

and using (23)

$$
\begin{aligned}
\mathrm{KL}(\mathbb{P}_B, \mathbb{P}_{B'}) = \frac{1}{2\,\sigma^2}\|B - B'\|_2^2 &\le \frac{\tau^2\, n_1\, s^{2/q}}{\sigma^2} \\
&\le \tau^2\, n_1\, \log(1 + n_2) \\
&\le \alpha \log|\Omega|
\end{aligned}
\tag{25}
$$

for some $0 < \alpha < 1/8$ if $0 < \tau < 1$ is chosen sufficiently small.

Case (2): $1 \le k \le n_2/2$. We take $\bar{\delta} = \left(\frac{s}{S}\right)^{1/q}$. For any two distinct B, B' in Ω,

$$
\begin{aligned}
\|B - B'\|_2^2 &\ge \frac{n_1 \tau^2\,(S+1)}{9}\,\left(\frac{s}{S}\right)^{2/q} \\
&\ge \frac{n_1\tau^2}{9}\,(s)^{2/q}\left(s\,\sigma^{-q}\,\left(\log\left(1 + \frac{n_2}{k}\right)\right)^{-q/2}\right)^{1-2/q} \\
&\ge \frac{n_1\tau^2}{9}\,s\,\sigma^{2-q}\,\left(\log\left(1 + \frac{n_2}{k}\right)\right)^{1-q/2} \\
&\ge \frac{n_1\tau^2}{9}\,s\,\sigma^{2-q}\,\left(\log\left(1 + n_2\, s^{-1}\,\sigma^q\right)\right)^{1-q/2}.
\end{aligned}
\tag{26}
$$

By Lemma 1, we have that

$$
\begin{aligned}
\log|\Omega| &\ge C\, n_1\, S\, \log\left(1 + \frac{n_2}{S}\right) \\
&\ge \frac{C\, n_1}{2}\, s\,\sigma^{-q}\left(\log\left(1 + n_2\, s^{-1}\,\sigma^q\right)\right)^{1-q/2}
\end{aligned}
$$

and

$$
\begin{aligned}
\mathrm{KL}(\mathbb{P}_B, \mathbb{P}_{B'}) = \frac{1}{2\,\sigma^2}\|B - B'\|_2^2 &\le \frac{\tau^2\, n_1}{\sigma^2} s^{2/q}\, S^{1-2/q} \\
&\le \frac{\tau^2\, n_1}{\sigma^2} s^{2/q}\,\left(s\,\sigma^{-q}\,\left(\log\left(1 + n_2\, s^{-1}\,\sigma^q\right)\right)^{-q/2}\right)^{1-2/q} \\
&\le \tau^2\, n_1\,\sigma^{-q}\,\left(\log\left(1 + n_2\, s^{-1}\,\sigma^q\right)\right)^{1-q/2} \\
&\le \alpha\log|\Omega|
\end{aligned}
\tag{27}
$$

for some $0 < \alpha < 1/8$ if $0 < \tau < 1$ is chosen sufficiently small.

Case (3): $k > n_2/2$. Since $k > n_2/2$, the inequality (21) is violated for $k = n_2/2$, so that

$$
s \ge \frac{n_2\,\sigma^q}{2}. \tag{28}
$$

In this case $S = n_2/2$ and, using (28), we can take $\bar{\delta} = \frac{n_2\,\sigma^q}{2}$. We have that for any two distinct B, B' in Ω,

$$
\|B - B'\|_2^2 \ge \frac{\tau^2 n_1\, n_2\,\sigma^2}{18}. \tag{29}
$$

On the other hand, by Lemma 1, we have that

$$\log|\Omega| \geq C\, n_1\, n_2$$

and

$$\begin{aligned} \mathrm{KL}(\mathbb{P}_B, \mathbb{P}_{B'}) &= \frac{1}{2\,\sigma^2}\|B - B'\|_2^2 \leq \frac{\tau^2\, n_1\, n_2}{2} \\ &\leq \alpha \log|\Omega| \end{aligned} \tag{30}$$

for some $0 < \alpha < 1/8$ if $0 < \tau < 1$ is chosen sufficiently small.

Now the statement of the Theorem 2 follows from (24) - (25), (26) - (27), (29) - (30) and the Theorem 2.5 in [21].

D Proof of Theorem 3.

