Probab. Theory Relat. Fields (2017) 169:523–564
DOI 10.1007/s00440-016-0736-y

Robust matrix completion

Olga Klopp[1] · Karim Lounici[2] · Alexandre B. Tsybakov[3]

Received: 25 December 2014 / Revised: 10 August 2016 / Published online: 24 August 2016

Abstract This paper considers the problem of estimation of a low-rank matrix when most of its entries are not observed and some of the observed entries are corrupted. The observations are noisy realizations of a sum of a low-rank matrix, which we wish to estimate, and a second matrix having a complementary sparse structure such as elementwise sparsity or columnwise sparsity. We analyze a class of estimators obtained as solutions of a constrained convex optimization problem combining the nuclear norm penalty and a convex relaxation penalty for the sparse constraint. Our assumptions allow for simultaneous presence of random and deterministic patterns in the sampling scheme. We establish rates of convergence for the low-rank component from partial and corrupted observations in the presence of noise and we show that these rates are minimax optimal up to logarithmic factors.

Mathematics Subject Classification 62G05 · 62J02 · 62G35

1 Introduction

In the recent years, there have been a considerable interest in statistical inference for high-dimensional matrices. One particular problem is matrix completion where one observes only a small number $N \ll m_1 m_2$ of the entries of a high-dimensional $m_1 \times m_2$ matrix L_0 of rank r and aims at inferring the missing entries. In general,

✉ Olga Klopp
kloppolga@math.cnrs.fr

1 CREST and MODAL'X, University Paris Ouest, 92001 Nanterre, France

2 School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0160, USA

3 CREST-ENSAE, UMR CNRS 9194, 3, Av. Pierre Larousse, 92240 Malakoff, France

recovery of a matrix from a small number of observed entries is impossible, but, if the unknown matrix has low rank, then accurate and even exact recovery is possible. In the noiseless setting, [7,14,22] established the following remarkable result: assuming that the matrix L_0 satisfies some low coherence condition, this matrix can be recovered exactly by a constrained nuclear norm minimization with high probability from only $N \gtrsim r \max\{m_1, m_2\} \log^2(m_1 + m_2)$ entries observed uniformly at random. A more common situation in applications corresponds to the noisy setting in which the few available entries are corrupted by noise. Noisy matrix completion has been in the focus of several recent studies (see, e.g., [5,12,17,18,20,21,23]).

The matrix completion problem is motivated by a variety of applications. An important question in applications is whether or not matrix completion procedures are robust to corruptions. Suppose that we observe noisy entries of $A_0 = L_0 + S_0$ where L_0 is an unknown low-rank matrix and S_0 corresponds to some gross/malicious corruptions. We wish to recover L_0 but we observe only few entries of A_0 and, among those, a fraction happens to be corrupted by S_0. Of course, we do not know which entries are corrupted. It has been shown empirically that uncontrolled and potentially adversarial gross errors affecting only a small portion of observations can be particularly harmful. For example, Xu et al. [16] showed that a very popular matrix completion procedure using nuclear norm minimization can fail dramatically even if S_0 contains only a single nonzero column. It is particularly relevant in applications to recommendation systems where malicious users try to manipulate the outcome of matrix completion algorithms by introducing spurious perturbations S_0. Hence, there is a need for new matrix completion techniques that are robust to the presence of corruptions S_0.

With this motivation, we consider the following setting of *robust matrix completion*. Let $A_0 \in \mathbb{R}^{m_1 \times m_2}$ be an unknown matrix that can be represented as a sum $A_0 = L_0 + S_0$ where L_0 is a low-rank matrix and S_0 is a matrix with some low complexity structure such as entrywise sparsity or columnwise sparsity. We consider the observations (X_i, Y_i), $i = 1, \dots, N$, satisfying the trace regression model

$$Y_i = \mathrm{tr}(X_i^T A_0) + \xi_i, \quad i = 1, \dots, N, \tag{1}$$

where tr(M) denotes the trace of matrix M. Here, the noise variables ξ_i are independent and centered, and X_i are $m_1 \times m_2$ matrices taking values in the set

$$\mathcal{X} = \{e_j(m_1)e_k^T(m_2), 1 \le j \le m_1, 1 \le k \le m_2\}, \tag{2}$$

where $e_l(m)$, $l = 1, \dots, m$, are the canonical basis vectors in $\mathbb{R}^m$. Thus, we observe some entries of matrix A_0 with random noise. Based on the observations (X_i, Y_i), we wish to obtain accurate estimates of the components L_0 and S_0 in the high-dimensional setting $N \ll m_1 m_2$. Throughout the paper, we assume that $(X_1, \dots, X_n)$ is independent of $(\xi_1, \dots, \xi_n)$.

We assume that the set of indices i of our N observations is the union of two disjoint components Ω and $\tilde{\Omega}$. The first component Ω corresponds to the "non-corrupted" noisy entries of L_0, i.e., to the observations, for which the entry of S_0 is zero. The second set $\tilde{\Omega}$ corresponds to the observations, for which the entry of S_0 is nonzero. Given an observation, we do not know whether it belongs to the corrupted or non-corrupted part of the observations and we have $|\Omega| + |\tilde{\Omega}| = N$, where $|\Omega|$ and $|\tilde{\Omega}|$ are non-random numbers of non-corrupted and corrupted observations, respectively.

A particular case of this setting is the matrix decomposition problem where $N = m_1 m_2$, i.e., we observe all entries of A_0. Several recent works consider the matrix decomposition problem, mostly in the noiseless setting, $\xi_i \equiv 0$. Chandrasekaran et al. [8] analyzed the case when the matrix S_0 is sparse, with small number of non-zero entries. They proved that exact recovery of (L_0, S_0) is possible with high probability under additional identifiability conditions. This model was further studied by Hsu et al. [15] who give milder conditions for the exact recovery of (L_0, S_0). Also in the noiseless setting, Candes et al. [6] studied the same model but with positions of corruptions chosen uniformly at random. Xu et al. [16] studied a model, in which the matrix S_0 is columnwise sparse with sufficiently small number of non-zero columns. Their method guarantees approximate recovery for the non-corrupted columns of the low-rank component L_0. Agarwal et al. [1] consider a general model, in which the observations are noisy realizations of a linear transformation of A_0. Their setup includes the matrix decomposition problem and some other statistical models of interest but does not cover the matrix completion problem. Agarwal et al. [1] state a general result on approximate recovery of the pair (L_0, S_0) imposing a "spikiness condition" on the low-rank component L_0. Their analysis includes as particular cases both the entrywise corruptions and the columnwise corruptions.

The robust matrix completion setting, when $N < m_1 m_2$, was first considered by Candes et al. [6] in the noiseless case for entrywise sparse S_0. Candes et al. [6] assumed that the support of S_0 is selected uniformly at random and that N is equal to $0.1 m_1 m_2$ or to some other fixed fraction of $m_1 m_2$. Chen et al. [9] considered also the noiseless case but with columnwise sparse S_0. They proved that the same procedure as in [8] can recover the non-corrupted columns of L_0 and identify the set of indices of the corrupted columns. This was done under the following assumptions: the locations of the non-corrupted columns are chosen uniformly at random; L_0 satisfies some sparse/low-rank incoherence condition; the total number of corrupted columns is small and a sufficient number of non-corrupted entries is observed. More recently, Chen et al. [10] and Li [27] considered noiseless robust matrix completion with entrywise sparse S_0. They proved exact recovery of the low-rank component under an incoherence condition on L_0 and some additional assumptions on the number of corrupted observations.

To the best of our knowledge, the present paper is the first study of robust matrix completion with noise. Our analysis is general and covers in particular the cases of columnwise sparse corruptions and entrywise sparse corruptions. It is important to note that we do not require strong assumptions on the unknown matrices, such as the incoherence condition, or additional restrictions on the number of corrupted observations as in the noiseless case. This is due to the fact that we do not aim at exact recovery of the unknown matrix. We emphasize that we do not need to know the rank of L_0 nor the sparsity level of S_0. We do not need to observe all entries of A_0 either. We only need to know an upper bound on the maximum of the absolute values of the entries of L_0 and S_0. Such information is often available in applications; for example, in recommendation systems, this bound is just the maximum rating. Another important point is that our method allows us to consider quite general and unknown sampling distribution. All the previous works on noiseless robust matrix completion assume the uniform sampling distribution. However, in practice the observed entries are not guaranteed to follow the uniform scheme and the sampling distribution is not exactly known.

We establish oracle inequalities for the cases of entrywise sparse and columnwise sparse S_0. For example, in the case of columnwise corruptions, we prove the following bound on the normalized Frobenius error of our estimator $(\hat{L}, \hat{S})$ of (L_0, S_0): with high probability

$$\frac{\|\hat{L} - L_0\|_2^2}{m_1 m_2} + \frac{\|S_0 - \hat{S}\|_2^2}{m_1 m_2} \lesssim \frac{r \max(m_1, m_2) + |\tilde{\Omega}|}{|\Omega|} + \frac{s}{m_2}$$

where the symbol $\lesssim$ means that the inequality holds up to a multiplicative absolute constant and a factor, which is logarithmic in m_1 and m_2. Here, r denotes the rank of L_0, and s is the number of corrupted columns. Note that, when the number of corrupted columns s and the proportion of corrupted observations $|\tilde{\Omega}|/|\Omega|$ are small, this bound implies that $O(r \max(m_1, m_2))$ observations are enough for successful and robust to corruptions matrix completion. We also show that, both under the columnwise corruptions and entrywise corruptions, the obtained rates of convergence are minimax optimal up to logarithmic factors.

This paper is organized as follows. Section 2.1 contains the notation and definitions. We introduce our estimator in Sect. 2.2 and we state the assumptions on the sampling scheme in Sect. 2.3. Section 3 presents a general upper bound for the estimation error. In Sects. 4 and 5, we specialize this bound to the settings with columnwise corruptions and entrywise corruptions, respectively. In Sect. 6, we prove that our estimator is minimax rate optimal up to a logarithmic factor. The Appendix contains the proofs.

2 Preliminaries

2.1 Notation and definitions

2.1.1 General notation

For any set I, $|I|$ denotes its cardinality and $\bar{I}$ its complement. We write $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$.

For a matrix A, A^i is its ith column and A_{ij} is its (i, j)th entry. Let $I \subset \{1, \dots m_1\} \times \{1, \dots m_2\}$ be a subset of indices. Given a matrix A, we denote by A_I its restriction on I, that is, $(A_I)_{ij} = A_{ij}$ if $(i, j) \in I$ and $(A_I)_{ij} = 0$ if $(i, j) \notin I$. In what follows, $\mathbf{Id}$ denotes the matrix of ones, i.e., $\mathbf{Id}_{ij} = 1$ for any (i, j) and $\mathbf{0}$ denotes the zero matrix, i.e., $\mathbf{0}_{ij} = 0$ for any (i, j).

For any $p \geq 1$, we denote by $\|\cdot\|_p$ the usual l_p−norm. Additionally, we use the following matrix norms: $\|A\|_*$ is the nuclear norm (the sum of singular values), $\|A\|$ is the operator norm (the largest singular value), $\|A\|_\infty$ is the largest absolute value of the entries:

$$\|A\|_\infty = \max_{1 \leq j \leq m_1, 1 \leq k \leq m_2} |A_{jk}|,$$

the norm $\|A\|_{2,1}$ is the sum of l_2 norms of the columns of A and $\|A\|_{2,\infty}$ is the largest l_2 norm of the columns of A:

$$\|A\|_{2,1} = \sum_{k=1}^{m_2} \|A^k\|_2 \quad \text{and} \quad \|A\|_{2,\infty} = \max_{1\le k\le m_2} \|A^k\|_2.$$

The inner product of matrices A and B is defined by $\langle A, B\rangle = \mathrm{tr}(AB^\top)$.

2.1.2 Notation related to corruptions

We first introduce the index sets $\mathcal{I}$ and $\tilde{\mathcal{I}}$. These are subsets of $\{1, \dots, m_1\} \times \{1, \dots, m_2\}$ that are defined differently for the settings with columnwise sparse and entrywise sparse corruption matrix S_0.

For the columnwise sparse matrix S_0, we define

$$\tilde{\mathcal{I}} = \{1, \dots, m_1\} \times J \tag{3}$$

where $J \subset \{1, \dots, m_2\}$ is the set of indices of the non-zero columns of S_0. For the entrywise sparse matrix S_0, we denote by $\tilde{\mathcal{I}}$ the set of indices of the non-zero elements of S_0. In both settings, $\mathcal{I}$ denotes the complement of $\tilde{\mathcal{I}}$.

Let $\mathcal{R} : \mathbb{R}^{m_1\times m_2} \to \mathbb{R}_+$ be a norm that will be used as a regularizer relative to the corruption matrix S_0. The associated dual norm is defined by the relation

$$\mathcal{R}^*(A) = \sup_{\mathcal{R}(B)\le 1} \langle A, B\rangle. \tag{4}$$

Let $|A|$ denote the matrix whose entries are the absolute values of the entries of matrix A. The norm $\mathcal{R}(\cdot)$ is called *absolute* if it depends only on the absolute values of the entries of A:

$$\mathcal{R}(A) = \mathcal{R}(|A|).$$

For instance, the l_p-norm and the $\|\cdot\|_{2,1}$-norm are absolute. We call $\mathcal{R}(\cdot)$ *monotonic* if $|A| \le |B|$ implies $\mathcal{R}(A) \le \mathcal{R}(B)$. Here and below, the inequalities between matrices are understood as entry-wise inequalities. Any absolute norm is monotonic and vice versa (see, e.g., [3]).

2.1.3 Specific notation

- We set $d = m_1 + m_2$, $m = m_1 \wedge m_2$, and $M = m_1 \vee m_2$.
- Let $\{\epsilon_i\}_{i=1}^n$ be a sequence of i.i.d. Rademacher random variables. We define the following random variables called the stochastic terms:

$$\Sigma_R = \frac{1}{n}\sum_{i\in\Omega} \epsilon_i X_i, \quad \Sigma = \frac{1}{N}\sum_{i\in\Omega} \xi_i X_i, \quad \text{and} \quad W = \frac{1}{N}\sum_{i\in\Omega} X_i.$$

- We denote by r the rank of matrix L_0.
- We denote by N the number of observations, and by $n = |\Omega|$ the number of non-corrupted observations. The number of corrupted observations is $|\tilde{\Omega}| = N - n$. We set $æ = N/n$.
- We use the generic symbol C for positive constants that do not depend on n, m_1, m_2, r, s and can take different values at different appearances.

2.2 Convex relaxation for robust matrix completion

For the usual matrix completion, i.e., when the corruption matrix $S_0 = \mathbf{0}$, one of the most popular methods of solving the problem is based on constrained nuclear norm minimization. For example, the following constrained matrix Lasso estimator is introduced in [18]:

$$\hat{A} \in \underset{\|A\|_\infty \le \mathbf{a}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle X_i, A \rangle)^2 + \lambda \|A\|_* \right\},$$

where $\lambda > 0$ is a regularization parameter and $\mathbf{a}$ is an upper bound on $\|L_0\|_\infty$.

To account for the presence of non-zero corruptions S_0, we introduce an additional norm-based penalty that should be chosen depending on the structure of S_0. We consider the following estimator $(\hat{L}, \hat{S})$ of the pair (L_0, S_0):

$$(\hat{L}, \hat{S}) \in \underset{\substack{\|L\|_\infty \le \mathbf{a} \\ \|S\|_\infty \le \mathbf{a}}}{\arg\min} \left\{ \frac{1}{N} \sum_{i=1}^{N} (Y_i - \langle X_i, L + S \rangle)^2 + \lambda_1 \|L\|_* + \lambda_2 \mathcal{R}(S) \right\}. \tag{5}$$

Here $\lambda_1 > 0$ and $\lambda_2 > 0$ are regularization parameters and $\mathbf{a}$ is an upper bound on $\|L_0\|_\infty$ and $\|S_0\|_\infty$. Note that this definition and all the proofs can be easily adapted to the setting with two different upper bounds for $\|L_0\|_\infty$ and $\|S_0\|_\infty$ as it can be the case in some applications. Thus, the results of the paper extend to this case as well.

For the following two key examples of sparsity structure of S_0, we consider specific regularizers $\mathcal{R}$.

- *Example 1.* Suppose that S_0 is *columnwise sparse*, that is, it has a small number $s < m_2$ of non-zero columns. We use the $\|\cdot\|_{2,1}$-norm regularizer for such a sparsity structure: $\mathcal{R}(S) = \|S\|_{2,1}$. The associated dual norm is $\mathcal{R}^*(S) = \|S\|_{2,\infty}$.
- *Example 2.* Suppose now that S_0 is *entrywise sparse*, that is, that it has $s \ll m_1 m_2$ non-zero entries. The usual choice of regularizer for such a sparsity structure is the l_1 norm: $\mathcal{R}(S) = \|S\|_1$. The associated dual norm is $\mathcal{R}^*(S) = \|S\|_\infty$.

In these two examples, the regularizer $\mathcal{R}$ is *decomposable* with respect to a properly chosen set of indices I. That is, for any matrix $A \in \mathbb{R}^{m_1 \times m_2}$ we have

$$\mathcal{R}(A) = \mathcal{R}(A_I) + \mathcal{R}(A_{\bar{I}}). \tag{6}$$

For instance, the $\|\cdot\|_{2,1}$-norm is decomposable with respect to any set I such that

$$I = \{1, \ldots, m_1\} \times J \tag{7}$$

where $J \subset \{1, \ldots, m_2\}$. The usual l_1 norm is decomposable with respect to any subset of indices I.

2.3 Assumptions on the sampling scheme and on the noise

In the literature on the usual matrix completion ($S_0 = \mathbf{0}$), it is commonly assumed that the observations X_i are i.i.d. For robust matrix completion, it is more realistic to assume the presence of two subsets in the observed X_i. The first subset $\{X_i,\ i \in \Omega\}$ is a collection of i.i.d. random matrices with some unknown distribution on

$$\mathcal{X}' = \{e_j(m_1)e_k^T(m_2), (j,k) \in \mathcal{I}\}. \tag{8}$$

These X_i's are of the same type as in the usual matrix completion. They are the X-components of non-corrupted observations (recall that the entries of S_0 corresponding to indices in $\mathcal{I}$ are equal to zero). On this non-corrupted part of observations, we require some assumptions on the sampling distribution (see Assumptions 1, 2, 5, and 9 below).

