SHARP VARIABLE SELECTION OF A SPARSE SUBMATRIX IN A HIGH-DIMENSIONAL NOISY MATRIX

CRISTINA BUTUCEA^{1,2}, YURI I. INGSTER[†] AND IRINA A. SUSLINA³

Abstract. We observe a $N \times M$ matrix of independent, identically distributed Gaussian random variables which are centered except for elements of some submatrix of size $n \times m$ where the mean is larger than some a > 0. The submatrix is sparse in the sense that n/N and m/M tend to 0, whereas n, m, N and M tend to infinity. We consider the problem of selecting the random variables with significantly large mean values, as was also considered by [M. Kolar, S. Balakrishnan, A. Rinaldo and A. Singh, *NIPS* (2011)]. We give sufficient conditions on a as a function of n, m, N and M and construct a uniformly consistent procedure in order to do sharp variable selection. We also prove the minimax lower bounds under necessary conditions which are complementary to the previous conditions. The critical values a^* separating the necessary and sufficient conditions are sharp (we show exact constants), whereas [M. Kolar, S. Balakrishnan, A. Rinaldo and A. Singh, *NIPS* (2011)] only prove rate optimality and focus on suboptimal computationally feasible selectors. Note that rate optimality in this problem leaves out a large set of possible parameters, where we do not know whether consistent selection is possible.

Mathematics Subject Classification. 62G05, 62G20.

Received February 18, 2014. Revised May 22, 2014.

1. INTRODUCTION

Large data sets of random variables appear nowadays in many applied fields such as signal processing, biology and, in particular, genomics, finance *etc.* In genomic studies of cancer we may require to detect sample-variable associations see [30]. Our problem further addresses the question: if such an association is detected can we estimate the sample components and the particular variables involved in this association? This problem is also known as biclustering and has recently received a lot of attention.

We observe random variables that form an $N \times M$ matrix $\mathbf{Y} = \{Y_{ij}\}_{i=1,\dots,N,j=1,\dots,M}$:

$$Y_{ij} = s_{ij} + \xi_{ij}, \quad i = 1, \dots, N, \quad j = 1, \dots, M,$$
(1.1)

where $\{\xi_{ij}\}\$ are i.i.d. random variables and $s_{ij} \in \mathbb{R}$, for all $i \in \{1, \ldots, N\}$, $j \in \{1, \ldots, M\}$. The error terms ξ_{ij} are assumed to be distributed as standard Gaussian $\mathcal{N}(0, 1)$ random variables.

[†] deceased.

Keywords and phrases. Estimation, minimax testing, large matrices, selection of sparse signal, sharp selection bounds, variable selection.

¹ Université Paris-Est, LAMA (UMR 8050), UPEMLV, UPEC, CNRS, 77454, Marne-la-Vallée, France.

 $^{^2\,}$ CREST, Timbre J340 3, av. Pierre Larousse, 92240 Malakoff Cedex, France. cristina.butucea@u-pem.fr

³ St. Petersburg National Research University of Information Technologies, Mechanics and Optics, 49 Kronverkskiy pr., 197101 St. Petersburg, Russia.

C. BUTUCEA ET AL.

We assume that our data have zero mean, except for the elements of a submatrix, identified by a sparse collection of n rows and m columns, where the mean is significantly positive. More precisely, denote by

$$\mathcal{C}_{nm} = \{ C = A \times B \subset \{1, \dots, N\} \times \{1, \dots, M\}, \operatorname{Card}(A) = n, \operatorname{Card}(B) = m \},$$
(1.2)

the collection of subsets of n rows and m columns out of a matrix of size $N \times M$. Our assumption means that our data have mean $s_{ij} = 0$ except for elements in a submatrix of size $n \times m$, indexed by a set C_0 in \mathcal{C}_{nm} , where $s_{ij} \geq a$, for some a > 0.

The model is that, for some a > 0 which may depend on n, m, N and M,

there exists $C_0 \in \mathcal{C}_{nm}$ such that $s_{ij} = 0$, if $(i, j) \notin C_0$, and $s_{ij} \ge a$, if $(i, j) \in C_0$. (1.3)

Let $\mathcal{S}_{nm,a}$ be the collection of all matrices $S = S_C, C \in \mathcal{C}_{nm}$ that satisfy (1.3). Our model implies also that there exists some C_0 in \mathcal{C}_{nm} such that $S = S_{C_0}$ belongs to $\mathcal{S}_{nm,a}$.

We discuss here only significantly positive means of our random variables. The problem of selecting the variables with significantly negative means can be treated in the same way, by replacing variables Y_{ij} with $-Y_{ij}$.

Denote by P_C the probability measure that corresponds to observations (1.1) with matrix $S = S_C$ $\{s_{ij}\}_{i=1,\dots,N, j=1,\dots,M}, s_{ij} = 0 \text{ if } (i,j) \notin C, s_{ij} \geq a > 0 \text{ if } (i,j) \in C.$ We also denote $P_0 = P_{C_0}$ and E_0 the expected value with respect to the measure P_0 .

For simplicity, we denote $Y_C = \sum_{(i,j)\in C} Y_{ij}$, $\xi_C = \sum_{(i,j)\in C} \xi_{ij}$ and so on. Our goal is to propose a consistent estimator of C_0 , that is to select the variables in the large matrix of size $N \times M$ where the mean values are significantly positive. Our approach is to find the boundary values of a > 0. as function of n, m, N and M, where consistent selection is possible and separate them from the cases where consistent selection is not possible anymore.

We are interested here in sparse matrices, *i.e.* the case when n is much smaller than N and m is much smaller than M.

We study here the variable selection problem in a matrix from a minimax point of view. A selector is any measurable function of the observations, $\hat{C} = \hat{C}(\{Y_{ij}\})$ taking values in \mathcal{C}_{nm} . For such a selector $\hat{C} = \hat{C}(Y)$, Y = $\{Y_{ij}\}$ we denote the maximal risk by

$$R_{nm,a}(\hat{C}) = \sup_{S_{C_0} \in \mathcal{S}_{nm,a}} P_{C_0}(\hat{C}(Y) \neq C_0).$$

We define the minimax risk as

$$R_{nm,a} = \inf_{\hat{C}} R_{nm,a}(\hat{C}).$$

From now on, we assume in the asymptotics that $N \to \infty$, $M \to \infty$ and $n = n_{NM} \to \infty$, $n \ll N$, m = $m_{NM} \to \infty, \ m \ll M$. Other assumptions will be given later.

We say that a selector is consistent in the minimax sense, if $R_{nm,a}(\hat{C}) \to 0$.

For sequences of real numbers $\{u_n\}_{n>1}$ and real positive numbers $\{v_n\}_{n>1}$, we say that they are asymptotically equivalent, $u_n \sim v_n$, if $\lim_{n\to\infty} u_n/v_n = 1$. Moreover, we say that the sequences are asymptotically of the same order, $u_n \asymp v_n$, if there exist two constants $0 < c \leq C < \infty$ such that $c \leq \liminf_{n \to \infty} u_n / v_n$ and $\limsup_{n \to \infty} u_n / v_n \le C.$

We suppose that a > 0 is unknown. The aim of this paper is to give asymptotically sharp boundaries for the minimax risk for selecting the submatrix. It means that, first, we are interested in the conditions on $a = a_{NM}$ which guarantee the possibility of selection *i.e.*, the fact that $R_{nm,a} \to 0$. We construct the selecting procedure

$$\hat{C}^{\star}(Y) = \arg \max_{C \in \mathcal{C}_{nm}} Y_C, \tag{1.4}$$

where we recall that $Y_C = \sum_{(i,j) \in C} Y_{ij}$. We investigate the upper bounds of the minimax selection risk of this procedure. Second, we describe conditions on a for which we have the impossibility of selection, *i.e.*, the maximal

risk $R_{nm,a} \rightarrow 1$. These results are called the lower bounds. The two sets of conditions are complementary in a sense that violation of the upper bound conditions imply either impossibility of selection or even that of testing (as we compare later on to the sharp minimax rates for testing in [9]).

Kolar *et al.* [24] considered the same problem. They give upper bounds of convergence for the same scan procedure and they show lower bounds which are rate optimal but not sharp. The authors also propose three other procedures of rows and columns selection and they prove suboptimal upper bounds for these selectors. The authors implement and show how the suboptimal procedures behave numerically.

Let us stress out the fact that having rate optimality in this problem means that a large set of values a are not covered neither by the upper nor by the lower bounds. That means that for such values of a we do not know whether consistent selection is possible or not.

Therefore, we devote our efforts here to providing sharp minimax rates, that should be used as a benchmark for the behaviour of any computationally feasible selector.

Different sparsity assumptions for matrices were imagined. In a regression setup, matrix completion can be seen as reconstruction of a matrix S from few observed linear functionals of the matrix. More generally, the trace regression problem considers estimation of the matrix S from observations (X_i, Y_i) in the model $Y_i = tr(X_i^{\top}S) + \xi_i$. In particular, matrix completion is obtained when X_i has one element 1 and all others are null.

In the setting without noise ($\xi_i = 0$ for all *i*) and under the incoherence property saying that the singular vectors of the matrix are sufficiently spread on the unit sphere, [13, 18, 27] studied exact recovery. The same problem, in a noisy setting was studied by [12, 23] with the Frobenius norm. The more general trace regression problem was studied by [28] with Schatten-*p* norms and by [25] with nuclear norm penalization.