This proof essentially follows the scheme suggested in [4] by adding an extension to the case of sub-Gaussian noise. Let $A \in \mathbb{R}^{n_1 \times n_2}$ be a fixed, but arbitrary matrix. Define for all $1 \leq r \leq n_1 n_2$

$$\mathcal{B}_r = \left\{ \bar{A} = A' - A \in \mathbb{R}^{n_1 \times n_2} \ : \ \|A'\|_0 = r \right\}.$$

Let $\{J_k\}$, $k = 1, \dots, \binom{n_1 n_2}{r}$ be all the sets of matrix indices (i,j) of cardinality r. Define

$$\mathcal{B}_{r,k} = \left\{ \bar{A} = (\bar{a}_{ij}) \in \mathcal{B}_r \ : \ a'_{ij} \neq 0 \iff (i,j) \in J_k \right\}$$

where $a'_{ij} = \bar{a}_{ij} + a_{ij}$. We have that $\dim(\mathcal{B}_{r,k}) \leq r$. Let $\Pi_{r,k}(B)$ denote the projection of the matrix B onto $\mathcal{B}_{r,k}$ and $\mathrm{pen}(A) = \lambda \|A\|_0 \log\left(\dfrac{e\, n_1\, n_2}{|A|_0 \vee 1}\right)$. By the definition of $\hat{M}$, for any $A \in \mathbb{R}^{n_1 \times n_2}$,

$$\|Y - \hat{M}\|_2^2 + \mathrm{pen}(\hat{M}) \leq \|Y - A\|_2^2 + \mathrm{pen}(A).$$

Rewriting this inequality yields

$$\begin{aligned} \|M - \hat{M}\|_2^2 + \mathrm{pen}(\hat{M}) &\leq \|M - A\|_2^2 + 2 \sum_{(i,j)} \xi_{ij} (\hat{M} - A)_{ij} + \mathrm{pen}(A) \\ &\leq \|M - A\|_2^2 + 2 \left(\sum_{(i,j)} \xi_{ij} \frac{(\hat{M} - A)_{ij}}{\|\hat{M} - A\|_2} \right) \|\hat{M} - A\|_2 + \mathrm{pen}(A). \end{aligned}$$

For $B = (b_{ij}) \in \mathbb{R}^{n_1 \times n_2}$ we set $V(B) = \sum_{(i,j)} \frac{\xi_{ij}\, b_{ij}}{\|B\|_2}$, then for any $a > 1$

$$\left(1 - \frac{1}{a}\right) \|M - \hat{M}\|_2^2 + \mathrm{pen}(\hat{M}) \leq \left(1 + \frac{1}{a}\right) \|M - A\|_2^2 + 2aV^2(\hat{M} - A) + \mathrm{pen}(A). \tag{31}$$

Next, since $\mathbb{R}^{n_1\times n_2} = \bigcup_{r=0}^{n_1 n_2} \bigcup_{k=1}^{\binom{n_1 n_2}{r}} \mathcal{B}_{r,k}$, we obtain

$$2aV^2(\hat{M}-A)-\mathrm{pen}(\hat{M}) \leq \max_{0\leq r\leq n_1 n_2} \max_{0\leq k\leq \binom{n_1 n_2}{r}} \max_{\bar{A}\in\mathcal{B}_{r,k}} \left\{2aV^2(\bar{A}) - \mathrm{pen}(\bar{A}+A)\right\}.$$

Note that for $r=0$ we have that $\mathcal{B}_0(A) = \{-A\}$ and

$$2aV^2(-A) - \mathrm{pen}(-A+A) = 2aV^2(A).$$

Let $J_{\bar{A}}$ denotes the sparsity pattern of $\bar{A} = (\bar{a}_{ij})$, i.e.

$$J_{\bar{A}} = \{(i,j)\in\mathbb{N}_{n_1\times n_2} \ : \ \bar{a}_{ij}\neq 0\},$$

then for any $\bar{A}\in\mathcal{B}_{r,k}$

$$V^2(\bar{A}) = \left(\sum_{(i,j)\in J_{\bar{A}}} \frac{\xi_{ij}\,\bar{a}_{ij}}{\|\bar{A}\|_2}\right)^2 \leq \|\Pi_{r,k}(E)\|_2^2.$$

This together with (31) imply

$$\begin{aligned}\|M-\hat{M}\|_2^2 \leq{}& \frac{a+1}{a-1}\|M-A\|_2^2 + \frac{a}{a-1}\mathrm{pen}(A) + \frac{2a^2}{a-1}V^2(A)\\ &+ \frac{a}{a-1}\left[\max_{1\leq r\leq n_1 n_2} \max_{0\leq k\leq \binom{n_1 n_2}{r}} \left\{2a\|\Pi_{r,k}(E)\|_2^2 - \lambda r \log\left(\frac{e\,n_1\,n_2}{r}\right)\right\}\right].\end{aligned} \tag{32}$$

By Assumption 1, the errors ξ_{ij} are sub-gaussian. We will use the following tail bounds in order to control the last term in (32).

