Bernoulli **25**(4B), 2019, 3883–3911
https://doi.org/10.3150/19-BEJ1114

Structured matrix estimation and completion

OLGA KLOPP[1,3,*], YU LU[2,**], ALEXANDRE B. TSYBAKOV[3,†] and HARRISON H. ZHOU[2,‡]

[1]*ESSEC Business School, Cergy-Pontoise, France. E-mail:* [*]*kloppolga@math.cnrs.fr*
[2]*Yale University, New Haven, USA. E-mail:* [**]*yu.lu@yale.edu;* [‡]*huibin.zhou@yale.edu*
[3]*ENSAE, UMR CNRS 9194, Palaiseau, France. E-mail:* [†]*alexandre.tsybakov@ensae.fr*

We study the problem of matrix estimation and matrix completion under a general framework. This framework includes several important models as special cases such as the Gaussian mixture model, mixed membership model, bi-clustering model and dictionary learning. We establish the optimal convergence rates in a minimax sense for estimation of the signal matrix under the Frobenius norm and under the spectral norm. As a consequence of our general result we obtain minimax optimal rates of convergence for various special models.

Keywords: matrix completion; matrix estimation; minimax optimality; mixture model; stochastic block model

1. Introduction

Over the past decade, there have been considerable interest in statistical inference for high-dimensional matrices. A fundamental model in this context is the matrix de-noising model, under which one observes a matrix $\theta^* + W$ where θ^* is an unknown non-random $n \times m$ matrix of interest, and W is a random noise matrix. The aim is to estimate θ^* from such observations. Often in applications a part of elements of M is missing. The problem of reconstructing the signal matrix θ^* given partial observations of its entries is known as matrix completion problem. There has been an important research in the past years devoted to accurate matrix completion methods.

In general, the signal θ^* cannot be recovered consistently from noisy and possibly missing observations. If we only know that θ^* is an arbitrary $n \times m$ matrix, the guaranteed error of estimating θ^* from noisy observations can be prohibitively high. However, if θ^* has an additional structure one can expect to estimate it with high accuracy from a moderate number of noisy observations. The algorithmic and analytical tractability of the problem depends on the type of adopted structural model. A popular assumption in the matrix completion literature is that the unknown matrix θ^* is of low rank or can be well approximated by a low rank matrix. Significant progresses have been made on low rank matrix estimation and completion problems, see, for example, [8–10,18,22,29,31,34,35]. However, in several applications, the signal matrix θ^* can have other than just low rank structure. Some examples are as follows.

- *Biology*. The biological data are sometimes expected to have clustering structures. For example, in the gene microarray data, a large number of gene expression levels are measured under different experimental conditions. It has been observed in the experiments that there

1350-7265 © 2019 ISI/BS

is a bi-clustering structure on the genes [14]. This means that, besides being of low rank, the gene microarray data can be rearranged to approximately have a block structure.

- *Computer Vision*. To capture higher-level features in natural images, it is common to represent data as a sparse linear combination of basis elements [36] leading to sparse coding models. Unlike the principle component analysis that looks for low rank decompositions, sparse coding learns useful representations with number of basis vectors, which is often greater than the dimension of the data.
- *Networks*. In network models, such as social networks or citation networks, the links between objects are usually governed by the underlying community structures. To capture such structures, several block models have been recently proposed with the purpose of explaining the network data [2,25,28].

While there are some successful algorithmic advancements on adapting new structures in these specific applications, not much is known on the fundamental limits of statistical inference for the corresponding models. A few exceptions are the stochastic block model [20,33] and the bi-clustering model [19]. However, many other structures of signal matrix are not analyzed.

The aim of this paper is to study a general framework of estimating structured matrices. We consider a unified model that includes Gaussian mixture model, mixed membership model [2], bi-clustering model [24], and dictionary learning as special cases. We first study the optimal convergence rates in a minimax sense for estimation of the signal matrix under the Frobenius norm and under the spectral norm from complete observations on the sparsity classes of matrices. Then, we investigate this problem in the partial observations regime (structured matrix completion problem) and study the minimax optimal rates under the same norms. We also establish accurate oracle inequalities for the suggested methods.

2. Notation

This section provides a brief summary of the notation used throughout this paper. Let A, B be matrices in $\mathbb{R}^{n\times m}$.

- For a matrix A, A_{ij} is its (i,j)th entry, $A_{\cdot j}$ is its jth column and $A_{i\cdot}$ is its ith row.
- The scalar product of two matrices A, B of the same dimensions is denoted by $\langle A, B\rangle = \mathrm{tr}(A^T B)$.
- We denote by $\|A\|_2$ the Frobenius norm of A and by $\|A\|_\infty$ the largest absolute value of its entries: $\|A\|_\infty = \max_{i,j} |A_{ij}|$. The spectral norm of A is denoted by $\|A\|$.
- For $x \in \mathbb{R}^k$, we denote by $\|x\|_0$ its l_0-norm (the number of non-zero components of x), and by $\|x\|_q$ its l_q-norm, $1 \le q \le \infty$.
- We denote by $\|A\|_{0,\infty}$ the largest l_0-norm of the rows of $A \in \mathbb{R}^{n\times k}$:

$$\|A\|_{0,\infty} = \max_{1\le i\le p} \|A_{i\cdot}\|_0.$$

- For any $i \in \mathbb{N}$, we write for brevity $[i] = \{1, \dots, i\}$.
- Given a matrix $A = (A_{ij}) \in \mathbb{R}^{n\times m}$, and a set of indices $I \subset [n] \times [m]$, we define the restriction of A on I as a matrix A_I with elements $(A_I)_{ij} = A_{ij}$ if $(i,j) \in I$ and $(A_I)_{ij} = 0$ otherwise.

- The notation $\mathbf{I}_{k\times k}$ and $\mathbf{0}_{k\times l}$ (abbreviated to $\mathbf{I}$ and $\mathbf{0}$ when there is no ambiguity) stands for the $k \times k$ identity matrix and the $k \times l$ matrix with all entries 0, respectively.
- We denote by $|S|$ the cardinality of a finite set S, by $\lfloor x \rfloor$ the integer part of $x \in \mathbb{R}$, and by $\lceil x \rceil$ the smallest integer greater than $x \in \mathbb{R}$.
- We denote by $\mathcal{N}_\varepsilon(\mathcal{A})$ the ε-covering number, under the Frobenius norm, of a set $\mathcal{A}$ of matrices.

3. General model and examples

Assume that we observe a matrix $Y = (Y_{ij}) \in \mathbb{R}^{n\times m}$ with entries

$$Y_{ij} = E_{ij}\big(\theta^*_{ij} + \xi_{ij}\big), \quad i = 1, \dots, n, i = 1, \dots, m, \tag{3.1}$$

where θ^*_{ij} are the entries of the unknown matrix of interest $\theta^* = (\theta^*_{ij}) \in \mathbb{R}^{n\times m}$, the values ξ_{ij} are independent random variables representing the noise, and E_{ij} are i.i.d. Bernoulli variables with parameter $p \in (0, 1]$ such that (E_{ij}) is independent of (ξ_{ij}). Model (3.1) is called the matrix completion model. Under this model, an entry of matrix θ^* is observed with noise (independently of the other entries) with probability p, and it is not observed with probability $1 - p$. We can equivalently write (3.1) in the form

$$Y/p = \theta^* + W, \tag{3.2}$$

where W is a matrix with entries

$$W_{ij} = \theta^*_{ij}(E_{ij} - p)/p + \xi_{ij}E_{ij}/p.$$

The model with complete noisy observations is a special case of (3.1) (and equivalently of (3.3)) corresponding to $p = 1$. In this case, $W_{ij} = \xi_{ij}$.

Model (3.1) is called the matrix completion model. Under this model, an entry of matrix θ^* is observed with noise (independently of the other entries) with probability p, and it is not observed with probability $1 - p$. We can equivalently write (3.1) in the form

$$Y/p = \theta^* + W, \tag{3.3}$$

where W is a matrix with entries

$$W_{ij} = \theta^*_{ij}(E_{ij} - p)/p + \xi_{ij}E_{ij}/p.$$

We denote by $\mathbb{P}_{\theta^*}$ the probability distribution of Y satisfying (3.1) and by $\mathbb{E}_{\theta^*}$ the corresponding expectation. When there is no ambiguity, we abbreviate $\mathbb{P}_{\theta^*}$ and $\mathbb{E}_{\theta^*}$ to $\mathbb{P}$ and $\mathbb{E}$, respectively.

We assume that ξ_{ij} are independent zero mean sub-Gaussian random variables. The sub-Gaussian property means that the following assumption is satisfied.

Assumption 1. There exists $\sigma > 0$ such that, for all $(i, j) \in [n] \times [m]$,

$$\forall \lambda \in \mathbb{R}, \quad \mathbb{E}\exp(\lambda \xi_{ij}) \leq \exp\bigl(\lambda^2 \sigma^2/2\bigr).$$

We assume that the signal matrix θ^* is structured, that is, it can be factorized using sparse factors. Specifically, let s_n, k_n, s_m, k_m be integers such $0 \leq s_n \leq k_n$ and $0 \leq s_m \leq k_m$. We assume that

$$\theta^* \in \Theta(s_n, s_m) \subset \mathbb{R}^{n \times m},$$

where

$$\Theta(s_n, s_m) = \bigl\{\theta = XBZ^T : X \in \mathcal{A}_{s_n}, B \in \mathbb{R}^{k_n \times k_m} \text{ and } Z \in \mathcal{A}_{s_m}\bigr\}.$$

Here, for $s_n = 0$ we assume that $n = k_n$ and the set $\mathcal{A}_{s_n}$ is a set containing only one element, which is the $n \times n$ identity matrix, and for $1 \leq s_n \leq k_n$,

$$\mathcal{A}_{s_n} = \mathcal{A}_{s_n}(n, k_n) = \bigl\{A \in \mathcal{D}_n^{n \times k_n}, \|A_{i\cdot}\|_0 \leq s_n, \text{for all } i \in [n]\bigr\}, \tag{3.4}$$

where the set $\mathcal{D}_n$ is a subset of $\mathbb{R}$ called an alphabet. The set $\mathcal{A}_{s_m}$ is defined analogously by replacing n by m. We will also consider the class $\Theta_*(s_n, s_m)$ defined analogously to $\Theta(s_n, s_m)$, with the only difference that the inequality in (3.4) is replaced by the equality.

Choosing different values of s_n, k_n, s_m, k_m, and different alphabets we obtain several well-known examples of matrix structures.

- Mixture Model:

$$\Theta_{\mathrm{MM}} = \bigl\{\theta \in \mathbb{R}^{n \times m} : \theta = XB \text{ for some } B \in \mathbb{R}^{k \times m} \\ \text{and } X \in \{0, 1\}^{n \times k} \text{ with } \|X_{i\cdot}\|_0 = 1, \forall i \in [n]\bigr\}.$$

  Here $s_n = 1$ and $s_m = 0$.
- Sparse Dictionary Learning:

$$\Theta_{\mathrm{SDL}} = \bigl\{\theta = BZ^T \in \mathbb{R}^{d \times n} : B \in \mathbb{R}^{d \times k}, Z \in \mathbb{R}^{n \times k} \text{ with } \|Z_{i\cdot}\|_0 \leq s, \forall i \in [n]\bigr\}.$$

  In this example, $s_n = 0$ and $s_m = s$.
- Stochastic Block Model:

$$\Theta_{\mathrm{SBM}} = \bigl\{\theta = ZBZ^T \in \mathbb{R}^{n \times n} : B \in [0, 1]^{k \times k}, Z \in \{0, 1\}^{n \times k} \text{ with } \|Z_{i\cdot}\|_0 = 1, \forall i \in [n]\bigr\}.$$

  Here $s_n = s_m = 1$.
- Mixed Membership Model:

$$\Theta_{\mathrm{MMM}} = \bigl\{\theta = ZBZ^T \in \mathbb{R}^{n \times n} : B \in [0, 1]^{k \times k}, Z \in [0, 1]^{n \times k}, \\ \text{with } \|Z_{i\cdot}\|_1 = 1, \|Z_{i\cdot}\|_0 \leq s, \text{for all } i \in [n]\bigr\}.$$

  In this example, $s_n = s_m = s$.

- Bi-clustering Model:

$$\Theta_{\mathrm{Bi}} = \big\{\theta = XBZ^T \in \mathbb{R}^{n\times m} : B \in [0,1]^{k_n\times k_m}, X \in \{0,1\}^{n\times k_n}, Z \in \{0,1\}^{m\times k_m} \\ \text{with } \|X_{i\cdot}\|_0 = 1, \forall i \in [n], \|Z_{i\cdot}\|_0 = 1, \forall i \in [m]\big\}.$$

  Here $s_n = s_m = 1$.