The second subset $\{X_i,\ i \in \tilde{\Omega}\}$ is a collection of matrices with values in

$$\mathcal{X}'' = \{e_j(m_1)e_k^T(m_2), (j,k) \in \tilde{\mathcal{I}}\}.$$

These are the X-components of corrupted observations. Importantly, we *make no assumptions* on how they are sampled. Thus, for any $i \in \tilde{\Omega}$, we have that the index of the corresponding entry belongs to $\tilde{\mathcal{I}}$ and we make no further assumption. If we take the example of recommendation systems, this partition into $\{X_i,\ i \in \Omega\}$ and $\{X_i,\ i \in \tilde{\Omega}\}$ accounts for the difference in behavior of normal and malicious users.

As there is no hope for recovering the unobserved entries of S_0, one should consider only the estimation of the restriction of S_0 to $\tilde{\Omega}$. This is equivalent to assume that we estimate the whole S_0 when all unobserved entries of S_0 are equal to zero, cf. [9]. This assumption will be done throughout the paper.

For $i \in \Omega$, we suppose that X_i are i.i.d realizations of a random matrix X having distribution Π on the set $\mathcal{X}'$. Let $\pi_{jk} = \mathbb{P}(X = e_j(m_1)e_k^T(m_2))$ be the probability to observe the (j,k)th entry. One of the particular settings of this problem is the case of the uniform on $\mathcal{X}'$ distribution Π. It was previously considered in the context of noiseless robust matrix completion, see, e.g., [9]. We consider here a more general sampling model. In particular, we suppose that any non-corrupted element is sampled with positive probability:

Assumption 1 There exists a positive constant $\mu \geq 1$ such that, for any $(j,k) \in \mathcal{I}$,

$$\pi_{jk} \geq (\mu|\mathcal{I}|)^{-1}.$$

If Π is the of uniform distribution on $\mathcal{X}'$ we have $\mu = 1$. For $A \in \mathbb{R}^{m_1 \times m_2}$ set

$$\|A\|_{L_2(\Pi)}^2 = \mathbb{E}(\langle A, X\rangle^2).$$

Assumption 1 implies that

$$\|A\|_{L_2(\Pi)}^2 \geq (\mu\,|\mathcal{I}|)^{-1}\,\|A_{\mathcal{I}}\|_2^2. \tag{9}$$

Denote by $\pi_{\cdot k} = \sum_{j=1}^{m_1} \pi_{jk}$ the probability to observe an element from the kth column and by $\pi_{j\cdot} = \sum_{k=1}^{m_2} \pi_{jk}$ the probability to observe an element from the jth row. The following assumption requires that no column and no row is sampled with too high probability.

Assumption 2 There exists a positive constant $L \geq 1$ such that

$$\max_{i,j}(\pi_{\cdot k}, \pi_{j\cdot}) \leq L/m.$$

This assumption will be used in Theorem 1 below. In Sects. 4 and 5, we apply Theorem 1 to the particular cases of columnwise sparse and entrywise sparse corruptions. There, we will need more restrictive assumptions on the sampling distribution (see Assumptions 5 and 9).

We assume below that the noise variables ξ_i are sub-gaussian:

Assumption 3 There exist positive constants σ and c_1 such that

$$\max_{i=1,\ldots,n} \mathbb{E}\exp(\xi_i^2/\sigma^2) < c_1.$$

3 Upper bounds for general regularizers

In this section we state our main result which applies to a general convex program (5) where $\mathcal{R}$ is an absolute norm and a decomposable regularizer. In the next sections, we consider in detail two particular choices, $\mathcal{R}(\cdot) = \|\cdot\|_1$ and $\mathcal{R}(\cdot) = \|\cdot\|_{2,1}$. Introduce the notation:

$$\Psi_1 = \mu^2\, m_1 m_2\, r\,(æ^2\lambda_1^2 + \mathbf{a}^2\,(\mathbb{E}\,(\|\Sigma_R\|))^2) + \mathbf{a}^2\,\mu\sqrt{\frac{\log(d)}{n}},$$

$$\Psi_2 = \mu\,\mathbf{a}\,\mathcal{R}(\mathbf{Id}_{\tilde{\Omega}})\left(\frac{\lambda_2\,\mathbf{a}}{\lambda_1}\mathbb{E}\,(\|\Sigma_R\|) + æ\lambda_2 + \mathbf{a}\,\mathbb{E}\left(\mathcal{R}^*(\Sigma_R)\right)\right),$$

$$\Psi_3 = \frac{\mu\,|\tilde{\Omega}|\left(\mathbf{a}^2 + \sigma^2\log(d)\right)}{N}\left(\frac{\mathbf{a}\,\mathbb{E}\,(\|\Sigma_R\|)}{\lambda_1} + \frac{\mathbf{a}\,\mathbb{E}\,(\mathcal{R}^*(\Sigma_R))}{\lambda_2} + æ\right) + \frac{\mathbf{a}^2|\tilde{\mathcal{I}}|}{m_1 m_2},$$

$$\Psi_4 = \mu\, \mathbf{a}^2 \sqrt{\frac{\log(d)}{n}} + \mu\, \mathbf{a}\, \mathcal{R}(\mathbf{Id}_{\tilde{\Omega}})[æ\, \lambda_2 + \mathbf{a}\, \mathbb{E}(\mathcal{R}^*(\Sigma_R))] \\ + \left[\frac{\mathbf{a}\, \mathbb{E}\,(\mathcal{R}^*(\Sigma_R))}{\lambda_2} + æ\right] \frac{\mu\, |\tilde{\Omega}|(\mathbf{a}^2 + \sigma^2 \log(d))}{N} \tag{10}$$

where $d = m_1 + m_2$.

Theorem 1 *Let $\mathcal{R}$ be an absolute norm and a decomposable regularizer. Assume that $\|L_0\|_\infty \le \mathbf{a}$, $\|S_0\|_\infty \le \mathbf{a}$ for some constant $\mathbf{a}$ and let Assumptions 1–3 be satisfied. Let $\lambda_1 > 4\,\|\Sigma\|$, and $\lambda_2 \ge 4\,(\mathcal{R}^*(\Sigma) + 2\mathbf{a}\mathcal{R}^*(W))$. Then, with probability at least $1 - 4.5\, d^{-1}$,*

$$\frac{\|L_0 - \hat{L}\|_2^2}{m_1 m_2} + \frac{\|S_0 - \hat{S}\|_2^2}{m_1 m_2} \le C\, \{\Psi_1 + \Psi_2 + \Psi_3\} \tag{11}$$

where C is an absolute constant. Moreover, with the same probability,

$$\frac{\|\hat{S}_{\mathcal{I}}\|_2^2}{|\mathcal{I}|} \le C\Psi_4. \tag{12}$$

The term Ψ_1 in (11) corresponds to the estimation error associated with matrix completion of a rank r matrix. The second and the third terms account for the error induced by corruptions. In the next two sections we apply Theorem 1 to the settings with the entrywise sparse and columnwise sparse corruption matrices S_0.

4 Columnwise sparse corruptions

In this section, we assume that that S_0 has at most s non-zero columns, and $s \le m_2/2$. We use here the $\|\cdot\|_{2,1}$-norm regularizer $\mathcal{R}$. Then, the convex program (5) takes form

$$(\hat{L}, \hat{S}) \in \mathop{\arg\min}_{\substack{\|L\|_\infty \le \mathbf{a} \\ \|S\|_\infty \le \mathbf{a}}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (Y_i - \langle X_i, L + S\rangle)^2 + \lambda_1 \|L\|_1 + \lambda_2 \|S\|_{2,1} \right\}. \tag{13}$$

Since S_0 has at most s non-zero columns, we have $|\tilde{\mathcal{I}}| = m_1 s$. Furthermore, by the Cauchy–Schwarz inequality, $\|\mathbf{Id}_{\tilde{\Omega}}\|_{2,1} \le \sqrt{s|\tilde{\Omega}|}$. Using these remarks we replace Ψ_2, Ψ_3 and Ψ_4 by the larger quantities

$$\Psi_2' = \mu\, \mathbf{a} \sqrt{s|\tilde{\Omega}|} \left(\frac{\mathbf{a}\, \lambda_2}{\lambda_1} \mathbb{E}\,(\|\Sigma_R\|) + æ\lambda_2 + \mathbf{a}\, \mathbb{E}\|\Sigma_R\|_{2,\infty} \right),$$

$$\Psi_3' = \frac{\mu\, |\tilde{\Omega}| \left(\mathbf{a}^2 + \sigma^2 \log(d)\right)}{N} \left(\frac{\mathbf{a}\mathbb{E}\,(\|\Sigma_R\|)}{\lambda_1} + \frac{\mathbf{a}\mathbb{E}\|\Sigma_R\|_{2,\infty}}{\lambda_2} + æ \right) + \frac{\mathbf{a}^2 s}{m_2},$$

$$\Psi_4' = \mu\,\mathbf{a}^2\sqrt{\frac{\log(d)}{n}} + \mu\,\mathbf{a}\sqrt{s|\tilde{\Omega}|}\left[æ\,\lambda_2 + \mathbf{a}\,\mathbb{E}\|\Sigma_R\|_{2,\infty}\right] + \left[\frac{\mathbf{a}\,\mathbb{E}\|\Sigma_R\|_{2,\infty}}{\lambda_2} + æ\right]\frac{\mu\,|\tilde{\Omega}|\left(\mathbf{a}^2 + \sigma^2\log(d)\right)}{N}.$$

Specializing Theorem 1 to this case yields the following corollary.

Corollary 4 *Assume that* $\|L_0\|_\infty \le \mathbf{a}$ *and* $\|S_0\|_\infty \le \mathbf{a}$. *Let the regularization parameters* (λ_1, λ_2) *satisfy*

$$\lambda_1 > 4\,\|\Sigma\| \quad \textit{and} \quad \lambda_2 \ge 4\left(\|\Sigma\|_{2,\infty} + 2\mathbf{a}\|W\|_{2,\infty}\right).$$

Then, with probability at least $1 - 4.5\,d^{-1}$, *for any solution* $(\hat{L}, \hat{S})$ *of the convex program* (13) *with such regularization parameters* (λ_1, λ_2) *we have*

$$\frac{\|L_0 - \hat{L}\|_2^2}{m_1 m_2} + \frac{\|S_0 - \hat{S}\|_2^2}{m_1 m_2} \le C\,\left\{\Psi_1 + \Psi_2' + \Psi_3'\right\}.$$

where C *is an absolute constant. Moreover, with the same probability,*

$$\frac{\|\hat{S}_{\mathcal{I}}\|_2^2}{|\mathcal{I}|} \le C\Psi_4'.$$

In order to get a bound in a closed form, we need to obtain suitable upper bounds on the stochastic terms Σ, Σ_R and W. We derive such bounds under an additional assumption on the column marginal sampling distribution. Set $\pi_{\cdot,k}^{(2)} = \sum_{j=1}^{m_1}\pi_{jk}^2$.

Assumption 5 There exists a positive constant $\gamma \ge 1$ such that

$$\max_k \pi_{\cdot,k}^{(2)} \le \frac{\gamma^2}{|\mathcal{I}|\,m_2}.$$

This condition prevents the columns from being sampled with too high probability and guarantees that the non-corrupted observations are well spread out among the columns. Assumption 5 is clearly less restrictive than assuming that Π is uniform as it was done in the previous work on noiseless robust matrix completion. In particular, Assumption 5 is satisfied when the distribution Π is approximately uniform, i.e., when $\pi_{jk} \asymp \frac{1}{m_1(m_2 - s)}$. Note that Assumption 5 implies the following milder condition on the marginal sampling distribution:

$$\max_k \pi_{\cdot k} \le \frac{\sqrt{2}\,\gamma}{m_2}. \tag{14}$$

Condition (14) is sufficient to control $\|\Sigma\|_{2,\infty}$ and $\|\Sigma_R\|_{2,\infty}$ while to we need a stronger Assumption 5 to control $\|W\|_{2,\infty}$.

The following lemma gives the order of magnitude of the stochastic terms driving the rates of convergence.

Lemma 6 *Let the distribution* Π *on* $\mathcal{X}'$ *satisfy Assumptions 1, 2 and 5. Let also Assumption 3 hold. Assume that* $N \le m_1 m_2$, $n \le |\mathcal{I}|$, *and* $\log m_2 \ge 1$. *Then, there exists an absolute constant* $C > 0$ *such that, for any* $t > 0$, *the following bounds on the norms of the stochastic terms hold with probability at least* $1 - e^{-t}$, *as well as the associated bounds in expectation.*

(i)
$$\|\Sigma\| \le C\sigma \max\left(\sqrt{\frac{L(t+\log d)}{æ\, Nm}}, \frac{(\log m)(t+\log d)}{N}\right) \quad \textit{and}$$
$$\mathbb{E}\,\|\Sigma_R\| \le C\left(\sqrt{\frac{L\log(d)}{nm}} + \frac{\log^2 d}{N}\right);$$

(ii)
$$\|\Sigma\|_{2,\infty} \le C\sigma\left(\sqrt{\frac{\gamma(t+\log(d))}{æNm_2}} + \frac{t+\log d}{N}\right) \quad \textit{and}$$
$$\mathbb{E}\,\|\Sigma\|_{2,\infty} \le C\sigma\left(\sqrt{\frac{\gamma\log(d)}{æNm_2}} + \frac{\log d}{N}\right);$$

(iii)
$$\|\Sigma_R\|_{2,\infty} \le C\left(\sqrt{\frac{\gamma(t+\log(d))}{nm_2}} + \frac{t+\log d}{n}\right) \quad \textit{and}$$
$$\mathbb{E}\,\|\Sigma_R\|_{2,\infty} \le C\left(\sqrt{\frac{\gamma\log(d)}{nm_2}} + \frac{\log d}{n}\right);$$

(iv)
$$\|W\|_{2,\infty} \le C\left(\frac{\gamma(t+\log m_2)^{1/4}}{\sqrt{æNm_2}}\left(1+\sqrt{\frac{m_2(t+\log m_2)}{n}}\right)^{1/2} + \frac{t+\log m_2}{N}\right)$$
$$\mathbb{E}\|W\|_{2,\infty} \le C\left(\frac{\gamma\log^{1/4}(d)}{\sqrt{æNm_2}}\left(1+\sqrt{\frac{m_2\log d}{n}}\right)^{1/2} + \frac{\log d}{N}\right).$$

Let

$$n^* = 2\log(d)\left(\frac{m_2}{\gamma} \vee \frac{m\log^2 m}{L}\right). \tag{15}$$

Recall that $æ = \frac{N}{n} \ge 1$. If $n \ge n^*$, using the bounds given by Lemma 6, we can chose the regularization parameters λ_1 and λ_2 in the following way:

$$\lambda_1 = C\,(\sigma \vee \mathbf{a})\sqrt{\frac{L\,\log(d)}{Nm}} \quad \text{and} \quad \lambda_2 = C\,\gamma\,(\sigma \vee \mathbf{a})\sqrt{\frac{\log(d)}{Nm_2}}, \tag{16}$$

where $C > 0$ is a large enough numerical constant.

With this choice of the regularization parameters, Corollary 4 implies the following result.

Corollary 7 *Let the distribution* Π *on* $\mathcal{X}'$ *satisfy Assumptions 1, 2 and 5. Let Assumption 3 hold and* $\|L_0\|_\infty \leq \mathbf{a}$, $\|S_0\|_\infty \leq \mathbf{a}$. *Assume that* $N \leq m_1 m_2$ *and* $n^* \leq n$. *Then, with probability at least* $1 - 6/d$ *for any solution* $(\hat{L}, \hat{S})$ *of the convex program (13) with the regularization parameters* (λ_1, λ_2) *given by (16), we have*

$$\frac{\|L_0 - \hat{L}\|_2^2}{m_1 m_2} + \frac{\|S_0 - \hat{S}\|_2^2}{m_1 m_2} \leq C_{\mu,\gamma,L} (\sigma \vee \mathbf{a})^2 \log(d)\, æ \frac{r\,M + |\tilde{\Omega}|}{n} + \frac{\mathbf{a}^2 s}{m_2} \tag{17}$$

where $C_{\mu,\gamma,L} > 0$ *can depend only on* μ, γ, L. *Moreover, with the same probability,*

$$\frac{\|\hat{S}_{\mathcal{I}}\|_2^2}{|\mathcal{I}|} \leq C_{\mu,\gamma,L} \frac{æ(\sigma \vee \mathbf{a})^2\, |\tilde{\Omega}|\, \log(d)}{n} + \frac{\mathbf{a}^2 s}{m_2}.$$

Remarks 1. The upper bound (17) can be decomposed into two terms. The first term is proportional to rM/n. It is of the same order as in the case of the usual matrix completion, see [18,20]. The second term accounts for the corruption. It is proportional to the number of corrupted columns s and to the number of corrupted observations $|\tilde{\Omega}|$. This term vanishes if there is no corruption, i.e., when $S_0 = \mathbf{0}$.

2. If all entries of A_0 are observed, i.e., the matrix decomposition problem is considered, the bound (17) is analogous to the corresponding bound in [1]. Indeed, then $|\tilde{\Omega}| = sm_1$, $N = m_1 m_2$, $æ \leq 2$ and we get

$$\frac{\|L_0 - \hat{L}\|_2^2}{m_1 m_2} + \frac{\|S_0 - \hat{S}\|_2^2}{m_1 m_2} \lesssim \&(\sigma \vee \mathbf{a})^2 \left(\frac{r\,M}{m_1 m_2} + \frac{s}{m_2} \right).$$

The estimator studied in [1] for matrix decomposition problem is similar to our program (13). The difference between these estimators is that in (13) the minimization is over $\|\cdot\|_\infty$-balls while the program of [1] uses the minimization over $\|\cdot\|_{2,\infty}$-balls and requires the knowledge of a bound on the norm $\|L_0\|_{2,\infty}$ of the unknown matrix L_0.

