Bernoulli **20**(3), 2014, 1029–1058
DOI: 10.3150/12-BEJ487

High-dimensional covariance matrix estimation with missing observations

KARIM LOUNICI

School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0160, USA.
E-mail: klounici@math.gatech.edu

In this paper, we study the problem of high-dimensional covariance matrix estimation with missing observations. We propose a simple procedure computationally tractable in high-dimension and that does not require imputation of the missing data. We establish non-asymptotic sparsity oracle inequalities for the estimation of the covariance matrix involving the Frobenius and the spectral norms which are valid for any setting of the sample size, probability of a missing observation and the dimensionality of the covariance matrix. We further establish minimax lower bounds showing that our rates are minimax optimal up to a logarithmic factor.

Keywords: covariance matrix; Lasso; low-rank matrix estimation; missing observations; non-commutative Bernstein inequality; optimal rate of convergence

1. Introduction

Let $X, X_1, \dots, X_n \in \mathbb{R}^p$ be i.i.d. zero mean vectors with unknown covariance matrix $\Sigma = \mathbb{E}[X \otimes X]$. Our objective is to estimate the unknown covariance matrix Σ when the vectors $X_1, \dots, X_n$ are partially observed, that is, when some of their components are not observed. More precisely, we consider the following framework. Denote by $X_i^{(j)}$ the jth component of the vector X_i. We assume that each component $X_i^{(j)}$ is observed independently of the others with probability $\delta \in (0, 1]$. Note that δ can be easily estimated by the proportion of observed entries. Therefore, we assume from now on that δ is known. Note also that the case $\delta = 1$ corresponds to the standard case of fully observed vectors. Let $(\delta_{i,j})_{1 \le i \le n, 1 \le j \le p}$ be a sequence of i.i.d. Bernoulli random variables with parameter δ and independent from $X_1, \dots, X_n$. We observe n i.i.d. random vectors $Y_1, \dots, Y_n \in \mathbb{R}^p$ whose components satisfy

$$Y_i^{(j)} = \delta_{i,j} X_i^{(j)}, \qquad 1 \le i \le n, 1 \le j \le p. \tag{1.1}$$

We can think of the $\delta_{i,j}$ as masked variables. If $\delta_{i,j} = 0$, then we cannot observe the jth component of X_i and the default value 0 is assigned to $Y_i^{(j)}$. Our goal is then to estimate Σ given the partial observations $Y_1, \dots, Y_n$.

The statistical problem of covariance estimation with missing observations is fundamental in multivariate statistics since it is often used as the first step to retrieve information in numerous applications where datasets with missing observations are common, for example:

1. Climate studies: n is the number of time points and p the number of observations stations, which may sometimes fail to produce an observation due to instrument malfunction. As a consequence, the generated datasets usually contain missing values.
2. Gene expression micro-arrays: n is the number of measurements and p the number of tested genes. Despite the improvement of gene expression techniques, the generated datasets frequently contain missing values with up to 90% of genes affected.
3. Cosmology: n is the number of images produced by a telescope and p is the number of pixels per image. With the development of very large telescopes and wide sky surveys, the generated datasets are huge but usually contain missing observations due to partial sky coverage or defective pixels.

One simple strategy to deal with missing data is to exclude from the analysis any variable for which observations are missing, thus restricting the analysis to a subset of fully observed variables. In gene expression data where 90% of the genes are affected by missing values, we would be left with too few variables to perform a legitimate statistical analysis. Also, discarding variables with very few missing observations is a waste of available information. Existing procedures involve complex imputation techniques to fill in the missing values through computationally intensive implementation of the EM algorithm (see [30] and the references cited therein for more details). In this paper, we propose a simple procedure computationally tractable in high-dimension that does not require imputing missing observations or discarding any available observation to recover the covariance matrix Σ.