Our problem is highly connected to the detection problem (or testing) which was considered in this setting by Butucea and Ingster [9]. A more general setup, where each observation is replaced by a smooth signal was considered by Butucea and Gayraud [10]. We can apply our results to their setup in order to select the signals with significant energy (norm larger than a). Related testing problems were considered by [2,4].

When comparing the critical values for submatrix selection and for testing we note a gap between these two values. Indeed, there are values of the thresholding value a^* where consistent detection is possible, but no consistent submatrix selector can be found. This phenomenon was long known, see *e.g.* [21].

The problem of choosing a submatrix in a Gaussian random matrix has been previously studied by Sun and Nobel [29], who are interested in the largest square submatrix in Y under the null hypothesis such that its average is larger than some fixed threshold. An algorithm of choosing such submatrices was previously introduced in Shabalin *et al.* [30]. It was successfully implemented for the detection problem in [9] and an interesting open problem would be to prove consistency and convergence rates of this heuristic algorithm, which is beyond our scope here.

The plan of the paper is as follows. In Section 2 we state the main results of this paper: the upper bounds for the selection procedure \hat{C}^* under conditions on a, as well as inconsistency property of this procedure under complementary conditions on a, and, finally, lower bounds for variable selection. We compare these results with the results for detection in [9]. We give results for the vector case (m = M = 1). In Section 3 we prove the upper bounds for the selection of variables, that is a bound from below on a, in which $R_{nm,a}(\hat{C}^*) =$ $\sup_{S_{C_0}} P_{C_0}(\hat{C}^* \neq C_0) \rightarrow 0$. In Section 4 we prove lower bounds for variable selection, that is, a bound from above on the parameter a which imply that the minimax estimation risk $R_{nm,a}$ tends to 1. Two techniques provide the sharp lower bounds. One method is classical for nonparametric estimation, while the other makes a generalization of a well-known result to testing $L \geq 2$ hypotheses: the minimax risk is larger than the risk of the maximum likelihood estimator. Finally, Section 5 gives leads for extensions of present results.

2. Main results

Let

$$N \to \infty, n \to \infty, p = n/N \to 0; \quad M \to \infty, m \to \infty, q = m/M \to 0.$$
 (2.1)

C. BUTUCEA ET AL.

We suppose that a > 0 is unknown. The aim of this paper is to give asymptotically sharp boundaries for variable selection in a sparse high-dimensional matrix. Our approach is to give, on the one hand, sufficient asymptotic conditions on a such that the probability of wrongly selecting the variables in C_0 tends to 0 and, on the other hand, conditions under which no consistent selection is possible.

First, we are interested in the conditions on $a = a_{nmNM}$ which guarantee consistent variable selection, *i.e.*, the fact that we construct the selector \hat{C}^* in (1.4) and prove that $R_{nm,a}(\hat{C}^*) \to 0$. The selector \hat{C}^* is scanning the large $N \times M$ matrix and maximizes the sum of the inputs over all $n \times m$ submatrices.

The key quantities appearing in next theorems are

$$B = B_{n,m,N,M} = \min\{A_1, A_2, A\}, \text{ where } A = \frac{a\sqrt{nm}}{\sqrt{2(n\log(p^{-1}) + m\log(q^{-1}))}},$$

$$A_1 = \frac{a\sqrt{m}}{\sqrt{2}(\sqrt{\log(n)} + \sqrt{\log(N-n)})}, \qquad A_2 = \frac{a\sqrt{n}}{\sqrt{2}(\sqrt{\log(m)} + \sqrt{\log(M-m)})}.$$
(2.2)

We shall distinguish the case when B = A that we shall call of severe sparsity, from the case when $B = A_1$ or $B = A_2$ that we call of moderate sparsity, see the examples we discuss later on in order to justify this change of behaviour.

2.1. Sharp variable selection

The following Theorem gives sufficient conditions for the boundary $a = a_{n,m,N,M}$ such that selection is consistent uniformly over the class $S_{nm,a}$. The selector which attains these bounds is $\hat{C}^{\star}(Y)$ defined by (1.4).

Note that $P_{C_0}(\hat{C}^{\star}(Y) \neq C_0)$ does not depend on $C_0 = C_0(N, M, n, m, a)$. Therefore, for any C_0 we have

$$R_{nm,a}(\hat{C}^{\star}) = \max_{S_{C_0} \in \mathcal{S}_{nm,a}} P_{C_0}(\hat{C}^{\star}(Y) \neq C_0) = P_{C_0}(\hat{C}^{\star}(Y) \neq C_0)$$

Theorem 2.1 (Upper bounds). Assume (2.1) and assume $B = B_{n,m,N,M}$ defined by (2.2) is such that

$$\liminf B_{n,m,N,M} > 1, \tag{2.3}$$

then the selector \hat{C}^{\star} given by (1.4) is consistent, that is

$$R_{nm,a}(\hat{C}^{\star}) = P_{C_0}(\hat{C}^{\star} \neq C_0) \to 0.$$

Proof is given in Section 3.

Condition (2.3) is equivalent to saying that

 $\liminf A > 1$ and $\liminf A_1 > 1$ and $\liminf A_2 > 1$.

The following proposition says that $\liminf A_1 > 1$ and $\liminf A_2 > 1$ are necessary conditions for the consistency (in the minimax sense) of the selector \hat{C}^* of C_0 .

Proposition 2.2. Assume (2.1) and let the selector \hat{C}^{\star} be the selector given by (1.4). If

$$\limsup A_1 < 1 \text{ or } \limsup A_2 < 1$$

then, for any C_0 such that $S_{C_0} \in S_{nm,a}$,

$$P_{C_0}(\hat{C}^\star \neq C_0) \to 1.$$

Proof is given in Section 4.2.

Let us note that conditions on terms like A_1 and A_2 did not appear for the detection problem. Indeed, for variable selection, if we can select the columns but not the rows or the other way around, we cannot select the submatrix. Hence, additional and unavoidable terms A_1 and A_2 .

In the following theorem we give a sufficient condition on a under which consistent selection of C_0 is impossible uniformly over the set $S_{nm,a}$. These are the minimax lower bounds for variable selection. **Theorem 2.3.** Assume (2.1). If, moreover, $B = B_{n,m,N,M}$ defined by (2.2) is such that

$$\limsup B_{n,m,N,M} < 1, \tag{2.4}$$

then there is no consistent selection of C_0 uniformly over $\mathcal{S}_{nm,a}$, that is

$$\inf_{\hat{C}} \sup_{S_{C_0} \in \mathcal{S}_{nm, a}} P_{C_0}(\hat{C}(Y) \neq C_0) \to 1,$$

asymptotically, where the infimum is taken over all measurable functions $\hat{C} = \hat{C}(Y)$.

Proof of this theorem is given in Sections 4.1 and 4.2.

Theorems 2.1 and 2.3 imply that the critical value for a is

$$a^{*} \sim \max\left\{\frac{\sqrt{2\log(n)} + \sqrt{2\log(N-n)}}{\sqrt{m}}, \frac{\sqrt{2\log(m)} + \sqrt{2\log(M-m)}}{\sqrt{n}}, \frac{\sqrt{2(n\log(N/n) + m\log(M/m))}}{\sqrt{nm}}\right\}.$$
(2.5)

By critical value we mean in the sense that, for a such that $\liminf a/a^* > 1$, there is an estimator which is uniformly consistent, while, for a such that $\limsup a/a^* < 1$, no uniformly consistent estimator exists.

Example 2.4. Let us consider the particular case where the matrix and the submatrix are square (N = M and n = m) and, moreover, such that

$$\frac{\log(n)}{\log(N)} = \frac{\log(m)}{\log(M)} \to 0.$$

Then, $\log(n(N-n)) \sim \log(N-n) \sim \log(N)$ and $\log(m(M-m)) \sim \log(M-m) \sim \log(M)$ which imply that $A_1 = A_2 \ge A$ and, therefore, B = A. Therefore, the case B = A is more generally associated to the setup of severe sparsity.

Example 2.5. If we consider the particular case where $n = N^P$ and $m = M^Q$ grow polynomially, for some fixed P, Q in (0, 1), the critical value becomes

$$(a^*)^2 \sim \max\left\{\frac{2(1+\sqrt{P})^2\log(N)}{m}, \frac{2(1+\sqrt{Q})^2\log(M)}{n}, \frac{2(1-P)\log(N)}{m} + \frac{2(1-Q)\log(M)}{n}\right\}.$$

If, moreover, n = m and N = M, we get $(a^*)^2 \sim \max\{2(1 + \sqrt{P})^2, 4(1 - P)\}\log(N)/n$. So, the amount of sparsity depends on whether P is larger or smaller than 1/9. In this particular example, we have moderate sparsity, $B = A_1 = A_2 \leq A$, as soon as $P \geq 1/9$.

Let us stress the fact that our results are doubly sharp: first, the sets of values a are complementary in our upper and lower bounds, and second, the limit of the estimation risk is either 0 in the upper bounds or 1 in the lower bounds. Kolar *et al.* [24] give only rates up to constants. This implies that for some values of a, their result do not allow to conclude about the behaviour of the procedure, while our results leave no gaps.

Remark 2.6. We have investigated the upper limits of the selector \hat{C}^* under the assumption that $s_{ij} = a$, $(i, j) \in C_0$. It follows that, when $s_{ij} \ge a$, $(i, j) \in C_0$, statements of upper bounds stated in this section are valid.