Note that the classes Θ_{SBM} and Θ_{MMM} are not exactly equal to but rather subclasses of $\Theta(1,1)$ and $\Theta(s,s)$, respectively.

Statistical properties of inference methods under the general model (3.1) are far from being understood. Some results were obtained in particular settings such as the Mixture Model and Stochastic Block Model.

Gaussian mixture models provide a useful framework for several machine learning problems such as clustering, density estimation and classification. There is a quite long history of research on mixtures of Gaussians. We mention only some of this work including methods for estimating mixtures such as pairwise distances [15,16], spectral methods [26,40] or the method of moments [5,13]. Most of these papers are concerned with construction of computationally efficient methods but do not address the issue of statistical optimality. In [3] authors provide precise information theoretic bounds on the clustering accuracy and sample complexity of learning a mixture of two isotropic Gaussians in high dimensions under small mean separation.

The Stochastic Block Model is a useful benchmark for the task of recovering community structure in graph data. More generally, any sufficiently large graph behaves approximately like a stochastic block model for some k, which can be large. The problem of estimation of the probability matrix θ^* in the stochastic block model under the Frobenius norm was considered by several authors [6,11,12,41,42] but convergence rates obtained there are suboptimal. More recently, minimax optimal rates of estimation were obtained by Gao et al. [20] in the dense case and by Klopp et al. [33] in the sparse case.

Recently, a related problem to ours was studied by Soni et al. [38]. These authors consider the case when the matrix to be estimated is the product of two matrices, one of which, called a sparse factor, has a small number of non-zero entries (in contrast to this, we assume row-sparsity). The estimator studied in [38] is a sieve maximum likelihood estimator penalized by the l_0-norm of the sparse factor where the sieve is chosen as a specific countable set.

4. Results for the case of finite alphabets

We start by considering the case of finite alphabets $\mathcal{D}_n$ and $\mathcal{D}_m$ and complete observations, that is $p = 1$. In this section, we establish the minimax optimal rates of estimation of θ^* under the Frobenius norm and we show that they are attained by the least squares estimator

$$\hat{\theta} \in \arg\min_{\theta\in\Theta} \|Y - \theta\|_2^2, \tag{4.1}$$

where Θ is a suitable class of structured matrices. We first derive an upper bound on the risk of this estimator uniformly over the classes $\Theta = \Theta(s_n, s_m)$. The following theorem provides an

oracle inequality for the Frobenius risk of $\hat{\theta}$. Here and in what follows, we adopt the convention that $0 \log \frac{x}{0} = 0$ for any $x > 0$. We also set for brevity

$$d = n + m, \qquad r_n = n \wedge k_n, \qquad r_m = m \wedge k_m.$$

Theorem 1. *Let Assumption 1 hold, and let $p = 1$. If the sets $\mathcal{D}_n$ and $\mathcal{D}_m$ are finite, there exists a constant $C > 0$ depending only on the cardinalities of $\mathcal{D}_n$ and $\mathcal{D}_m$ such that, for all $\theta^* \in \mathbb{R}^{n \times m}$ and all $\varepsilon > 0$, the risk of the estimator (4.1) satisfies*

$$\mathbb{E}_{\theta^*}\left\{\left\|\hat{\theta} - \theta^*\right\|_2^2\right\} \le (1+\varepsilon) \inf_{\bar{\theta} \in \Theta(s_n, s_m)} \left\|\bar{\theta} - \theta^*\right\|_2^2 + \frac{C\sigma^2}{\varepsilon}(R_X + R_B + R_Z),$$

where $R_X = nr_m \wedge ns_n \log \frac{ek_n}{s_n}$, $R_B = r_n r_m$, $R_Z = mr_n \wedge ms_m \log \frac{ek_m}{s_m}$.

This theorem is proved in Appendix A.

Note that if the set $\mathcal{A}_{s_n}$ and/or $\mathcal{A}_{s_m}$ in the definition of $\Theta(s_n, s_m)$ contains only the identity matrix, the corresponding term R_X and/or R_Z disappears from the upper bound of Theorem 1.

In Theorem 1, the true signal θ^* can be arbitrary. By assuming that $\theta^* \in \Theta(s_n, s_m)$, we immediately deduce from Theorem 1 that the following bound holds.

Corollary 2. *Under the assumptions of Theorem 1,*

$$\sup_{\theta \in \Theta(s_n, s_m)} \mathbb{E}_\theta\left\{\|\hat{\theta} - \theta\|_2^2\right\} \le C\sigma^2(R_X + R_B + R_Z)$$

for a constant $C > 0$ depending only on the cardinalities of $\mathcal{D}_n$ and $\mathcal{D}_m$.

The next theorem provides a lower bound showing that the convergence rate of Corollary 2 is minimax optimal. This lower bound is valid for the general matrix completion model (3.1). In what follows, the notation $\inf_{\hat{\vartheta}}$ stands for the infimum over all estimators $\hat{\vartheta}$ taking values in $\mathbb{R}^{n \times m}$.

Theorem 3. *Let the entries W_{ij} of matrix W in model (3.3) be independent random variables with Gaussian distribution $\mathcal{N}(0, \sigma^2)$, and let the alphabets $\mathcal{D}_n$ and $\mathcal{D}_m$ contain the set $\{0, 1\}$. There exists an absolute constant $C > 0$ such that*

$$\inf_{\hat{\vartheta}} \sup_{\theta \in \Theta(s_m, s_n)} \mathbb{P}_\theta\left\{\|\hat{\vartheta} - \theta\|_2^2 \ge \frac{C\sigma^2}{p}(R_X + R_B + R_Z)\right\} \ge 0.1, \tag{4.2}$$

and

$$\inf_{\hat{\vartheta}} \sup_{\theta \in \Theta(s_m, s_n)} \mathbb{E}_\theta \|\hat{\vartheta} - \theta\|_2^2 \ge \frac{C\sigma^2}{p}(R_X + R_B + R_Z). \tag{4.3}$$

Furthermore, the same inequalities hold with $\Theta_*(s_m, s_n)$ in place of $\Theta(s_m, s_n)$ if $s_n \in \{0, 1\}$ and $s_m \in \{0, 1\}$.

The proof of Theorem 3 is given in Appendix B.1.

The three ingredients R_X, R_B, and R_Z of the optimal rate are coming from the ignorance of X, B and Z respectively. Note that when $s_n = k_n \le (n \wedge m)$ and $s_m = 0$ we obtain $\frac{\sigma^2}{p} k_n(n+m)$ which coincides with the classical minimax rates in low-rank matrix de-noising problems.

The proof is based on constructing subsets of Θ by fixing two of these parameters to get each of the three terms. The choice of B when fixing the pairs (X, B) and (Z, B) is based on a probabilistic method, namely, Lemma 17. Similar techniques have been used in [33] to prove the lower bounds for sparse graphon estimation, and in [20].

Remark 1. Formally, the lower bounds established in Theorem 3 do not apply directly to Stochastic Block and Mixture Membership Models as these sets are sub-classes of $\Theta(1, 1)$ and $\Theta(s, s)$. But, by a closer inspection of the proof of Theorem 3, we can see that it also covers the case of SBM and (with a slight modification) it can be applied to the Mixed Membership Model.

Remark 2. Theorem 3 can be extended to more general sub-Gaussian distributions under an additional Kullback–Leibler divergence assumption. Assume that there is a constant c such that the distribution of Y in model (3.1) satisfies

$$\mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) \le \frac{cp}{2\sigma^2} \left\|\theta - \theta'\right\|_2^2.$$

Let the alphabets $\mathcal{D}_n$ and $\mathcal{D}_m$ contain the set $\{0, 1\}$. Then there exists an absolute constant $C > 0$ such that

$$\inf_{\hat{\vartheta}} \sup_{\theta \in \Theta(s_m, s_n)} \mathbb{P}_\theta \left\{ \|\hat{\vartheta} - \theta\|_2^2 \ge \frac{C\sigma^2}{p}(R_X + R_B + R_Z) \right\} \ge 0.1,$$

and

$$\inf_{\hat{\vartheta}} \sup_{\theta \in \Theta(s_m, s_n)} \mathbb{E}_\theta \|\hat{\vartheta} - \theta\|_2^2 \ge \frac{C\sigma^2}{p}(R_X + R_B + R_Z).$$

The proof of this result is similar to that of Theorem 3 and only needs to replace the equality in (B.3) by inequality. In addition, the lower bounds hold with $\Theta_*(s_m, s_n)$ in place of $\Theta(s_m, s_n)$ if $s_n \in \{0, 1\}$ and $s_m \in \{0, 1\}$.

Remark 3. We summarize the minimax rates for some examples introduced in Section 3 in Table 1. The case of Θ_{SBM} is due to [20].

Table 1. Minimax rates

Θ_{MM}	$\min\{n\log(ek) + km, nm\}$
Θ_{SDL}	$\min\{ns\log(ek/s) + kd, nd\}$
Θ_{SBM}	$n\log(ek) + k^2$
Θ_{Bi}	$\min\{n\log(ek_n) + m\log(ek_m) + k_n k_m, nk_m + m\log(ek_m), mk_n + n\log(ek_n), nm\}$
Θ_{MMM}	$sn\log(\frac{ek}{s}) + k^2$

5. Optimal rates in the spectral norm

In this section, we derive the optimal rates of convergence of estimators of θ^* when the error is measured in the spectral norm. Interestingly, our results imply that these optimal rates coincide with those obtained for estimation of matrices with no structure. That is, the additional structure that we consider in the present paper does not have any impact on the rate of convergence of the minimax risk when the error is measured in the spectral norm.

The lower bound under the spectral norm can be obtained as a corollary of the lower bound under the Frobenius norm given by Theorem 3.

Corollary 4. *Under the assumptions of Theorem 3, there exists a absolute constant $C' > 0$ such that*

$$\inf_{\hat{\vartheta}} \sup_{\theta \in \Theta(s_m, s_n)} \mathbb{P}_\theta \left\{ \|\hat{\vartheta} - \theta\|^2 \geq \frac{C'\sigma^2}{p}(n \vee m) \right\} \geq 0.1,$$

and

$$\inf_{\hat{\vartheta}} \sup_{\theta \in \Theta(s_m, s_n)} \mathbb{E}_\theta \|\hat{\vartheta} - \theta\|^2 \geq \frac{C'\sigma^2}{p}(n \vee m).$$

The proof of this corollary is given in Appendix B.2.

To get matching upper bounds we can use the soft thresholding estimator introduced in [34] or the hard thresholding estimator proposed in [30]. These papers deal with the completion problem for low rank matrices in the context of trace regression model, which is a slightly different setting.

Here, we consider the hard thresholding estimator. Set

$$Y' = Y/p.$$

The singular value decomposition of matrix Y' has the form

$$Y' = \sum_{j=1}^{\operatorname{rank}(Y')} \sigma_j(Y') u_j(Y') v_j(Y')^T, \tag{5.1}$$

where $\operatorname{rank}(Y')$ is the rank of Y', $\sigma_j(Y')$ are the singular values of Y' indexed in the decreasing order, and $u_j(Y')$ (respectively, $v_j(Y')$) are the left (respectively, the right) singular vectors of Y'. The hard thresholding estimator is defined by the formula

$$\tilde{\theta} = \sum_{j:\sigma_j(Y') \geq \lambda} \sigma_j(Y') u_j(Y') v_j(Y')^T, \tag{5.2}$$

where $\lambda > 0$ is the regularization parameter. In this section, we assume that the noise variables W_{ij} are bounded as stated in the next assumption.

Assumption 2. For all i, j we have $\mathbb{E}(W_{ij}) = 0$, $\mathbb{E}(W_{ij}^2) = \sigma^2$ and there exists a positive constant $b > 0$ such that

$$\max_{i,j} |W_{ij}| \le b.$$

A more general case of sub-Gaussian noise can be treated as well; in this case, we can work on the event $\mathcal{E}_b$ where $\|W\|_\infty$ is bounded by a suitable constant b and show that the probability of the complement of $\mathcal{E}_b$ is small.

The following theorem gives the upper bound on the estimation error of the hard thresholding estimator (5.2).