3. Suppose that the number of corrupted columns is small ($s \ll m_2$). Then, Corollary 7 guarantees, that the prediction error of our estimator is small whenever the number of non-corrupted observations n satisfies the following condition

$$n \gtrsim (m_1 \vee m_2)\mathrm{rank}(L_0) + |\tilde{\Omega}| \tag{18}$$

where $|\tilde{\Omega}|$ is the number of corrupted observations. This quantifies the sample size sufficient for successful (robust to corruptions) matrix completion. When the rank r of L_0 is small and $s \ll m_2$, the right hand side of (18) is considerably smaller than the total number of entries $m_1 m_2$.

4. By changing the numerical constants, one can obtain that the upper bound (17) is valid with probability $1 - 6d^{-\alpha}$ for any $\alpha \geq 1$.

5 Entrywise sparse corruptions

We assume now that S_0 has s non-zero entries but they do not necessarily lay in a small subset of columns. We will also assume that $s \leq \frac{m_1 m_2}{2}$. We use now the l_1-regularizer $\mathcal{R}$. Then the convex program (5) takes the form

$$(\hat{L}, \hat{S}) \in \underset{\substack{\|L\|_\infty \leq \mathbf{a} \\ \|S\|_\infty \leq \mathbf{a}}}{\arg\min} \left\{ \frac{1}{N} \sum_{i=1}^{N} (Y_i - \langle X_i, L + S \rangle)^2 + \lambda_1 \|L\|_* + \lambda_2 \|S\|_1 \right\}. \tag{19}$$

The support $\tilde{\mathcal{I}} = \{(j,k) \,:\, (S_0)_{jk} \neq 0\}$ of the non-zero entries of S_0 satisfies $|\tilde{\mathcal{I}}| = s$. Also, $\|\mathbf{Id}_{\tilde{\Omega}}\|_1 = |\tilde{\Omega}|$ so that Ψ_2, Ψ_3, and Ψ_4 take form

$$\Psi_2'' = \mu\, \mathbf{a}\, |\tilde{\Omega}| \left(\frac{\mathbf{a}\, \lambda_2}{\lambda_1} \mathbb{E}\left(\|\Sigma_R\| \right) + æ\lambda_2 + \mathbf{a}\, \mathbb{E}\|\Sigma_R\|_{2,\infty} \right),$$

$$\Psi_3'' = \frac{\mu\, |\tilde{\Omega}| (\mathbf{a}^2 + \sigma^2 \log(d))}{N} \left(\frac{\mathbf{a}\mathbb{E}\left(\|\Sigma_R\| \right)}{\lambda_1} + \frac{\mathbf{a}\mathbb{E}\|\Sigma_R\|_{2,\infty}}{\lambda_2} + æ \right) + \frac{\mathbf{a}^2 s}{m_1 m_2},$$

$$\Psi_4'' = \mu\, \mathbf{a}^2 \sqrt{\frac{\log(d)}{n}} + \mu\, \mathbf{a}\, |\tilde{\Omega}| \left[æ\, \lambda_2 + \mathbf{a}\, \mathbb{E}\|\Sigma_R\|_{2,\infty} \right] + \left[\frac{\mathbf{a}\, \mathbb{E}\|\Sigma_R\|_{2,\infty}}{\lambda_2} + æ \right] \frac{\mu\, |\tilde{\Omega}| \left(\mathbf{a}^2 + \sigma^2 \log(d) \right)}{N}.$$

Specializing Theorem 1 to this case yields the following corollary:

Corollary 8 *Assume that* $\|L_0\|_\infty \leq \mathbf{a}$ *and* $\|S_0\|_\infty \leq \mathbf{a}$. *Let the regularization parameters* (λ_1, λ_2) *satisfy*

$$\lambda_1 > 4\,\|\Sigma\| \quad \textit{and} \quad \lambda_2 \geq 4\left(\|\Sigma\|_\infty + 2\mathbf{a}\|W\|_\infty \right).$$

Then, with probability at least $1 - 4.5\, d^{-1}$, *for any solution* $(\hat{L}, \hat{S})$ *of the convex program* (19) *with such regularization parameters* (λ_1, λ_2) *we have*

$$\frac{\|L_0 - \hat{L}\|_2^2}{m_1 m_2} + \frac{\|S_0 - \hat{S}\|_2^2}{m_1 m_2} \leq C\, \left\{ \Psi_1 + \Psi_2'' + \Psi_3'' \right\}$$

where C *is an absolute constant. Moreover, with the same probability,*

$$\frac{\|\hat{S}_{\mathcal{I}}\|_2^2}{|\mathcal{I}|} \leq C \Psi_4''.$$

In order to get a bound in a closed form we need to obtain suitable upper bounds on the stochastic terms Σ, Σ_R and W. We provide such bounds under the following additional assumption on the sampling distribution.

Assumption 9 There exists a positive constant $\gamma \geq 1$ such that

$$\max_{i,j} \pi_{ij} \leq \frac{\mu_1}{|\mathcal{I}|}.$$

This assumption prevents any entry from being sampled too often and guarantees that the observations are well spread out over the non-corrupted entries. Assumptions 1 and 9 imply that the sampling distribution Π is approximately uniform in the sense that $\pi_{jk} \asymp \frac{1}{|\mathcal{I}|}$. In particular, since $|\mathcal{I}| \leq \frac{m_1 m_2}{2}$, Assumption 9 implies Assumption 2 for $L = 2\mu_1$.

Lemma 10 *Let the distribution Π on $\mathcal{X}'$ satisfy Assumptions 1, and 9. Let also Assumption 3 hold. Then, there exists an absolute constant $C > 0$ such that, for any $t > 0$, the following bounds on the norms of the stochastic terms hold with probability at least $1 - e^{-t}$, as well as the associated bounds in expectation.*

(i)
$$\|W\|_\infty \leq C\left(\frac{\mu_1}{æ m_1 m_2} + \sqrt{\frac{\mu_1\,(t+\log d))}{æ N m_1 m_2}} + \frac{t+\log d}{N}\right) \quad \textit{and}$$
$$\mathbb{E}\|W\|_\infty \leq C\left(\frac{\mu_1}{æ m_1 m_2} + \sqrt{\frac{\mu_1 \log d}{æ N m_1 m_2}} + \frac{\log d}{N}\right);$$

(ii)
$$\|\Sigma\|_\infty \leq C\sigma\left(\sqrt{\frac{\mu_1(t+\log d)}{æ N m_1 m_2}} + \frac{t+\log d}{N}\right) \quad \textit{and}$$
$$\mathbb{E}\,\|\Sigma\|_\infty \leq C\sigma\left(\sqrt{\frac{\mu_1 \log d}{æ N m_1 m_2}} + \frac{\log d}{N}\right);$$

(iii)
$$\|\Sigma_R\|_\infty \leq C\left(\sqrt{\frac{\mu_1(t+\log d)}{n\, m_1 m_2}} + \frac{t+\log d}{n}\right) \quad \textit{and}$$
$$\mathbb{E}\,\|\Sigma_R\|_\infty \leq C\left(\sqrt{\frac{\mu_1 \log d}{n\, m_1 m_2}} + \frac{\log d}{n}\right).$$

Using Lemma 6(i), and Lemma 10, under the conditions

$$\frac{m_1 m_2 \log d}{\mu_1} \geq n \geq \frac{2\, m \log(d) \log^2(m)}{L} \tag{20}$$

we can choose the regularization parameters λ_1 and λ_2 in the following way:

$$\lambda_1 = C(\sigma \vee \mathbf{a})\sqrt{\frac{\mu_1 \log(d)}{Nm}} \qquad \text{and}$$
$$\lambda_2 = C(\sigma \vee \mathbf{a})\frac{\log(d)}{N}. \tag{21}$$

With this choice of the regularization parameters, Corollary 8 and Lemma 10 imply the following result.

Corollary 11 *Let the distribution Π on $\mathcal{X}'$ satisfy Assumptions 1, and 9. Let Assumption 3 hold and $\|L_0\|_\infty \leq \mathbf{a}$, $\|S_0\|_\infty \leq \mathbf{a}$. Assume that $N \leq m_1 m_2$ and that condition (20) holds. Then, with probability at least $1 - 6/d$ for any solution $(\hat{L}, \hat{S})$ of the convex program (19) with the regularization parameters (λ_1, λ_2) given by (21), we have*

$$\frac{\|L_0 - \hat{L}\|_2^2}{m_1 m_2} + \frac{\|S_0 - \hat{S}\|_2^2}{m_1 m_2} \leq C_{\mu,\mu_1} æ(\sigma \vee \mathbf{a})^2 \log(d) \frac{r\,M + |\tilde{\Omega}|}{n} + \frac{\mathbf{a}^2 s}{m_1 m_2} \tag{22}$$

where $C_{\mu,\mu_1} > 0$ can depend only on μ and μ_1. Moreover, with the same probability

$$\frac{\|\hat{S}_{\mathcal{I}}\|_2^2}{|\mathcal{I}|} \leq C_{\mu,\mu_1} \frac{æ(\sigma \vee \mathbf{a})^2 |\tilde{\Omega}| \log(d)}{n} + \frac{\mathbf{a}^2 s}{m_1 m_2}.$$

Remarks 1. As in the columnwise sparsity case, we can recognize two terms in the upper bound (22). The first term is proportional to $r\,M/n$. It is of the same order as the rate of convergence for the usual matrix completion, see [18,20]. The second term accounts for the corruptions and is proportional to the number s of nonzero entries in S_0 and to the number of corrupted observations $|\tilde{\Omega}|$. We will prove in Sect. 6 below that these error terms are of the correct order up to a logarithmic factor.

2. If $s \ll n < m_1 m_2$, the bound (22) implies that one can estimate a low-rank matrix from a nearly minimal number of observations, even when a part of the observations has been corrupted.
3. If all entries of A_0 are observed, i.e., the matrix decomposition problem is considered, the bound (22) is analogous to the corresponding bound in [1]. Indeed, then $|\tilde{\Omega}| \leq s$, $N = m_1 m_2$, $æ \leq 2$ and we get

$$\frac{\|L_0 - \hat{L}\|_2^2}{m_1 m_2} + \frac{\|S_0 - \hat{S}\|_2^2}{m_1 m_2} \lesssim \&(\sigma \vee \mathbf{a})^2 \left(\frac{r\,M}{m_1 m_2} + \frac{s}{m_1 m_2} \right).$$

6 Minimax lower bounds

In this section, we prove the minimax lower bounds showing that the rates attained by our estimator are optimal up to a logarithmic factor. We will denote by $\inf_{(\hat{L},\hat{S})}$ the infimum over all pairs of estimators $(\hat{L}, \hat{S})$ for the components L_0 and S_0 in the decomposition $A_0 = L_0 + S_0$ where both $\hat{L}$ and $\hat{S}$ take values in $\mathbb{R}^{m_1 \times m_2}$. For any $A_0 \in \mathbb{R}^{m_1 \times m_2}$, let $\mathbb{P}_{A_0}$ denote the probability distribution of the observations $(X_1, Y_1, \ldots, X_n, Y_n)$ satisfying (1).

We begin with the case of columnwise sparsity. For any matrix $S \in \mathbb{R}^{m_1 \times m_2}$, we denote by $\|S\|_{2,0}$ the number of nonzero columns of S. For any integers $0 \leq r \leq \min(m_1, m_2)$, $0 \leq s \leq m_2$ and any $\mathbf{a} > 0$, we consider the class of matrices

$$\begin{aligned} \mathcal{A}_{GS}(r, s, \mathbf{a}) = \big\{ & A_0 = L_0 + S_0 \in \mathbb{R}^{m_1 \times m_2} : \operatorname{rank}(L_0) \leq r,\ \|L_0\|_\infty \leq \mathbf{a}, \\ & \text{and } \|S_0\|_{2,0} \leq s\,,\ \|S_0\|_\infty \leq \mathbf{a} \big\} \end{aligned} \tag{23}$$

and define

$$\psi_{GS}(N, r, s) = (\sigma \wedge \mathbf{a})^2 \left(\frac{Mr + |\tilde{\Omega}|}{n} + \frac{s}{m_2} \right).$$

The following theorem gives a lower bound on the estimation risk in the case of columnwise sparsity.

Theorem 2 *Suppose that* $m_1, m_2 \geq 2$. *Fix* $\mathbf{a} > 0$ *and integers* $1 \leq r \leq \min(m_1, m_2)$ *and* $1 \leq s \leq m_2/2$. *Let Assumption* 9 *be satisfied. Assume that* $Mr \leq n$, $|\tilde{\Omega}| \leq sm_1$ *and* $æ \leq 1 + s/m_2$. *Suppose that the variables* ξ_i *are i.i.d. Gaussian* $\mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$, *for* $i = 1, \ldots, n$. *Then, there exist absolute constants* $\beta \in (0, 1)$ *and* $c > 0$, *such that*

$$\inf_{(\hat{L},\hat{S})} \sup_{(L_0,S_0)\in \mathcal{A}_{GS}(r,s,\mathbf{a})} \mathbb{P}_{A_0}\left(\frac{\|\hat{L} - L_0\|_2^2}{m_1 m_2} + \frac{\|\hat{S} - S_0\|_2^2}{m_1 m_2} > c\psi_{GS}(N, r, s) \right) \geq \beta.$$

We turn now to the case of entrywise sparsity. For any matrix $S \in \mathbb{R}^{m_1 \times m_2}$, we denote by $\|S\|_0$ the number of nonzero entries of S. For any integers $0 \leq r \leq \min(m_1, m_2)$, $0 \leq s \leq m_1 m_2/2$ and any $\mathbf{a} > 0$, we consider the class of matrices

$$\begin{aligned} \mathcal{A}_S(r, s, a) = \big\{ & A_0 = L_0 + S_0 \in \mathbb{R}^{m_1 \times m_2} : \operatorname{rank}(L_0) \leq r, \\ & \|S_0\|_0 \leq s,\ \|L_0\|_\infty \leq \mathbf{a}, \|S_0\|_\infty \leq \mathbf{a} \} \end{aligned}$$

and define

$$\psi_S(N, r, s) = (\sigma \wedge \mathbf{a})^2 \left\{ \frac{Mr + |\tilde{\Omega}|}{n} + \frac{s}{m_1 m_2} \right\}.$$

We have the following theorem for the lower bound in the case of entrywise sparsity.

Theorem 3 *Assume that* $m_1, m_2 \geq 2$. *Fix* $\mathbf{a} > 0$ *and integers* $1 \leq r \leq \min(m_1, m_2)$ *and* $1 \leq s \leq m_1 m_2/2$. *Let Assumption* 9 *be satisfied. Assume that* $Mr \leq n$ *and there exists a constant* $\rho > 0$ *such that* $|\tilde{\Omega}| \leq \rho\, r\, M$. *Suppose that the variables* ξ_i *are i.i.d. Gaussian* $\mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$, *for* $i = 1, \ldots, n$. *Then, there exist absolute constants* $\beta \in (0, 1)$ *and* $c > 0$, *such that*

$$\inf_{(\hat{L},\hat{S})} \sup_{(L_0,S_0)\in \mathcal{A}_S(r,s,\mathbf{a})} \mathbb{P}_{A_0}\left(\frac{\|\hat{L} - L_0\|_2^2}{m_1 m_2} + \frac{\|\hat{S} - S_0\|_2^2}{m_1 m_2} > c\psi_S(N, r, s) \right) \geq \beta. \tag{24}$$

Acknowledgements The work of O. Klopp was conducted as part of the project Labex MME-DII (ANR11-LBX-0023-01). The work of K. Lounici was supported in part by Simons Grant 315477 and by NSF Career Grant DMS-1454515. The work of A.B. Tsybakov was supported by GENES and by the French National Research Agency (ANR) under the Grants IPANEMA (ANR-13-BSH1-0004-02), Labex ECODEC (ANR—11-LABEX-0047), ANR—11-IDEX-0003-02, and by the "Chaire Economie et Gestion des Nouvelles Données", under the auspices of Institut Louis Bachelier, Havas-Media and Paris-Dauphine. The authors want to thank the anonymous referee for his extremely valuable remarks.

Appendix A: Proofs of Theorem 1 and of Corollary 7

A.1: Proof of Theorem 1

The proofs of the upper bounds have similarities with the methods developed in [18] for noisy matrix completion but the presence of corruptions in our setting requires a new approach, in particular, for proving "restricted strong convexity property" (Lemma 15) which is the main difficulty in the proof.

Recall that our estimator is defined as

$$(\hat{L},\hat{S}) \in \underset{\substack{\|L\|_\infty \le \mathbf{a} \\ \|S\|_\infty \le \mathbf{a}}}{\arg\min} \left\{ \frac{1}{N}\sum_{i=1}^{N} (Y_i - \langle X_i, L+S\rangle)^2 + \lambda_1 \|L\|_* + \lambda_2 \mathcal{R}(S) \right\}$$

and our goal is to bound from above the Frobenius norms $\|L_0 - \hat{L}\|_2^2$ and $\|S_0 - \hat{S}\|_2^2$.

(1) Set $\mathcal{F}(L,S) = \frac{1}{N}\sum_{i=1}^{N} (Y_i - \langle X_i, L+S\rangle)^2 + \lambda_1\|L\|_* + \lambda_2\mathcal{R}(S)$, $\Delta L = L_0 - \hat{L}$ and $\Delta S = S_0 - \hat{S}$. Using the inequality $\mathcal{F}(\hat{L},\hat{S}) \le \mathcal{F}(L_0, S_0)$ and (1) we get

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} (\langle X_i, \Delta L + \Delta S\rangle + \xi_i)^2 + \lambda_1\|\hat{L}\|_* + \lambda_2\mathcal{R}(\hat{S}) \\ \le \frac{1}{N}\sum_{i=1}^{N} \xi_i^2 + \lambda_1\|L_0\|_* + \lambda_2\mathcal{R}(S_0).\end{aligned}$$

After some algebra this implies

$$\begin{aligned}\frac{1}{N}\sum_{i\in\Omega} \langle X_i, \Delta L + \Delta S\rangle^2 \le & \underbrace{\frac{2}{N}\sum_{i\in\tilde{\Omega}} |\langle \xi_i X_i, \Delta L + \Delta S\rangle| - \frac{1}{N}\sum_{i\in\tilde{\Omega}} \langle X_i, \Delta L + \Delta S\rangle^2}_{\mathbf{I}} \\ & + \underbrace{2\,|\langle \Sigma, \Delta L\rangle| + \lambda_1 \left(\|L_0\|_* - \|\hat{L}\|_*\right)}_{\mathbf{II}} \\ & + \underbrace{2\,|\langle \Sigma, \Delta S_{\mathcal{I}}\rangle| + \lambda_2 \left(\mathcal{R}(S_0) - \mathcal{R}(\hat{S})\right)}_{\mathbf{III}} \end{aligned} \tag{25}$$

where $\Sigma = \frac{1}{N}\sum_{i\in\Omega} \xi_i X_i$ and we have used the equality $\langle \Sigma, \Delta S\rangle = \langle \Sigma, \Delta S_{\mathcal{I}}\rangle$. We now estimate each of the three terms on the right hand side of (25) separately. This will be done on the random event

$$\mathcal{U} = \left\{ \max_{1\le i\le N} |\xi_i| \le C_*\sigma\sqrt{\log d} \right\} \tag{26}$$

where $C_* > 0$ is a suitably chosen constant. Using a standard bound on the maximum of sub-gaussian variables and the constraint $N \leq m_1 m_2$ we get that there exists an absolute constant $C_* > 0$ such that $\mathbb{P}(\mathcal{U}) \geq 1 - \frac{1}{2d}$. In what follows, we take this constant C_* in the definition of $\mathcal{U}$.