Indeed, the random part of the expansion $Y_C - Y_{C_0}$ is independent of s_{ij} . The absolute value of the deterministic part (the difference of expectations) attains its minimum when $s_{ij} = a$.

C. BUTUCEA ET AL.

2.2. Variable selection vs. detection

Let us compare the result in Theorems 2.1 and 2.3 with the upper bounds and the lower bounds for detection of a set C_0 where our observations have significant means, *i.e.* above threshold *a*. The testing problem for our model can be stated as

$$H_0: s_{ij} = 0$$
 for all (i, j)

and we call P_0 the likelihood in this case, against the alternative

$$H_1$$
: there exists $C_0 \in \mathcal{C}_{nm}$ such that $S = S_{C_0} \in \mathcal{S}_{nm,a}$.

Let us recall the following theorems.

Theorem 2.7 (Upper bounds for detection, see [9]). Assume (2.1) and let a be such that at least one of the following conditions hold

$$a^2 nmpq = \frac{(anm)^2}{NM} \to \infty$$
 or $\liminf A > 1.$

Then distinguishability is possible, i.e.

$$\inf_{\psi(Y)} \left(P_0(\psi(Y) = 1) + \sup_{S_{C_0} \in \mathcal{S}_{nm,a}} P_{C_0}(\psi(Y) = 0) \right) \to 0.$$

where the infimum is taken over all measurable functions ψ taking values in $\{0,1\}$.

It was also shown in [9], that the asymptotically optimal test procedure ψ^* combines the scan statistic based on our \hat{C}^* with a linear statistic which sums all observations $Y = \{Y_{ij}\}_{i,j}$. The test procedure ψ^* rejects the null hypothesis as soon as either the linear or the scan test rejects.

Theorem 2.8 (Lower bounds for detection, see [9]). Assume (2.1) and

$$n\log(p^{-1}) \simeq m\log(q^{-1}), \quad \frac{\log\log(p^{-1})}{\log(q^{-1})} \to 0, \quad \frac{\log\log(q^{-1})}{\log(p^{-1})} \to 0.$$
 (2.6)

Moreover, assume that

$$a^2 nmpq = \frac{(anm)^2}{NM} \to 0$$
 and $\limsup A < 1.$

Then, consistent detection is impossible, that is

$$\inf_{\psi(Y)} \left(P_0(\psi(Y) = 1) + \sup_{S_{C_0} \in \mathcal{S}_{nm,a}} P_{C_0}(\psi(Y) = 0) \right) \to 1,$$

where the infimum is taken over all measurable functions ψ taking values in $\{0,1\}$.

We deduce that there is a gap between least conditions for testing that C_0 exists and selection of the actual variables $(i, j) \in C_0$ (estimation of C_0). In Table 1 we summarize possible cases were consistent selection and/or consistent testing is possible or not. It is understood that the linear statistic will detect a submatrix with small entries if the size of the submatrix is large enough. If the linear statistic detects, our results basically mean that we cannot select the submatrix in that case.

We prove the following lemma.

Lemma 2.9. Assume (2.1). If a is such that

 $\limsup A < 1$ and $\liminf A_1 > 1$ and $\liminf A_2 > 1$

then $a^2 nmpq \rightarrow 0$.

| Selection \setminus Test | Yes | No |
|----------------------------|---|-----------------------------|
| Yes | $\liminf B > 1$ | — |
| | 1) $\limsup A < 1$ | Under (2.6) for the test: |
| | and $a^2 nmpq \to \infty$ | $\limsup A < 1$ |
| No. | | and $a^2 nmpq \to 0$ |
| | 2) $\liminf A > 1$ and | |
| | $(\limsup A_1 < 1 \text{ or } \limsup A_2 < 1)$ | |

TABLE 1. Conditions for minimax variable selection and/or minimax testing.

This lemma is useful for further discussion of the conditions when no variable selection is possible, as given in Table 1. Indeed, the lower bounds for variable selection in Theorem 2.3 say that no consistent selection is possible as soon as $\limsup B < 1$.

This set implies, on the one hand, that $\limsup A < 1$ and in this case testing is possible if $a^2 nmpq \to \infty$, testing is impossible under (2.6) if $a^2 nmpq \to 0$.

On the other hand, it might happen that $\liminf A > 1$ which is sufficient for testing and that either $\limsup A_1 < 1$ or $\limsup A_2 < 1$.

Moreover, the conditions of Lemma 2.9 are included in the set of conditions where no selection and no testing can be performed, while the case 1) in Table 1 is included in the set

 $\limsup A < 1$ and, either $\limsup A_1 < 1$ or $\limsup A_2 < 1$.

Example 2.10. Let us consider an example of setup where consistent detection is possible, but no consistent selection can be done. Take $N = n^2$, $M = \log(n)$, $m = \log \log(n)$ (and, for instance, $a^2 = \log(n)/\log \log(n)$). For all a such that $a^2 \gg \log(n)/(\log \log(n))^2$ as $n \to \infty$, we have $a^2 nmpq = a^2(\log \log(n))^2/\log(n) \to \infty$. Therefore, on the one hand, distinguishability holds, see Theorem (2.7), *i.e.* we can construct a particular test procedure ψ^* such that

$$P_0(\psi^{\star}(Y) = 1) + \sup_{S_{C_0} \in \mathcal{S}_{nm,a}} P_{C_0}(\psi^{\star}(Y) = 0) \to 0.$$

On the other hand,

$$\frac{a^2m}{2(\sqrt{\log(n)} + \sqrt{\log(N-n)})^2} = \frac{a^2\log\log(n)}{(2+\sqrt{2})^2\log(n)}(1+o(1)) < 1,$$

for all a such that $a^2 < (1 - \delta)(2 + \sqrt{2})^2 \log(n) / \log \log(n)$, $\delta > 0$. By Theorem 2.3, no consistent selection is possible in this case.

2.3. Vector case

Sparsity assumptions were introduced for vectors. There is a huge amount of literature for estimating, testing and variable selection for (sparse or not sparse) vectors, since the pioneering work by Donoho *et al.* [16]. See also [1,6] in the context of false discovery rate for multiple testing, [3,8,11] and references therein, for estimation and testing issues. Note that in the vector case, variable selection was studied for the regression model, see *e.g.* [5,7,14,26,32] and references therein. Sharp minimax rates for testing can be summed up following the works by Donoho and Jin [15], Ingster [19] and Ingster and Suslina [22].

From our results concerning variable selection can also be proven for the vector case, that is for the gaussian independent, observations

$$X_i = s_i + \xi_i, \quad i = 1, \dots, N,$$

where $s_i \ge a$ for all *i* in a set A_0 of *n* elements and $s_i = 0$ otherwise. We suppose $n, N \to \infty$ such that $n/N \to 0$. Similarly, we can show the following result. Theorem 2.11 (Upper bounds). In the previous model, if

$$\liminf \frac{a}{\sqrt{2\log(N)} + \sqrt{2\log(n)}} > 1,$$

then the estimator $\hat{A}^{\star} = \arg \max_A \sum_{i \in A} X_i$ is such that

$$\sup_{A_0} P_{A_0}(\hat{A}^* \neq A_0) \to 0.$$

Lower bounds. If

$$\limsup \frac{a}{\sqrt{2\log(N)} + \sqrt{2\log(n)}} < 1,$$

then

$$\inf_{\hat{A}} \sup_{A_0} P_{A_0}(\hat{A} \neq A_0) \to 1.$$

The critical value is $a^* = \sqrt{2 \log N} + \sqrt{2 \log(n)}$. It is equivalent to $\sqrt{2 \log N}$ if $\log(n) / \log(N) \to 0$ and $a^* = \sqrt{2}(1 + \sqrt{1 - \beta})\sqrt{\log N}$ if $N = n^{\beta}$ for some $\beta \in (0, 1)$. This result agrees with sharp results in [20] (see Sect. 3.1, Rem. 2 and references therein).

Let us stress the fact that the particular case we study here is fundamentally different from the vector setup. Indeed, an additional regime is observed according to the sparsity structure of the submatrix (severe or moderate) and it cannot be obtained from previous results for vectors by, say, vectorizing the matrix.