Theorem 5. *Assume that $\|\theta^*\|_\infty \le \theta_{\mathrm{mx}}$ and let Assumption 2 hold. Let $\lambda = c(b + \theta_{\mathrm{mx}})\sqrt{\frac{n \vee m}{p}}$ where $c > 0$ is a sufficiently large absolute constant. Assume that $p \ge \log(n + m)/(n \vee m)$. Then, with $\mathbb{P}_{\theta^*}$ probability at least $1 - 2/(n + m)$, the hard thresholding estimator $\tilde{\theta}$ satisfies*

$$\|\tilde{\theta} - \theta^*\|^2 \le C(b + \theta_{\mathrm{mx}})^2 \frac{n \vee m}{p},$$

where $C > 0$ is an absolute constant.

The proof of Theorem 5 is close to the argument in [30]. It is given in Appendix C.

6. A general oracle inequality under incomplete observations

The aim of this section is to present a general theorem about the behavior of least squares estimators in the setting with incomplete observations. This theorem will be applied in the next section to obtain an analog of the upper bound of Theorem 1 for general alphabets. To state the theorem, it does not matter whether we consider a vector or matrix setting. Therefore, in this section, we will deal with the vector model. Assume that we observe a vector $Y = (Y_1, \dots, Y_N)$ with entries

$$Y_i = E_i\big(\theta_i^* + \xi_i\big), \quad i = 1, \dots, N, \tag{6.1}$$

for some unknown $\theta^* = (\theta_1^*, \dots, \theta_N^*)$. Our goal is to estimate θ^*. Here, ξ_i are independent random noise variables, and E_i are i.i.d. Bernoulli variables with parameter $p \in (0, 1]$ such that $(E_1, \dots, E_N)$ is independent of $(\xi_1, \dots, \xi_N)$.

When p is known we can equivalently write (6.1) in the form

$$Y' = \theta^* + W, \tag{6.2}$$

where now W is a vector with entries

$$W_i = \theta_i^*(E_i - p)/p + \xi_i E_i/p,$$

and $Y' = Y/p$. In this section, we denote by $\mathbb{P}_{\theta^*}$ the probability distribution of Y' satisfying (6.2).

Consider the least squares estimator of θ^*:

$$\hat{\theta} \in \arg\min_{\theta \in \Theta} \left\| Y' - \theta \right\|_2^2, \tag{6.3}$$

where Θ is a subset of $\mathbb{R}^N$. For some element θ_0 of $\arg\min_{\theta\in\Theta} \|\theta - \theta^*\|_2^2$ we set $\Theta_1 = \{\theta \in \Theta : \|\theta - \theta_0\|_2 \le 1\}$.

Set

$$\varepsilon_0 = \frac{1}{2}\big(\inf\big\{\varepsilon \in (0,1] : N\varepsilon^2 > \log \mathcal{N}_\varepsilon(\Theta_1)\big\} + \sup\big\{\varepsilon \in (0,1] : N\varepsilon^2 < \log \mathcal{N}_\varepsilon(\Theta_1)\big\}\big).$$

Since $\mathcal{N}_\varepsilon(\Theta_1)$ is a decreasing left-continuous function of $\varepsilon \in (0,1]$, we have

$$\frac{1}{2}\log \mathcal{N}_{\varepsilon_0}(\Theta_1) \le N\varepsilon_0^2 \le \log \mathcal{N}_{\varepsilon_0}(\Theta_1). \tag{6.4}$$

Theorem 6. *Let ξ_i be independent random variables satisfying $\mathbb{E}e^{\lambda\xi_i} \le e^{\lambda^2\sigma^2/2}$ for some $\sigma > 0$ and all $\lambda \in \mathbb{R}$. Assume that there exists a constant θ_{mx} such that $\|\theta\|_\infty \le \theta_{\mathrm{mx}}$ for all $\theta \in \Theta$. Then, for any $\theta^* \in \mathbb{R}^N$, with $\mathbb{P}_{\theta^*}$-probability at least $1 - 4/\mathcal{N}_{\varepsilon_0}(\Theta_1) - \exp(-pN/6)$, the least squares estimator (6.3) satisfies the oracle inequality*

$$\left\|\hat{\theta} - \theta^*\right\|_2^2 \le 3 \inf_{\theta\in\Theta} \left\|\theta - \theta^*\right\|_2^2 + C\frac{\theta_{\mathrm{mx}}^2 + \sigma^2}{p} N\varepsilon_0^2,$$

where $C > 0$ is an absolute constant.

The proof of this theorem is given in Appendix D.

Note that Theorem 6 has no assumption on the true signal θ^*. Using Theorem 6 with $\theta^* \in \Theta$ we immediately deduce that

$$\inf_{\theta\in\Theta} \mathbb{P}_\theta\left(\|\hat{\theta} - \theta\|_2^2 \le C\frac{\theta_{\mathrm{mx}}^2 + \sigma^2}{p} N\varepsilon_0^2\right) \ge 1 - 4/\mathcal{N}_{\varepsilon_0}(\Theta_1) - \exp(-pN/6),$$

where $C > 0$ is an absolute constant.

Theorem 6 shows that the rate of convergence of the least squares estimator is determined by the value of ε_0 satisfying the global entropy condition (6.4). This quantity is the critical covering radius that appeared in the literature in different contexts, see, for example, [43]. In particular, this critical radius has been shown to determine the minimax optimal rates in nonparametric estimation problems. However, it may lead to slightly suboptimal rates (with deterioration by a logarithmic factor) for parametric estimation problems.

7. Structured matrix completion with general alphabets

For the structured matrix completion over infinite alphabets we consider the following parameter spaces:

$$\widetilde{\Theta}(s_n, s_m) = \big\{\theta = XBZ^T : X \in \widetilde{\mathcal{A}}_n, \|B\|_\infty \le B_{\max}, Z \in \widetilde{\mathcal{A}}_m, \|\theta\|_\infty \le \theta_{\mathrm{mx}}\big\}.$$

Here, $B_{\max}$ and θ_{mx} are positive constants, and for $1 \le s_n \le k_n$,

$$\widetilde{\mathcal{A}}_n = \left\{ A \in \mathcal{D}_n^{n \times k_n} : \|A_{i\cdot}\|_0 \le s_n, \text{ for all } i \in [n] \text{ and } \|A\|_\infty \le 1 \right\}.$$

If $s_n = 0$, we assume that $n = k_n$ and we define $\widetilde{\mathcal{A}}_{s_n}$ as the set containing only one element, which is the $n \times n$ identity matrix.

The difference from the class $\Theta(s_n, s_m)$ is only in the fact that the elements of matrix $\theta \in \widetilde{\Theta}(s_n, s_m)$ and those of the corresponding factor matrices X, B, Z are assumed to be uniformly bounded. This assumption is natural in many situations, for example, in the Stochastic Block Model or in recommendation systems, where the entries of the matrix are ratings. We introduce the bounds of the entries of the factor matrices in order to fix ambiguities associated with the factorization structure.

A key ingredient in applying Theorem 6 to this particular case is to find the covering number $\log \mathcal{N}_\varepsilon(\Theta_1)$ when $\Theta = \widetilde{\Theta}(s_n, s_m)$. For any $\Theta \subset \mathbb{R}^{n \times m}$, any $\theta_0 \in \Theta$, and any $u > 0$, set

$$\Theta_u = \left\{ \theta \in \Theta : \|\theta - \theta_0\|_2 \le u \right\}.$$

The following result is proved in Appendix E.

Proposition 7. *For any $\theta_0 \in \widetilde{\Theta}(s_n, s_m)$, $0 < \varepsilon < 1$, and $u \le 1$ we have*

$$\log \mathcal{N}_\varepsilon\big(\widetilde{\Theta}_u(s_n, s_m)\big) \le R_1(\varepsilon) \wedge R_2(\varepsilon) \wedge R_3(\varepsilon) \wedge R_4(\varepsilon),$$

where

$$\begin{aligned}
R_1(\varepsilon) &= n s_n \log \frac{e k_n}{s_n} + m s_m \log \frac{e k_m}{s_m} \\
&\qquad + (n s_n + m s_m) \log \frac{6 B_{\max} \sqrt{mn} s_m s_n}{\varepsilon} + r_n r_m \log \frac{9u}{\varepsilon}, \\
R_2(\varepsilon) &= n r_m \log \frac{6u}{\varepsilon} + m s_m \log \frac{e k_m}{s_m} + m s_m \log \frac{2 B_{\max} \sqrt{mn} s_m s_n}{\varepsilon}, \\
R_3(\varepsilon) &= m r_n \log \frac{6u}{\varepsilon} + n s_n \log \frac{e k_n}{s_n} + n s_n \log \frac{2 B_{\max} \sqrt{mn} s_m s_n}{\varepsilon}, \\
R_4(\varepsilon) &= mn \log \frac{3u}{\varepsilon}.
\end{aligned}$$

Theorem 6, together with Proposition 7, imply implies the following upper bound on the estimation error in structured matrix completion.

Corollary 8. *Consider model (3.1). Let Assumption 1 hold. Then, for any $\theta^* \in \mathbb{R}^{n \times m}$, the least squares estimator (4.1) with $\Theta = \widetilde{\Theta}(s_n, s_m)$ satisfies the inequality*

$$\left\|\hat{\theta} - \theta^*\right\|_2^2 \le 3 \inf_{\theta \in \widetilde{\Theta}(s_n, s_m)} \left\|\theta - \theta^*\right\|_2^2 + C \frac{\theta_{\mathrm{mx}}^2 + \sigma^2}{p} \big(R_1(\varepsilon_0) \wedge R_2(\varepsilon_0) \wedge R_3(\varepsilon_0) \wedge R_4(\varepsilon_0)\big)$$

with $\mathbb{P}_{\theta^*}$-probability at least

$$1 - \exp\bigl(-c\bigl(R_1(\varepsilon_0) \wedge R_2(\varepsilon_0) \wedge R_3(\varepsilon_0) \wedge R_4(\varepsilon_0)\bigr)\bigr) - \exp(-pmn/18),$$

where, $C, c > 0$ are absolute constants.

Note that Proposition 7 and (6.4) imply that $\varepsilon_0 \geq c'\sqrt{(m+n)/mn}$ for some numerical constant c'. Then, we have that for the general scheme of matrix completion and general alphabets the upper bound given by Corollary 8 departs from the lower bound of Theorem 3 by a logarithmic factor:

Corollary 9. *Let the assumptions of Corollary 8 be satisfied. Then the least squares estimator (4.1) with $\Theta = \widetilde{\Theta}(s_n, s_m)$ satisfies the inequality*

$$\inf_{\theta \in \widetilde{\Theta}(s_n, s_m)} \mathbb{P}_\theta\left(\|\hat{\theta} - \theta\|_2^2 \leq C\frac{\theta_{\mathrm{mx}}^2 + \sigma^2}{p}\bigl[\log(n+m) + \log(s_n s_m)\bigr](R_X + R_B + R_Z)\right)$$
$$\geq 1 - \exp\bigl(-c\bigl[\log(n+m) + \log(s_n s_m)\bigr](R_X + R_B + R_Z)\bigr) - \exp(-pmn/18),$$

where $C, c > 0$ are absolute constants.

Remark 4. Comparing the upper bound provided by Corollary 9 to the lower bound of Theorem 3 we see that, in addition to the logarithmic factor, we have $\theta_{\mathrm{mx}}^2 + \sigma^2$ instead of σ^2.

Remark 5. When $s_n = k_n = k$ and $s_m = 0$, the estimator defined in (4.1) reduces to the top-k SVD of the noisy observation Y and the estimator defined in (6.3) becomes the top-k SVD of the matrix Y'. This estimator was largely studied in the literature for low-rank matrix denoising and low-rank completion problems (see, e.g., [7,17,21,23]). Corollary 8 implies that it is minimax optimal up to the logarithmic factor $\log(nm)$, see [34]. In the case of bounded noise (Assumption 2), proceeding as in the proof of Theorem 5 we can get rid of this logarithmic factor. Note also that Corollary 9 does not imply exact recovery when $\sigma \to 0$. Exact matrix completion requires so called "incoherence conditions" (see, e.g., [10]). When such conditions are not imposed, exact recovery might not be possible with a few observed entries.