We start by estimating **I**. On the event $\mathcal{U}$, we get

$$\mathbf{I} \leq \frac{1}{N} \sum_{i \in \tilde{\Omega}} \xi_i^2 \leq \frac{C\sigma^2 |\tilde{\Omega}| \log(d)}{N}. \tag{27}$$

Now we estimate **II**. For a linear vector subspace S of a euclidean space, let P_S denote the orthogonal projector on S and let $S^\perp$ denote the orthogonal complement of S. For any $A \in \mathbb{R}^{m_1 \times m_2}$, let $u_j(A)$ and $v_j(A)$ be the left and right orthonormal singular vectors of A, respectively . Denote by $S_1(A)$ the linear span of $\{u_j(A)\}$, and by $S_2(A)$ the linear span of $\{v_j(A)\}$. We set

$$\mathbf{P}_A^\perp(B) = P_{S_1^\perp(A)} B P_{S_2^\perp(A)} \quad \text{and} \quad \mathbf{P}_A(B) = B - \mathbf{P}_A^\perp(B).$$

By definition of $\mathbf{P}_{L_0}^\perp$, for any matrix B the singular vectors of $\mathbf{P}_{L_0}^\perp(B)$ are orthogonal to the space spanned by the singular vectors of L_0. This implies that $\left\| L_0 + \mathbf{P}_{L_0}^\perp(\Delta L) \right\|_* = \|L_0\|_* + \left\| \mathbf{P}_{L_0}^\perp(\Delta L) \right\|_*$. Thus,

$$\begin{aligned}
\|\hat{L}\|_* &= \|L_0 + \Delta L\|_* \\
&= \left\| L_0 + \mathbf{P}_{L_0}^\perp(\Delta L) + \mathbf{P}_{L_0}(\Delta L) \right\|_* \\
&\geq \left\| L_0 + \mathbf{P}_{L_0}^\perp(\Delta L) \right\|_* - \left\| \mathbf{P}_{L_0}(\Delta L) \right\|_* \\
&= \|L_0\|_* + \left\| \mathbf{P}_{L_0}^\perp(\Delta L) \right\|_* - \left\| \mathbf{P}_{L_0}(\Delta L) \right\|_*,
\end{aligned}$$

which yields

$$\|L_0\|_* - \|\hat{L}\|_* \leq \left\| \mathbf{P}_{L_0}(\Delta L) \right\|_* - \left\| \mathbf{P}_{L_0}^\perp(\Delta L) \right\|_*. \tag{28}$$

Using (28) and the duality between the nuclear and the operator norms, we obtain

$$\mathbf{II} \leq 2\|\Sigma\| \|\Delta L\|_* + \lambda_1 \left(\left\| \mathbf{P}_{L_0}(\Delta L) \right\|_* - \left\| \mathbf{P}_{L_0}^\perp(\Delta L) \right\|_* \right).$$

The assumption that $\lambda_1 \geq 4\|\Sigma\|$ and the triangle inequality imply

$$\mathbf{II} \leq \frac{3}{2}\lambda_1 \left\| \mathbf{P}_{L_0}(\Delta L) \right\|_* \leq \frac{3}{2}\lambda_1 \sqrt{2r}\, \|\Delta L\|_2 \tag{29}$$

where $r = \operatorname{rank}(L_0)$ and we have used that $\operatorname{rank}(\mathbf{P}_{L_0}(\Delta L)) \leq 2 \operatorname{rank}(L_0)$.

For the third term in (25), we use the duality between the $\mathcal{R}$ and $\mathcal{R}^*$, and the identity $\Delta S_{\mathcal{I}} = -\hat{S}_{\mathcal{I}}$:

$$\mathbf{III} \leq 2\mathcal{R}^*(\Sigma)\mathcal{R}(\hat{S}_{\mathcal{I}}) + \lambda_2(\mathcal{R}(S_0) - \mathcal{R}(\hat{S})).$$

This and the assumption that $\lambda_2 \geq 4\mathcal{R}^*(\Sigma)$ imply

$$\mathbf{III} \leq \lambda_2 \mathcal{R}(S_0). \tag{30}$$

Plugging (29), (30) and (27) in (25) we get that, on the event $\mathcal{U}$,

$$\frac{1}{n}\sum_{i\in\Omega} \langle X_i, \Delta L + \Delta S\rangle^2 \leq \frac{3\,æ\,\lambda_1}{\sqrt{2}}\sqrt{r}\,\|\Delta L\|_2 + æ\lambda_2\mathcal{R}(S_0) + \frac{C\sigma^2|\tilde{\Omega}|\log(d)}{n} \tag{31}$$

where $æ = N/n$.

(2) Second, we will show that a kind of restricted strong convexity holds for the random sampling operator given by (X_i) on a suitable subset of matrices. In words, we prove that the observation operator captures a substantial component of any pair of matrices L, S belonging to a properly chosen *constrained set* (cf. Lemma 15(ii) below for the exact statement). This will imply that, with high probability,

$$\frac{1}{n}\sum_{i\in\Omega} \langle X_i, \Delta L + \Delta S\rangle^2 \geq \|\Delta L + \Delta S\|^2_{L_2(\Pi)} - \mathcal{E} \tag{32}$$

with an appropriate residual $\mathcal{E}$, whenever we prove that $(\Delta L, \Delta S)$ belongs to the constrained set. This will be a substantial element of the remaining part of the proof. The result of the theorem will then be deduced by combining (31) and (32).

We start by defining our constrained set. For positive constants δ_1 and δ_2, we first introduce the following set of matrices where ΔS should lie:

$$\mathcal{B}(\delta_1, \delta_2) = \{B \in \mathbb{R}^{m_1\times m_2} \,:\, \|B\|^2_{L_2(\Pi)} \leq \delta_1^2 \text{ and } \mathcal{R}(B) \leq \delta_2\}. \tag{33}$$

The constants δ_1 and δ_2 define the constraints on the $L_2(\Pi)$-norm and on the sparsity of the component S. The error term $\mathcal{E}$ in (32) depends on δ_1 and δ_2. We will specify the suitable values of δ_1 and δ_2 for the matrix ΔS later. Next, we define the following set of pairs of matrices:

$$\mathcal{D}(\tau, \kappa) = \left\{(A, B) \in \mathbb{R}^{m_1\times m_2} \,:\, \|A + B\|^2_{L_2(\Pi)} \geq \sqrt{\frac{64\,\log(d)}{\log(6/5)\;n}},\right.$$
$$\left.\|A + B\|_\infty \leq 1,\, \|A\|_* \leq \sqrt{\tau}\,\|A_{\mathcal{I}}\|_2 + \kappa\right\}$$

where κ and $\tau < m_1 \wedge m_2$ are some positive constants. This will be used for $A = \Delta L$ and $B = \Delta S$. If the $L_2(\Pi)$-norm of the sum of two matrices is too small, the right hand side of (32) is negative. The first inequality in the definition of $\mathcal{D}(\tau, \kappa)$ prevents from this. Condition $\|A\|_* \leq \sqrt{\tau}\,\|A_{\mathcal{I}}\|_2 + \kappa$ is a relaxed form of the condition $\|A\|_* \leq \sqrt{\tau}\,\|A\|_2$ satisfied by matrices with rank τ. We will show that, with high probability, the matrix ΔL satisfies this condition with $\tau = C\,\mathrm{rank}(L_0)$ and a small κ. To prove it, we need the bound $\mathcal{R}(B) \leq \delta_2$ on the corrupted part.

Finally, define our *constrained set* as the intersection

$$\mathcal{D}(\tau, \kappa) \cap \left\{\mathbb{R}^{m_1 \times m_2} \times \mathcal{B}(\delta_1, \delta_2)\right\}.$$

We now return to the proof of the theorem. To prove (11), we bound separately the norms $\|\Delta L\|_2$ and $\|\Delta S\|_2$. Note that

$$\begin{aligned}\|\Delta L\|_2^2 &\leq \|\Delta L_{\mathcal{I}}\|_2^2 + \|\Delta L_{\tilde{\mathcal{I}}}\|_2^2 \leq \|\Delta L_{\mathcal{I}}\|_2^2 + 4\mathbf{a}^2|\tilde{\mathcal{I}}| \\ &\leq \mu|\mathcal{I}|\|\Delta L_{\mathcal{I}}\|_{L_2(\Pi)}^2 + 4\mathbf{a}^2|\tilde{\mathcal{I}}| \end{aligned} \tag{34}$$

and similarly,

$$\|\Delta S\|_2^2 \leq \mu|\mathcal{I}|\|\Delta S_{\mathcal{I}}\|_{L_2(\Pi)}^2 + 4\mathbf{a}^2|\tilde{\mathcal{I}}|.$$

In view of these inequalities, it is enough to bound the quantities $\|\Delta S_{\mathcal{I}}\|_{L_2(\Pi)}^2$ and $\|\Delta L_{\mathcal{I}}\|_2^2$. A bound on $\|\Delta S_{\mathcal{I}}\|_{L_2(\Pi)}^2$ with the rate as claimed in (11) is given in Lemma 14 below. In order to bound $\|\Delta L_{\mathcal{I}}\|_{L_2(\Pi)}^2$ (or $\|\Delta L_{\mathcal{I}}\|_2^2$ according to cases), we will need the following argument.

Case 1 Suppose that $\|\Delta L + \Delta S\|_{L_2(\Pi)}^2 < 16\mathbf{a}^2\sqrt{\dfrac{64\,\log(d)}{\log(6/5)\;n}}$. Then a straightforward inequality

$$\|\Delta L + \Delta S\|_{L_2(\Pi)}^2 \geq \frac{1}{2}\|\Delta L\|_{L_2(\Pi)}^2 - \|\Delta S\|_{L_2(\Pi)}^2 \tag{35}$$

together with Lemma 14 below implies that, with probability at least $1 - 2.5/d$,

$$\|\Delta L\|_{L_2(\Pi)}^2 \leq \Delta_1 \tag{36}$$

where

$$\begin{aligned}\Delta_1 = C\Psi_4/\mu = C\Bigg\{&\mathbf{a}^2\sqrt{\frac{\log(d)}{n}} + \mathbf{a}\,\mathcal{R}(\mathbf{Id}_{\tilde{\Omega}})\left[\text{æ}\,\lambda_2 + \mathbf{a}\,\mathbb{E}\left(\mathcal{R}^*(\Sigma_R)\right)\right] \\ &+ \left(\frac{\mathbf{a}\,\mathbb{E}\left(\mathcal{R}^*(\Sigma_R)\right)}{\lambda_2} + \text{æ}\right)\frac{|\tilde{\Omega}|\left(\mathbf{a}^2 + \sigma^2\log(d)\right)}{N}\Bigg\}.\end{aligned}$$

Note also that $\Psi_4 \le C(\Psi_1 + \Psi_2 + \Psi_3)$. In view of (34), (36) and of fact that $|\mathcal{I}| \le m_1 m_2$, the bound on $\|\Delta L\|_2^2$ stated in the theorem holds with probability at least $1 - 2.5/d$.

Case 2 Assume now that $\|\Delta L + \Delta S\|_{L_2(\Pi)}^2 \ge 16\mathbf{a}^2 \sqrt{\dfrac{64\,\log(d)}{\log(6/5)\; n}}$. We will show that in this case and with an appropriate choice of δ_1, δ_2, τ and κ, the pair $\frac{1}{4\mathbf{a}}(\Delta L, \Delta S)$ belongs to the intersection $\mathcal{D}(\tau,\kappa) \cap \{\mathbb{R}^{m_1\times m_2} \times \mathcal{B}(\delta_1,\delta_2)\}$.

Lemma 13 below and (27) imply that, on the event $\mathcal{U}$,

$$\begin{aligned} \|\Delta L\|_* &\le 4\sqrt{2r}\|\Delta L\|_2 + \frac{\lambda_2\,\mathbf{a}}{\lambda_1}\mathcal{R}(\mathbf{Id}_{\tilde{\Omega}}) + \frac{C\sigma^2|\tilde{\Omega}|\log(d)}{N\lambda_1} \\ &\le 4\sqrt{2r}\|\Delta L_{\mathcal{I}}\|_2 + 8\mathbf{a}\sqrt{2r|\tilde{\mathcal{I}}|} + \frac{\lambda_2\,\mathbf{a}}{\lambda_1}\mathcal{R}(\mathbf{Id}_{\tilde{\Omega}}) + \frac{C\sigma^2|\tilde{\Omega}|\log(d)}{N\lambda_1}. \end{aligned} \tag{37}$$

Lemma 14 yields that, with probability at least $1 - 2.5\,d^{-1}$,

$$\frac{\Delta S}{4\mathbf{a}} \in \mathcal{B}\left(\frac{\sqrt{\Delta_1}}{4\mathbf{a}}, 2\mathcal{R}(\mathbf{Id}_{\tilde{\Omega}}) + \frac{C|\tilde{\Omega}|\left(\mathbf{a}^2 + \sigma^2\log(d)\right)}{4\mathbf{a}N\lambda_2}\right) = \bar{\mathcal{B}}.$$

This property and (37) imply that $\frac{1}{4\mathbf{a}}(\Delta L, \Delta S) \in \mathcal{D}(\tau,\kappa) \cap \{\mathbb{R}^{m_1\times m_2} \times \bar{\mathcal{B}}\}$ with probability at least $1 - 2.5\,d^{-1}$, where

$$\tau = 32r \quad \text{and} \quad \kappa = 2\sqrt{2r|\tilde{\mathcal{I}}|} + \frac{\lambda_2}{4\lambda_1}\mathcal{R}(\mathbf{Id}_{\tilde{\Omega}}) + \frac{C\sigma^2|\tilde{\Omega}|\log(d)}{4\mathbf{a}\,N\lambda_1}.$$

Therefore, we can apply Lemma 15(ii). From Lemma 15(ii) and (31) we obtain that, with probability at least $1 - 4.5\,d^{-1}$,

$$\frac{1}{2}\|\Delta L + \Delta S\|_{L_2(\Pi)}^2 \le \frac{3\,æ\,\lambda_1}{\sqrt{2}}\sqrt{r}\,\|\Delta L\|_2 + C\mathcal{E} \tag{38}$$

where

$$\begin{aligned} \mathcal{E} &= \mu\,\mathbf{a}^2\,r\,|\mathcal{I}|\,(\mathbb{E}\,(\|\Sigma_R\|))^2 + 8\mathbf{a}^2\sqrt{2r|\tilde{\mathcal{I}}|}\,\mathbb{E}\,(\|\Sigma_R\|) \\ &\quad + \lambda_2\mathcal{R}(\mathbf{Id}_{\tilde{\Omega}})\left(\frac{\mathbf{a}^2\mathbb{E}\,(\|\Sigma_R\|)}{\lambda_1} + \mathbf{a}\,æ\right) \\ &\quad + \frac{|\tilde{\Omega}|\left(\mathbf{a}^2 + \sigma^2\log(d)\right)}{N}\left(\frac{\mathbf{a}\,\mathbb{E}\,(\|\Sigma_R\|)}{\lambda_1} + \frac{\mathbf{a}\,\mathbb{E}\,(\mathcal{R}^*(\Sigma_R))}{\lambda_2} + æ\right) + \Delta_1. \end{aligned} \tag{39}$$

Using an elementary argument and then (34) we find

$$\begin{aligned}\frac{3\,æ}{\sqrt{2}}\lambda_1\sqrt{r}\,\|\Delta L\|_2 &\leq \frac{9\,æ^2\,\mu\,m_1\,m_2\,r\,\lambda_1^2}{2}+\frac{\|\Delta L\|_2^2}{4\mu m_1 m_2}\\ &\leq \frac{9\,æ^2\,\mu\,m_1\,m_2\,r\,\lambda_1^2}{2}+\frac{\|\Delta L_{\mathcal{I}}\|_2^2}{4\mu m_1 m_2}+\frac{\mathbf{a}^2|\tilde{\mathcal{I}}|}{\mu\,m_1 m_2}.\end{aligned}$$

This inequality and (38) yield

$$\|\Delta L+\Delta S\|_{L_2(\Pi)}^2 \leq \frac{9\,æ^2\,\mu\,m_1\,m_2\,r\,\lambda_1^2}{4}+\frac{\|\Delta L_{\mathcal{I}}\|_2^2}{4\mu m_1 m_2}+\frac{\mathbf{a}^2|\tilde{\mathcal{I}}|}{\mu\,m_1 m_2}+C\mathcal{E}.$$

Using again (35), Lemma 14, (9) and the bound $|\mathcal{I}|\leq m_1 m_2$ we obtain

$$\frac{\|\Delta L_{\mathcal{I}}\|_2^2}{\mu m_1 m_2}\leq C\left\{æ^2\,\mu\,m_1\,m_2\,r\,\lambda_1^2+\frac{\mathbf{a}^2|\tilde{\mathcal{I}}|}{\mu\,m_1 m_2}+\mathcal{E}\right\}.$$

This and the inequality $\sqrt{2r|\tilde{\mathcal{I}}|}\,\mathbb{E}\left(\|\Sigma_R\|\right)\leq \dfrac{|\tilde{\mathcal{I}}|}{\mu\,m_1 m_2}+\mu\,m_1 m_2\,r\,\left(\mathbb{E}\left(\|\Sigma_R\|\right)\right)^2$ imply that, with probability at least $1-4.5\,d^{-1}$,

$$\frac{\|\Delta L_{\mathcal{I}}\|_2^2}{m_1 m_2}\leq C\left\{\Psi_1+\Psi_2+\Psi_3\right\}. \tag{40}$$

In view of (40) and (34), $\|\Delta L\|_2^2$ is bounded by the right hand side of (11) with probability at least $1-4.5\,d^{-1}$. Finally, inequality (12) follows from Lemma 14, (9) and the identity $\Delta S_{\mathcal{I}}=-\hat{S}_{\mathcal{I}}$.