3. Upper bounds

Proof of Theorem 2.1. Note that

$$P_{C_0}(\hat{C}^{\star} \neq C_0) = P_{C_0}(\max_{C \in \mathcal{C}_{nm}} Y_C - Y_{C_0} > 0).$$

We shall split the sets C according to the size of their common elements with the true underlying $C_0 = A_0 \times B_0$. Let us denote by $\mathcal{C}_{nm,kl}$ the collection of sets $C = A \times B$ such that k be the number of elements in $A \cap A_0$ and l the number of elements in $B \cap B_0$. Then,

$$P_{C_0}(\hat{C}^{\star} \neq C_0) = P_{C_0} \left(\max_{k=0,\dots,n} \max_{l=0,\dots,m} \max_{C \in \mathcal{C}_{nm,kl}} Y_C - Y_{C_0} > 0 \right)$$

$$\leq P_{C_0} \left(\max_{k=0,\dots,n} \max_{l=0,\dots,m} \max_{C \in \mathcal{C}_{nm,kl}} (\xi_{C \setminus C_0} - \xi_{C_0 \setminus C} - a(nm - kl)) > 0 \right).$$

From now, we fix $0 < \delta < 1$ and separate two cases: when $kl < (1 - \delta)nm$ and when $kl \ge (1 - \delta)nm$. As δ will be chosen small, it means that we treat differently the cases where the matrix C overlaps C_0 but weakly (or not at all) and where the matrices overlap almost entirely. We write and deal successively with each term in

$$P_{C_0}(\hat{C}^{\star} \neq C_0) \le P_{C_0} \left(\max_{\substack{k,l \ kl < (1-\delta)nm}} \max_{C \in \mathcal{C}_{nm,kl}} (\xi_{C \setminus C_0} - \xi_{C_0 \setminus C} - a(nm - kl)) > 0 \right)$$
(3.1)

$$+P_{C_0}\left(\max_{k,l}\max_{kl\ge (1-\delta)nm}\max_{C\in\mathcal{C}_{nm,kl}}(\xi_{C\setminus C_0}-\xi_{C_0\setminus C}-a(nm-kl))>0\right).$$
(3.2)

3.1. Weak intersection

Let us fix k and l such that $kl < (1 - \delta)nm$ for some $0 < \delta < 1$. Equivalently, we have $nm - kl > \delta nm$. In this case, we shall bound the probability in (3.1) as follows

$$P_{C_{0}}\left(\max_{k,l}\max_{kl<(1-\delta)nm}\max_{C\in\mathcal{C}_{nm,kl}}(\xi_{C\setminus C_{0}}-\xi_{C_{0}\setminus C}-a(nm-kl))>0\right)$$

$$\leq \sum_{k=0}^{n}\sum_{l=0}^{m}I_{kl<(1-\delta)nm}P_{C_{0}}\left(\max_{C\in\mathcal{C}_{nm,kl}}\xi_{C\setminus C_{0}}+\max_{C\in\mathcal{C}_{nm,kl}}\xi_{C\cap C_{0}}-\xi_{C_{0}}\geq a(nm-kl)\right)$$

$$\leq \sum_{k=0}^{n}\sum_{l=0}^{m}I_{kl<(1-\delta)nm}\left(T_{1,kl}+T_{2,kl}+T_{3,kl}\right),$$

where we denote by $I_{kl < (1-\delta)nm}$ the indicator function of the set where $kl < (1-\delta)nm$ and by

$$T_{1,kl} = P_{C_0} \left(\max_{C \in \mathcal{C}_{nm,kl}} \xi_{C \setminus C_0} > (1 - \delta_1) a(nm - kl) \right)$$
$$T_{2,kl} = P_{C_0} \left(\max_{C \in \mathcal{C}_{nm,kl}} \xi_{C \cap C_0} > \frac{\delta_1}{2} a(nm - kl) \right)$$
$$T_{3,kl} = P_{C_0} \left(-\xi_{C_0} > \frac{\delta_1}{2} a(nm - kl) \right),$$

for some $0 < \delta_1 < 1$.

Before continuing the proof, recall that, if n, N tend to infinity, such that $n/N \to 0$, we have

$$\log\left(\binom{N-n}{n-k}\right) \sim (n-k)\log\left(\frac{N-n}{n-k}\right) + (N-2n+k)\log\left(\frac{N-n}{N-2n+k}\right)$$
$$\sim (n-k)\log\left(\frac{N-n}{n-k}\right)(1+o(1)) \tag{3.3}$$

and

$$\log\left(\binom{n}{k}\right) \le \min\left\{(n-k)\log\left(\frac{ne}{n-k}\right), k\log\left(\frac{ne}{k}\right)\right\},\$$

for all k = 1, ..., n - 1 and $\log\binom{n}{n} = 0$.

In order to give an upper bound for $T_{1,kl}$, we shall distinguish the case where $k < (1 - \delta)n$ and l = m (the case k = n and $l < (1 - \delta)m$ is treated similarly) from the case $kl < (1 - \delta)nm$, k < n and l < m. On the one hand, if $k < (1 - \delta)n$ and l = m, we write, for a generic standard gaussian random variable Z (which might change later on):

$$T_{1,km} \leq P_{C_0} \left(\max_{A \in \mathcal{C}_{n,k}} \xi_{(A \setminus A_0) \times B_0} > (1 - \delta_1) a(n - k) m \right)$$

$$\leq {\binom{N-n}{n-k}} P(Z > (1 - \delta_1) a \sqrt{(n-k)m})$$

$$\leq \exp\left(-\frac{(1 - \delta_1)^2}{2} a^2 (n - k) m + \log\left(\binom{N-n}{n-k}\right) \right) \right),$$

where we use repeatedly that $P(Z > u) \leq \exp(-u^2/2)$, for all $u \geq 0$. Now, use (3.3) to get

$$T_{1,km} \le \exp\left(-(n-k)\left(\frac{(1-\delta_1)^2}{2}a^2m - \log\left(\frac{N-n}{n-k}\right)(1+o(1))\right)\right).$$

By assumption (2.3) we can say that

$$\min\left\{\frac{a^2 n m}{2(n \log(p^{-1}) + m \log(q^{-1}))}, \frac{a^2 m}{2(\sqrt{\log(N-n)} + \sqrt{\log(n)})^2}, \frac{a^2 n}{2(\sqrt{\log(M-m)} + \sqrt{\log(m)})^2}\right\} \ge 1 + \alpha,$$
(3.4)

for some fixed small $\alpha > 0$. Therefore, if $\delta_1 > 0$ is small enough, we have some $\alpha_1 > 0$ such that

$$\frac{(1-\delta_1)^2}{2}a^2m \ge (1+\alpha_1)\left(\log\left((N-n)n\right)\right) > \log\left(\frac{N-n}{n-k}\right)(1+o(1)) + \log(n),\tag{3.5}$$

asymptotically. Indeed, it is sufficient that $(1 - \delta_1)^2 (1 + \alpha) \ge 1 + \alpha_1$.

We get

$$T_{1,km} \le \exp(-(n-k)\log(n)).$$

We conclude that

$$\sum_{k:(n-k)>\delta n} T_{1,km} \le n \max_{k:(n-k)>\delta n} \{\exp(-(n-k)\log(n))\} < n^{-\delta n+1} = o(1)$$

On the other hand, if $kl < (1 - \delta)nm$, k < n and l < m, note first that the maximum is taken over all C in $C_{nm,kl}$, but only the rows and columns outside C_0 actually play a role over the sum $\xi_{C \setminus C_0}$. There are $\binom{N-n}{n-k} \cdot \binom{M-m}{m-l} \cdot \binom{n}{k} \cdot \binom{m}{l}$ different values of this sum. We write:

$$T_{1,kl} \leq \binom{N-n}{n-k} \cdot \binom{M-m}{m-l} \cdot \binom{n}{k} \cdot \binom{m}{l} P\left(Z > (1-\delta_1)a\sqrt{nm-kl}\right)$$

$$\leq \binom{N-n}{n-k} \cdot \binom{M-m}{m-l} \cdot \binom{n}{k} \cdot \binom{m}{l} \exp\left(-\frac{(1-\delta_1)^2}{2}a^2(nm-kl)\right)$$

$$\leq \exp\left(-\frac{(1-\delta_1)^2}{2}a^2(nm-kl) + \log\left(\binom{N-n}{n-k}\binom{M-m}{m-l}\binom{n}{k}\binom{m}{l}\right)\right).$$
(3.6)

As we have n, m, N, M tend to infinity, then

$$\begin{split} \log\left(\binom{N-n}{n-k}\cdot\binom{M-m}{m-l}\cdot\binom{n}{k}\cdot\binom{m}{l}\right) &\leq \left((n-k)\log\left(\frac{N-n}{n-k}\right) + (m-l)\log\left(\frac{M-m}{m-l}\right)\right)(1+o(1)) \\ &+ (n-k)\log\left(\frac{ne}{n-k}\right) + (m-l)\log\left(\frac{me}{m-l}\right) \\ &\leq \left((n-k)\log\left(\frac{N-n}{n}\right) + (m-l)\log\left(\frac{M-m}{m}\right)\right)(1+o(1)) \\ &+ \left((n-k)\log\left(\frac{n}{n-k}\right) + (m-l)\log\left(\frac{m}{m-l}\right)\right)(1+o(1)) \\ &+ (n-k)\log\left(\frac{ne}{n-k}\right) + (m-l)\log\left(\frac{me}{m-l}\right). \end{split}$$

Let us see that (N - n)/n = N/n(1 + o(1)) and that

$$(n-k)\log\left(\frac{n^2e}{(n-k)^2}\right) = n\left(1-\frac{k}{n}\right)\left(1-2\log\left(1-\frac{k}{n}\right)\right) \le \frac{2}{\sqrt{e}}n,$$

as $x(1 - 2\log(x)) \le 2/\sqrt{e}$ for all x in [0, 1].

Let us denote $\mathcal{X} := n \log(p^{-1})$ and $\mathcal{Y} := m \log(q^{-1})$. We have

$$\log\binom{N-n}{n-k} \cdot \binom{M-m}{m-l} \cdot \binom{n}{k} \cdot \binom{m}{l} \leq \left(\left(1-\frac{k}{n}\right)\mathcal{X} + \left(1-\frac{l}{m}\right)\mathcal{Y} + \frac{2}{\sqrt{e}}(n+m)\right)(1+o(1)).$$

Analogously to (3.5) we have

$$\frac{(1-\delta_1)^2}{2}a^2nm \ge (1+\alpha_1)\left(\mathcal{X}+\mathcal{Y}\right),$$

asymptotically.