Lemma 3. *Let Assumption 1 be satisfied. Then, there exists absolute constants $c_0, c_1, c_2, c_3 > 0$ such that for $K_1 = K_0\,K^2$ with $K_0 > 0$ large enough*

$$\mathbb{P}\left[\max_{1\leq r\leq n_1 n_2} \max_{0\leq k\leq \binom{n_1 n_2}{r}} \left\{\|\Pi_{r,k}(E)\|_2^2 - K_1 r\log\left(\frac{e\,n_1\,n_2}{r}\right)\right\} \geq \Delta\right] \leq c_1 \exp\left\{-\frac{c_2\,\Delta^2}{K^2}\right\}, \tag{33}$$

$$\mathbb{E}\left[\max_{1\leq r\leq n_1 n_2} \max_{0\leq k\leq \binom{n_1 n_2}{r}} \left\{\|\Pi_{r,k}(E)\|_2^2 - K_1 r\log\left(\frac{e\,n_1\,n_2}{r}\right)\right\}\right] \leq c_0\,K^2 \tag{34}$$

and

$$\mathbb{P}\left[V^2(A) - K_1\|A\|_0 \geq \Delta\right] \leq 2\exp\left\{-\frac{c_3\,\Delta^2}{K^2}\right\} \tag{35}$$

Now (14) follows from Lemma 3 and (32).

To prove (13), note that by Lemma 3 and (32), for $\lambda = 2a\, K_0\, K^2$ there exist numerical constants $C, C_1, C_2 > 0$ such that

$$\begin{aligned}
&\mathbb{P}\left(\|M - \hat{M}\|_2^2 \geq \inf_{A \in \mathbb{R}^{n_1 \times n_2}} \left\{ \frac{a+1}{a-1} \|M - A\|_2^2 + C\, \|A\|_0 \log\left(\frac{e\, n_1\, n_2}{\|A\|_0} \right) \right\} + \frac{2a^2}{a-1} \Delta \right) \\
&\leq \mathbb{P}\left(\left[\max_{1 \leq r \leq n_1 n_2} \max_{0 \leq k \leq \binom{n_1 n_2}{r}} \left\{ \|\Pi_{r,k}(E)\|_2^2 - K_1 r \log\left(\frac{e\, n_1\, n_2}{r} \right) \right\} \right] \geq \Delta/2 \right) \\
&\qquad\qquad\qquad + \mathbb{P}\left(V^2(A) - K_1 \|A\|_0 \geq \Delta/2 \right) \\
&\leq C_1 \exp\left\{ -C_2 \frac{\Delta}{K^2} \right\}
\end{aligned}$$

which proves (13).

E Proof of Lemma 3

We have that

$$\begin{aligned}
p_\Delta &\stackrel{\text{def}}{=} \mathbb{P}\left[\max_{1 \leq r \leq n_1 n_2} \max_{0 \leq k \leq \binom{n_1 n_2}{r}} \left\{ \|\Pi_{r,k}(E)\|_2^2 - K_1 r \log\left(\frac{e\, n_1\, n_2}{r} \right) \right\} \geq \Delta \right] \\
&\leq \sum_{r=1}^{n_1 n_2} \sum_{k=1}^{\binom{n_1 n_2}{r}} \mathbb{P}\left[\|\Pi_{r,k}(E)\|_2^2 \geq \Delta + K_1 r \log\left(\frac{e\, n_1\, n_2}{r} \right) \right] \\
&\leq \sum_{r=1}^{n_1 n_2} \binom{n_1 n_2}{r} \mathbb{P}\left[\mathbb{Z}_r \geq \Delta + K_1 r \log\left(\frac{e\, n_1\, n_2}{r} \right) - 2rK^2 \right]
\end{aligned}$$

where $\mathbb{Z}_r = \sum_{i=1}^{r} \xi_i^2 - \mathbb{E}(\xi_i^2)$ and $\xi_1, \ldots, \xi_r$ are i.i.d. random variables satisfying Assumption 1. Note that ξ_i^2 are sub-exponential random variables with $\|\xi_i^2\|_{\psi_1} \leq 2\, K^2$. Applying Bernstein-type inequality (see, e.g., Proposition 5.16 in [24]) and using that $\binom{n_1 n_2}{r} \leq \left(\frac{e\, n_1\, n_2}{r} \right)^r$ we get

$$\begin{aligned}
p_\Delta &\leq 2 \sum_{r=1}^{n_1 n_2} \binom{n_1 n_2}{r} \exp\left\{ -C_2 \left(K_0\, r\, \log\left(\frac{e\, n_1\, n_2}{r} \right) + \frac{\Delta}{2\, K^2} \right) \right\} \\
&= 2 \exp\left\{ -\frac{C_2\, \Delta}{K^2} \right\} \sum_{r=1}^{n_1 n_2} \left(\frac{e\, n_1\, n_2}{r} \right)^r \exp\left\{ -C_2\, K_0\, r\, \log\left(\frac{e\, n_1\, n_2}{r} \right) \right\}.
\end{aligned}$$