Lemma 12 *Assume that* $\lambda_2\geq 4\left(\mathcal{R}^*(\Sigma)+2\mathbf{a}\mathcal{R}^*(W)\right)$. *Then, we have*

$$\mathcal{R}(\Delta S_{\mathcal{I}})\leq 3\mathcal{R}(\Delta S_{\tilde{\Omega}})+\frac{1}{N\lambda_2}\left[4\mathbf{a}^2|\tilde{\Omega}|+\sum_{i\in\tilde{\Omega}}\xi_i^2\right]$$

Proof Let $\partial\|\cdot\|_*$, and $\partial\mathcal{R}$ denote the subdifferentials of $\|\cdot\|_*$ and of $\mathcal{R}$, respectively. By the standard condition for optimality over a convex set (see [2, Chapter 4, Section 2, Corollary 6]), we have

$$\begin{aligned}&-\frac{2}{N}\sum_{i=1}^{N}(Y_i-\langle X_i,\hat{L}+\hat{S}\rangle)\langle X_i,L+S-\hat{L}-\hat{S}\rangle\\ &+\lambda_1\langle\partial\|\hat{L}\|_*,L-\hat{L}\rangle+\lambda_2\langle\partial\mathcal{R}(\hat{S}),S-\hat{S}\rangle\geq 0\end{aligned} \tag{41}$$

for all feasible pairs (L, S). In particular, for $(\hat{L}, S_0)$ we obtain

$$-\frac{2}{N}\sum_{i=1}^{N}(Y_i - \langle X_i, \hat{L}+\hat{S}\rangle)\langle X_i, \Delta S\rangle + \lambda_2\langle \partial\mathcal{R}(\hat{S}), \Delta S\rangle \geq 0,$$

which implies

$$\begin{aligned}
&-\frac{2}{N}\sum_{i=1}^{N}\langle X_i, \Delta S\rangle^2 - \frac{2}{N}\sum_{i\in\tilde{\Omega}}\langle X_i, \Delta L\rangle\langle X_i, \Delta S\rangle - \frac{2}{N}\sum_{i\in\tilde{\Omega}}\xi_i\langle X_i, \Delta S\rangle \\
&-\frac{2}{N}\sum_{i\in\Omega}\langle X_i, \Delta L\rangle\langle X_i, \Delta S\rangle - 2\langle \Sigma, \Delta S\rangle + \lambda_2\langle\partial\mathcal{R}(\hat{S}), \Delta S\rangle \geq 0.
\end{aligned}$$

Using the elementary inequality $2ab \leq a^2 + b^2$ and the bound $\|\Delta L\|_\infty \leq 2\mathbf{a}$ we find

$$\begin{aligned}
&-\frac{2}{N}\sum_{i=1}^{N}\langle X_i, \Delta S\rangle^2 - \frac{2}{N}\sum_{i\in\tilde{\Omega}}\langle X_i, \Delta L\rangle\langle X_i, \Delta S\rangle - \frac{2}{N}\sum_{i\in\tilde{\Omega}}\xi_i\langle X_i, \Delta S\rangle \\
&\leq \frac{1}{N}\sum_{i\in\tilde{\Omega}}\langle X_i, \Delta L\rangle^2 + \frac{1}{N}\sum_{i\in\tilde{\Omega}}\xi^2 \\
&\leq \frac{4\mathbf{a}^2|\tilde{\Omega}|}{N} + \frac{1}{N}\sum_{i\in\tilde{\Omega}}\xi_i^2.
\end{aligned}$$

Combining the last two displays we get

$$\begin{aligned}
\lambda_2\langle\partial\mathcal{R}(\hat{S}), \hat{S}-S_0\rangle &\leq 2\left|\left\langle\frac{1}{N}\sum_{i\in\Omega}\langle X_i, \Delta L\rangle X_i, \Delta S\right\rangle\right| + 2\left|\langle\Sigma, \Delta S\rangle\right| \\
&\quad + \frac{4\mathbf{a}^2|\tilde{\Omega}|}{N} + \frac{1}{N}\sum_{i\in\tilde{\Omega}}\xi_i^2 \\
&\leq 2\mathcal{R}^*\left(\frac{1}{N}\sum_{i\in\Omega}\langle X_i, \Delta L\rangle X_i\right)\mathcal{R}(\Delta S) + 2\mathcal{R}^*(\Sigma)\mathcal{R}(\Delta S) \\
&\quad + \frac{4\mathbf{a}^2|\tilde{\Omega}|}{N} + \frac{1}{N}\sum_{i\in\tilde{\Omega}}\xi_i^2.
\end{aligned} \tag{42}$$

By Lemma 18,

$$\mathcal{R}^*\left(\frac{1}{N}\sum_{i\in\Omega}\langle X_i, \Delta L\rangle X_i\right) \leq 2\mathbf{a}\mathcal{R}^*(W) \tag{43}$$

where $W = \frac{1}{N}\sum_{i\in\Omega} X_i$. On the other hand, the convexity of $\mathcal{R}(\cdot)$ and the definition of subdifferential imply

$$\mathcal{R}(S_0) \geq \mathcal{R}(\hat{S}) + \langle \partial\mathcal{R}(\hat{S}), \Delta S\rangle. \tag{44}$$

Plugging (43) and (44) in (42) we obtain

$$\lambda_2(\mathcal{R}(\hat{S}) - \mathcal{R}(S_0)) \leq 4\mathbf{a}\mathcal{R}^*(W)\mathcal{R}(\Delta S) + 2\mathcal{R}^*(\Sigma)\mathcal{R}(\Delta S) + \frac{4\mathbf{a}^2|\tilde{\Omega}|}{N} + \frac{1}{N}\sum_{i\in\tilde{\Omega}}\xi_i^2.$$

Next, the decomposability of $\mathcal{R}(\cdot)$, the identity $(S_0)_{\mathcal{I}} = 0$ and the triangle inequality yield

$$\begin{aligned}\mathcal{R}(S_0 - \Delta S) - \mathcal{R}(S_0) &= \mathcal{R}\left((S_0 - \Delta S)_{\tilde{\mathcal{I}}}\right) + \mathcal{R}\left((S_0 - \Delta S)_{\mathcal{I}}\right) - \mathcal{R}\left((S_0)_{\tilde{\mathcal{I}}}\right)\\ &\geq \mathcal{R}\left((\Delta S)_{\mathcal{I}}\right) - \mathcal{R}\left((\Delta S)_{\tilde{\mathcal{I}}}\right).\end{aligned}$$

Since $\lambda_2 \geq 4\left(2\mathbf{a}\mathcal{R}^*(W) + \mathcal{R}^*(\Sigma)\right)$ the last two displays imply

$$\begin{aligned}&\lambda_2\left(\mathcal{R}\left((\Delta S)_{\mathcal{I}}\right) - \mathcal{R}\left((\Delta S)_{\tilde{\mathcal{I}}}\right)\right)\\ &\quad\leq \frac{\lambda_2}{2}\left(\mathcal{R}\left(\Delta S_{\tilde{\mathcal{I}}}\right) + \mathcal{R}\left((\Delta S)_{\mathcal{I}}\right)\right) + \frac{4\mathbf{a}^2|\tilde{\Omega}|}{N} + \frac{1}{N}\sum_{i\in\tilde{\Omega}}\xi_i^2.\end{aligned}$$

Thus,

$$\mathcal{R}\left(\Delta S_{\mathcal{I}}\right) \leq 3\mathcal{R}\left(\Delta S_{\tilde{\mathcal{I}}}\right) + \frac{1}{N\lambda_2}\left[4\mathbf{a}^2|\tilde{\Omega}| + \sum_{i\in\tilde{\Omega}}\xi_i^2\right]. \tag{45}$$

Since we assume that all unobserved entries of S_0 are zero, we have $(S_0)_{\tilde{\mathcal{I}}} = (S_0)_{\tilde{\Omega}}$. On the other hand, $S_{\tilde{\mathcal{I}}} = \hat{S}_{\tilde{\Omega}}$ as $\mathcal{R}(\cdot)$ is a monotonic norm. Indeed, adding to S a non-zero element on the non-observed part increases $\mathcal{R}(S)$ but does not modify $\frac{1}{N}\sum_{i=1}^N\left(Y_i - \langle X_i, L+S\rangle\right)^2$. To conclude, we have $\Delta S_{\tilde{\mathcal{I}}} = \Delta S_{\tilde{\Omega}}$, which together with (45), implies the Lemma. □

Lemma 13 *Suppose that* $\lambda_1 \geq 4\|\Sigma\|$ *and* $\lambda_2 \geq 4\mathcal{R}^*(\Sigma)$. *Then,*

$$\left\|\mathbf{P}_{L_0}^{\perp}(\Delta L)\right\|_* \leq 3\left\|\mathbf{P}_{L_0}(\Delta L)\right\|_* + \frac{\lambda_2\,\mathbf{a}}{\lambda_1}\mathcal{R}(\mathbf{Id}_{\tilde{\Omega}}) + \frac{1}{N\lambda_1}\sum_{i\in\tilde{\Omega}}\xi_i^2.$$

Proof Using (41) for $(L, S) = (L_0, S_0)$ we obtain

$$\begin{aligned}&-\frac{2}{N}\sum_{i=1}^N\langle X_i, \Delta S + \Delta L\rangle^2 - \frac{2}{N}\sum_{i\in\tilde{\Omega}}\langle \xi_i X_i, \Delta L + \Delta S\rangle\\ &-2\left\langle \Sigma, (\Delta S)_{\mathcal{I}}\right\rangle - 2\left\langle \Sigma, \Delta L\right\rangle + \lambda_1\langle \partial\|\hat{L}\|_*, \Delta L\rangle + \lambda_2\langle\partial\mathcal{R}(\hat{S}), \Delta S\rangle \geq 0.\end{aligned} \tag{46}$$

The convexity of $\|\cdot\|_*$ and of $\mathcal{R}(\cdot)$ and the definition of the subdifferential imply

$$\|L_0\|_* \geq \|\hat{L}\|_* + \langle \partial\|\hat{L}\|_*, \Delta L\rangle$$
$$\mathcal{R}(S_0) \geq \mathcal{R}(\hat{S}) + \langle \partial\mathcal{R}(\hat{S}), \Delta S\rangle.$$

Together with (46), this yields

$$\lambda_1(\|\hat{L}\|_* - \|L_0\|_*) + \lambda_2(\mathcal{R}(\hat{S}) - \mathcal{R}(S_0)) \leq 2\|\Sigma\|\|\Delta L\|_* + 2\mathcal{R}^*(\Sigma)\mathcal{R}(\Delta S_{\mathcal{I}}) + \frac{1}{N}\sum_{i\in\tilde{\Omega}}\xi_i^2.$$

Using the conditions $\lambda_1 \geq 4\|\Sigma\|$, $\lambda_2 \geq 4\mathcal{R}^*(\Sigma)$, the triangle inequality and (28) we get

$$\begin{aligned}&\lambda_1\left(\left\|\mathbf{P}_{L_0}^{\perp}(\Delta L)\right\|_* - \left\|\mathbf{P}_{L_0}(\Delta L)\right\|_*\right) + \lambda_2(\mathcal{R}(\hat{S}) - \mathcal{R}(S_0))\\ &\quad\leq \frac{\lambda_1}{2}\left(\left\|\mathbf{P}_{L_0}^{\perp}(\Delta L)\right\|_* + \left\|\mathbf{P}_{L_0}(\Delta L)\right\|_*\right) + \frac{\lambda_2}{2}\mathcal{R}(\hat{S}_{\mathcal{I}}) + \frac{1}{N}\sum_{i\in\tilde{\Omega}}\xi_i^2.\end{aligned}$$

Since we assume that all unobserved entries of S_0 are zero, we obtain $\mathcal{R}(S_0) \leq \mathbf{a}\mathcal{R}(\mathbf{Id}_{\tilde{\Omega}})$. Using this inequality in the last display proves the lemma. □

Lemma 14 *Let $n > m_1$ and $\lambda_2 \geq 4\left(\mathcal{R}^*(\Sigma) + 2\mathbf{a}\mathcal{R}^*(W)\right)$. Suppose that the distribution Π on $\mathcal{X}'$ satisfies Assumptions 1 and 2. Let $\|S_0\|_\infty \leq \mathbf{a}$ for some constant $\mathbf{a}$ and let Assumption 3 be satisfied. Then, with probability at least $1 - 2.5\,d^{-1}$,*

$$\|\Delta S\|_{L_2(\Pi)}^2 \leq C\Psi_4/\mu, \tag{47}$$

and

$$\mathcal{R}(\Delta S) \leq 8\mathbf{a}\mathcal{R}(\mathbf{Id}_{\tilde{\Omega}}) + \frac{|\tilde{\Omega}|(4\mathbf{a}^2 + C\sigma^2\log(d))}{N\lambda_2}. \tag{48}$$

Proof Using the inequality $\mathcal{F}(\hat{L}, \hat{S}) \leq \mathcal{F}(\hat{L}, S_0)$ and (1) we obtain

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}(\langle X_i, \Delta L + \Delta S\rangle + \xi_i)^2 + \lambda_2\mathcal{R}(\hat{S})\\ &\quad\leq \frac{1}{N}\sum_{i=1}^{N}(\langle X_i, \Delta L\rangle + \xi_i)^2 + \lambda_2\mathcal{R}(S_0)\end{aligned}$$

which implies

$$\frac{1}{N}\sum_{i\in\Omega}\langle X_i,\Delta S\rangle^2+\frac{1}{N}\sum_{i\in\tilde\Omega}\langle X_i,\Delta S\rangle^2+\frac{2}{N}\sum_{i\in\tilde\Omega}\langle X_i,\Delta L\rangle\langle X_i,\Delta S\rangle+\frac{2}{N}\sum_{i\in\tilde\Omega}\langle \xi_i X_i,\Delta S\rangle$$
$$+\frac{2}{N}\sum_{i\in\Omega}\langle X_i,\Delta L\rangle\left\langle X_i,\Delta S_{\mathcal{I}}\right\rangle+2\left\langle\Sigma,\Delta S_{\mathcal{I}}\right\rangle+\lambda_2\mathcal{R}(\hat S)\le\lambda_2\mathcal{R}(S_0).$$

From Lemma 18 and the duality between $\mathcal{R}$ and $\mathcal{R}^*$ we obtain

$$\frac{1}{N}\sum_{i\in\Omega}\langle X_i,\Delta S\rangle^2\le 2(2\mathbf{a}\,\mathcal{R}^*(W)+\mathcal{R}^*(\Sigma))\mathcal{R}(\Delta S_{\mathcal{I}})+\lambda_2(\mathcal{R}(S_0)-\mathcal{R}(\hat S))$$
$$+\frac{2}{N}\sum_{i\in\tilde\Omega}\langle X_i,\Delta L\rangle^2+\frac{2}{N}\sum_{i\in\tilde\Omega}\xi^2.$$

Since here $\Delta S_{\mathcal{I}}=-\hat S_{\mathcal{I}}$ and $\lambda_2\ge 4\left(\mathcal{R}^*(\Sigma)+2\mathbf{a}\mathcal{R}^*(W)\right)$ it follows that

$$\frac{1}{N}\sum_{i\in\Omega}\langle X_i,\Delta S\rangle^2\le\lambda_2\mathcal{R}\left(S_0\right)+\frac{2}{N}\sum_{i\in\tilde\Omega}\langle X_i,\Delta L\rangle^2+\frac{2}{N}\sum_{i\in\tilde\Omega}\xi^2. \tag{49}$$

Now, Lemma 12 and the bound $\|\Delta S\|_\infty\le 2\mathbf{a}$ imply that, on the event $\mathcal{U}$ defined in (26),

$$\begin{aligned}\mathcal{R}(\Delta S)&\le 4\mathcal{R}(\Delta S_{\tilde\Omega})+\frac{|\tilde\Omega|(4\mathbf{a}^2+C\sigma^2\log(d))}{N\lambda_2}\\&\le 8\mathbf{a}\mathcal{R}(\mathbf{Id}_{\tilde\Omega})+\frac{|\tilde\Omega|(4\mathbf{a}^2+C\sigma^2\log(d))}{N\lambda_2}.\end{aligned} \tag{50}$$

Thus, (48) is proved. To prove (47), consider the following two cases.
Case I $\|\Delta S\|^2_{L_2(\Pi)}<4\mathbf{a}^2\sqrt{\frac{64\log(d)}{\log(6/5)\,n}}$. Then (47) holds trivially.
Case II $\|\Delta S\|^2_{L_2(\Pi)}\ge 4\mathbf{a}^2\sqrt{\frac{64\log(d)}{\log(6/5)\,n}}$. Then inequality (50) and the bound $\|\Delta S\|_\infty\le 2\mathbf{a}$ imply that, on the event $\mathcal{U}$,

$$\frac{\Delta S}{2\mathbf{a}}\in\mathcal{C}\left(4\,\mathcal{R}(\mathbf{Id}_{\tilde\Omega})+\frac{|\tilde\Omega|\left(8\mathbf{a}^2+C\sigma^2\log(d)\right)}{2\mathbf{a}\,N\lambda_2}\right)$$

where, for any $\delta>0$, the set $\mathcal{C}(\delta)$ is defined as:

$$\mathcal{C}(\delta)=\left\{A\in\mathbb{R}^{m_1\times m_2}:\|A\|_\infty\le 1,\ \|A\|^2_{L_2(\Pi)}\ge\sqrt{\frac{64\,\log(d)}{\log\left(6/5\right)\,n}},\ \mathcal{R}(A)\le\delta\right\}. \tag{51}$$

Thus, we can apply Lemma 15(i) below. In view of this lemma, the inequalities (49), (27), $\|\Delta L\|_\infty \leq 2\mathbf{a}$ and $\mathcal{R}(S_0) \leq \mathbf{a}\mathcal{R}(\mathbf{Id}_{\tilde{\mathcal{I}}})$ imply that (47) holds with probability at least $1 - 2.5\, d^{-1}$. □

Lemma 15 *Let the distribution Π on $\mathcal{X}'$ satisfy Assumptions 1 and 2. Let $\delta, \delta_1, \delta_2, \tau$, and κ be positive constants. Then, the following properties hold.*

(i) *With probability at least* $1 - \dfrac{2}{d}$,

$$\frac{1}{n}\sum_{i\in\Omega}\langle X_i, S\rangle^2 \geq \frac{1}{2}\|S\|^2_{L_2(\Pi)} - 8\delta\mathbb{E}(\mathcal{R}^*(\Sigma_R))$$

for any $S \in \mathcal{C}(\delta)$.