Finally, we get, for large enough n, m, N, M

$$\begin{split} &-\frac{(1-\delta_1)^2 a^2}{2}(nm-kl) + \log\left(\binom{N-n}{n-k} \cdot \binom{M-m}{m-l} \cdot \binom{n}{k} \cdot \binom{m}{l}\right) \\ &\leq -\alpha_1 \left(1 - \frac{kl}{nm}\right) (\mathcal{X} + \mathcal{Y}) \\ &- \left(1 - \frac{kl}{nm}\right) (\mathcal{X} + \mathcal{Y}) + \left(\left(1 - \frac{k}{n}\right) \mathcal{X} + \left(1 - \frac{l}{m}\right) \mathcal{Y} + \frac{2}{\sqrt{e}}(n+m)\right) (1+o(1)) \\ &\leq -\frac{\alpha_1}{2} \left(1 - \frac{kl}{nm}\right) (\mathcal{X} + \mathcal{Y}) + \frac{k}{n} \left(\frac{l}{m} - 1\right) \mathcal{X} + \frac{l}{m} \left(\frac{k}{n} - 1\right) \mathcal{Y} + \frac{2}{\sqrt{e}}(n+m)(1+o(1)) \\ &\leq -\frac{\alpha_1}{2} \delta(\mathcal{X} + \mathcal{Y}) + \frac{2}{\sqrt{e}}(n+m)(1+o(1)). \end{split}$$

Therefore, we replace this bound in (3.6) and get

$$\sum_{k=0}^{n} \sum_{l=0}^{m} I_{kl < (1-\delta)nm} T_{1,kl} \le 2 \exp\left(-\frac{\alpha_1}{2}\delta(n\log(p^{-1}) + m\log(q^{-1})) + \frac{2}{\sqrt{e}}(n+m)(1+o(1)) + \log(nm)\right) = o(1).$$

For $T_{2,kl}$, only the common elements of C and C_0 play a role on the random variable $\xi_{C\cap C_0}$ and there are $\binom{n}{k} \cdot \binom{m}{l}$ such choices. Note that we cannot have here neither k = 0 nor l = 0, as $T_{2,kl} = 0$ in this cases. Therefore,

$$\begin{split} \sum_{k=1}^{n} \sum_{l=1}^{m} I_{kl < (1-\delta)nm} T_{2,kl} &\leq \sum_{k=1}^{n} \sum_{l=1}^{m} \binom{n}{k} \cdot \binom{m}{l} P\left(Z > \frac{\delta_1 a(nm-kl)}{2\sqrt{kl}}\right) \\ &\leq \sum_{k=1}^{n} \sum_{l=1}^{m} \binom{n}{k} \cdot \binom{m}{l} P\left(Z > \frac{\delta_1 \delta anm}{2\sqrt{(1-\delta)nm}}\right) \\ &\leq \sum_{k=1}^{n} \sum_{l=1}^{m} \exp\left(-\frac{\delta_1^2 \delta^2 a^2 nm}{8(1-\delta)} + k \log\left(\frac{ne}{k}\right) + l \log\left(\frac{me}{l}\right)\right) \\ &\leq \exp\left(-\frac{\delta_1^2 \delta^2 a^2 nm}{8(1-\delta)} + n + m + \log(nm)\right) = o(1). \end{split}$$

Here, we have used the fact that $x \log(x^{-1})$ is bounded from above by e^{-1} for all $x \in [0, 1]$ and used it for x = k/(ne) and for x = l/(me), respectively. Use (3.4) in order to conclude.

Finally, for $T_{3,kl}$, we write that $-\xi_{C_0}/\sqrt{nm}$ behaves like some standard Gaussian random variable Z and get

$$\sum_{k=0}^{n} \sum_{l=0}^{m} T_{3,kl} \le \sum_{k=0}^{n} \sum_{l=0}^{m} \exp\left(-\frac{\delta_1^2 a^2 (nm-kl)^2}{8nm}\right)$$
$$\le \exp\left(-\frac{\delta_1^2 \delta^2 a^2}{8} nm + \log(nm)\right) = o(1),$$

as a^2nm tends to infinity faster than $\log(nm)$ due to (3.4) in our setup.

In conclusion, the probability in (3.1) tends to 0:

$$P_{C_0}\left(\max_{k,l}\max_{kl<(1-\delta)nm}\max_{C\in\mathcal{C}_{nm,kl}}(\xi_{C\setminus C_0} - \xi_{C_0\setminus C} - a(nm-kl)) > 0\right) = o(1).$$
(3.7)

3.2. Large intersection

Let us fix k and l such that $kl \ge (1 - \delta)nm$, or, equivalently, $nm - kl \le \delta nm$. Note that it implies both $k \ge (1 - \delta_1)n$ and $l \ge (1 - \delta_1)m$ for some δ_1 depending on δ small as $\delta \to 0$. The case n = k and m = l gives an event with 0 probability.

We decompose as follows

$$\begin{aligned} \xi_{C\backslash C_0} - \xi_{C_0\backslash C} &= \left(\xi_{(A\backslash A_0)\times B_0} - \xi_{(A_0\backslash A)\times B_0}\right) + \left(\xi_{A_0\times (B\backslash B_0)} - \xi_{A_0\times (B_0\backslash B)}\right) \\ &+ \left(\xi_{(A\backslash A_0)\times (B\backslash B_0)} - \xi_{(A\backslash A_0)\times (B_0\backslash B)} + \xi_{(A_0\backslash A)\times (B_0\backslash B)} - \xi_{(A_0\backslash A)\times (B\backslash B_0)}\right) \\ &= S_1 + S_2 + S_3, \text{ say.}\end{aligned}$$

We shall bound from above as follows

$$\begin{split} &P_{C_0}\left(\max_{k\geq (1-\delta_1)n}\max_{l\geq (1-\delta_1)m}\max_{C\in\mathcal{C}_{nm,kl}}\left(\xi_{C\backslash C_0}-\xi_{C_0\backslash C}-a(nm-kl)\right)>0\right)\\ &\leq P_{C_0}\left(\max_{k\geq (1-\delta_1)n}\max_{l\geq (1-\delta_1)m}\max_{A\in\mathcal{C}_{n,k}}\left(S_1-(1-\tilde{\delta})a(n-k)\frac{m+l}{2}\right)>0\right)\\ &+P_{C_0}\left(\max_{k\geq (1-\delta_1)n}\max_{l\geq (1-\delta_1)m}\max_{B\in\mathcal{C}_{m,kl}}\left(S_2-(1-\tilde{\delta})a(m-l)\frac{n+k}{2}\right)>0\right)\\ &+P_{C_0}\left(\max_{k\geq (1-\delta_1)n}\max_{l\geq (1-\delta_1)m}\max_{C\in\mathcal{C}_{nm,kl}}\left(S_3-\tilde{\delta}a(nm-kl)\right)>0\right),\end{split}$$

where $C_{n,k}$ is the set of *n* rows in 1,..., *N* having *k* values in common with A_0 and similarly for $C_{m,l}$ set of *m* columns in 1,..., *M* having *l* values in common with B_0 . Moreover, the previous sum can be bounded from above by

$$\sum_{k \ge (1-\delta_1)n} P_{C_0} \left(\max_{A \in \mathcal{C}_{n,k}} S_1 > (1-\tilde{\delta})a(n-k)m(1-\delta_1/2) \right) \\ + \sum_{l \ge (1-\delta_1)m} P_{C_0} \left(\max_{B \in \mathcal{C}_{m,l}} S_2 > (1-\tilde{\delta})a(m-l)n(1-\delta_1/2) \right) \\ + \sum_{k \ge (1-\delta_1)n} \sum_{l \ge (1-\delta_1)m} P_{C_0} \left(\max_{C \in \mathcal{C}_{nm,kl}} S_3 > \tilde{\delta}a(nm-kl) \right) \\ = \sum_{k \ge (1-\delta_1)n} U_{1,k} + \sum_{l \ge (1-\delta_1)m} U_{2,l} + \sum_{k \ge (1-\delta_1)n} \sum_{l \ge (1-\delta_1)m} U_{3,kl} \text{ say,}$$

Let us now deal with $U_{1,kl}$. Note, first, that the case k = n gives probability 0. For $(1 - \delta_1/2)n \le k \le n - 1$, we put $p_{n,N} = \sqrt{\log(N-n)}/(\sqrt{\log(N-n)} + \sqrt{\log(n)})$ and $q_{n,N} = 1 - p_{n,N}$,

$$U_{1,k} \le P_{C_0} \left(\max_{A \in \mathcal{C}_{n,k}} \xi_{(A \setminus A_0) \times B_0} > (1 - \delta)(1 - \delta_1/2)a(n - k)mp_{n,N} \right) \\ + P_{C_0} \left(\max_{A \in \mathcal{C}_{n,k}} (-\xi_{(A_0 \setminus A) \times B_0}) > (1 - \delta)(1 - \delta_1/2)a(n - k)mq_{n,N} \right)$$

and, for some independent standard gaussian r.v. Z_1 and Z_2 , using $l \ge (1 - \delta_1)m$

$$\begin{aligned} U_{1,k} &\leq \binom{N-n}{n-k} P(Z_1 > (1-\delta)(1-\delta_1/2)p_{n,N}a\sqrt{(n-k)m}) \\ &+ \binom{n}{k} P(Z_2 > (1-\delta)(1-\delta_1/2)q_{n,N}a\sqrt{(n-k)m}) \\ &\leq \exp\left(-\frac{(1-\tilde{\delta})^2}{2}\frac{a^2m(n-k)\log(N-n)}{\left(\sqrt{\log(N-n)} + \sqrt{\log(n)}\right)^2} + \log\left(\binom{N-n}{n-k}\right)\right) \\ &+ \exp\left(-\frac{(1-\tilde{\delta})^2}{2}\frac{a^2m(n-k)\log(n)}{\left(\sqrt{\log(N-n)} + \sqrt{\log(n)}\right)^2} + \log(C_n^k)\right), \end{aligned}$$

with $1 - \tilde{\delta} = (1 - \delta)(1 - \delta_1/2)$. Note that $\log\binom{N-n}{n-k} \leq (n-k)\log(N-n)(1 + o(1))$ and that $\log(C_n^k) \leq (n-k)\log(n)(1 + o(1))$. We obtain

$$U_{1,k} \le \exp\left(-(n-k)\log(N-n)\left(\frac{(1-\tilde{\delta})^2}{2}\frac{a^2m}{(\sqrt{\log(N-n)}+\sqrt{\log(n)})^2} - (1+o(1))\right)\right) + \exp\left(-(n-k)\log(n)\left(\frac{(1-\tilde{\delta})^2}{2}\frac{a^2m}{(\sqrt{\log(N-n)}+\sqrt{\log(n)})^2} - (1+o(1))\right)\right).$$