8. Adaptation to unknown sparsity

The estimators considered above require the knowledge of the degrees of sparsity s_n and s_m of θ^*. In this section, we suggest a method that does not require such a knowledge and thus it is adaptive to the unknown degree of sparsity. Our approach will be to estimate θ^* using a sparsity penalized least squares estimator. Let

$$\mathcal{X} = \bigcup_{s_n=1}^{k_n} \bigcup_{s_m=1}^{k_m} \widetilde{\Theta}(s_n, s_m) \tag{8.1}$$

and set

$$R(s_n, s_m) = \big[nr_m \log(6\sqrt{n \wedge m})\big] \wedge \big[ns_n \log\big(k_n s_m (n \wedge m)\big)\big] + \big[mr_n \log(6\sqrt{n \wedge m})\big] \wedge \big[ms_m \log\big(k_n s_m (n \wedge m)\big)\big] + r_n r_m \log(9\sqrt{n \wedge m}). \tag{8.2}$$

For any $\theta = XBZ^T \in \mathcal{X}$ let

$$R(\theta) = R\big(\|X\|_{0,\infty}, \|Z\|_{0,\infty}\big). \tag{8.3}$$

In the following, Ω denotes the random set of observed indices (i, j) in model (3.1). In this section, we denote by $\hat{\theta}$ the following estimator

$$\hat{\theta} \in \operatorname*{arg\,min}_{\theta = XBZ^T \in \mathcal{X}} \big\{\|Y - \theta_\Omega\|_2^2 + \lambda R(\theta)\big\}, \tag{8.4}$$

where $\lambda > 0$ is a regularization parameter. Note that this estimator does not require the knowledge of p. The following theorem proved in Appendix F gives an upper bound on the estimation error of $\hat{\theta}$.

Theorem 10. *Assume that* $nm \log(3\sqrt{n \wedge m}) \geq 6 \log(k_n k_m)$ *and* $d \geq 10$. *Let* $\lambda = 8(\sigma \vee \theta_{\mathrm{mx}})^2$. *Then, for any* $\theta^* \in \mathbb{R}^{n \times m}$, *with* $\mathbb{P}_{\theta^*}$*-probability at least* $1 - 5\exp(-d/10) - 2\exp(-pnm)$ *the estimator* (8.4) *satisfies*

$$\big\|\hat{\theta} - \theta^*\big\|_2^2 \leq C \inf_{\theta \in \mathcal{X}} \left\{ \big\|\theta - \theta^*\big\|_2^2 + \frac{(\sigma \vee \theta_{\mathrm{mx}})^2}{p} R(\theta) \right\},$$

where $C > 0$ *is an absolute constant.*

Theorem 10 implies that for the general scheme of matrix completion and general alphabets we obtain the following upper bound which departs from the lower bound of Theorem 3 by a logarithmic factor:

$$\inf_{\theta \in \mathcal{X}} \mathbb{P}_\theta \left(\|\hat{\theta} - \theta\|_2^2 \leq C \frac{(\sigma \vee \theta_{\mathrm{mx}})^2}{p} \big[\log(n \wedge m) + \log(s_n s_m)\big](R_X + R_B + R_Z) \right) \geq 1 - 5\exp(-d/6) - 2\exp(-pnm),$$

where $C > 0$ is an absolute constant.

We finish this section by two remarks.

1. *Structured matrix estimation*. In the case of complete observations, that is $p = 1$, the estimator (8.4) coincides with the following estimator

$$\hat{\theta} \in \operatorname*{arg\,min}_{\theta = XBZ^T \in \mathcal{X}} \big\{\|Y - \theta\|_2^2 + \lambda R(\theta)\big\}. \tag{8.5}$$

Then, one can show that, with high probability, the following upper bound on the estimation error holds

$$\left\|\hat{\theta}-\theta^*\right\|_2^2 \leq C\sigma^2\big[\log(n\wedge m)+\log(s_n s_m)\big](R_X+R_B+R_Z).$$

Here we do not need an upper bound on $\|\theta^*\|_\infty$. At the same time, the estimator (8.5) is adaptive to the sparsity parameter (s_n, s_m).

2. *Sparse Factor Model*. Sparse Factor Model is studied in [38]. With our notation, it corresponds to a particular case of $n = k_n$ and X being the identity matrix with the difference that we consider row-sparse matrix Z while Z is assumed component-wise sparse in [38]. Convergence rates obtained in [38] are of the order $p^{-1}(nk_m + k_m m)$ (up to a logarithmic factor). This is greater then the upper bound given by Theorem 10 which, in this setting, is of the order $p^{-1}[n(k_m \wedge m) + s_m m]$.

Appendix A: Proof of Theorem 1

Set

$$\bar{R}_1(n,m) = ns_n \log\left(\frac{ek_n|\mathcal{D}_n|}{s_n}\right) + r_n m,$$

$$\bar{R}_2(n,m) = ns_n \log\left(\frac{ek_n|\mathcal{D}_n|}{s_n}\right) + ms_m \log\left(\frac{ek_m|\mathcal{D}_m|}{s_m}\right) + r_n r_m.$$

Since $\hat{\theta}$ is the least squares estimator on $\Theta(s_n, s_m)$, and $Y = \theta^* + W$, we have that for any $\bar{\theta} \in \Theta(s_n, s_m)$,

$$\left\|\hat{\theta}-\theta^*\right\|_2^2 \leq \left\|\bar{\theta}-\theta^*\right\|_2^2 + 2\langle\hat{\theta}-\bar{\theta}, W\rangle. \tag{A.1}$$

Now we use the following lemma proved in Appendix G.1.

Lemma 11. *Let $W \in \mathbb{R}^{n\times m}$ be a random matrix with independent σ-sub-Gaussian entries. Introduce the notation*

$$U^*_{\bar{\theta}} = \sup_{\theta\in\Theta(s_n,s_m),\theta\neq\bar{\theta}} \frac{\langle\theta-\bar{\theta}, W\rangle^2}{\|\theta-\bar{\theta}\|_2^2}.$$

For any $t > 0$, the following inequalities hold, where $C > 0$ is an absolute constant:

(i)

$$\sup_{\bar{\theta}\in\Theta(s_n,s_m)} \mathbb{P}\big\{U^*_{\bar{\theta}} \geq 3\sigma^2\big(\bar{R}_1(n,m)+t\big)\big\} \leq e^{-t}, \qquad \sup_{\bar{\theta}\in\Theta(s_n,s_m)} \mathbb{E}\big(U^*_{\bar{\theta}}\big) \leq C\sigma^2\bar{R}_1(n,m),$$

(ii)

$$\sup_{\bar{\theta}\in\Theta(s_n,s_m)} \mathbb{P}\big\{U^*_{\bar{\theta}} \geq 3\sigma^2\big(\bar{R}_2(n,m)+t\big)\big\} \leq e^{-t}, \qquad \sup_{\bar{\theta}\in\Theta(s_n,s_m)} \mathbb{E}\big(U^*_{\bar{\theta}}\big) \leq C\sigma^2\bar{R}_2(n,m),$$

(iii)

$$\sup_{\bar\theta\in\Theta(s_n,s_m)} \mathbb{P}\big\{U^*_{\bar\theta} \ge 3\sigma^2(nm+t)\big\} \le e^{-t}, \qquad \sup_{\bar\theta\in\Theta(s_n,s_m)} \mathbb{E}\big(U^*_{\bar\theta}\big) \le C\sigma^2 nm.$$

Applying Lemma 11(i) to (A.1), for any $\varepsilon > 0$, we get that

$$\mathbb{E}\big\{\big\|\hat\theta-\theta^*\big\|_2^2\big\} \le (1+\varepsilon)\big\|\theta^*-\bar\theta\big\|_2^2 + \frac{C\sigma^2}{\varepsilon}\bar R_1(n,m). \tag{A.2}$$

On the other hand, applying Lemma 11(i) to $\langle\hat\theta-\bar\theta, W\rangle = \langle(\hat\theta-\bar\theta)^T, W^T\rangle$ with n replaced by m, for all $\varepsilon > 0$, we get

$$\mathbb{E}\big\{\big\|\hat\theta-\theta^*\big\|_2^2\big\} \le (1+\varepsilon)\big\|\theta^*-\bar\theta\big\|_2^2 + \frac{C\sigma^2}{\varepsilon}\bar R_1(m,n). \tag{A.3}$$

Finally using Lemma 11(ii) and (iii) we get

$$\mathbb{E}\big\{\big\|\hat\theta-\theta^*\big\|_2^2\big\} \le (1+\varepsilon)\big\|\theta^*-\bar\theta\big\|_2^2 + \frac{C\sigma^2}{\varepsilon}\bar R_2(n,m) \tag{A.4}$$

and

$$\mathbb{E}\big\{\big\|\hat\theta-\theta^*\big\|_2^2\big\} \le (1+\varepsilon)\big\|\theta^*-\bar\theta\big\|_2^2 + \frac{C\sigma^2}{\varepsilon}nm. \tag{A.5}$$

Inequalities (A.2)–(A.5) imply that for all $\varepsilon > 0$ and all $\bar\theta \in \Theta(s_n, s_m)$

$$\mathbb{E}\big\{\big\|\hat\theta-\theta^*\big\|_2^2\big\} \le (1+\varepsilon)\big\|\theta^*-\bar\theta\big\|_2^2 + \frac{C\sigma^2}{\varepsilon}\bar R_3(n,m),$$

where $\bar R_3(n,m) = \min\{\bar R_1(n,m), \bar R_1(m,n), \bar R_2(n,m), nm\}$. Taking the supremum over $\bar\theta \in \Theta(s_n, s_m)$ and simplifying the expression for $\bar R_3(n,m)$, we obtain the result of Theorem 1.

Appendix B: Proof of lower bounds

B.1. Proof of Theorem 3

Lower bound with the terms R_X and R_Z. We only prove the lower bound with the term R_Z by fixing $X = X_0$ and $B = B_0$, where X_0 and B_0 are matrices specified below. The bound with R_X is analogous. Fix

$$X_0 = \begin{cases} [\mathbf{I}_{k_n\times k_n}, \mathbf{0}]^T, & \text{if } n \ge k_n, \\ [\mathbf{I}_{n\times n}, \mathbf{0}], & \text{otherwise.} \end{cases}$$

By Lemma 16, for $k_m \ge 2$, we can find $S_0 \subseteq \{0,1\}^{k_m}$ with the following properties:

(i) $\log|S_0| \ge c_1^* s_m \log\frac{ek_m}{s_m}$,

(ii) $c_2^* s_m \le \|a\|_0 \le s_m$ for all $a \in S_0$, and $\|a\|_0 = s_m$ for all $a \in S_0$ if $s_m \le k_m/2$,
(iii) $\|a-b\|_2^2 \ge c_3^* s_m$ for all $a, b \in S_0$ such that $a \ne b$,

where $c_j^* > 0$, $j = 1, 2, 3$, are absolute constants.

Assume first that $k_m \ge 2$ and $\min\{\frac{r_n}{97}, c_1^* s_m \log \frac{ek_m}{s_m}\} \ge \log 8$. Then, choose an arbitrary subset $\mathcal{S} \subseteq S_0$ of cardinality $|\mathcal{S}| = \lfloor \exp(\min\{\frac{r_n}{97}, c_1^* s_m \log \frac{ek_m}{s_m}\})\rfloor$ where $r_n = n \wedge k_n$ and we denote by $\lfloor x \rfloor$ the integer part of x. Since $\log|\mathcal{S}| \le r_n/96$, Lemma 17 implies that there exists a matrix $Q \in \{-1, 1\}^{r_n \times k_m}$ such that, for any $a, b \in \mathcal{S}$,

$$\frac{r_n}{2}\|a-b\|_2^2 \le \|Qa - Qb\|_2^2 \le \frac{3r_n}{2}\|a-b\|_2^2. \tag{B.1}$$

For this Q, let

$$B_0 = [\delta Q, \mathbf{0}_{(k_n - r_n)\times k_m}]^T$$

with $\delta > 0$ to be specified below. Define $\mathcal{Z} = \{Z \in \{0,1\}^{m \times k_m}, Z_{i\cdot} \in \mathcal{S} \text{ for all } i \in [m]\}$ and $T_Z = \{\theta = X_0 B_0 Z^T, Z \in \mathcal{Z}\}$. We have $T_Z \subseteq \Theta(s_n, s_m)$ and $\log|T_Z| = \log|\mathcal{Z}| = m \log|\mathcal{S}|$.