(ii) *With probability at least* $1 - \dfrac{2}{d}$,

$$\begin{aligned}\frac{1}{n}\sum_{i\in\Omega}\langle X_i, L+S\rangle^2 \geq \frac{1}{2}\|L+S\|^2_{L_2(\Pi)} &- \{360\mu\,|\mathcal{I}|\,\tau\,(\mathbb{E}\,(\|\Sigma_R\|))^2 \\ &+4\delta_1^2 + 8\delta_2\,\mathbb{E}(\mathcal{R}^*(\Sigma_R)) + 8\kappa\mathbb{E}\,(\|\Sigma_R\|)\}\end{aligned}$$

for any pair $(L, S) \in \mathcal{D}(\tau,\kappa)\cap\{\mathbb{R}^{m_1\times m_2}\times\mathcal{B}(\delta_1,\delta_2)\}$.

Proof We give a unified proof of (i) and (ii). Let $A = S$ for (i) and $A = L + S$ for (ii). Set

$$\mathcal{E} = \begin{cases} 8\delta\mathbb{E}\,(\mathcal{R}^*(\Sigma_R)) & \text{for (i)} \\ 360\mu\,|\mathcal{I}|\,\tau\,(\mathbb{E}\,(\|\Sigma_R\|))^2 + 4\delta_1^2 + 8\delta_2\,\mathbb{E}\,(\mathcal{R}^*(\Sigma_R)) + 8\kappa\mathbb{E}\,(\|\Sigma_R\|) & \text{for (ii)}\end{cases}$$

and

$$\mathcal{C} = \begin{cases}\mathcal{C}(\delta) & \text{for (i)} \\ \mathcal{D}(\tau,\kappa)\cap(\mathbb{R}^{m_1\times m_2}\times\mathcal{B}(\delta_1,\delta_2)) & \text{for (ii).}\end{cases}$$

To prove the lemma it is enough to show that the probability of the random event

$$\mathcal{B} = \left\{\exists\, A\in\mathcal{C} \text{ such that } \left|\frac{1}{n}\sum_{i\in\Omega}\langle X_i, A\rangle^2 - \|A\|^2_{L_2(\Pi)}\right| > \frac{1}{2}\|A\|^2_{L_2(\Pi)} + \mathcal{E}\right\}$$

is smaller than $2/d$. In order to estimate the probability of $\mathcal{B}$, we use a standard peeling argument. Set $\nu = \sqrt{\dfrac{64\,\log(d)}{\log(6/5)\;n}}$ and $\alpha = \dfrac{6}{5}$. For $l \in \mathbb{N}$, define

$$S_l = \{A\in\mathcal{C} : \alpha^{l-1}\nu \leq \|A\|^2_{L_2(\Pi)} \leq \alpha^l\nu\}.$$

If the event $\mathcal{B}$ holds, there exist $l \in \mathbb{N}$ and a matrix $A \in \mathcal{C} \cap S_l$ such that

$$\begin{aligned} \left| \frac{1}{n} \sum_{i \in \Omega} \langle X_i, A \rangle^2 - \|A\|^2_{L_2(\Pi)} \right| &> \frac{1}{2} \|A\|^2_{L_2(\Pi)} + \mathcal{E} \\ &> \frac{1}{2} \alpha^{l-1} \nu + \mathcal{E} \\ &= \frac{5}{12} \alpha^l \nu + \mathcal{E}. \end{aligned} \tag{52}$$

For each $l \in \mathbb{N}$, consider the random event

$$\mathcal{B}_l = \left\{ \exists\, A \in \mathcal{C}'(\alpha^l \nu) \;:\; \left| \frac{1}{n} \sum_{i \in \Omega} \langle X_i, A \rangle^2 - \|A\|^2_{L_2(\Pi)} \right| > \frac{5}{12} \alpha^l \nu + \mathcal{E} \right\}$$

where

$$\mathcal{C}'(T) = \{ A \in \mathcal{C} \;:\; \|A\|^2_{L_2(\Pi)} \leq T \}, \quad \forall T > 0.$$

Note that $A \in S_l$ implies that $A \in \mathcal{C}'(\alpha^l \nu)$. This and (52) grant the inclusion $\mathcal{B} \subset \cup_{l=1}^{\infty} \mathcal{B}_l$. By Lemma 16, $\mathbb{P}(\mathcal{B}_l) \leq \exp(-c_5\, n\, \alpha^{2l} \nu^2)$ where $c_5 = 1/128$. Using the union bound we find

$$\begin{aligned} \mathbb{P}(\mathcal{B}) &\leq \sum_{l=1}^{\infty} \mathbb{P}(\mathcal{B}_l) \\ &\leq \sum_{l=1}^{\infty} \exp(-c_5\, n\, \alpha^{2l}\, \nu^2) \\ &\leq \sum_{l=1}^{\infty} \exp(-(2\, c_5\, n\, \log(\alpha)\, \nu^2) l) \end{aligned}$$

where we have used the inequality $e^x \geq x$. We finally obtain, for $\nu = \sqrt{\dfrac{64 \log(d)}{\log(6/5)\; n}}$,

$$\mathbb{P}(\mathcal{B}) \leq \frac{\exp\left(-2\, c_5\, n\, \log(\alpha)\, \nu^2\right)}{1 - \exp\left(-2\, c_5\, n\, \log(\alpha)\, \nu^2\right)} = \frac{\exp(-\log(d))}{1 - \exp(-\log(d))}.$$

□

Let

$$Z_T = \sup_{A \in \mathcal{C}'(T)} \left| \frac{1}{n} \sum_{i \in \Omega} \langle X_i, A \rangle^2 - \|A\|^2_{L_2(\Pi)} \right|.$$

Lemma 16 *Let the distribution* Π *on* $\mathcal{X}'$ *satisfy Assumptions 1 and 2. Then,*

$$\mathbb{P}\left(Z_T > \frac{5}{12}T + \mathcal{E}\right) \leq \exp(-c_5\, n\, T^2)$$

where $c_5 = \dfrac{1}{128}$.

Proof We follow a standard approach: first we show that Z_T concentrates around its expectation and then we bound from above the expectation. Since $\|A\|_\infty \leq 1$ for all $A \in \mathcal{C}'(T)$, we have $|\langle X_i, A\rangle| \leq 1$. We use first a Talagrand type concentration inequality, cf. [4, Theorem 14.2], implying that

$$\mathbb{P}\left(Z_T \geq \mathbb{E}\left(Z_T\right) + \frac{1}{9}\left(\frac{5}{12}T\right)\right) \leq \exp(-c_5\, n\, T^2) \tag{53}$$

where $c_5 = \dfrac{1}{128}$. Next, we bound the expectation $\mathbb{E}\left(Z_T\right)$. By a standard symmetrization argument (see e.g. [19, Theorem 2.1]) we obtain

$$\begin{aligned}\mathbb{E}\left(Z_T\right) &= \mathbb{E}\left(\sup_{A\in\mathcal{C}'(T)}\left|\frac{1}{n}\sum_{i\in\Omega}\langle X_i, A\rangle^2 - \mathbb{E}(\langle X, A\rangle^2)\right|\right)\\ &\leq 2\mathbb{E}\left(\sup_{A\in\mathcal{C}'(T)}\left|\frac{1}{n}\sum_{i\in\Omega}\epsilon_i\,\langle X_i, A\rangle^2\right|\right)\end{aligned}$$

where $\{\epsilon_i\}_{i=1}^n$ is an i.i.d. Rademacher sequence. Then, the contraction inequality (see e.g. [19]) yields

$$\mathbb{E}\left(Z_T\right) \leq 8\mathbb{E}\left(\sup_{A\in\mathcal{C}'(T)}\left|\frac{1}{n}\sum_{i\in\Omega}\epsilon_i\,\langle X_i, A\rangle\right|\right) = 8\mathbb{E}\left(\sup_{A\in\mathcal{C}'(T)}|\langle\Sigma_R, A\rangle|\right)$$

where $\Sigma_R = \dfrac{1}{n}\sum_{i=1}^n \epsilon_i X_i$. Now, to obtain a bound on $\mathbb{E}(\sup_{A\in\mathcal{C}'(T)}|\langle\Sigma_R, A\rangle|)$ we will consider separately the cases $\mathcal{C} = \mathcal{C}(\delta)$ and $\mathcal{C} = \mathcal{D}(\tau,\kappa)\cap\{\mathbb{R}^{m_1\times m_2}\times\mathcal{B}(\delta_1,\delta_2)\}$.
Case I $A \in \mathcal{C}(\delta)$ and $\|A\|^2_{L_2(\Pi)} \leq T$. By the definition of $\mathcal{C}(\delta)$ we have $\mathcal{R}(A) \leq \delta$. Thus, by the duality between $\mathcal{R}$ and $\mathcal{R}^*$,

$$\mathbb{E}\left(Z_T\right) \leq 8\mathbb{E}\left(\sup_{\mathcal{R}(A)\leq\delta}|\langle\Sigma_R, A\rangle|\right) \leq 8\delta\,\mathbb{E}\left(\mathcal{R}^*(\Sigma_R)\right).$$

This and the concentration inequality (53) imply

$$\mathbb{P}\left(Z_T > \frac{5}{12}T + \mathcal{E}\right) \leq \exp(-c_5\, n\, T^2)$$

with $c_5 = \dfrac{1}{128}$ and $\mathcal{E} = 8\delta\,\mathbb{E}\left(\mathcal{R}^*(\Sigma_R)\right)$ as stated.

Case II $A = L + S$ where $(L, S) \in \mathcal{D}(\tau, \kappa)$, $S \in \mathcal{B}(\delta_1, \delta_2)$, and $\|L + S\|^2_{L_2(\Pi)} \le T$. Then, by the definition of $\mathcal{B}(\delta_1, \delta_2)$, we have $\mathcal{R}(S) \le \delta_2$. On the other hand, the definition of $\mathcal{D}(\tau, \kappa)$ yields

$$\|L\|_* \le \sqrt{\tau}\|L_{\mathcal{I}}\|_2 + \kappa$$

and

$$\|L\|_{L_2(\Pi)} \le \|L + S\|_{L_2(\Pi)} + \|S\|_{L_2(\Pi)} \le \sqrt{T} + \delta_1.$$

The last two inequalities imply

$$\|L\|_* \le \sqrt{\mu\,|\mathcal{I}|\,\tau}(\sqrt{T} + \delta_1) + \kappa := \Gamma_1.$$

Therefore we can write

$$\begin{aligned}\mathbb{E}\left(\sup_{A\in\mathcal{C}'(T)} |\langle\Sigma_R, A\rangle|\right) &\le 8\mathbb{E}\left(\sup_{\|L\|_*\le\Gamma_1} |\langle\Sigma_R, L\rangle| + \sup_{\mathcal{R}(S)\le\delta_2} |\langle\Sigma_R, S\rangle|\right)\\ &\le 8\left\{\Gamma_1\,\mathbb{E}\left(\|\Sigma_R\|\right) + \delta_2\,\mathbb{E}\left(\mathcal{R}^*(\Sigma_R)\right)\right\}.\end{aligned}$$

Combining this bound with the following elementary inequalities:

$$\begin{gathered}\frac{1}{9}\left(\frac{5}{12}T\right) + 8\sqrt{\mu\,|\mathcal{I}|\,\tau\,T}\,\mathbb{E}\left(\|\Sigma_R\|\right) \le \left(\frac{1}{9} + \frac{8}{9}\right)\frac{5}{12}T + 44\mu\,|\mathcal{I}|\,\tau\,\left(\mathbb{E}\left(\|\Sigma_R\|\right)\right)^2,\\ \delta_1\sqrt{\mu\,|\mathcal{I}|\,\tau},\mathbb{E}\left(\|\Sigma_R\|\right) \le \mu\,|\mathcal{I}|\,\tau\,\left(\mathbb{E}\left(\|\Sigma_R\|\right)\right)^2 + \frac{\delta_1^2}{2}\end{gathered}$$

and using the concentration bound (53) we obtain

$$\mathbb{P}\left(Z_T > \frac{5}{12}T + \mathcal{E}\right) \le \exp(-c_5\,n\,T^2)$$

with $c_5 = \dfrac{1}{128}$ and

$$\mathcal{E} = 360\mu\,|\mathcal{I}|\,\tau\,\left(\mathbb{E}\left(\|\Sigma_R\|\right)\right)^2 + 4\delta_1^2 + 8\delta_2\,\mathbb{E}\left(\mathcal{R}^*(\Sigma_R)\right) + 8\kappa\mathbb{E}\left(\|\Sigma_R\|\right) \quad (54)$$

as stated. ☐

A.2: Proof of Corollary 7

With λ_1 and λ_2 given by (16) we obtain

$$\Psi_1 = \mu^2 æ^2 (\sigma \vee \mathbf{a})^2 \frac{M\, r\, \log d}{N},$$
$$\Psi_2' \leq \mu^2 æ^2 (\sigma \vee \mathbf{a})^2 \log(d) \frac{|\tilde{\Omega}|}{N} + \frac{\mathbf{a}^2 s}{m_2},$$
$$\Psi_3' = \frac{\mu\, æ |\tilde{\Omega}| (\mathbf{a}^2 + \sigma^2 \log(d))}{N} + \frac{\mathbf{a}^2 s}{m_2}$$
$$\Psi_4' \leq \frac{\mu\, æ^2 |\tilde{\Omega}| (\mathbf{a}^2 + \sigma^2 \log(d))}{N} + \mathbf{a}^2 \sqrt{\frac{\log(d)}{n}} + \frac{\mathbf{a}^2 s}{m_2}.$$

Appendix B: Proof of Theorems 2 and 3

Note that the assumption $æ \leq 1 + s/m_2$ implies that

$$\frac{|\tilde{\Omega}|}{n} \leq \frac{s}{m_2}. \tag{55}$$

Assume w.l.o.g. that $m_1 \geq m_2$. For a $\gamma \leq 1$, define

$$\tilde{\mathcal{L}} = \left\{ \tilde{L} = (l_{ij}) \in \mathbb{R}^{m_1 \times r} : l_{ij} \in \left\{0, \gamma(\sigma \wedge \mathbf{a}) \left(\frac{rM}{n}\right)^{1/2}\right\}, \quad \forall 1 \leq i \leq m_1,\ 1 \leq j \leq r \right\},$$

and consider the associated set of block matrices

$$\mathcal{L} = \{ L = (\tilde{L} \quad \cdots \quad \tilde{L} \quad O) \in \mathbb{R}^{m_1 \times m_2} : \tilde{L} \in \tilde{\mathcal{L}} \},$$

where O denotes the $m_1 \times (m_2 - r\lfloor m_2/(2r) \rfloor)$ zero matrix, and $\lfloor x \rfloor$ is the integer part of x.

We define similarly the set of matrices

$$\tilde{\mathcal{S}} = \{ \tilde{S} = (s_{ij}) \in \mathbb{R}^{m_1 \times s} : s_{ij} \in \left\{0, \gamma(\sigma \wedge \mathbf{a})\right\}, \quad \forall 1 \leq i \leq m_1,\ 1 \leq j \leq s \},$$

and

$$\mathcal{S} = \{ S = (\tilde{O} \quad \tilde{S}) \in \mathbb{R}^{m_1 \times m_2} : \tilde{S} \in \tilde{\mathcal{S}} \},$$

where $\tilde{O}$ is the $m_1 \times (m_2 - s)$ zero matrix. We now set

$$\mathcal{A} = \{ A = L + S \,:\, L \in \mathcal{L},\ S \in \mathcal{S} \}.$$

Remark 2 In the case $m_1 < m_2$, we only need to change the construction of the low rank component of the test set. We first introduce a matrix $\tilde{L} = \left(\bar{L} \middle| O\right) \in \mathbb{R}^{r \times m_2}$

where $\bar{L} \in \mathbb{R}^{r\times(m_2/2)}$ with entries in $\{0, \gamma(\sigma \wedge \mathbf{a})(\frac{rM}{n})^{1/2}\}$ and then we replicate this matrix to obtain a block matrix L of size $m_1 \times m_2$

$$L = \begin{pmatrix} \tilde{L} \\ \hline \vdots \\ \hline \tilde{L} \\ \hline O \end{pmatrix}.$$

By construction, any element of $\mathcal{A}$ as well as the difference of any two elements of $\mathcal{A}$ can be decomposed into a low rank component L of rank at most r and a group sparse component S with at most s nonzero columns. In addition, the entries of any matrix in $\mathcal{A}$ take values in $[0, a]$. Thus, $\mathcal{A} \subset \mathcal{A}_{GS}(r, s, \mathbf{a})$.