Contemporary datasets are often huge with both large sample size n and dimension p and typically $p \gg n$. Consequently, a question of considerable practical interest is to perform dimension reduction, that is finding a good low-dimensional approximation for these huge datasets. This recent paradigm where high-dimensional objects of interest admit in fact a small intrinsic dimension has produced spectacular results in several fields. For instance, in compressed sensing, it is possible to recover s-sparse vectors of dimension p with only $n = \mathcal{O}(s \log(\mathrm{e}p/s))$ measurements provided these measurements are carried out properly (see [4,9,12,20,22,24] and the references cited therein for more details). An analogous result holds in matrix completion where approximate or exact recovery of a low-rank matrix $A \in \mathbb{R}^{p \times p}$ via nuclear norm minimization is possible with as few as $\mathcal{O}(pr \log^2 p)$ observed entries where r is the rank of A, under various sets of conditions on the sampling operator and the matrix of interest A. See [10,11,15,18,21,23,26,27] for more details. See also [5,19] for rank minimization approach.

A popular dimension reduction technique for covariance matrices is Principal Component Analysis (PCA), which exploits the spectrum of the sample covariance matrix. In the high-dimensional setting, [16] showed that the standard PCA procedure is bound to fail since the sample covariance spectrum is too spread out. Several alternatives have been studied in the literature to provide better estimates of the covariance matrix in the high-dimensional setting. A popular approach in Gaussian graphical models consists in estimating the inverse of the covariance matrix (called concentration matrix) since it admits a naturally sparse (or approximately sparse) structure if the dependence graph is itself sparse. See [2,7,25,35] and the references cited therein for more details. A limitation of this approach is that it does not apply to low rank matrices Σ since the concentration matrix does not exist in this case. Another popular approach assumes that the unknown covariance matrix is sparse in the sense that most of the entries are exactly or approximately zero and then proposes to perform either entrywise thresholding or tapering of the

sample covariance matrix [3,6,8,13,28,29]. Note that the sparsity notion adopted in this approach is not adapted to strongly correlated datasets with dense covariance matrix.

In random matrix theory, [14,16,17] and the references cited therein investigate the asymptotic distribution of the sample covariance matrix eigenvalues for different settings of n and p. See also [33] for a very nice survey of existing non-asymptotic results on the spectral norm deviation of the sample covariance matrix from its population counterpart. In this paper, we adopt this approach and we will provide further details as we present our results.

Note that the results derived in the works cited above do not cover datasets with missing observations. For instance, when the data contains no missing observation ($\delta = 1$), [33] established a non-asymptotic control on the stochastic deviation $\|\Sigma_n - \Sigma\|_\infty$ of the empirical covariance matrix $\Sigma_n = \frac{1}{n}\sum_{i=1}^n X_i \otimes X_i$ provided some tail conditions are satisfied by the common distribution of $X_1, \dots, X_n$. Exploiting these results, it is possible to establish oracle inequalities for the covariance version of the matrix Lasso estimator

$$\hat{\Sigma}^L = \operatorname*{argmin}_{S \in \mathcal{S}_p} \|\Sigma_n - S\|_2^2 + \lambda \|S\|_1, \tag{1.2}$$

where $\mathcal{S}_p$ is the set of $p \times p$ positive-semidefinite symmetric matrices, $\|S\|_2$ and $\|S\|_1$ are respectively the Frobenius and nuclear norm of S and $\lambda > 0$ is a regularization parameter that should be chosen of the order of magnitude of $\|\Sigma_n - \Sigma\|_\infty$ (note here that $\|S\|_1 = \operatorname{tr}(S)$ for any $S \in \mathcal{S}_p$). This estimator is the covariance version of the matrix Lasso estimator initially introduced in the matrix regression framework, see [23,27] and the references cited therein. To the best of our knowledge, the procedure (1.2) has not been studied in the covariance estimation problem.