We use (3.4), for small enough δ

$$(1 - \tilde{\delta})^2 a^2 m \ge (1 + 2\alpha_2) 2(\sqrt{\log(n)} + \sqrt{\log(N - n)})^2,$$

for some $\alpha_2 > 0$ and this means

$$\frac{(1-\tilde{\delta})^2}{2} \frac{a^2 m}{(\sqrt{\log(N-n)} + \sqrt{\log(n)})^2} - (1+o(1)) \ge 2\alpha_2 - o(1) \ge \alpha_2.$$

Finally,

$$\sum_{(1-\delta_1)n \le k < n} U_{1,k} \le \sum_{(1-\delta_1)n \le k < n} (e^{-\alpha_2 \log(N-n)(n-k)} + e^{-\alpha_2 \log(n)(n-k)})$$
$$\le \sum_{1 \le j \le \delta_1 n} \left(e^{-\alpha_2 \log(N-n)j} + e^{-\alpha_2 \log(n)j} \right)$$
$$= \left(e^{-\alpha_2 \log(N-n)} + e^{-\alpha_2 \log(n)} \right)(1+o(1)) = o(1).$$

The term $U_{2,l}$ is similar.

As for the last term, $U_{3,kl}$, we compare each sum in S_3 to $\tilde{\delta}a(nm-kl)/4$. The most difficult (the largest) upper bound is for the first sum, as it gives the largest number of choices $\binom{N-n}{n-k}\binom{M-m}{m-l}$. Note that this term

is 0 if k = n or l = m. Therefore, we only explain this term, for $k \le n-1$ and $l \le m-1$,

$$\begin{aligned} U_{31,kl} &= P_{C_0} \left(\max_{C \in \mathcal{C}_{nm,kl}} \xi_{(A \setminus A_0) \times (B \setminus B_0)} > \frac{\tilde{\delta}}{4} a(nm-kl) \right) \\ &\leq \binom{N-n}{n-k} \binom{M-m}{m-l} \exp\left(-\frac{(\tilde{\delta}/4)^2}{2} \frac{a^2(nm-kl)^2}{(n-k)(m-l)} \right) \\ &\leq \exp\left(-\frac{(\tilde{\delta}/4)^2 a^2(n(m-l)P_{k,n} + (n-k)mP_{l,m})^2}{2(n-k)(m-l)} + (n-k)\log(N-n) + (m-l)\log(M-m) \right), \end{aligned}$$

where $P_{k,n} = 1 - (n-k)/(2n)$ and $P_{l,m} = 1 - (m-l)/(2m)$. Recall that $n-k \leq \delta_1 n$ and that $m-l \leq \delta_1 m$. We get

$$U_{31,kl} \le \exp\left(-\frac{(\tilde{\delta}/4)^2 a^2}{2} \left(H + 2nm P_{k,n} P_{l,m}\right) + \delta_1 (n \log(N-n) + m \log(M-m))\right),$$

where

$$H = \frac{n^2}{n-k}(m-l)P_{k,n}^2 + (n-k)\frac{m^2}{m-l}P_{l,n}^2 \ge \frac{1}{\delta_1}\left(nP_{k,n}^2 + mP_{l,n}^2\right)$$

Recall that $P_{k,n} \ge 1 - \delta_1/2$ and $P_{l,m} \ge 1 - \delta_1/2$. We get for $(\tilde{\delta}/4)^2 = \delta_1$:

$$U_{31,kl} \le \exp\left(-\frac{a^2}{2}\left(nP_{k,n}^2 + mP_{l,n}^2\right) - \delta_1(a^2 nm P_{k,n} P_{l,m} - (n\log(N-n) + m\log(M-m))\right),$$

with

$$a^{2}nmP_{k,n}P_{l,m} \ge (1-\delta_{1}/2)^{2} \left(\frac{1}{2}a^{2}nm + \frac{1}{2}a^{2}nm\right)$$

$$\ge (1-\delta_{1}/2)^{2}(1+\alpha) \left(n\log(n(N-n)) + m\log(m(M-m))\right),$$

by (3.4). By taking δ_1 small enough, we may find $\delta_2 > 0$ such that $(1 - \delta_1/2)^2(1 + \alpha) \ge 1 + \delta_2$. This is enough to conclude that

$$a^{2}nmP_{k,n}P_{l,m} - (n\log(N-n) + m\log(M-m)) > 0$$

and that

$$U_{31,kl} \le \exp\left(-\frac{a^2}{2}(n+m)(1-\delta_1/2)^2\right)$$

$$\le \exp\left(-(1-\delta_1/2)^2(1+\alpha)(\log(m(M-m)) + \log(n(N-n)))\right)$$

$$\le \exp\left(-(1+\delta_2)(\log(m(M-m)) + \log(n(N-n)))\right).$$

In conclusion,

$$\sum_{(1-\delta_1)n \le k < n} \sum_{(1-\delta_1)m \le l < m} U_{31,kl} \le \exp\left(-(1+\delta_2)\log((M-m)(N-n)) - \delta_2\log(nm)\right) = o(1).$$

Here, we have proven that

$$P_{C_0}\left(\max_{k, l} \max_{kl \ge (1-\delta)nm} \max_{C \in \mathcal{C}_{nm,kl}} (\xi_{C \setminus C_0} - \xi_{C_0 \setminus C} - a(nm - kl)) > 0\right) = o(1).$$
(3.8)

From (3.8) and (3.7) we deduce that the probability $P_{C_0}(\hat{C}^* \neq C_0)$ tends to 0 and this concludes the proof of the upper bounds.

4. Lower bounds

Let (2.1) and (2.4). We shall call the case when B = A the case of severe sparsity, while the case where either $B = A_1$ or $B = A_2$ will be designated by moderately sparse cases. Let us first consider a set Θ of matrices having size $N \times M$ and containing S_C , for all $C \in \mathcal{C}_{nm}$, such that $[S_C]_{ij} = a \cdot I((i, j) \in C)$. This set is on the border of $\mathcal{S}_{nm,a}$, as we replace $[S_C]_{ij} \ge a$ with equality, for all $(i, j) \in C$. The set Θ has $L = \binom{N}{n} \cdot \binom{M}{m}$ elements. Let P_0 denote the likelihood of $N \times M$ standard gaussian observations and, as previously, P_C the likelihood of our observations under parameter S_C . The minimax risk is bounded from below by the minimax risk over Θ :

$$\inf_{\hat{C}} \sup_{S_C \in \mathcal{S}_{nm,a}} P_C(\hat{C}(Y) \neq C) \ge \inf_{\hat{C}} \sup_{S_C \in \Theta} P_C(\hat{C}(Y) \neq C).$$

4.1. Severe sparsity

Proof of Theorem 2.3 for severely sparse case. In this case, we shall apply Theorem 2.4 in [31]: if there exists $\tau > 0$ and $0 < \alpha < 1$ such that

$$\frac{1}{L}\sum_{S_C\in\Theta} P_C\left(\frac{dP_0}{dP_C} \ge \tau\right) \ge 1 - \alpha$$

then

$$\inf_{\hat{C}} \sup_{S_C \in \Theta} P_C(\hat{C}(Y) \neq C) \ge \frac{\tau L}{1 + \tau L} (1 - \alpha)$$

In our model, the likelihood ratio is

$$\frac{dP_0}{dP_C} = \exp\left(-aY_C + \frac{a^2nm}{2}\right). \tag{4.1}$$

This implies that

$$P_C\left(\frac{dP_0}{dP_C} \ge \tau\right) = P_C\left(-aY_C + \frac{a^2nm}{2} \ge \log(\tau)\right)$$
$$= P_0\left(-\frac{1}{\sqrt{nm}}\xi_C - \frac{a\sqrt{nm}}{2} \ge \frac{\log(\tau)}{a\sqrt{nm}}\right)$$
$$= P\left(Z \ge \frac{\log(\tau)}{a\sqrt{nm}} + \frac{a\sqrt{nm}}{2}\right),$$

where Z is standard gaussian. Let $z_{1-\alpha}$ be the quantile of probability $1-\alpha$ of a standard gaussian distribution, such that $P(Z \ge -z_{1-\alpha}) = 1-\alpha$. In order to check (4.1), we need $\log(\tau) \le -a^2 nm/2 - z_{1-\alpha}a\sqrt{nm}$.