We first establish a lower bound of the order rM/n. Let $\tilde{\mathcal{A}} \subset \mathcal{A}$ be such that for any $A = L+S \in \tilde{\mathcal{A}}$ we have $S = \mathbf{0}$. The Varshamov–Gilbert bound (cf. Lemma 2.9 in [25]) guarantees the existence of a subset $\mathcal{A}^0 \subset \tilde{\mathcal{A}}$ with cardinality $\mathrm{Card}(\mathcal{A}^0) \geq 2^{(rM)/8}+1$ containing the zero $m_1 \times m_2$ matrix $\mathbf{0}$ and such that, for any two distinct elements A_1 and A_2 of $\mathcal{A}^0$,

$$||A_1 - A_2||_2^2 \geq \frac{Mr}{8}\left(\gamma^2(\sigma \wedge \mathbf{a})^2\frac{Mr}{n}\right)\left\lfloor\frac{m_2}{r}\right\rfloor \geq \frac{\gamma^2}{16}(\sigma \wedge \mathbf{a})^2\, m_1 m_2\, \frac{Mr}{n}. \tag{56}$$

Since $\xi_i \sim \mathcal{N}(0, \sigma^2)$ we get that, for any $A \in \mathcal{A}^0$, the Kullback–Leibler divergence $K\left(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_A\right)$ between $\mathbb{P}_{\mathbf{0}}$ and $\mathbb{P}_A$ satisfies

$$K\left(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_A\right) = \frac{|\Omega|}{2\sigma^2}\|A\|_{L_2(\Pi)}^2 \leq \frac{\mu_1\gamma^2\, Mr}{2} \tag{57}$$

where we have used Assumption 9. From (57) we deduce that the condition

$$\frac{1}{\mathrm{Card}(\mathcal{A}^0) - 1}\sum_{A\in\mathcal{A}^0} K(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_A) \leq \frac{1}{16}\log\left(\mathrm{Card}(\mathcal{A}^0) - 1\right) \tag{58}$$

is satisfied if $\gamma > 0$ is chosen as a sufficiently small numerical constant. In view of (56) and (58), the application of Theorem 2.5 in [25] implies

$$\inf_{(\hat{L},\hat{S})}\ \sup_{(L_0,S_0)\in\mathcal{A}_{GS}(r,s,\mathbf{a})} \mathbb{P}_{A_0}\left(\frac{\|\hat{L} - L_0\|_2^2}{m_1 m_2} + \frac{\|\hat{S} - S_0\|_2^2}{m_1 m_2} > \frac{C(\sigma \wedge \mathbf{a})^2\, Mr}{n}\right) \geq \beta \tag{59}$$

for some absolute constants $\beta \in (0, 1)$.

We now prove the lower bound relative to the corruptions. Let $\bar{\mathcal{A}} \subset \mathcal{A}$ such that for any $A = L + S \in \bar{\mathcal{A}}$ we have $L = \mathbf{0}$. The Varshamov–Gilbert bound (cf. Lemma 2.9 in [25]) guarantees the existence of a subset $\mathcal{A}^0 \subset \bar{\mathcal{A}}$ with cardinality $\mathrm{Card}(\mathcal{A}^0) \geq 2^{(sm_1)/8} + 1$ containing the zero $m_1 \times m_2$ matrix $\mathbf{0}$ and such that, for any two distinct elements A_1 and A_2 of $\mathcal{A}^0$,

$$||S_1 - S_2||_2^2 \geq \frac{sm_1}{8}(\gamma^2(\sigma \wedge \mathbf{a})^2) = \frac{\gamma^2 (\sigma \wedge \mathbf{a})^2 s}{8m_2} m_1 m_2. \tag{60}$$

For any $A \in \mathcal{A}_0$, the Kullback–Leibler divergence between $\mathbb{P}_{\mathbf{0}}$ and $\mathbb{P}_A$ satisfies

$$K\left(\mathbb{P}_{\mathbf{0}}, \mathbb{P}_A\right) = \frac{|\tilde{\Omega}|}{2\sigma^2}\gamma^2(\sigma \wedge \mathbf{a})^2 \leq \frac{\gamma^2 m_1 s}{2}$$

which implies that condition (58) is satisfied if $\gamma > 0$ is chosen small enough. Thus, applying Theorem 2.5 in [25] we get

$$\inf_{(\hat{L},\hat{S})} \sup_{(L_0,S_0)\in \mathcal{A}_{GS}(r,s,\mathbf{a})} \mathbb{P}_{A_0}\left(\frac{\|\hat{L} - L_0\|_2^2}{m_1 m_2} + \frac{\|\hat{S} - S_0\|_2^2}{m_1 m_2} > \frac{C(\sigma \wedge \mathbf{a})^2 s}{m_2}\right) \geq \beta \tag{61}$$

for some absolute constant $\beta \in (0, 1)$. Theorem 2 follows from inequalities (55), (59) and (61).

The proof of Theorem 3 follows the same lines as that of Theorem 2. The only difference is that we replace $\tilde{S}$ by the following set

$$\{S = (s_{ij}) \in \mathbb{R}^{m_1 \times m_2} : s_{ij} \in \{0, \gamma(\sigma \wedge \mathbf{a})\}, \quad \forall 1 \leq i \leq m_1,\ \lfloor m_2/2 \rfloor + 1 \leq j \leq m_2\}.$$

We omit further details here.

Appendix C: Proof of Lemma 6

Part (i) of Lemma 6 is proved in Lemmas 5 and 6 in [18].

Proof of (ii) For the sake of brevity, we set $X_i(j,k) = \langle X_i, e_j(m_1)e_k(m_2)^\top \rangle$. By definition of Σ and $\|\cdot\|_{2,\infty}$, we have

$$\|\Sigma\|_{2,\infty}^2 = \max_{1\leq k\leq m_2} \sum_{j=1}^{m_1} \left(\frac{1}{N}\sum_{i\in\Omega} \xi_i X_i(j,k)\right)^2.$$

For any fixed k, we have

$$\sum_{j=1}^{m_1}\left(\frac{1}{N}\sum_{i\in\Omega}\xi_i X_i(j,k)\right)^2 = \frac{1}{N^2}\sum_{i_1,i_2\in\Omega}\xi_{i_1}\xi_{i_2}\sum_{j=1}^{m_1}X_{i_1}(j,k)X_{i_2}(j,k)$$
$$= \Xi^\top A_k \Xi, \tag{62}$$

where $\Xi = (\xi_1,\dots,\xi_n)^\top$ and $A_k \in \mathbb{R}^{|\Omega|\times|\Omega|}$ with entries

$$a_{i_1 i_2}(k) = \frac{1}{N^2}\sum_{j=1}^{m_1}X_{i_1}(j,k)X_{i_2}(j,k).$$

We freeze the X_i and we apply the version of Hanson–Wright inequality in [24] to get that there exists a numerical constant C such that with probability at least $1-e^{-t}$

$$|\Xi^\top A_k\Xi - \mathbb{E}[\Xi^\top A_k \Xi|X_i]| \le C\sigma^2\left(\|A_k\|_2\sqrt{t} + \|A_k\|t\right). \tag{63}$$

Next, we note that

$$\|A_k\|_2^2 = \sum_{i_1,i_2} a^2_{i_1 i_2}(k) \le \frac{1}{N^4}\sum_{i_1 i_2}\left(\sum_{j_1=1}^{m_1}X^2_{i_1}(j_1,k)\right)\left(\sum_{j_1=1}^{m_1}X^2_{i_2}(j_1,k)\right)$$
$$\le \frac{1}{N^4}\left[\sum_{i_1}\sum_{j_1=1}^{m_1}X^2_{i_1}(j_1,k)\right]^2 = \left[\frac{1}{N^2}\sum_{i_1}\sum_{j_1=1}^{m_1}X_{i_1}(j_1,k)\right]^2,$$

where we have used the Cauchy–Schwarz inequality in the first line and the relation $X_i^2(j,k) = X_i(j,k)$.

Note that $Z_i(k) := \sum_{j=1}^{m_1} X_i(j,k)$ follows a Bernoulli distribution with parameter $\pi_{\cdot k}$ and consequently $Z(k) = \sum_{i\in\Omega} Z_i(k)$ follows a Binomial distribution $B(|\Omega|, \pi_{\cdot k})$. We apply Bernstein's inequality (see, e.g., [4, page 486]) to get that, for any $t>0$,

$$\mathbb{P}\left(|Z(k) - \mathbb{E}[Z(k)]| \ge 2\sqrt{|\Omega|\pi_{\cdot k}t} + t\right) \le 2e^{-t}.$$

Consequently, we get with probability at least $1-2e^{-t}$ that

$$\|A_k\|_2^2 \le \left(\frac{|\Omega|\pi_{\cdot k} + 2\sqrt{|\Omega|\pi_{\cdot k}t} + t}{N^2}\right)^2$$

and, using $\|A_k\| \le \|A_k\|_2$, that

$$\|A_k\| \le \frac{|\Omega|\pi_{\cdot k} + 2\sqrt{|\Omega|\pi_{\cdot k}t} + t}{N^2}.$$

Note also that

$$\mathbb{E}[\Xi^\top A_k \Xi \big| X_i] = \frac{\sigma^2}{N^2} Z(k).$$

Combining the last three displays with (63) we get, up to a rescaling of the constants, with probability at least $1 - e^{-t}$ that

$$\sum_{j=1}^{m_1} \left(\frac{1}{N} \sum_{i \in \Omega_r} \xi_i X_i(j,k) \right)^2 \leq C \frac{\sigma^2}{N^2} \left(|\Omega| \pi_{\cdot k} + 2\sqrt{|\Omega| \pi_{\cdot k} t} + t \right) (1 + \sqrt{t} + t).$$

Replacing t by $t + \log m_2$ in the above display and using the union bound gives that, with probability at least $1 - e^{-t}$,

$$\begin{aligned} \|\Sigma\|_{2,\infty} &\leq C \frac{\sigma}{N} \left(|\Omega| \pi_{\cdot k} + 2\sqrt{|\Omega| \pi_{\cdot k} (t + \log m_2)} + (t + \log m_2) \right)^{1/2} \\ &\quad \times (1 + \sqrt{t + \log m_2} + t + \log m_2)^{1/2} \\ &= C \frac{\sigma}{N} \left(\sqrt{|\Omega| \pi_{\cdot k}} + \sqrt{t + \log m_2} \right) \left(1 + \sqrt{t + \log m_2} \right). \end{aligned}$$

Assuming that $\log m_2 \geq 1$ we get with probability at least $1 - e^{-t}$ that

$$\|\Sigma\|_{2,\infty} \leq C \frac{\sigma}{N} \left(\sqrt{|\Omega| \pi_{\cdot k} (t + \log m_2)} + (t + \log m_2) \right).$$

Using (14), we get that there exists a numerical constant $C > 0$ such with probability at least $1 - e^{-t}$

$$\|\Sigma\|_{2,\infty} \leq C \frac{\sigma}{N} \left(\sqrt{\frac{\gamma^{1/2} n (t + \log m_2)}{m_2}} + (t + \log m_2) \right).$$

Finally, we use Lemma 17 to obtain the required bound on $\mathbb{E}\|\Sigma\|_{2,\infty}$.

Proof of (iii) We follow the same lines as in the proof of part (ii) above. The only difference is to replace ξ_i by ϵ_i, σ by 1 and N by n.

Proof of (iv) We need to establish the bound on

$$\|W\|_{2,\infty}^2 = \max_{1 \leq k \leq m_2} \sum_{j=1}^{m_1} \left(\frac{1}{N} \sum_{i \in \Omega} X_i(j,k) \right)^2.$$

For any fixed k, we have

$$\sum_{j=1}^{m_1}\left(\frac{1}{N}\sum_{i\in\Omega}X_i(j,k)\right)^2=\frac{1}{N^2}\sum_{i\in\Omega}\sum_{j=1}^{m_1}X_i^2(j,k)+\frac{1}{N^2}\sum_{i_1\neq i_2}\sum_{j=1}^{m_1}X_{i_1}(j,k)X_{i_2}(j,k).$$

The first term on the right hand side of the last display can be written as

$$\frac{1}{N^2}\sum_{i\in\Omega}\sum_{j=1}^{m_1}X_i^2(j,k)=\frac{1}{N^2}\sum_{i\in\Omega}\sum_{j=1}^{m_1}X_i(j,k)=\frac{Z(k)}{N^2}.$$

Using the concentration bound on $Z(k)$ in the proof of part (ii) above, we get that, with probability at least $1-e^{-t}$,

$$\frac{1}{N^2}\sum_{i\in\Omega}\sum_{j=1}^{m_1}X_i^2(j,k)\leq\frac{|\Omega|}{N^2}\pi_{\cdot k}+2\frac{\sqrt{|\Omega|\pi_{\cdot k}t}}{N^2}+\frac{t}{N^2}. \tag{64}$$

Next, the random variable

$$U_2=\frac{1}{N^2}\sum_{i_1\neq i_2}\sum_{j=1}^{m_1}[X_{i_1}(j,k)X_{i_2}(j,k)-\pi_{j,k}^2]$$

is a U-statistic of order 2. We use now a Bernstein-type concentration inequality for U-statistics. To this end, we set $X_i(\cdot,k)=(X_i(1,k),\ldots,X_i(m_1,k))^\top$ and

$$h(X_{i_1}(\cdot,k),X_{i_2}(\cdot,k))=\sum_{j=1}^{m_1}[X_{i_1}(j,k)X_{i_2}(j,k)-\pi_{j,k}^2].$$

Let $e_0(m_1)=\mathbf{0}_{m_1}$ be the zero vector in $\mathbb{R}^{m_1}$. Note that $X_i(\cdot,k)$ takes values in $\{e_j(m_1),\ 0\leq j\leq m_1\}$. For any function $g:\{e_j(m_1),\ 0\leq j\leq m_1\}^2\to\mathbb{R}$, we set $\|g\|_{L^\infty}=\max_{0\leq j_1,j_2\leq m_1}|g(e_{j_1}(m_1),e_{j_2}(m_1))|$.

We will need the following quantities to control the tail behavior of U_2

$$\begin{aligned}
\mathbf{A}&=\|h\|_{L^\infty},\\
\mathbf{B}^2&=\max\left\{\left\|\sum_{i_1}\mathbb{E}h^2(X_{i_1}(\cdot,k),\cdot)\right\|_{L^\infty},\left\|\sum_{i_2}\mathbb{E}h^2(\cdot,X_{i_2}(\cdot,k))\right\|_{L^\infty}\right\},\\
\mathbf{C}&=\sum_{i_1\neq i_2}\mathbb{E}[h^2(X_{i_1}(\cdot,k),X'_{i_2}(\cdot,k))]\quad\text{and}
\end{aligned}$$

$$\mathbf{D} = \sup \Bigg\{ \mathbb{E} \sum_{i_1 \neq i_2} h\left[X_{i_1}(\cdot,k),\, X'_{i_2}(\cdot,k)\right] f_{i_1}[X_{i_1}(\cdot,k)] g_{i_2}[X'_{i_2}(\cdot,k)],$$

$$\mathbb{E} \sum_{i_1} f_{i_1}^2(X_{i_1}(\cdot,k)) \leq 1, \mathbb{E} \sum_{i_2} g_{i_2}^2(X'_{i_2}(\cdot,k)) \leq 1 \Bigg\},$$

where $X'_i(\cdot,k)$ are independent replications of $X_i(\cdot,k)$ and $f, g : \mathbb{R}^{m_1} \to \mathbb{R}$.

We now evaluate the above quantities in our particular setting. It is not hard to see that $\mathbf{A} = \max\{\pi_{\cdot k}^{(2)},\, 1 - \pi_{\cdot k}^{(2)}\} \leq 1$ where $\pi_{\cdot k}^{(2)} = \sum_{j=1}^{m_1} \pi_{jk}^2$. We also have that

$$\begin{aligned}
\mathbf{C} &= \sum_{i_1 \neq i_2} \left[\mathbb{E}\left[\left\langle X_{i_1}(\cdot,k),\, X'_{i_2}(\cdot,k)\right\rangle^2\right] - \left(\sum_{j=1}^{m_1} \pi_{jk}^2\right)^2 \right] \\
&= |\Omega|(|\Omega|-1) \left[\mathbb{E}\left[\left\langle X_{i_1}(\cdot,k),\, X'_{i_2}(\cdot,k)\right\rangle\right] - \left(\sum_{j=1}^{m_1} \pi_{jk}^2\right)^2 \right] \\
&= |\Omega|(|\Omega|-1) \left[\sum_{j=1}^{m_1} \pi_{jk}^2 - \left(\sum_{j=1}^{m_1} \pi_{jk}^2\right)^2 \right] \leq |\Omega|(|\Omega|-1)\pi_{\cdot k}^{(2)},
\end{aligned}$$

where we have used in the second line that $\langle X_{i_1}(\cdot,k),\, X'_{i_2}(\cdot,k)\rangle^2 = \langle X_{i_1}(\cdot,k),\, X'_{i_2}(\cdot,k)\rangle$ since $\langle X_{i_1}(\cdot,k),\, X'_{i_2}(\cdot,k)\rangle$ takes values in $\{0, 1\}$.

We now derive a bound on $\mathbf{D}$. By Jensen's inequality, we get

$$\sum_i \sqrt{\mathbb{E}\left[f_i^2(X_i(\cdot,k))\right]} \leq |\Omega|^{1/2} \sqrt{\mathbb{E}\left[\sum_i f_i^2(X_i(\cdot,k))\right]} \leq |\Omega|^{1/2}$$

where we used the bound $\mathbb{E}[\sum_i f_i^2(X_i(\cdot,k))] \leq 1$. Thus, the Cauchy–Schwarz inequality implies

$$\begin{aligned}
\mathbf{D} &\leq \sum_{i_1 \neq i_2} \mathbb{E}\left[h^2(X_{i_1}, X'_{i_2})\right] \mathbb{E}^{1/2}\left[f_{i_1}^2(X_{i_1}(\cdot,k))\right] \mathbb{E}^{1/2}\left[g_{i_2}^2(X'_{i_2}(\cdot,k))\right] \\
&\leq \max_{i_1 \neq i_2} \left\{\mathbb{E}^{1/2}\left[h^2(X_{i_1}, X'_{i_2})\right]\right\} \sum_{i_1, i_2} \mathbb{E}^{1/2}\left[f_{i_1}^2(X_{i_1}(\cdot,k))\right] \mathbb{E}^{1/2}\left[g_{i_2}^2(X'_{i_2}(\cdot,k))\right] \\
&\leq \max_{i_1 \neq i_2} \left\{\mathbb{E}^{1/2}\left[h^2(X_{i_1}, X'_{i_2})\right]\right\} |\Omega| \\
&\leq |\Omega| \left(\sum_{j=1}^{m_1} \pi_{jk}^2\right)^{1/2} = |\Omega| \left[\pi_{\cdot k}^{(2)}\right]^{1/2},
\end{aligned}$$

where we have used the fact that $\mathbb{E}[h^2(X_{i_1}, X'_{i_2})] \leq \sum_{j=1}^{m_1} \pi_{jk}^2$ following from an argument similar to that used to bound $\mathbf{C}$.