When the data contains missing observations ($\delta < 1$), we no longer have access to Σ_n. Given the observations $Y_1, \dots, Y_n$, we can build the following empirical covariance matrix

$$\Sigma_n^{(\delta)} = \frac{1}{n}\sum_{i=1}^n Y_i \otimes Y_i.$$

In this case, a naive approach to derive oracle inequalities consists in computing the matrix Lasso estimator (1.2) with Σ_n replaced by $\Sigma_n^{(\delta)}$. Unfortunately this approach is bound to fail since $\Sigma_n^{(\delta)}$ is not a good estimator of Σ when $\delta < 1$. Indeed, some elementary algebra gives that $\mathbb{E}(\Sigma_n^{(\delta)}) = \Sigma^{(\delta)}$ with

$$\Sigma^{(\delta)} = \left(\delta - \delta^2\right) \operatorname{diag}(\Sigma) + \delta^2 \Sigma,$$

where $\operatorname{diag}(\Sigma)$ is the $p \times p$ diagonal matrix obtained by putting all the non-diagonal entries of Σ to zero (see Section 5.8 below for the details of the computation). When $\delta = 1$, we see that $\Sigma^{(1)} = \Sigma$ and $\Sigma_n^{(1)} = \Sigma_n$. However, when observations are missing ($\delta < 1$), $\Sigma^{(\delta)}$ can be very far from Σ. Hence, $\Sigma_n^{(\delta)}$ will be a poor estimator of Σ since it concentrates around its mean $\Sigma^{(\delta)}$ under suitable tail conditions on the distribution of X. Consequently, the stochastic deviation $\|\Sigma_n^{(\delta)} - \Sigma\|_\infty$ will be too large and the matrix Lasso estimator (1.2) with Σ_n replaced by $\Sigma_n^{(\delta)}$, which requires λ to be of the order of magnitude of $\|\Sigma_n^{(\delta)} - \Sigma\|_\infty$, will perform poorly since its rate of estimation grows with λ.

We present now our reconstruction procedure based on the following simple observation

$$\Sigma = \left(\delta^{-1} - \delta^{-2}\right) \operatorname{diag}\left(\Sigma^{(\delta)}\right) + \delta^{-2}\Sigma^{(\delta)} \qquad \forall 0 < \delta \leq 1. \tag{1.3}$$

Therefore, we can define the following unbiased estimator of Σ when the data set contains missing observations

$$\tilde{\Sigma}_n = \left(\delta^{-1} - \delta^{-2}\right) \operatorname{diag}\left(\Sigma_n^{(\delta)}\right) + \delta^{-2}\Sigma_n^{(\delta)}. \tag{1.4}$$

Our estimator is then solution of the following penalized empirical risk minimization problem:

$$\hat{\Sigma}^{\lambda} = \operatorname*{argmin}_{S \in \mathcal{S}_p} \|\tilde{\Sigma}_n - S\|_2^2 + \lambda \|S\|_1, \tag{1.5}$$

where $\lambda > 0$ is a regularization parameter to be tuned properly. We note that this simple procedure can be computed efficiently in high-dimension since $\hat{\Sigma}^{\lambda}$ is solution of a convex minimization problem. The optimal choice of the tuning parameter λ is of the order of magnitude of the stochastic deviation $\|\tilde{\Sigma}_n - \Sigma\|_{\infty}$. Therefore, in order to order to establish sharp oracle inequalities for (1.5), we need to first study the deviations of $\|\tilde{\Sigma}_n - \Sigma\|_{\infty}$. This analysis is more difficult as compared to the study of $\|\Sigma_n - \Sigma\|_{\infty}$ since we need to derive the sharp scaling of $\|\tilde{\Sigma}_n - \Sigma\|_{\infty}$ with δ.