For any matrices $\theta = X_0 B_0 Z^T \in T_Z$, and $\bar\theta = X_0 B_0 \bar Z^T \in T_Z$ we have

$$\|\theta - \bar\theta\|_2^2 = \delta^2 \left\| QZ^T - Q\bar Z^T \right\|_2^2 = \delta^2 \sum_{i=1}^m \left\| QZ_{i\cdot}^T - Q\bar Z_{i\cdot}^T \right\|_2^2.$$

Using (B.1) and property (iii) of S_0, we find

$$\|\theta - \bar\theta\|_2^2 \ge \frac{r_n \delta^2}{2} \sum_{i=1}^m \|Z_{i\cdot} - \bar Z_{i\cdot}\|_2^2 \ge \frac{c_3^* r_n \delta^2 m s_m}{2}. \tag{B.2}$$

On the other hand, Lemma 15 together with (B.1) implies that the Kullback–Leibler divergence between $\mathbb{P}_\theta$ and $\mathbb{P}_{\bar\theta}$ satisfies

$$\mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_{\bar\theta}) = \frac{p}{2\sigma^2}\|\theta - \bar\theta\|_2^2 \le \frac{3pr_n\delta^2}{4\sigma^2} \sum_{i=1}^m \|Z_{i\cdot} - \bar Z_{i\cdot}\|_2^2 \le \frac{3pr_n\delta^2 m s_m}{2\sigma^2}. \tag{B.3}$$

If we choose now $\delta^2 = \frac{C_0\sigma^2}{pr_n s_m}\log|\mathcal{S}|$ for some absolute constant $C_0 > 0$ small enough, then (B.2), (B.3), Theorem 2.5 in [39] and the fact that $|\mathcal{S}| \ge 8$ imply that

$$\inf_{\hat\vartheta} \sup_{\theta \in T_Z} \mathbb{P}_\theta \left\{ \|\hat\vartheta - \theta\|_2^2 \ge C_1 \frac{\sigma^2}{p}\left(mr_n \wedge ms_m \log\frac{ek_m}{s_m}\right)\right\} \ge 0.7 \tag{B.4}$$

for some absolute constant $C_1 > 0$. This yields the term of the lower bound containing R_Z in the case when $k_m \ge 2$ and $\min\{\frac{r_n}{97}, c_1^* s_m \log \frac{ek_m}{s_m}\} \ge \log 8$. In the complementary case, when $k_m = 1$ or $\min\{\frac{r_n}{97}, c_1^* s_m \log \frac{ek_m}{s_m}\} < \log 8$, the value R_Z is smaller than $C\sigma^2 m/p$ for an absolute constant $C > 0$. Thus, in this case, it suffices to prove the lower bound of order m/p. To do this, let the

matrices X_0 and B_0 be such that their $(1,1)$th entry is equal to 1 and all other entries are 0, and consider the set of matrices Z such that their first column is a binary vector in $\{0,1\}^m$ and all other columns are 0. This defines a set of matrices $\theta = X_0 B_0 Z^T$ contained in $\Theta(s_n, s_m)$, which is isometric, under the Frobenius norm, to the set of binary vectors $\{0,1\}^m$ equipped with the Euclidean norm. Therefore, a lower bound of order m/p follows in a standard way as for vector estimation problem. We omit further details.

Analogously, by permuting n and m, we obtain that

$$\inf_{\hat{\vartheta}} \sup_{\theta \in \Theta(s_n, s_m)} \mathbb{P}_\theta \left\{ \|\hat{\vartheta} - \theta\|_2^2 \geq C_1 \frac{\sigma^2}{p} \left(n r_m \wedge n s_n \log \frac{e k_n}{s_n} \right) \right\} \geq 0.7, \tag{B.5}$$

which yields the term of the lower bound containing R_X.

Lower bound with the term R_B. To obtain the term containing R_B in the lower bound (4.2), we fix $X = X_0$ and $Z = Z_0$ where

$$X_0 = \begin{cases} [\mathbf{I}_{k_n \times k_n}, \mathbf{0}]^T, & \text{if } n \geq k_n, \\ [\mathbf{I}_{n \times n}, \mathbf{0}], & \text{otherwise,} \end{cases} \quad \text{and} \quad Z_0 = \begin{cases} [\mathbf{I}_{k_m \times k_m}, \mathbf{0}]^T, & \text{if } m \geq k_m, \\ [\mathbf{I}_{m \times m}, \mathbf{0}], & \text{otherwise.} \end{cases}$$

We first note that if $r_n r_m < 16$, the lower bound with term R_B is trivially obtained by distinguishing between two matrices. For $r_n r_m \geq 16$, by vectorizing a $r_n \times r_m$ matrix into a $r_n r_m$ dimensional vector, and applying the Varshamov–Gilbert bound [39], Lemma 2.9, we obtain that there exists a subset $\mathcal{B} \subseteq \{0,1\}^{r_n \times r_m}$ such that for any $Q, \bar{Q} \in \mathcal{B}$,

$$\|Q - \bar{Q}\|_2^2 = \sum_{i,j} \mathbf{1}\{Q_{ij} \neq \bar{Q}_{ij}\} \geq \frac{r_n r_m}{8}$$

and $\log |\mathcal{B}| \geq \frac{r_n r_m}{8}$. We define

$$T_B = \left\{ \theta = X_0 B Z_0^T, B = \delta \begin{bmatrix} Q & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, Q \in \mathcal{B} \right\}.$$

Clearly, $T_B \subseteq \Theta(s_n, s_m)$. For any $\theta = X_0 B Z_0^T \in T_B$, $\bar{\theta} = X_0 \bar{B} Z_0^T \in T_B$, we have

$$\|\theta - \bar{\theta}\|_2^2 = \|B - \bar{B}\|_2^2 = \delta^2 \|Q - \bar{Q}\|_2^2 \geq \frac{r_n r_m \delta^2}{8}.$$

Lemma 15 implies that $\mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_{\bar{\theta}}) = \frac{p}{2\sigma^2} \|\theta - \bar{\theta}\|_2^2 \leq \frac{p \delta^2 r_n r_m}{2\sigma^2}$. Choosing $\delta^2 = C_0' \sigma^2 / p$ for some constant $C_0' > 0$ small enough and using Theorem 2.5 in [39] we obtain

$$\inf_{\hat{\vartheta}} \sup_{\theta \in T_B} \mathbb{P} \left\{ \|\hat{\vartheta} - \theta\|_2^2 \geq C_2 \frac{\sigma^2}{p} r_n r_m \right\} \geq 0.7 \tag{B.6}$$

for some absolute constant $C_2 > 0$.

Combining (B.4), (B.5) and (B.6) proves the lower bound (4.2). The bound (4.3) follows from (4.2) and Markov's inequality.

Finally, the lower bounds for the classes $\Theta_*(s_n, s_m)$ with $s_n, s_m \in \{0, 1\}$ are proved analogously. It suffices to note that, if $s_m = 0$, there is no matrix Z in the definition of the class and thus there is no term R_Z. If $s_m = 1$, we follow the above argument corresponding to R_Z with the only difference that, by (ii) of Lemma 16, we can grant the exact equality $\|a\|_0 = 1$ for all $a \in S_0$ and $k_m \geq 2$. We omit further details.

B.2. Proof of Corollary 4

Define

$$\Gamma_2 = \big\{\theta = XBZ^T : X \in \mathcal{A}_{s_n}(n, 2), B \in \mathbb{R}^{2\times 2} \text{ and } Z \in \mathcal{A}_{s_m}(m, 2)\big\}.$$

For any $\theta = XBZ^T \in \Gamma_2$, let

$$\tilde{X} = [X, \mathbf{0}_{n\times(k_n-2)}], \qquad \tilde{Z} = [Z, \mathbf{0}_{m\times(k_m-2)}], \qquad \tilde{B} = \begin{bmatrix} B & \mathbf{0}_{2\times(k_m-2)} \\ \mathbf{0}_{(k_n-2)\times 2} & \mathbf{0} \end{bmatrix}.$$

We have $\theta = XBZ^T = \tilde{X}\tilde{B}\tilde{Z}^T \in \Theta(s_n, s_m)$, which implies that $\Gamma_2 \subseteq \Theta(s_n, s_m)$. Thus, for any $t > 0$, and any estimator $\hat{\vartheta} \in \mathbb{R}^{n\times m}$ we have

$$\sup_{\theta\in\Theta(s_n,s_m)} \mathbb{P}_\theta\big(\|\hat{\vartheta} - \theta\|^2 \geq t\big) \geq \sup_{\theta\in\Gamma_2} \mathbb{P}_\theta\big(\|\hat{\vartheta} - \theta\|^2 \geq t\big). \tag{B.7}$$

For an estimator $\hat{\vartheta} \in \mathbb{R}^{n\times m}$, let $\hat{\vartheta}_2 \in \mathbb{R}^{n\times m}$ be the closest matrix to $\hat{\vartheta}$ in the Frobenius norm among all matrices of rank at most 2. Since $\theta \in \Gamma_2$ is of rank at most 2, we have $\|\hat{\vartheta} - \hat{\vartheta}_2\| \leq \|\hat{\vartheta} - \theta\|$ and

$$\|\theta - \hat{\vartheta}_2\|_2^2 \leq 4\|\theta - \hat{\vartheta}_2\|^2 \leq 8\big(\|\theta - \hat{\vartheta}\|^2 + \|\hat{\vartheta} - \hat{\vartheta}_2\|^2\big) \leq 16\|\theta - \hat{\vartheta}\|^2.$$

Thus,

$$\mathbb{P}_\theta\big(\|\hat{\vartheta} - \theta\|^2 \geq t\big) \geq \mathbb{P}_\theta\big(\|\hat{\vartheta}_2 - \theta\|_2^2 \geq 16t\big)$$

for any $\theta \in \Gamma_2$ and any estimator $\hat{\vartheta}$. The last inequality and (B.7) imply

$$\inf_{\hat{\vartheta}} \sup_{\theta\in\Theta(s_n,s_m)} \mathbb{P}_\theta\big(\|\hat{\vartheta} - \theta\|^2 \geq t\big) \geq \inf_{\hat{\vartheta}} \sup_{\theta\in\Gamma_2} \mathbb{P}_\theta\big(\|\hat{\vartheta} - \theta\|_2^2 \geq 16t\big).$$

The result of Corollary 4 follows now by choosing $t = C_3\frac{\sigma^2}{p}(n + m)$ for some constant $C_3 > 0$ and using Theorem 3 with $k_m = k_n = 2$.

Appendix C: Proof of Theorem 5

Note that $Y' - \theta^* = Y/p - \theta^* = W$. It is straightforward to see that if $\lambda \geq \|W\|$, then

$$\big\|\tilde{\theta} - \theta^*\big\| \leq 2\lambda.$$

Thus, to prove the theorem, it suffices to show that for $\lambda = c(b+\theta_{\mathrm{mx}})\sqrt{\frac{n\vee m}{p}}$ with $c>0$ large enough we have $\lambda \geq \|W\|$ with high probability.

Since $W_{ij} = \theta_{ij}(E_{ij}-p)/p + \xi_{ij}E_{ij}/p$, we have

$$\|W\| \leq p^{-1}\big(\|\Sigma_1\| + \|\Sigma_2\|\big), \tag{C.1}$$

where $\Sigma_1 \in \mathbb{R}^{n\times m}$ is a matrix with entries $\xi_{ij}E_{ij}$ and $\Sigma_2 \in \mathbb{R}^{n\times m}$ is a matrix with entries $\theta_{ij}(E_{ij}-p)$. The second term in (C.1) is controlled using the following bound on the spectral norms of random matrices.

Proposition 12 ([4]). *Let A be an $n\times m$ matrix whose entries A_{ij} are independent centered bounded random variables. Then, for any $0<\varepsilon\leq 1/2$ there exists an absolute constant c_ε depending only on ε such that, for every $t>0$,*

$$\mathbb{P}\big\{\|A\| \geq (1+\varepsilon)2\sqrt{2}(\sigma_1\vee\sigma_2)+t\big\} \leq (n\wedge m)\exp\left(-\frac{t^2}{c_\varepsilon\sigma_*^2}\right),$$

where

$$\sigma_1 = \max_i \sqrt{\sum_j \mathbb{E}\big[A_{ij}^2\big]}, \qquad \sigma_2 = \max_j \sqrt{\sum_i \mathbb{E}\big[A_{ij}^2\big]}, \qquad \sigma_* = \max_{ij}|A_{ij}|.$$

We now apply Proposition 12 with $A_{ij} = (E_{ij}-p)\theta_{ij}$. Then

$$\sigma_1 \leq \theta_{\mathrm{mx}}\sqrt{np}, \qquad \sigma_2 \leq \theta_{\mathrm{mx}}\sqrt{mp} \quad \text{and} \quad \sigma_* \leq \theta_{\mathrm{mx}}.$$

Using these bounds and taking in Proposition 12 the values $\varepsilon = 1/2$ and $t = \sqrt{c_{1/2}}\theta_{\mathrm{mx}}\log(n+m)$ we obtain that there exists an absolute constant $c^*>0$ such that

$$\|\Sigma_2\| \leq 3\theta_{\mathrm{mx}}\sqrt{2(n\vee m)p} + c^*\theta_{\mathrm{mx}}\sqrt{2\log(n+m)}$$

with probability at least $1-1/(n+m)$. Similarly, there exists an absolute constants $c^*>0$ such that, with probability at least $1-1/(n+m)$,

$$\|\Sigma_1\| \leq 3\sigma\sqrt{2(n\vee m)p} + c^* b\sqrt{2\log(n+m)}.$$

Using these remarks, the assumption $p \geq \log(n+m)/(n\vee m)$, and (C.1) we obtain that the choice $\lambda = c(b+\theta_{\mathrm{mx}})\sqrt{\frac{n\vee m}{p}}$ with $c>0$ large enough implies the inequality $\lambda \geq \|W\|$ with probability at least $1-2/(n+m)$.