Finally, we get a bound on $\mathbf{B}$. Set $\pi_{0,k} = 1 - \pi_{\cdot,k}$. Note first that

$$\left\| \sum_{i_1} \mathbb{E}h^2(X_{i_1}(\cdot, k), \cdot) \right\|_{L^\infty} = |\Omega| \max_{0 \leq j' \leq m_1} \left\{ \sum_{j=0}^{m_1} h^2(e_j(m_1), e_{j'}(m_1)) \pi_{jk} \right\}$$
$$\leq |\Omega| (\pi_{\cdot k}^{(2)})^2 + |\Omega| \max_{1 \leq j' \leq m_1} \pi_{j',k}.$$

By symmetry, we obtain the same bound on $\| \sum_{i_2} \mathbb{E}h^2(\cdot, X_{i_2}(\cdot, k)) \|_{L^\infty}$. Thus we have

$$\mathbf{B} \leq |\Omega|^{1/2} \left(\pi_{\cdot k}^{(2)} + \max_{1 \leq j' \leq m_1} \pi_{j',k}^{1/2} \right).$$

Set now $U_2 = \sum_{i_1 \neq i_2} h(X_{i_1}(\cdot, k), X_{i_2}(\cdot, k))$. We apply a decoupling argument (See for instance Theorem 3.4.1 page 125 in [11]) to get that there exists a constant $C > 0$, such that for any $u > 0$

$$\mathbb{P}\left(\sum_{i_1 \neq i_2} h\left(X_{i_1}(\cdot, k), X_{i_2}(\cdot, k)\right) \geq u \right) \leq C \mathbb{P}\left(\sum_{i_1 \neq i_2} h(X_{i_1}(\cdot, k), X'_{i_2}(\cdot, k)) \geq u/C \right),$$

where $X'_i(\cdot, k)$ is independent of $X_i(\cdot, k)$ and has the same distribution as $X_i(\cdot, k)$. Next, Theorem 3.3 in [13] gives that, for any $u > 0$,

$$\mathbb{P}\left(\sum_{i_1 \neq i_2} h(X_{i_1}(\cdot, k), X'_{i_2}(\cdot, k)) \geq u \right) \leq C \exp\left[-\frac{1}{C} \min\left(\frac{u^2}{\mathbf{C}^2}, \frac{u}{\mathbf{D}}, \frac{u^{2/3}}{\mathbf{B}^{2/3}}, \frac{u^{1/2}}{\mathbf{A}^{1/2}} \right) \right],$$

for some absolute constant $C > 0$. Combining the last display with our bounds on $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$, we get that for any $t > 0$, with probability at least $1 - 2e^{-t}$,

$$\begin{aligned}
\left| \frac{1}{N^2} \sum_{i_1 \neq i_2} \sum_{j=1}^{m_1} X_{i_1}(j, k) X_{i_2}(j, k) \right| &\leq \frac{|\Omega|(|\Omega| - 1)}{N^2} \pi_{\cdot k}^{(2)} \\
&\quad + \frac{C}{N^2} \left(\mathbf{C} t^{1/2} + \mathbf{D} t + \mathbf{B} t^{3/2} + \mathbf{A} t^2 \right) \\
&\leq \frac{|\Omega|(|\Omega| - 1)}{N^2} \pi_{\cdot k}^{(2)} \\
&\quad + C \left[\frac{|\Omega|(|\Omega| - 1)}{N^2} \pi_{\cdot k}^{(2)} t^{1/2} + \frac{|\Omega|}{N^2} \left(\pi_{\cdot k}^{(2)} \right)^{1/2} t \right. \\
&\quad \left. + \frac{|\Omega|^{1/2}}{N^2} \left(\pi_{\cdot k}^{(2)} + \max_{1 \leq j' \leq m_1} \pi_{j',k}^{1/2} \right) t^{3/2} + \frac{t^2}{N^2} \right],
\end{aligned}$$

where $C > 0$ is a numerical constant. Combining the last display with (64) we get that, for any $t > 0$ with probability at least $1 - 3e^{-t}$,

$$
\begin{aligned}
\sum_{j=1}^{m_1}\left(\frac{1}{N}\sum_{i\in\Omega}X_i(j,k)\right)^2 \leq\ & \frac{|\Omega|(|\Omega|-1)}{N^2}\pi_{\cdot k}^{(2)} \\
& + C\left[\left(\frac{|\Omega|(|\Omega|-1)}{N^2}\pi_{\cdot k}^{(2)} + \frac{2\sqrt{|\Omega|\pi_{\cdot k}}}{N^2}\right)t^{1/2}\right. \\
& + \frac{|\Omega|}{N^2}\pi_{\cdot k} + \left(\frac{|\Omega|}{N^2}\left(\pi_{\cdot k}^{(2)}\right)^{1/2} + \frac{1}{N^2}\right)t \\
& \left. + \frac{|\Omega|^{1/2}}{N^2}\left(\pi_{\cdot k}^{(2)} + \max_{1\leq j'\leq m_1}\pi_{j',k}^{1/2}\right)t^{3/2} + \frac{t^2}{N^2}\right].
\end{aligned}
$$

Set $\pi_{\max} = \max_{1\leq k\leq m_2}\{\pi_{\cdot k}\}$ and $\pi_{\max}^{(2)} = \max_{1\leq k\leq m_2}\{\pi_{\cdot k}^{(2)}\}$. Using the union bound and up to a rescaling of the constants, we get that, with probability at least $1 - e^{t}$,

$$
\begin{aligned}
\|W\|_{2,\infty}^2 \leq\ & \frac{|\Omega|(|\Omega|-1)}{N^2}\pi_{\max}^{(2)} \\
& + C\left[\left(\frac{|\Omega|(|\Omega|-1)}{N^2}\pi_{\max}^{(2)} + \frac{2\sqrt{|\Omega|\pi_{\max}}}{N^2}\right)(t+\log m_2)^{1/2}\right. \\
& + \frac{|\Omega|}{N^2}\pi_{\max} + \frac{|\Omega|}{N^2}\left(\pi_{\max}^{(2)}\right)^{1/2}(t+\log m_2) \\
& \left. + \frac{|\Omega|^{1/2}}{N^2}\left(\pi_{\max}^{(2)} + \max_{j,k}\{\pi_{jk}^{1/2}\}\right)(t+\log m_2)^{3/2} + \frac{(t+\log m_2)^2}{N^2}\right].
\end{aligned}
$$

Recall that $|\Omega| = n$ and $æ = N/n$. Assumption 5 and the fact that $n \leq |\mathcal{I}|$ imply that there exists a numerical constant $C > 0$ such that, with probability at least $1 - e^{-t}$,

$$
\|W\|_{2,\infty}^2 \leq C\left(\frac{\gamma^2}{æ N m_2}\left(\sqrt{t+\log m_2} + (t+\log m_2)\sqrt{\frac{m_2}{n}}\right) + \frac{(t+\log m_2)^2}{N^2}\right)
$$

where we have used that $\pi_{j,k} \leq \pi_{\cdot k} \leq \sqrt{2}\gamma/m_2$. Finally, the bound on the expectation $\mathbb{E}\|W\|_{2,\infty}$ follows from this result and Lemma 17.

Appendix D: Proof of Lemma 10

With the notation $X_i(j,k) = \langle X_i, e_j(m_1)e_k(m_2)^\top\rangle$ we have

$$
\|\Sigma\|_\infty = \max_{1\leq j\leq m_1, 1\leq k\leq m_2}\left|\frac{1}{N}\sum_{i\in\Omega}\xi_i X_i(j,k)\right|.
$$

Under Assumption 3, the Orlicz norm $\|\xi_i\|_{\psi_2} = \inf\{x > 0 : \mathbb{E}[(\xi_i/x)^2] \le e\}$ satisfies $\|\xi_i\|_{\psi_2} \le c\sigma$ for some numerical constant $c > 0$ and all i. This and the relation (See Lemma 5.5 in [26][1])

$$\mathbb{E}[|\xi_i|^\ell] \le \frac{\ell}{2}\Gamma\left(\frac{\ell}{2}\right)\|\xi_i\|_{\psi_2}^\ell, \quad \forall \ell \ge 1,$$

imply that $N^{-\ell}\mathbb{E}[|\xi_i|^\ell X_i^\ell(j,k)] = N^{-\ell}\mathbb{E}[X_i(j,k)]\mathbb{E}[|\xi_i|^\ell] \le (\ell!/2)c^2 v(c\sigma/N)^{\ell-2}$ for all $\ell \ge 2$ and $v = \frac{\sigma^2\mu_1}{N^2 m_1 m_2}$, where we have used the independence between ξ_i and X_i, and Assumption 9. Thus, for any fixed (j,k), we have

$$\sum_{i\in\Omega}\mathbb{E}\left[\frac{1}{N^2}\xi_i^2 X_i^2(j,k)\right] \le |\Omega|\frac{c^2\sigma^2\mu_1}{N^2 m_1 m_2} = \frac{c^2\mu_1\sigma^2}{æ N m_1 m_2} =: v_1,$$

and

$$\sum_{i\in\Omega}\mathbb{E}\left[\frac{1}{N^\ell}|\xi_i|^\ell X_i^\ell(j,k)\right] \le \frac{\ell!}{2}v_1\left(\frac{c\sigma}{N}\right)^{\ell-2}.$$

Thus, we can apply Bernstein's inequality (see, e.g. [4, page 486]), which yields

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i\in\Omega}\xi_i X_i(j,k)\right| > C\left(\sqrt{\frac{\mu_1\sigma^2 t}{æ N m_1 m_2}} + \frac{\sigma t}{N}\right)\right) \le 2e^{-t}$$

for any fixed (j,k). Replacing here t by $t + \log(m_1 m_2)$ and using the union bound we obtain

$$\mathbb{P}\left(\|\Sigma\|_\infty > C\left(\sqrt{\frac{\mu_1(t+\log(m_1 m_2))}{æ N m_1 m_2}} + \frac{(t+\log(m_1 m_2))}{N}\right)\right) \le 2e^{-t}.$$

The bound on $\mathbb{E}[\|\Sigma\|_\infty]$ in the statement of Lemma 10 follows from this inequality and Lemma 17. The same argument proves the bounds on $\|\Sigma_R\|_\infty$ and $\mathbb{E}\|\Sigma_R\|_\infty$ in the statement of Lemma 10. By a similar (and even somewhat simpler) argument, we also get that

$$\mathbb{P}\left(\|W - \mathbb{E}[W]\|_\infty > C\left(\sqrt{\frac{\mu_1(t+\log(m_1 m_2))}{æ N m_1 m_2}} + \frac{t+\log(m_1 m_2)}{N}\right)\right) \le 2e^{-t}$$

while Assumption 9 implies that $\|\mathbb{E}[W]\|_\infty \le \frac{\mu_1}{æ m_1 m_2}$.

[1] This statement actually appears as an intermediate step in the proof of this lemma.

Appendix E: Technical Lemmas

Lemma 17 *Let Y be a non-negative random variable. Let there exist $A \geq 0$, and $a_j > 0$, $\alpha_j > 0$ for $1 \leq j \leq m$, such that*

$$\mathbb{P}\left(Y > A + \sum_{j=1}^{m} a_j t^{\alpha_j}\right) \leq e^{-t}, \quad \forall t > 0.$$

Then

$$\mathbb{E}[Y] \leq A + \sum_{j=1}^{m} a_j \alpha_j \Gamma(\alpha_j),$$

where $\Gamma(\cdot)$ is the Gamma function.

Proof Using the change of variable $u = \sum_{j=1}^{m} a_j v^{\alpha_j}$ we get

$$\begin{aligned}
\mathbb{E}[Y] &= \int_0^{\infty} \mathbb{P}(Y > t)dt \leq A + \int_0^{\infty} \mathbb{P}(Y > A + u)du \\
&= A + \int_0^{\infty} \mathbb{P}(Y > A + \sum_{j=1}^{m} a_j v^{\alpha_j}) \left(\sum_{j=1}^{m} a_j \alpha_j v^{\alpha_j - 1}\right) dv \\
&\leq A + \int_0^{\infty} \left(\sum_{j=1}^{m} a_j \alpha_j v^{\alpha_j - 1}\right) e^{-v} dv = A + \sum_{j=1}^{m} a_j \alpha_j \Gamma(\alpha_j).
\end{aligned}$$

□

Lemma 18 *Assume that $\mathcal{R}$ is an absolute norm. Then*

$$\mathcal{R}^*\left(\frac{1}{N}\sum_{i\in\Omega} \langle X_i, \Delta L\rangle\, X_i\right) \leq 2\mathbf{a}\mathcal{R}^*(W)$$

where $W = \frac{1}{N}\sum_{i\in\Omega} X_i$.

Proof In view of the definition of $\mathcal{R}^*$,

$$\begin{aligned}
\frac{1}{2\mathbf{a}}\mathcal{R}^*\left(\frac{1}{N}\sum_{i\in\Omega} \langle X_i, \Delta L\rangle\, X_i\right) &= \sup_{\mathcal{R}(B)\leq 1} \left\langle \frac{1}{N}\sum_{i\in\Omega} \frac{\langle X_i, \Delta L\rangle}{2\mathbf{a}} X_i, B\right\rangle \\
&\leq \sup_{\mathcal{R}(B')\leq 1} \left\langle \frac{1}{N}\sum_{i\in\Omega} X_i, B'\right\rangle = \mathcal{R}^*(W),
\end{aligned}$$

where we have used the inequalities $\langle X_i, \Delta L\rangle \leq \|\Delta L\|_\infty \leq 2\mathbf{a}$, and the fact that $\mathcal{R}$ is an absolute norm. □

References

1. Agarwal, A., Negahban, S., Wainwright, M.J.: Noisy matrix decomposition via convex relaxation: optimal rates in high dimensions. Ann. Stat. **40**(2), 1171–1197 (2012)
2. Aubin, J.P., Ekeland, I.: Applied nonlinear analysis. In: Pure and Applied Mathematics. Wiley, New York (1984)
3. Bauer, F.L., Stoer, J., Witzgall, C.: Absolute and monotonic norms. Numer. Math. **3**, 257–264 (1961)
4. Buehlmann, P., van de Geer, S.: Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer, New York (2011)
5. Cai, T.T., Zhou, W.: Matrix completion via max-norm constrained optimization. doi:10.1007/978-1-4612-0537-1.201E
6. Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis? J. ACM **58**(1), 1–37 (2009)
7. Candès, E.J., Tao, T.: The power of convex relaxation: near-optimal matrix completion. IEEE Trans. Inf. Theory **56**(5), 2053–2080 (2010)
8. Chandrasekaran, V., Sanghavi, S., Parrilo, P.A., Willsky, A.S.: Rank-sparsity incoherence for matrix decomposition. SIAM J. Optim. **21**(2), 572–596 (2011)
9. Chen, Y., Huan, X., Caramanis, C., Sanghavi, S.: Robust matrix completion with corrupted columns. ICML, 873–880 (2011)
10. Chen, Y., Jalali, A., Sanghavi, S., Caramanis, C.: Low-rank matrix recovery from errors and erasures. IEEE Trans. Inf. Theory **59**(7), 4324–4337 (2013)
11. de la Peña, V.H., Giné, E.: Decoupling. Probability and Its Applications (New York). Springer, New York (1999) (from dependence to independence, randomly stopped processes. U-statistics and processes. Martingales and beyond)
12. Foygel, R., Srebro, N.: Concentration-based guarantees for low-rank matrix reconstruction. J. 24nd Annu. Conf. Learn. Theory (COLT) (2011)
13. Giné, E., Latała, R., Zinn, J.: High dimensional probability, II (Seattle, 1999). In: Progress in Probability. Exponential and Moment Inequalities for U-Statistics, vol. 47, pp. 13–38. Birkhäuser, Boston (2000)
14. Gross, D.: Recovering low-rank matrices from few coefficients in any basis. IEEE Trans. Inf. Theory **57**(3), 1548–1566 (2011)
15. Hsu, D., Kakade, S.M., Zhang, T.: Robust matrix decomposition with sparse corruptions. IEEE Trans. Inf. Theory **57**(11), 7221–7234 (2011)
16. Huan, X., Caramanis, C., Sanghavi, S.: Robust PCA via outlier pursuit. IEEE Trans. Inf. Theory **58**(5), 3047–3064 (2012)
17. Keshavan, R.H., Montanari, A., Sewoong, O.: Matrix completion from noisy entries. J. Mach. Learn. Res. **11**, 2057–2078 (2010)
18. Klopp, O.: Noisy low-rank matrix completion with general sampling distribution. Bernoulli **20**(1), 282–303 (2014)
19. Koltchinskii, V.: Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour (Saint-Flour Probability Summer School). Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. In: Lecture Notes in Mathematics, vol. 2033. Springer, Heidelberg (2011)
20. Koltchinskii, V., Lounici, K., Tsybakov, A.B.: Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. Ann. Stat. **39**(5), 2302–2329 (2011)
21. Negahban, S., Wainwright, M.J.: Restricted strong convexity and weighted matrix completion: optimal bounds with noise. J. Mach. Learn. Res. **13**, 1665–1697 (2012)
22. Recht, B.: A simpler approach to matrix completion. J. Mach. Learn. Res. **12**, 3413–3430 (2011)
23. Rohde, A., Tsybakov, A.: Estimation of high-dimensional low-rank matrices. Ann. Stat. **39**(2), 887–930 (2011)
24. Rudelson, M., Vershynin, R.: Hanson–Wright inequality and sub-Gaussian concentration. Electron. Commun. Probab. **18**(82), 9 (2013)
25. Tsybakov, A.B.: Introduction to nonparametric estimation. In: Springer Series in Statistics. Springer, New York (2009) (revised and extended from the 2004 French original, translated by Vladimir Zaiats)
26. Vershynin, R.: Introduction to the non-asymptotic analysis of random matrices. arXiv:1011.3027 (2010)
27. Xiaodong, L.: Compressed sensing and matrix completion with constant proportion of corruptions. Constr. Approx. **37**(1) (2013)