The rest of the paper is organized as follows. In Section 2, we recall some tools and definitions. In Section 3, we establish oracle inequalities for the Frobenius and spectral norms for our procedure (1.5) and also propose a data-driven choice of the regularization parameter. In Section 4, we establish minimax lower bounds for data with missing observations $\delta \in (0, 1]$, thus showing that our procedures are minimax optimal up to a logarithmic factor. Finally, Section 5 contains all the proofs of the paper.

We emphasize that the results of this paper are non-asymptotic in nature, hold true for any setting of n, p, δ, are minimax optimal (up to a logarithmic factor) and do not require the unknown covariance matrix Σ to be low-rank. We note also that to the best of our knowledge, there exists in the literature no minimax lower bound result for statistical estimation problems with missing observations.

2. Tools and definitions

2.1. Sub-exponential random vectors

We recall now the definition and some basic properties of sub-exponential random vectors.

Definition 1. *The ψ_α-norms of a real-valued random variable V are defined by*

$$\|V\|_{\psi_\alpha} = \inf\left\{u > 0 : \mathbb{E}\exp\left(|V|^\alpha / u^\alpha\right) \leq 2\right\}, \qquad \alpha \geq 1.$$

We say that a random variable V with values in $\mathbb{R}$ is sub-exponential if $\|V\|_{\psi_\alpha} < \infty$ for some $\alpha \geq 1$. If $\alpha = 2$, we say that V is sub-Gaussian.

We recall some well-known properties of sub-exponential random variables:

1. For any real-valued random variable V such that $\|V\|_\alpha < \infty$ for some $\alpha \geq 1$, we have

$$\mathbb{E}\big[|V|^m\big] \leq 2\frac{m}{\alpha}\Gamma\left(\frac{m}{\alpha}\right)\|V\|_{\psi_\alpha}^m \qquad \forall m \geq 1, \tag{2.1}$$

   where $\Gamma(\cdot)$ is the Gamma function.
2. If a real-valued random variable V is sub-Gaussian, then V^2 is sub-exponential. Indeed, we have

$$\big\|V^2\big\|_{\psi_1} \leq 2\|V\|_{\psi_2}^2. \tag{2.2}$$

Definition 2. *A random vector $X \in \mathbb{R}^p$ is sub-exponential if $\langle X, x\rangle$ are sub-exponential random variables for all $x \in \mathbb{R}^p$. The ψ_α-norms of a random vector X are defined by*

$$\|X\|_{\psi_\alpha} = \sup_{x\in\mathbb{R}^p:|x|_2=1} \big\|\langle X, x\rangle\big\|_{\psi_\alpha}, \qquad \alpha \geq 1.$$

We recall the Bernstein inequality for sub-exponential real-valued random variables (see, e.g., Corollary 5.17 in [33]).

Proposition 1. *Let $Y_1, \ldots, Y_n$ be independent centered sub-exponential random variables, and $K = \max_i \|Y_i\|_{\psi_1}$. Then for every $t \geq 0$, we have with probability at least $1 - \mathrm{e}^{-t}$*

$$\left|\frac{1}{n}\sum_{i=1}^n Y_i\right| \leq CK\left(\sqrt{\frac{t}{n}} \vee \frac{t}{n}\right),$$

where $C > 0$ is an absolute constant.

2.2. Some elements of matrix theory

Denote by $\mathcal{S}_p$ the set of $p \times p$ symmetric positive-semidefinite matrices. Any matrix $A \in \mathcal{S}_p$ admits the following spectral representation

$$A = \sum_{j=1}^r \sigma_j(A) u_j(A) \otimes u_j(A),$$

where $r = \mathrm{rank}(A)$ is the rank of A, $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_r(A) > 0$ are the non-zero eigenvalues of A and $u_1(A), \ldots, u_r(A) \in \mathbb{R}^p$ are the associated orthonormal eigenvectors (we also set $\sigma_{r+1}(A) = \cdots = \sigma_p(A) = 0$). The linear vector space L is the linear span of $\{u_1(A), \ldots, u_r(A)\}$ and is called support of A. We will denote respectively by P_L and $P_L^\perp$ the orthogonal projections onto L and $L^\perp$.