Appendix D: Proof of Theorem 6

Let $R = \sqrt{\frac{3(\theta_{\mathrm{mx}}^2+\sigma^2)}{p}N\varepsilon_0^2}$. We consider two cases separately.

Case 1: $\hat{\theta} \in \mathcal{G} \triangleq \{\theta \in \Theta, \|\theta - \theta_0\|_2 \leq 2^{s^*} R\}$. Then the desired result follows from the fact that $\|\hat{\theta} - \theta_0\|_2 \leq 2^{s^*} R$ and from the inequality

$$\left\|\hat{\theta} - \theta^*\right\|_2^2 \leq 2\left\|\theta_0 - \theta^*\right\|_2^2 + 2\|\hat{\theta} - \theta_0\|_2^2.$$

Case 2: $\hat{\theta} \notin \mathcal{G}$. The definition of the least squares estimator (6.3) implies $\|Y' - \hat{\theta}\|_2^2 \leq \|Y' - \theta_0\|_2^2$. Writing Y' as $\theta^* + W$ and rearranging, we obtain

$$\left\|\hat{\theta} - \theta^*\right\|_2^2 \leq \left\|\theta_0 - \theta^*\right\|_2^2 + 2\langle \hat{\theta} - \theta_0, W\rangle.$$

Since $\hat{\theta} \in \mathcal{G}^c$, Lemma 18 yields

$$\langle \hat{\theta} - \theta_0, W\rangle \leq \frac{1}{8}\|\theta - \theta_0\|_2^2 + 96R^2 \tag{D.1}$$

with probability greater than $1 - 4\exp(-\alpha R^2/2) - 2\exp(-pN/6)$ where $\alpha = \frac{p}{6(\theta_{\max}^2 + \sigma^2)}$. On the event where (D.1) holds,

$$\begin{aligned}\left\|\hat{\theta} - \theta^*\right\|_2^2 &\leq \left\|\theta_0 - \theta^*\right\|_2^2 + \frac{1}{4}\|\theta - \theta_0\|_2^2 + 192R^2 \\ &\leq \frac{3}{2}\left\|\theta_0 - \theta^*\right\|_2^2 + \frac{1}{2}\left\|\hat{\theta} - \theta^*\right\|_2^2 + 192R^2.\end{aligned}$$

This yields $\|\hat{\theta} - \theta^*\|_2^2 \leq 3\|\theta_0 - \theta^*\|_2^2 + 384R^2$ with probability greater than $1 - 4\exp(-\alpha R^2/2) - \exp(-pN/6)$.

Appendix E: Proof of Proposition 7

We start by proving the upper bound corresponding to $R_1(\varepsilon)$. Let $\bar{\mathcal{A}}_n(\delta)$ denote the δ-covering set of $\widetilde{\mathcal{A}}_n$ under the ℓ_∞ norm. For any given $\theta = XBZ^T \in \widetilde{\Theta}_u(s_n, s_m)$, there exist $X_0 \in \bar{\mathcal{A}}_n(\delta_1)$ and $Z_0 \in \bar{\mathcal{A}}_m(\delta_2)$ such that $\|X - X_0\|_\infty \leq \delta_1$ and $\|Z - Z_0\|_\infty \leq \delta_2$. For such X_0 and Z_0, let $\overline{\mathcal{T}}(X_0, Z_0)$ be the $\varepsilon/3$-covering set of $\mathcal{T}_u(X_0, Z_0)$ defined in Lemma 7 from the Supplementary Material [32]. For any $B \in \mathcal{T}_u(X_0, Z_0)$, there exists $B_0 \in \overline{\mathcal{T}}(X_0, Z_0)$ such that

$$\begin{aligned}&\left\|XBZ^T - X_0B_0Z_0^T\right\|_2 \\ &\quad\leq \left\|XBZ^T - X_0BZ^T\right\|_2 + \left\|X_0BZ^T - X_0BZ_0^T\right\|_2 + \left\|X_0BZ_0^T - X_0B_0Z_0^T\right\|_2 \\ &\quad\leq 2s_n\sqrt{nm}\|X - X_0\|_\infty\|BZ\|_\infty + 2s_m\sqrt{nm}\|X_0B\|_\infty\|Z - Z_0\|_\infty + \frac{\varepsilon}{3} \\ &\quad\leq 2B_{\max}\sqrt{nm}s_ns_m(\delta_1 + \delta_2) + \frac{\varepsilon}{3},\end{aligned}$$

where in the second inequality we have used Lemma 9 from the Supplementary Material [32] and the last inequality is due to the assumptions that $\|B\|_\infty \leq B_{\max}$, $\|X\|_\infty \leq 1$ and $\|Z\|_\infty \leq 1$.

Choosing $\delta_1 = \delta_2 = \varepsilon/(6B_{\max}\sqrt{nm}s_n s_m)$, we get that the set

$$\overline{\Theta}_u := \bigcup_{X_0 \in \overline{\mathcal{A}}_n(\delta_1), Z_0 \in \overline{\mathcal{A}}_m(\delta_2)} \overline{\mathcal{T}}(X_0, Z_0)$$

is an ε-covering set of $\widetilde{\Theta}_u(s_n, s_m)$. Then, Lemmas 7 and 8 from the Supplementary Material [32] imply

$$\begin{aligned}\log \mathcal{N}_\varepsilon\big(\widetilde{\Theta}_u(s_n, s_m)\big) &\le \log\big|\overline{\mathcal{A}}_n(\delta_1)\big| + \log\big|\overline{\mathcal{A}}_m(\delta_2)\big| + \max_{X_0, Z_0} \log\big|\overline{\mathcal{T}}(X_0, Z_0)\big| \\ &\le ns_n \log\frac{ek_n}{s_n} + ns_n \log\frac{6B_{\max}\sqrt{mn}s_m s_n}{\varepsilon} + r_n r_m \log\frac{9u}{\varepsilon} \\ &\quad + ms_m \log\frac{ek_m}{s_m} + ms_m \log\frac{6B_{\max}\sqrt{mn}s_m s_n}{\varepsilon}.\end{aligned}$$

To get upper bounds corresponding to $R_2(\varepsilon)$, $R_3(\varepsilon)$ and $R_4(\varepsilon)$ we define

$$\begin{aligned}\Theta_u^1 &= \big\{\theta = AZ^T, A \in [-s_n B_{max}, s_n B_{max}]^{n\times k_m}, Z \in \widetilde{\mathcal{A}}_m, \|\theta - \theta_0\|_2 \le u\big\}, \\ \Theta_u^2 &= \big\{\theta = XG^T, X \in \widetilde{\mathcal{A}}_n, G \in [-s_m B_{max}, s_m B_{max}]^{k_n\times m}, \|\theta - \theta_0\|_2 \le u\big\},\end{aligned}$$

and

$$\Theta_u^3 = \big\{\theta \in [-s_n s_m B_{max}, s_n s_m B_{max}]^{n\times m}, \|\theta - \theta_0\|_2 \le u\big\}.$$

It is easy to verify that $\widetilde{\Theta}_u(s_n, s_m) \subseteq \Theta_u^1$, $\widetilde{\Theta}_u(s_n, s_m) \subseteq \Theta_u^2$ and $\widetilde{\Theta}_u(s_n, s_m) \subseteq \Theta_u^3$. Using the same techniques as above, we obtain

$$\begin{aligned}\log \mathcal{N}_\varepsilon\big(\widetilde{\Theta}_u(s_n, s_m)\big) &\le nr_m \log\frac{6u}{\varepsilon} + ms_m \log\frac{ek_m}{s_m} + ms_m \log\frac{2B_{\max}\sqrt{mn}s_m s_n}{\varepsilon}, \\ \log \mathcal{N}_\varepsilon\big(\widetilde{\Theta}_u(s_n, s_m)\big) &\le mr_n \log\frac{6u}{\varepsilon} + ns_n \log\frac{ek_n}{s_n} + ns_n \log\frac{2B_{\max}\sqrt{mn}s_m s_n}{\varepsilon},\end{aligned}$$

and

$$\log \mathcal{N}_\varepsilon\big(\widetilde{\Theta}_u(s_n, s_m)\big) \le mn \log\frac{3u}{\varepsilon}.$$

Combining these bounds completes the proof of Proposition 7.

Appendix F: Proof of Theorem 10

Let $\mathcal{I} = \min_{(s_n, s_m)} R(s_n, s_m)$ and $\nu^2 = \frac{(\sigma \vee \theta_{\mathrm{mx}})^2}{p}\mathcal{I}$. By the definition of $R(s_n, s_m)$ in (8.2) we have $\mathcal{I} \ge d$. Note first that if $\|\hat{\theta} - \theta^*\|_2 \le 2^7\nu$, then Theorem 10 holds trivially. So, without loss of

generality, we can assume that $\hat{\theta} \in \mathcal{X}_\nu \triangleq \{\theta \in \mathcal{X} : \|\theta - \theta^*\|_2 > 2^7\nu\}$. By the definition (8.4) of the estimator $\hat{\theta} = \hat{X}\hat{B}\hat{Z}^T$ we have that for any $\theta = XBZ^T \in \mathcal{X}$

$$\|Y - \hat{\theta}_\Omega\|_2^2 + \lambda R(\hat{\theta}) \le \|Y - \theta_\Omega\|_2^2 + \lambda R(\theta)$$

which implies

$$\left\|\hat{\theta}_\Omega - \theta^*_\Omega\right\|_2^2 \le \left\|\theta_\Omega - \theta^*_\Omega\right\|_2^2 - 2\left\langle \xi_\Omega, \theta - \theta^*\right\rangle + \lambda R(\theta) + 2\left\langle \xi_\Omega, \hat{\theta} - \theta^*\right\rangle - \lambda R(\hat{\theta}), \tag{F.1}$$

where we set $\xi = (\xi_{ij})$. We will bound each term in (F.1) separately. Lemma 4 from the Supplementary Material [32] implies that with probability at least $1 - \exp(-pnm) - 2\exp(-d/10)$,

$$\left\langle \xi_\Omega, \hat{\theta} - \theta^*\right\rangle \le 2(\sigma \vee \theta_{\mathrm{mx}})^2 R(\hat{\theta}) + \frac{p}{8}\left\|\hat{\theta} - \theta^*\right\|_2^2. \tag{F.2}$$

To control $\langle \xi_\Omega, \theta - \theta^*\rangle$, we use Lemma 5 from the Supplementary Material [32] with $t = p\|\theta - \theta^*\|_2^2 + (\sigma \vee \theta_{\mathrm{mx}})^2 R(\theta)$. It follows that, with probability at least $1 - \exp(-d/2)$,

$$\left\langle \xi_\Omega, \theta - \theta^*\right\rangle \le (\sigma \vee \theta_{\mathrm{mx}})^2 R(\theta) + p\left\|\theta - \theta^*\right\|_2^2, \tag{F.3}$$

where we have used that $R(\theta) \ge d$. On the other hand, Lemma 1 from the Supplementary Material [32] implies that, with probability at least $1 - \exp(-pnm) - 2\exp(-d/6)$,

$$\left\|\hat{\theta}_\Omega - \theta^*_\Omega\right\|_2^2 + 4\theta_{\mathrm{mx}}^2 R(\hat{\theta}) \ge \frac{p}{2}\left\|\hat{\theta} - \theta^*\right\|_2^2. \tag{F.4}$$

Finally, using Lemma 3 from the Supplementary Material [32] with $a_{ij} = (\theta - \theta^*)_{ij}^2$ and $t = \frac{p}{2}\|\theta - \theta^*\|_2^2 + 4\theta_{\mathrm{mx}}^2 d$ we get that, with probability at least $1 - \exp(-d)$,

$$\left\|\theta_\Omega - \theta^*_\Omega\right\|_2^2 \le 4\theta_{\mathrm{mx}}^2 R(\theta) + \frac{3p}{2}\left\|\theta - \theta^*\right\|_2^2, \tag{F.5}$$

where we have used that $R(\theta) \ge d$. Plugging (F.2)–(F.5) in (F.1) we get

$$\begin{aligned}\frac{p}{4}\|\hat{\theta} - \theta^*\|_2^2 \le{}& \frac{5p}{2}\|\theta - \theta^*\|_2^2 + 8(\sigma \vee \theta_{\mathrm{mx}})^2 R(\hat{\theta}) + 6(\sigma \vee \theta_{\mathrm{mx}})^2 R(\theta)\\ &+ \lambda R(\theta) - \lambda R(\hat{\theta})\end{aligned}$$

with probability larger then $1 - 5\exp(-d/10) - 2\exp(-pnm)$ where we have used that $d \ge 10$. Taking here $\lambda = 8(\sigma \vee \theta_{\mathrm{mx}})^2$ finishes the proof.