Taking K_0 large enough we get

$$p_\Delta \leq 2 \exp\left\{ -\frac{C_2\, \Delta}{K^2} \right\} \sum_{r=1}^{\infty} \exp\left\{ -r \log 2 \right\} \leq C_1 \exp\left\{ -\frac{C_2\, \Delta}{K^2} \right\}.$$

This proves (33) and easily implies the bound on expectation value (34).

To proof (35), we apply Bernstein-type inequality to $V^2(A) = \sum_{(i,j)\in J_A} (\xi_{ij})^2$:

$$\mathbb{P}\left[\sum_{(i,j)\in J_A} \xi_{ij}^2 - \mathbb{E}\left(\xi_{ij}^2\right) \geq K_1\|A\|_0 - 2\|A\|_0\, K^2 + \Delta\right] \\ \leq \exp\left\{-C_2\left(K_0\,\|A\|_0 - \|A\|_0 + \frac{\Delta}{2\,K^2}\right)\right\} \leq 2\exp\left\{-\frac{C_2\,\Delta}{K^2}\right\}.$$

F Proof of Corollary 2.

We use Theorem 3. First, taking $A = 0$ in (15), we get

$$\begin{aligned}\|M - \hat{M}\|_2^2 &\leq \frac{a+1}{a-1}\|M\|_2^2 + \frac{2a^2}{a-1}\Delta \\ &\leq \frac{a+1}{a-1} n_1\, s^{2/q} + \frac{2a^2}{a-1}\Delta\end{aligned} \tag{36}$$

with probability at least $1 - 2\exp\left\{-\frac{C_2\,\Delta}{K^2}\right\}$.

Now, choosing $A = M$, we obtain that

$$\begin{aligned}\|M - \hat{M}\|_2^2 &\leq C\,K^2\,\|M\|_0\,\log\left(\frac{e\,n_1\,n_2}{\|M\|_0 \vee 1}\right) + \frac{2\,a^2}{a-1}\Delta \\ &\leq C\,K^2\,n_1\,n_2 + \frac{2\,a^2}{a-1}\Delta\end{aligned} \tag{37}$$

with probability at least $1 - 2\exp\left\{-\frac{C_2\,\Delta}{K^2}\right\}$.

Finally, Theorem 3 implies that for any $1 \leq s' \leq n_2/2$, all $a > 1$ and any $\Delta > 0$

$$\|M - \hat{M}\|_2^2 \leq \inf_{A\in\mathcal{A}(2s')} \frac{a+1}{a-1}\|M - A\|_2^2 + C\,K^2\,n_1\,s'\,\log\left(1 + \frac{n_2}{2\,s'}\right) + \frac{2a^2}{a-1}\Delta \tag{38}$$

with probability at least $1 - 2\exp\left\{-\frac{C_2\,\Delta}{K^2}\right\}$. Now we use the following lemma.

Lemma 4. *Let $1 \leq s' \leq n_2/2$ and $0 < q \leq 2$. For any $M \in \mathcal{A}(q, s)$, there exists $A \in \mathcal{A}(2s')$ such that*

$$\|M - A\|_2^2 \leq s^{2/q}\,(s')^{1-2/q} n_1. \tag{39}$$

For the proof of this lemma, see Lemma 7.2 in [22] (case $0 < q \leq 1$) and the proof of Lemma 7.4 in [22] (case $1 < q \leq 2$).

Now, (38) and Lemma 4 imply that for any $1 \leq s' \leq n_2/2$

$$\|M - \hat{M}\|_2^2 \leq C\left(K^2\,n_1\,s'\,\log\left(1 + \frac{n_2}{s'}\right) + s^{2/q}\,(s')^{1-2/q} n_1 + \Delta\right). \tag{40}$$

The terms depending on s' on the right side of (40) are balanced by choosing

$$s' = \left\lfloor c' \frac{s}{K^q} \left(\log\left(1 + n_2\, K^q s^{-1}\right)\right)^{-q/2} \right\rfloor$$

with suitable constant $c' > 0$. With this choice of s we get

$$\|M - \hat{M}\|_2^2 \leq C \left(n_1\, s\, K^{2-q} \left(\log\left(1 + n_2 \frac{K^q}{s}\right)\right)^{1-q/2} + \Delta \right). \qquad (41)$$

The inequalities (36), (37) and (41) imply the statement of the Corollary 2.