On the one hand, if $a\sqrt{nm} = O(1)$ we take τ as solution of the equation $\log(\tau) = -a^2 nm/2 - z_{1-\alpha}a\sqrt{nm}$. Therefore, we have $\tau \approx 1$ and then

$$\frac{\tau L}{1+\tau L}(1-\alpha) \ge (1-\alpha)^2 > 0, \quad \text{as } L \to \infty.$$

On the other hand, if $a\sqrt{nm} \to \infty$, we take $\tau^{-1} = L/\log(L)$, with $L = \binom{N}{n}\binom{M}{m}$, which gives $\tau L \to \infty$ and $\log(\tau^{-1}) \sim \log(L)$. We can prove that

$$\log(\tau^{-1}) \ge \frac{a^2 nm}{2} + z_{1-\alpha} a \sqrt{nm} = \frac{a^2 nm}{2} \left(1 + \frac{2z_{1-\alpha}}{a \sqrt{nm}} \right)$$

Indeed, we known that $\log(L) \sim n \log(p^{-1}) + m \log(q^{-1})$ and, by assumption (2.4),

$$\frac{a^2 nm}{2(n\log(p^{-1}) + m\log(q^{-1}))} \le 1 - \delta,$$

asymptotically, for some $\delta > 0$. It implies that

$$\frac{a^2 nm}{2\log(\tau^{-1})} \le \left(1 + \frac{2z_{1-\alpha}}{a\sqrt{nm}}\right)^{-1},$$

asymptotically. This gives the lower bound

$$\frac{\tau L}{1+\tau L}(1-\alpha) \ge (1-\alpha)^2 > 0.$$

As $\alpha > 0$ can be chosen arbitrarly small, we obtain the result

$$\inf_{\hat{C}} \sup_{S_C \in \Theta} P_C(\hat{C}(Y) \neq C) \to 1.$$

4.2. Moderate sparsity

Lemma 4.1. If η_1, \ldots, η_J are *i.i.d.* random variables with standard gaussian law, then

if
$$t < 1$$
, $P(\max_{j=1,\dots,J} \eta_j \ge t\sqrt{2\log(J)}) \to 1$, as $J \to \infty$,

and

if
$$t > 1$$
, $P(\max_{j=1,\dots,J} \eta_j \ge t\sqrt{2\log(J)}) \to 0$, as $J \to \infty$.

Proof. This Lemma is an obvious consequence of the limit behaviour of the normalized maximum of i.i.d. Gaussian random variables as follows:

$$V_J := \max_{j=1,\dots,J} \eta_j \sqrt{2\log(J)} - 2\log(J) + \frac{1}{2}\log(\log(J)) + \frac{1}{4}\log(4\pi) \to^d U,$$

where U has the Gumbel law with distribution function $P(U \le x) = \exp(-\exp(-x))$ for all real number x, see [17]. Therefore, if t < 1,

$$P\left(\max_{j=1,\dots,J}\eta_{j} \ge t\sqrt{2\log(J)}\right) = P\left(V_{J} \ge (t-1)2\log(J) + \frac{1}{2}\log(\log(J)) + \frac{1}{4}\log(4\pi)\right),$$

which tends to 1 when $J \to \infty$. The other limit is obtained by a similar argument.

Proof of Proposition 2.2. Let us assume that $\limsup A_1 < 1$ and treat the other case similarly This means that $A_1 \leq 1 - \alpha$, for some fixed $0 < \alpha < 1$. Equivalently, $a\sqrt{m} \leq (1 - \alpha)(\sqrt{2\log(n)} + \sqrt{2\log(N - n)})$.

In this case we shall reduce the set of matrices C to those matrices having the same columns as C_0 and n-1 rows in common with C_0 . Then we sum up each line over these columns and reduce the problem to the vector case. Thus,

$$P_{C_0}(\hat{C}^{\star} \neq C_0) = P_{C_0} \left(\max_{C \in \mathcal{C}_{nm}} Y_C - Y_{C_0} > 0 \right)$$

$$\geq P_{C_0} \left(\max_{C = A \times B_0} Y_C - Y_{C_0} > 0 \right)$$

$$\geq P_{C_0} \left(\max_{A} \sum_{A} Y_{i.} - \sum_{A_0} Y_{i.} > 0 \right),$$

where the maximum over A is taken over all sets of n rows having n-1 rows in common with A_0 and

$$Y_{i} := \sum_{j \in B_0} Y_{ij} = amI(i \in A_0) + \sum_{j \in B_0} \xi_{ij}.$$

Denote by $\eta_i = m^{-1/2} \sum_{j \in B_0} \xi_{ij}$ for i = 1, ..., N, which are i.i.d. random variables of standard gaussian law. Therefore, we get

$$\begin{aligned} P_{C_0}(\hat{C}^{\star} \neq C_0) &\geq P_{C_0}(\max_{i \notin A_0} \eta_i + \max_{k \in A_0}(-(\eta_k + a\sqrt{m})) > 0) \\ &\geq P_{C_0}(\max_{i \notin A_0} \eta_i + \max_{k \in A_0}(-\eta_k) > (1 - \alpha)(\sqrt{2\log(N - n)} + \sqrt{2\log(n)})) \\ &= 1 - P_{C_0}(\max_{i \notin A_0} \eta_i + \max_{k \in A_0}(-\eta_k) \le (1 - \alpha)(\sqrt{2\log(N - n)} + \sqrt{2\log(n)})), \end{aligned}$$

by the assumption on A_1 . Moreover

$$P_{C_0}(\max_{i \notin A_0} \eta_i + \max_{k \in A_0} (-\eta_k) \le (1 - \alpha)(\sqrt{2\log(N - n)} + \sqrt{2\log(n)}))$$

$$\le P_{C_0}(\max_{i \notin A_0} \eta_i \le (1 - \alpha)\sqrt{2\log(N - n)}) + P_{C_0}(\max_{k \in A_0} (-\eta_k) \le (1 - \alpha)\sqrt{2\log(n)}),$$

0. by Lemma 4.1.

which tends to 0, by Lemma 4.1.

Proof of Theorem 2.3 for moderately sparse case. In this case we check that the minimax risk is bounded from below by the risk of the maximum likelihood estimator \hat{C}^* and that its risk tends to 1 under our assumptions by Proposition 2.2. Let us see that

$$\begin{split} \inf_{\hat{C}} \sup_{S_{C} \in \Theta} P_{C}(\hat{C}(Y) \neq C) &\geq \inf_{\hat{C}} \frac{1}{L} \sum_{k=1}^{L} P_{C_{k}}(\hat{C}(Y) \neq C_{k}) \\ &\geq \inf_{\hat{C}} \left(1 - \frac{1}{L} \sum_{k=1}^{L} P_{C_{k}}(\hat{C}(Y) = C_{k}) \right) \\ &\geq 1 - \sup_{\hat{C}} \frac{1}{L} \sum_{k=1}^{L} E_{0} \left(I(\hat{C}(Y) = C_{k}) \frac{dP_{C_{k}}}{dP_{0}}(Y) \right), \end{split}$$

where $L = {\binom{N}{n}} {\binom{M}{m}}$ is the number of elements in Θ . In the previous supremum, we may replace the arbitrary measurable function $\hat{C}(Y)$ by a test function $\psi(Y)$ taking values in $1, \ldots, L$. The test maximising

$$\sup_{\psi(Y)} \frac{1}{L} \sum_{k=1}^{L} E_0 \left(I(\psi(Y) = k) \frac{dP_{C_k}}{dP_0}(Y) \right)$$

will choose k such that C_k has maximal likelihood: $\{Y : \frac{dP_{C_k}}{dP_0}(Y) \ge \frac{dP_{C_j}}{dP_0}(Y)$, for all $j = 1, \ldots, L\}$. Thus, we get the risk of a maximum likelihood estimator,

$$\inf_{\hat{C}} \sup_{S_{C} \in \Theta} P_{C}(\hat{C}(Y) \neq C) \ge 1 - \frac{1}{L} \sum_{k=1}^{L} P_{C_{k}}(\hat{C}^{\star}(Y) = C_{k})$$
$$\ge \frac{1}{L} \sum_{k=1}^{L} P_{C_{k}}(\hat{C}^{\star}(Y) \neq C_{k}),$$

which tends to 1 by Proposition 2.2.

Proof of Lemma 2.9. We want to prove that $a^2 nmpq \rightarrow 0$. We write that

$$a^{2}nmpq = a^{2} \cdot \frac{n^{2}}{N} \cdot \frac{m^{2}}{M} = a^{2}N^{2u^{2}-1}M^{2v^{2}-1} \le a^{2}(\max\{N,M\})^{2(u^{2}+v^{2}-1)},$$

where we denote $u^2 = \log(n) / \log(N)$ and $v^2 = \log(m) / \log(M)$. Note that $u^2, v^2 \in (0, 1)$.

Suppose now that N, M, n, m are large enough.

131

As $\limsup A < 1$ we have that

$$a^2 < 2\left(\frac{\log(N/n)}{m} + \frac{\log(M/m)}{n}\right) \tag{4.2}$$

and that gives

$$a^{2}nmpq \leq 2 \cdot (\max\{N,M\})^{2(u^{2}+v^{2}-1)} \left(\frac{\log(N/n)}{m} + \frac{\log(M/m)}{n}\right) \to 0$$

as soon as $\limsup(u^2 + v^2) < 1$.