Appendix G: Proofs of the lemmas

G.1. Lemmas for Theorem 1

Proof of Lemma 11. We start by proving (i). Note that for any fixed $X \in \mathcal{A}_n$, $\theta = XBZ^T$ belongs to a linear space of dimension not greater than $nm \wedge k_n m = r_n m$ as BZ^T belongs to

a linear space of dimension not greater than $k_n m$. Thus, $\theta - \theta^*$ belongs to a linear space of dimension not greater than $r_n m + 1$, which we denote by $W_{r_n m}(X)$. We have

$$\sup_{\theta \in \Theta(s_n, s_m), \theta \neq \theta^*} \frac{\langle \theta - \theta^*, W \rangle^2}{\|\theta - \theta^*\|_2^2} \leq \max_{X \in \mathcal{A}_n} U_X,$$

where, for a fixed $X \in \mathcal{A}_n$, we define

$$U_X = \sup_{\theta \in \Theta(s_n, s_m), \theta \neq \theta^*, \theta = XBZ^T} \frac{\langle \theta - \theta^*, W \rangle^2}{\|\theta - \theta^*\|_2^2} \leq \sup_{u \in W_{r_n m}(X): \|u\|_2 = 1} \langle u, W \rangle.$$

It follows from Lemma 13 that

$$\mathbb{P}\big\{U_X \geq \sigma^2\big(2(r_n m + 1) + 3v\big)\big\} \leq e^{-v}, \quad \forall v > 0.$$

Note that, for any $A \in \mathcal{A}_n$, there are at most s_n non-zero entries in each row of A. This implies that the number of different supports of matrix A is at most $\binom{k_n}{s_n}^n$. For those ns_n non-zero entries, there are at most $|\mathcal{D}_n|^{ns_n}$ choices. Then, we have

$$\log |\mathcal{A}_n| \leq n \log \binom{k_n}{s_n} + ns_n \log |\mathcal{D}_n| \leq ns_n \log\left(\frac{ek_n |\mathcal{D}_n|}{s_n}\right). \tag{G.1}$$

Applying the union bound and using (G.1) we get

$$\mathbb{P}\Big\{\max_{X \in \mathcal{A}_n} U_X \geq \sigma^2\big(2(r_n m + 1) + 3v\big)\Big\} \leq \left(\frac{ek_n |\mathcal{D}_n|}{s_n}\right)^{s_n n} e^{-v}, \quad \forall v > 0,$$

which yields the first result of Lemma 11. To get the bound on the expectation, we use the fact that for any non-negative random variable ξ and any $a > 0$

$$\mathbb{P}(\xi \geq a + t) \geq e^{-t}, \quad \forall t > 0$$

implies $\mathbb{E}\xi \leq a + 1$. The proof of (ii) follows the same lines fixing both X and Z. To prove (iii) we use that $\theta = XBZ^T$ belongs to a linear space of dimension not greater than nm. □

Lemma 13. *Let ξ be a σ-sub-Gaussian random vector in $\mathbb{R}^n$, and let $\mathcal{W}$ be a linear subspace of $\mathbb{R}^n$ with $\dim(\mathcal{W}) = d$. Consider the Euclidean ball $B(0,1) = \{u \in \mathcal{W} : \|u\|_2 \leq 1\}$. Then, for any $t > 0$,*

$$\mathbb{P}\Big(\max_{u \in B(0,1)} \big(u^T \xi\big)^2 \geq \sigma^2(d + 2\sqrt{dt} + 2t)\Big) \leq e^{-t}.$$

Proof. We have

$$\max_{u \in B(0,1)} \big(u^T \xi\big)^2 = \max_{u \in B(0,1)} \big(u^T P_{\mathcal{W}} \xi\big)^2 = \|P_{\mathcal{W}} \xi\|_2^2,$$

where $P_{\mathcal{W}}$ is the orthogonal projector onto $\mathcal{W}$. Applying the following lemma with $A = P_W$ yields the result.

Lemma 14 (Hsu et al. [27]). *Let ξ be a σ-sub-Gaussian random vector in $\mathbb{R}^n$, and let $A \in \mathbb{R}^{n\times n}$ be a matrix. Set $\Sigma = A^T A$. Then, for any $t > 0$,*

$$\mathbb{P}\Big(\|A\xi\|_2^2 \geq \sigma^2\big(\mathrm{tr}(\Sigma) + 2\sqrt{\mathrm{tr}\big(\Sigma^2\big)t} + 2\lambda_{\max}(\Sigma)t\big)\Big) \leq e^{-t},$$

where $\mathrm{tr}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ denote the trace and the maximal eigenvalue of Σ. □

G.2. Lemmas for Theorem 3

Lemma 15. *Assume that the noise variables W_{ij} in model (3.1) are i.i.d. Gaussian with distribution $\mathcal{N}(0, \sigma^2)$. Then, the Kullback–Leibler divergence between $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ has the form*

$$\mathrm{KL}(\mathbb{P}_\theta, \mathbb{P}_{\theta'}) = \frac{p}{2\sigma^2}\big\|\theta - \theta'\big\|_2^2.$$

Proof of this lemma is straightforward and it is therefore omitted.

Lemma 16. *Let $k \geq 2$ and $s \geq 1$ be integers, $s \leq k$. There exists a subset S_0 of the set of binary sequences $\{0, 1\}^k$ such that*

(i) $\log|S_0| \geq c_1^* s \log \frac{ek}{s}$,
(ii) $c_2^* s \leq \|a\|_0 \leq s$ *for all* $a \in S_0$, *and* $\|a\|_0 = s$ *for all* $a \in S_0$ *if* $s \leq k/2$,
(iii) $\|a - b\|_2^2 \geq c_3^* s$ *for all* $a, b \in S_0$ *such that* $a \neq b$,

where $c_j^* > 0$, $j = 1, 2, 3$, are absolute constants.

Proof. For $s \leq k/2$ the result follows from Lemma A.3 in [37]. For $k/2 < s \leq k$ and $k \geq 32$, we restrict the consideration only to binary sequences in $\{0, 1\}^k$ such that the first $m = \lceil k/4 \rceil$ elements can be either 0 or 1, the last $s - m$ elements are 1 and the remaining elements are 0. Then, (i)–(iii) follow from the Varshamov–Gilbert bound [39], Lemma 2.9, applied to the set of binary sequences of length m. For $k/2 < s \leq k$ and $k < 32$, the result is obvious. □

Lemma 17. *Let $\{a_1, a_2, \ldots, a_N\} \subseteq \{0, 1\}^k$. Let r be an integer satisfying $r > 96 \log N$. Then, there exists a matrix $Q \in \{-1, 1\}^{r\times k}$ such that for any $u, v \in [N]$,*

$$\frac{r}{2}\|a_u - a_v\|_2^2 \leq \|Qa_u - Qa_v\|_2^2 \leq \frac{3r}{2}\|a_u - a_v\|_2^2.$$

Proof. The result follows immediately from Johnson–Lindenstrauss lemma as stated in [1], Theorem 2, by taking there $\beta = 1$ and $\varepsilon = 1/2$. □

G.3. Lemmas for Theorem 6

Lemma 18. *Let* $R = \sqrt{\frac{3(\theta_{\mathrm{mx}}^2+\sigma^2)}{p} N\varepsilon_0^2}$ *and* $\Theta^R = \{\theta \in \Theta, \|\theta - \theta_0\|_2 \geq 2^{s^*} R\}$. *Then we have*

$$\mathbb{P}\left\{ \sup_{\theta\in\Theta^R} \left(\langle \theta - \theta_0, W\rangle - \frac{1}{8}\|\theta - \theta_0\|_2^2 \right) > 96R^2 \right\} \leq 4\exp\left(-\frac{\alpha R^2}{2} \right) + \exp(-pN/6),$$

where $\alpha = \frac{p}{6(\theta_{\mathrm{mx}}^2+\sigma^2)}$.

Proof of Lemma 18. Let $\mathcal{E} = \{\|W\|_2 \leq \sqrt{\frac{N}{2\alpha}}\}$ and for $s \geq 2$ let $\Theta_s^R = \{\theta \in \Theta^R, 2^s R \leq \|\theta - \theta_0\|_2 \leq 2^{s+1} R\}$. Then we have that $\Theta^R = \bigcup_{s=s^*}^{\infty} \Theta_s^R$. The union bound yields

$$\begin{aligned}
&\mathbb{P}\left\{ \sup_{\theta\in\Theta^R} \left(\langle \theta - \theta_0, W\rangle - \frac{1}{8}\|\theta - \theta_0\|_2^2 \right) > 96R^2, \mathcal{E} \right\} \\
&\quad \leq \sum_{s=s^*}^{\infty} \mathbb{P}\left\{ \sup_{\theta\in\Theta_s^R} \left(\langle \theta - \theta_0, W\rangle - \frac{1}{8}\|\theta - \theta_0\|_2^2 \right) > 96R^2, \mathcal{E} \right\} \\
&\quad \leq \sum_{s=s^*}^{\infty} \mathbb{P}\left\{ \sup_{\theta\in\Theta_s^R} \langle \theta - \theta_0, W\rangle \geq \left(2^{2s-3} + 96\right) R^2, \mathcal{E} \right\} \\
&\quad \leq \sum_{s=s^*}^{\infty} \mathbb{P}\left\{ \sup_{\theta\in\Theta_{2^{s+1}R}} \langle \theta - \theta_0, W\rangle \geq \left(2^{2s-3} + 96\right) R^2, \mathcal{E} \right\},
\end{aligned}$$

where the last step is due to the fact that $\Theta_s^R \subseteq \Theta_{2^{s+1}R}$. Now we are going to apply Lemma 19 with $D = 2^{s+1}R$, $\varepsilon = D\varepsilon_0$ and $t = (2^{2s-4} + 48)R^2$. It is easy to check that $t \in [DR, D^2]$ for all $s \geq s^*$. Then (G.2) yields

$$\sum_{s=s^*}^{\infty} \mathbb{P}\left\{ \sup_{\theta\in\Theta_{2^{s+1}R}} \langle \theta - \theta_0, W\rangle \geq \left(2^{2s-3} + 96\right) R^2, \mathcal{E} \right\} \leq \sum_{s=2}^{\infty} 2e^{-\frac{\alpha s R^2}{4}} \leq 4e^{-\frac{\alpha R^2}{2}}.$$

The last inequality holds when $\alpha R^2 \geq 4$. By Lemma 20, $P(\mathcal{E}^c) \leq \exp(-pN/6)$, and therefore we obtain the desired result. □

Lemma 19. *Suppose* ε *satisfies* $\sqrt{N}\varepsilon \leq 2D\sqrt{\log \mathcal{N}_\varepsilon(\Theta_D)}$. *Then, for* $D \geq \sqrt{2\log \mathcal{N}_\varepsilon(\Theta_D)/\alpha}$ *and for any* $t \in [D\sqrt{2\log \mathcal{N}_\varepsilon(\Theta_D)/\alpha}, D^2]$, *we have*

$$\mathbb{P}\left\{ \sup_{\theta\in\Theta_D} \langle \theta - \theta_0, W\rangle \geq 2t, \mathcal{E} \right\} \leq 2\exp\left(-\frac{\alpha t^2}{2D^2} \right), \tag{G.2}$$

where $\mathcal{E} = \{\|W\|_2 \leq \sqrt{\frac{N}{2\alpha}}\}$.