Acknowledgments

The authors want to thank L.A. Bassalygo for suggesting a shorter proof of Lemma 1. This work was supported by the French National Research Agency (ANR) under the grants ANR-13-BSH1-0004-02, ANR - 11-LABEX-0047, and by GENES. It was also supported by the "Chaire Economie et Gestion des Nouvelles Données", under the auspices of Institut Louis Bachelier, Havas-Media and Paris-Dauphine.

References

[1] Abramovich, F., Benjamini, Y., Donoho, D. L. and Johnstone, I. M. (2006), Adapting to unknown sparsity by controlling the false discovery rate, *Ann. Statist.*, **34 (2)**, 584–653.

[2] Bickel, P.J., Ritov, Y. and Tsybakov, A. (2009) Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, **37**(4), 1705 - 1732.

[3] Birgé, L. and Massart, P. (2001) Gaussian model selection, *J. Eur. Math. Soc. (JEMS)*, **3**, 203–268.

[4] Bunea, F., Tsybakov, A. and Wegkamp, M. (2004) Aggregation for regression learning. `arXiv:math/0410214`

[5] Bunea, F., Tsybakov, A. and Wegkamp, M. (2007) Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, **1**, 169 - 194.

[6] Cai, T., Ma, Z. and Wu, Y. (2012) Sparse PCA: Optimal rates and adaptive estimation. `arXiv:1211.1309`.

[7] Cai, T. and Zhou, H. (2012) Optimal rates of convergence for sparse covariance matrix estimation, *Ann. Statist.*, **4**0 (5), 2389–2420.

[8] Candès, E.J. and Recht, B. (2009) Exact matrix completion via convex optimization. *Fondations of Computational Mathematics*, **9(6)**, 717-772.

[9] Donoho, D. L. and Johnstone,I. M. (1994) Minimax risk over l_p-balls for l_q-error *Prob. Theory and Related Fields*, **99**, 277–303.

[10] Feller, W. (1968) *An Introduction to Probability Theory and its Applications*, Vol. I, 3rd edn. New York: Wiley.

[11] Johnstone, I.M. (2013) *Gaussian Estimation: Sequence and Wavelet Models.* Book draft, 2013.

[12] Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (1998). *Wavelets, Approximation and Statistical Applications. Lecture Notes in Statistics*, vol. **129**. Springer, New York.

[13] Ibragimov, I.A., Hasminskii, R.Z. (1981) *Statistical Estimation. Asymptotic Theory.* Springer, New York.

[14] Koltchinskii, V., Lounici, K. and Tsybakov, A. (2011) Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *Annals of Statistics*, **39(5)**, 2302-2329.

[15] Klopp, O. (2012) Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, to appear.

[16] Lounici, K. (2008) Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.*, **2**, 90 - 102.

[17] Rigollet, P. and Tsybakov, A. (2011) Exponential screening and optimal rates of sparse estimation. *Ann. Statist.*, **39 (2)**, 731–771.

[18] Rigollet, P., Tsybakov, A. B. (2012) Sparse estimation by exponential weighting. *Statistical Science* **27** 558–575.

[19] Rigollet, P., Tsybakov, A. B. (2012) Comment: "Minimax estimation of large covariance matrices under ℓ_1-norm", *Statist. Sinica*, **22 (4)**, 1358–1367.

[20] Rohde, A., Tsybakov, A.B. (2011) Estimation of high-dimensional low rank matrices. *Annals of Statistics*, **39**, 887- 930.

[21] Tsybakov, A. B. (2009). *Introduction to non-parametric estimation.* Springer Series in Statistics. New York, NY: Springer.

[22] Tsybakov, A.B. (2014) Aggregation and minimax optimality in high-dimensional estimation. In: *Proceedings of the International Congress of Mathematicians* (Seoul, 2014), **3**, 225 - 246.

[23] van de Geer, S. (2008) High-dimensional generalized linear models and the Lasso. *Ann. Statist.*, **36**, 614 - 645.

[24] Vershynin, R. (2012) *Introduction to the non-asymptotic analysis of random matrices.* In *Compressed Sensing, Theory and Applications*, ed. Y. Eldar and G. Kutyniok, Chapter 5. Cambridge University Press.