Let us further denote by $X = \log(N)/m$ and $Y = \log(M)/n$. Then, (4.2) writes also

$$a^{2} < 2((1-u^{2})X + (1-v^{2})Y).$$
(4.3)

On the other hand, $\liminf A_1 > 1$ implies that we can find $\delta_1 > 0$ such

$$a > (1+2\delta_1)\sqrt{\frac{2\log(N)}{m}} \left(u + \sqrt{1 + \frac{\log(1-n/N)}{\log(N)}}\right)$$
$$> (1+2\delta_1)\sqrt{2X} \left(u + 1 - \frac{1}{2\log(N)}\frac{n}{N}(1+o(1))\right)$$
$$> (1+\delta_1)\sqrt{2X}(u+1).$$

Similarly, we get from $\liminf A_2 > 1$ that $a > (1 + \delta_2)\sqrt{2Y(v+1)}$, for some $\delta_2 > 0$. Together with (4.3), we get, for $\delta = \max\{\delta_1, \delta_2\}$

$$2((1-u^2)X + (1-v^2)Y) > (1+\delta)2\max\{X(u+1)^2, Y(v+1)^2\}$$

Assume that $X(u+1)^2 \ge Y(v+1)^2$, then we write

$$(1 - v^2)Y > X[(1 + \delta)2u(u + 1) + \delta(1 - u^2)] > (1 + \delta)2u(u + 1)X$$

and therefore

$$(1+\delta)\frac{2u(u+1)}{1-v^2} < \frac{Y}{X} \le \frac{(u+1)^2}{(v+1)^2}.$$

This further gives $(1 + \Delta)(u + v) < 1$, where for any u > 0, v > 0 we have

$$\Delta = \frac{3uv + 2\delta u(v+1)}{u+v} = \delta + \frac{uv(3+2\delta) + \delta(u-v)}{u+v} > \frac{\delta}{2} > 0.$$

Similarly if $X(u+1)^2 < Y(v+1)^2$ we have $(1+\Delta_1)(u+v) < 1$, $\Delta_1 > \delta/2$, which implies $\limsup(u^2+v^2) < 1$. \Box

5. FINAL COMMENTS

Future extensions of the problem of sharp asymptotics for selection include several open problems.

First, a natural question is how to proceed when the variance of our observations is not available. Model (1.1) becomes $Y_{ij} = s_{ij} + \sigma \xi_{ij}$, i = 1, ..., N, j = 1, ..., M, for $\sigma > 0$ unknown. Following [9], we suggest to use the estimator

$$\hat{\sigma}^2 = \frac{1}{NM} \sum_{(i,j)\in C} Y_{ij}^2$$

of σ^2 , which has bias and variance given by

$$E\hat{\sigma}^2 - \sigma^2 = \sigma^2 G_{\sigma}, \ Var(\hat{\sigma}^2) = \frac{2\sigma^4}{NM}(1 + 2G_{\sigma}), \ \text{where} \ G_{\sigma} = \frac{1}{\sigma^2 NM} \sum_{(i,j)\in C} s_{ij}^2.$$

The same selection procedure should be employed with Y_{ij} replaced by $Y_{ij}/\hat{\sigma}$. We conjecture that, under some conditions on G_{σ} , all results should hold when we replace a by $a/\sqrt{\sigma}$ in the Definition 2.2 of A, A_1 and A_2 .

Next, let us consider two-sided variable selection problem, *i.e.* finding C_0 where the mean $|s_{ij}| \ge a$, for $(i, j) \in C_0$. A natural modification of the method is to use in this case the selecting procedure

$$\widetilde{C}^{\star}(Y) = \arg \max_{C \in \mathcal{C}_{nm}} T_C$$
, where $T_C = \frac{1}{2nm} \sum_{(i,j) \in C} (Y_{ij}^2 - 1)$.

We expect similar results to hold if we replace a by a^2 in A, A_1 and A_2 given by (2.2).

Other extensions include considering non Gaussian observations, but having distribution in the exponential family of distributions as well as the interesting case when the submatrix is supposed to be a block (adjacent rows and columns). In the latter case, we expect the selection procedure to be fast and quite easy to implement.

Acknowledgements. The authors want to thank the anonymous referees who helped improve significantly the presentation of the manuscript.

References

- F. Abramovich, Y. Benjamini, D.L. Donoho and I.M. Johnstone, Adapting to unknown sparsity by controlling the false discovery rate. Ann. Statist. 34 (2006) 559–1047.
- [2] E. Arias-Castro, E.J. Candès and A. Durand, Detection of an anomalous clusters in a network. Ann. Statist. 39 (2011) 278-304.
- [3] E. Arias-Castro, E.J. Candès and Y. Plan, Global Testing and Sparse Alternatives: ANOVA, Multiple Comparisons and the Higher Criticism. Preprint arXiv:1007.1434 (2010).
- [4] E. Arias-Castro, D.L. Donoho and X. Huo, Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Trans. Inform. Theory* 51 (2005) 2402–2425.
- [5] E. Arias-Castro and K. Lounici, Variable selection with exponential weights and l₀-penalization. Preprint arxiv:1208.2635 (2012).
- Y. Benjamini and Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Statist. Soc. Ser. B 57 (1995) 289–3300.
- [7] K. Bertin and G. Lecué, Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electron. J. Stat.* 2 (2008) 1224–1241.
- [8] P.J. Bickel, Y. Ritov and A.B. Tsybakov, Simultaneous analysis of Lasso and Dantzig selector. Ann. Statist. 37 (2009) 1705–1732.
- [9] C. Butucea and Yu.I. Ingster, Detection of a sparse submatrix of a high-dimensional noisy matrix. Bernoulli 19 (2013) 2652-2688.
- [10] C. Butucea and G. Gayraud, Sharp detection of smooth signals in a high-dimensional sparse matrix with indirect observations. Preprint arxiv:1301.4660 (2013).
- [11] T. Cai, J. Jin and M. Low, Estimation and confidence sets for sparse normal mixtures. Ann. Statist. 35 (2007) 2421–2449.
- [12] E.J. Candès and Y. Plan, Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. IEEE Trans. Inform. Theory. 57 (2011) 2342–2359
- [13] E.J. Candès and B. Recht, Exact matrix completion via convex optimization. Found. Comput. Math. 9 (2009) 717–772.
- [14] L. Comminges and A.S. Dalalyan, Tight conditions for consistency of variable selection in the context of high dimensionality. Ann. Statist. 40 (2012) 2359–2763.
- [15] D.L. Donoho and J. Jin, Higher criticism for detecting sparse heterogeneous mixtures. Ann. Statist. 32 (2004) 962–994.
- [16] D.L. Donoho, I.M. Johnstone, C. Hoch and A. Stern, Maximum entropy and the nearly black object. With Discussion. J. Roy. Statist. Soc., Ser. B. 54 (1992) 4181.
- [17] P. Embrechts, C. Klüppelberg and T. Mikosch, Modelling extremal events: for insurance and finance. Springer (1997).
- [18] D. Gross, Recovering low-rank matrices from few coefficients in any basis *IEEE*, *Information theory* 57 (2010) 1548–1566.
- [19] Yu.I. Ingster, Some problems of hypothesis testing leading to infinitely divisible distributions. Math. Methods Stat. 6 (1997) 47–69.
- [20] Yu.I. Ingster and N.A. Stepanova, Adaptive selection of sparse regression function components. Zapiski Nauchn. Sem. POMI ZAI (2012).
- [21] Yu.I. Ingster and I.A. Suslina, Nonparametric goodness-of-fit testing under gaussian models. Vol. 169 of Lect. Notes in Statist. Springer-Verlag, New York (2003).
- [22] Yu.I. Ingster and I.A. Suslina, On a detection of a signal of known shape in multichannel system. Zapiski Nauchn. Sem. POMI 294 (2002) 88–112, Transl. J. Math. Sci. 127 (2002) 1723–1736.
- [23] R.H. Keshavan, A. Montanari and S. Oh, Matrix completion from noisy entries. J. Mach. Learn. Res. 11 (2010) 2057–2078.

C. BUTUCEA $ET\ AL.$

- [24] M. Kolar, S. Balakrishnan, A. Rinaldo and A. Singh, Minimax localization of structural information in large noisy matrices. NIPS (2011).
- [25] V. Koltchinskii, K. Lounici and A.B. Tsybakov, Nuclear norm penalization and optimal rates for noisy low rank matrix completion. Ann. Statist. 39 (2011) 2302–2329.
- [26] J. Lafferty and L. Wasserman, Rodeo: sparse, greedy nonparametric regression. Ann. Statist. 36 (2008) 28-63.
- [27] B. Recht, A simpler approach to matrix completion. J. Machine Learning 12 (2011) 3413–3430.
- [28] A. Rohde and A.B. Tsybakov, Estimation of high-dimensional low-rank matrices, Ann. Statist. 39 (2011) 887-930.
- [29] X. Sun and A.B. Nobel, On the maximal size of Large-Average and ANOVA-fit Submatrices in a Gaussian Random Matrix. Bernoulli 19 (2013) 275–294.
- [30] A.A. Shabalin, V.J. Weigman, C.M. Perou and A.B. Nobel, Finding Large Average Submatrices in High Dimensional Data. Ann. Appl. Statist. 3 (2009) 985–1012.
- [31] A.B. Tsybakov, Introduction to nonparametric statistics. Springer Ser. Stat. Springer, New-York (2009).
- [32] N. Verzelen, Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. Electron. J. Stat. 6 (2012) 3890.