Proof of Lemma 19. Let $\mathcal{C}_D$ be an ε-covering set of Θ_D under the Frobenius norm. That is, for any $\theta \in \Theta_D$, there exists $\overline{\theta} \in \mathcal{C}_D$ such that $\|\theta - \overline{\theta}\|_2 \le \varepsilon$. Denote by $N \triangleq \mathcal{N}_\varepsilon(\Theta_D)$ the minimum cardinality of such $\mathcal{C}_D$. This yields

$$\begin{aligned}\langle \theta - \theta_0, W\rangle &= \langle \overline{\theta} - \theta_0, W\rangle + \langle \theta - \overline{\theta}, W\rangle \\ &\le \max_{\overline{\theta}\in\mathcal{C}_D} \langle \overline{\theta} - \theta_0, W\rangle + \varepsilon\|W\|_2,\end{aligned}$$

where we use Cauchy–Schwarz for the last inequality. On the event $\mathcal{E}$, we have that $\varepsilon\|W\|_2 \le \varepsilon\sqrt{\frac{N}{2\alpha}} \le D\sqrt{2\log(\mathcal{N}_\varepsilon(\Theta_D))/\alpha}$. It implies

$$\mathbb{P}\Big\{\sup_{\theta\in\Theta_D} \langle \theta - \theta_0, W\rangle \ge 2t, \mathcal{E}\Big\} \le \mathbb{P}\Big\{\max_{\overline{\theta}\in\mathcal{C}_D} \langle \overline{\theta} - \theta_0, W\rangle \ge t\Big\}$$

for $t \ge D\sqrt{2\log\mathcal{N}_\varepsilon(\Theta_D)/\alpha}$. By union bound and Lemma 20, the right hand side of the above inequality can be bounded from above by

$$\sum_{\overline{\theta}\in\mathcal{C}_D} \mathbb{P}\big\{\langle \overline{\theta} - \theta_0, W\rangle \ge t\big\} \le \exp\big(-\alpha \min\big\{t^2/D^2, t\big\} + \log\mathcal{N}_\varepsilon(\Theta_D)\big).$$

Then the desired result (G.2) holds when $D\sqrt{2\log N/\alpha} \le t \le D^2$. □

Lemma 20. *Let* $\alpha = \frac{p}{6(\theta_{\mathrm{mx}}^2+\sigma^2)}$. *Then, for any* $t > 0$ *and* $a \in \mathbb{R}^n$, *we have*

$$\mathbb{P}\big\{\langle a, W\rangle > t\big\} \le \exp\left(-\alpha \min\left\{\frac{t^2}{\|a\|^2}, \frac{\theta_{\mathrm{mx}} t}{\|a\|_\infty}\right\}\right), \tag{G.3}$$

and

$$\mathbb{P}\left\{\|W\|_2 \ge \sqrt{\frac{N}{2\alpha}}\right\} \le \exp(-pN/6). \tag{G.4}$$

Proof of Lemma 20. Since $W_i = \theta_i^* \frac{E_i - p}{p} + \xi_i \frac{E_i}{p}$, for $\lambda(\sigma \vee 2\theta_{\mathrm{mx}})\|a\|_\infty \le p$, we have

$$\mathbb{E}\big(e^{\lambda a_i W_i}\big) \le \mathbb{E}\Big(e^{\lambda a_i \theta_i^* \frac{E_i - p}{p}} e^{\frac{\lambda^2 a_i^2 \sigma^2 E_i}{2p^2}}\Big) \le e^{\frac{\lambda^2 a_i^2 \sigma^2}{2p}} e^{(\frac{\lambda a_i \theta_i^*}{p} + \frac{\lambda^2 a_i^2 \sigma^2}{2p^2})^2 p} \le e^{\frac{3\lambda^2 a_i^2 (\theta_{\mathrm{mx}}^2 + \sigma^2)}{2p}}.$$

Here the second inequality is due to the fact that $\mathbb{E}e^{\lambda(E_i - p)} \le e^{\lambda p}$ for $|\lambda| \le 1$. Now, following the Chernoff argument as in the proof of Lemma 5 from the Supplementary Material [32] we get (G.3).

To prove (G.4) note first that the variance of W_i satisfies

$$\mathbb{E}W_i^2 = \theta_i^2 \frac{1-p}{p} + \frac{\mathbb{E}\xi_i^2}{p} \le \frac{\theta_{\mathrm{mx}}^2 + \sigma^2}{p}.$$

Then we have

$$W_i^2 - \mathbb{E}W_i^2 = \big(\theta_i^*\big)^2 \frac{(E_i - p)(1-2p)}{p^2} + \frac{E_i}{p^2}\big(\xi_i^2 - \mathbb{E}\xi_i^2\big) + \frac{E_i - p}{p^2}\mathbb{E}\xi_i^2 + \frac{E_i(E_i - p)}{p^2} 2\theta_i^* \xi_i.$$

When $\lambda(\theta_{\mathrm{mx}}^2 + \sqrt{2}\sigma^2)/p^2 \le 1$, we obtain

$$\begin{aligned}\mathbb{E}\big(e^{\lambda(W_i^2 - \mathbb{E}W_i^2)}\big) &\le \mathbb{E}\big(e^{\lambda(\theta_i^*)^2 \frac{(E_i-p)(1-2p)}{p^2}} e^{\frac{2\lambda^2\sigma^4 E_i}{p^4}} e^{\frac{\lambda\sigma^2(E_i-p)}{p^2}} e^{\frac{2\lambda^2\theta_{\mathrm{mx}}^2\sigma^2}{p^4} E_{ij}(E_{ij}-p)^2}\big) \\ &\le e^{\frac{6\lambda^2(\theta_{\mathrm{mx}}^4 + \sigma^4)}{p^3}}.\end{aligned}$$

The Chernoff argument yields

$$\mathbb{P}\left\{\sum_{i=1}^N \big(W_i^2 - \mathbb{E}W_i^2\big) \ge t\right\} \le \exp\left\{-\lambda t + \frac{6\lambda^2(\theta_{\mathrm{mx}}^4 + \sigma^4)}{p^3} N\right\}.$$

For $t = \frac{2(\theta_{\mathrm{mx}}^2 + \sigma^2)}{p} N$, we choose $\lambda = \frac{p^2}{6(\theta_{\mathrm{mx}}^2 + \sigma^2)}$ to get the desired result. □

Acknowledgements

The work of O. Klopp was conducted as part of the project Labex MME-DII (ANR11-LBX-0023-01). 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), and Labex ECODEC (ANR-11-LABEX-0047).

Supplementary Material

Supplement to "Structured matrix estimation and completion" (DOI: 10.3150/19-BEJ1114 SUPP; .pdf). We provide the remaining proofs in the supplementary material [32].

References

[1] Achlioptas, D. (2003). Database-friendly random projections: Johnson–Lindenstrauss with binary coins. *J. Comput. System Sci.* **66** 671–687. Special issue on PODS 2001 (Santa Barbara, CA). MR2005771

[2] Airoldi, E.M., Blei, D.M., Fienberg, S.E. and Xing, E.P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

[3] Azizyan, M., Singh, A. and Wasserman, L. (2013). Minimax theory for high-dimensional Gaussian mixtures with sparse mean separation. In *Advances in Neural Information Processing Systems* (C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K.Q. Weinberger, eds.) **26** 2139–2147. Curran Associates.

[4] Bandeira, A.S. and van Handel, R. (2016). Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Ann. Probab.* **44** 2479–2506. MR3531673

[5] Belkin, M. and Sinha, K. (2010). Polynomial learning of distribution families. In 2010 *IEEE* 51*st Annual Symposium on Foundations of Computer Science—FOCS* 2010 103–112. Los Alamitos, CA: IEEE Computer Soc. MR3024780

[6] Borgs, C., Chayes, J. and Smith, A. (2015). Private graphon estimation for sparse graphs. In *Advances in Neural Information Processing Systems* 1369–1377.

[7] Bunea, F., She, Y. and Wegkamp, M.H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.* **39** 1282–1309. MR2816355

[8] Cai, T.T. and Zhou, W.-X. (2016). Matrix completion via max-norm constrained optimization. *Electron. J. Stat.* **10** 1493–1525. MR3507371

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

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

[11] Chan, S.H. and Airoldi, E.M. (2014). A consistent histogram estimator for exchangeable graph models. In *Proceedings of the* 31*st International Conference on Machine Learning* 208–216.

[12] Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *Ann. Statist.* **43** 177–214. MR3285604

[13] Chaudhuri, K., Dasgupta, S. and Vattani, A. (2009). Learning mixtures of gaussians using the k-means algorithm. Preprint. Available at arXiv:0912.0086.

[14] Cheng, Y. and Church, G.M. (2000). Biclustering of expression data. In *ISMB* **8** 93–103.

[15] Dasgupta, S. (1999). Learning mixtures of Gaussians. In 40*th Annual Symposium on Foundations of Computer Science* (*New York*, 1999) 634–644. Los Alamitos, CA: IEEE Computer Soc. MR1917603

[16] Dasgupta, S. and Schulman, L.J. (2000). A two-round variant of em for Gaussian mixtures. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, *UAI* '00 152–159. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

[17] Du, S.S., Wang, Y. and Singh, A. (2017). On the power of truncated SVD for general high-rank matrix estimation problems. In *Advances in Neural Information Processing Systems* 30 (*NIPS*) 2017 445–455.

[18] Foygel, R. and Srebro, N. (2011). Concentration-based guarantees for low-rank matrix reconstruction. In 24*nd Annual Conference on Learning Theory* (*COLT*).

[19] Gao, C., Lu, Y., Ma, Z. and Zhou, H.H. (2016). Optimal estimation and completion of matrices with biclustering structures. *J. Mach. Learn. Res.* **17** Paper No. 161, 29. MR3569248

[20] Gao, C., Lu, Y. and Zhou, H.H. (2015). Rate-optimal graphon estimation. *Ann. Statist.* **43** 2624–2652. MR3405606

[21] Giraud, C. (2015). *Introduction to High-Dimensional Statistics*. *Monographs on Statistics and Applied Probability* **139**. Boca Raton, FL: CRC Press. MR3307991

[22] Gross, D. (2011). Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory* **57** 1548–1566. MR2815834

[23] Hardt, M. (2014). Understanding alternating minimization for matrix completion. In 55*th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2014 651–660. Los Alamitos, CA: IEEE Computer Soc. MR3344916

[24] Hartigan, J.A. (1972). Direct clustering of a data matrix. *J. Amer. Statist. Assoc.* **67** 123–129.

[25] Holland, P.W., Laskey, K.B. and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088

[26] Hsu, D. and Kakade, S.M. (2013). Learning mixtures of spherical Gaussians: Moment methods and spectral decompositions. In *ITCS'*13*—Proceedings of the* 2013 *ACM Conference on Innovations in Theoretical Computer Science* 11–19. New York: ACM. MR3385380

[27] Hsu, D., Kakade, S.M. and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.* **17** no. 52, 6. MR2994877

[28] Karrer, B. and Newman, M.E.J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* (3) **83** 016107, 10. MR2788206

[29] Keshavan, R.H., Montanari, A. and Oh, S. (2010). Matrix completion from noisy entries. *J. Mach. Learn. Res.* **11** 2057–2078. MR2678022

[30] Klopp, O. (2011). Rank penalized estimators for high-dimensional matrices. *Electron. J. Stat.* **5** 1161–1183. MR2842903

[31] Klopp, O. (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli* **20** 282–303. MR3160583

[32] Klopp, O., Lu, Y., Tsybakov, A. and Zhou, H. (2019). Supplement to "Structured matrix estimation and completion." DOI:10.3150/19-BEJ1114SUPP.

[33] Klopp, O., Tsybakov, A.B. and Verzelen, N. (2017). Oracle inequalities for network models and sparse graphon estimation. *Ann. Statist.* **45** 316–354. MR3611494

[34] Koltchinskii, V., Lounici, K. and Tsybakov, A.B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39** 2302–2329. MR2906869

[35] Negahban, S. and Wainwright, M.J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *J. Mach. Learn. Res.* **13** 1665–1697. MR2930649

[36] Olshausen, B.A. and Field, D.J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vis. Res.* **37** 3311–3325.

[37] Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39** 731–771. MR2816337

[38] Soni, A., Jain, S., Haupt, J. and Gonella, S. (2016). Noisy matrix completion under sparse factor models. *IEEE Trans. Inform. Theory* **62** 3636–3661. MR3506754

[39] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation. Springer Series in Statistics*. New York: Springer. MR2724359

[40] Vempala, S. and Wang, G. (2004). A spectral algorithm for learning mixture models. *J. Comput. System Sci.* **68** 841–860. MR2059647

[41] Wolfe, P.J. and Olhede, S.C. (2013). Nonparametric graphon estimation. Preprint. Available at arXiv:1309.5936.

[42] Xu, J., Massoulié, L. and Lelarge, M. (2014). Edge label inference in generalized stochastic block models: From spectral theory to impossibility results. In *Conference on Learning Theory* 903–920. Barcelona, Spain.

[43] Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27** 1564–1599. MR1742500

Received September 2017 and revised February 2019