The Schatten q-norm of $A \in \mathcal{S}_p$ is defined by

$$\|A\|_q = \left(\sum_{j=1}^{p} |\sigma_j(A)|^q \right)^{1/q} \qquad \text{for } 1 \le q < \infty, \quad \text{and} \quad \|A\|_\infty = \sigma_1(A).$$

Note that the trace of any $S \in \mathcal{S}_p$ satisfies $\operatorname{tr}(S) = \|S\|_1$.

Recall the *trace duality* property:

$$\left|\operatorname{tr}\left(A^\top B\right)\right| \le \|A\|_1 \|B\|_\infty \qquad \forall A, B \in \mathbb{R}^{p \times p}.$$

We will also use the fact that the subdifferential of the convex function $A \mapsto \|A\|_1$ is the following set of matrices:

$$\partial \|A\|_1 = \left\{ \sum_{j=1}^{r} u_j(A) \otimes u_j(A) + P_L^\perp W P_L^\perp : \|W\|_\infty \le 1 \right\} \tag{2.3}$$

(cf. [34]).

We introduce now the notion of intrinsic dimension of a symmetric matrix. The intrinsic dimension of the matrix Σ can be measured through the effective rank

$$\mathbf{r}(\Sigma) := \frac{\operatorname{tr}(\Sigma)}{\|\Sigma\|_\infty}, \tag{2.4}$$

see Section 5.4.3 in [33]. Note that we always have $\mathbf{r}(\Sigma) \le \operatorname{rank}(\Sigma)$. In addition, we can possibly have $\mathbf{r}(\Sigma) \ll \operatorname{rank}(\Sigma)$ for approximately low-rank matrices Σ, that is matrices Σ with large rank but concentrated around a low-dimensional subspace. Consider for instance the covariance matrix Σ with eigenvalues $\sigma_1 = 1$ and $\sigma_2 = \cdots = \sigma_p = 1/p$, then $\mathbf{r}(\Sigma) = \frac{2p-1}{p} \ll p = \operatorname{rank}(\Sigma)$.

The following proposition is the matrix version of Bernstein's inequality for bounded random matrices [1] (see also Corollary 9.1 in [31]).

Proposition 2. *Let $Z_1, \dots, Z_n$ be symmetric independent random matrices in $\mathbb{R}^{p \times p}$ that satisfy $\mathbb{E}[Z_i] = 0$ and $\|Z_i\|_\infty \le U$ almost surely for some constant U and all $i = 1, \dots, n$. Define*

$$\sigma_Z = \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[Z_i^2\right] \right\|_\infty.$$

Then, for all $t > 0$, with probability at least $1 - \mathrm{e}^{-t}$ we have

$$\left\| \frac{Z_1 + \cdots + Z_n}{n} \right\|_\infty \le 2 \max\left\{ \sigma_Z \sqrt{\frac{t + \log(2p)}{n}}, U \frac{t + \log(2p)}{n} \right\}.$$

We also recall now that the l_q-norms of a vector $x = (x^{(1)}, \dots, x^{(p)})^\top \in \mathbb{R}^p$ is given by

$$|x|_q = \left(\sum_{j=1}^{p} \left|x^{(j)}\right|^q \right)^{1/q} \qquad \text{for } 1 \le q < \infty, \quad \text{and} \quad |x|_\infty = \max_{1 \le j \le p} \left|x^{(j)}\right|.$$

3. Oracle inequalities

We can now state the main result for the procedure (1.5).

Theorem 1. *Let $X_1, \dots, X_n$ be i.i.d. vectors in $\mathbb{R}^p$ with covariance matrix Σ. For any $p \geq 2, n \geq 1$, we have on the event $\lambda \geq 2\|\tilde{\Sigma}_n - \Sigma\|_\infty$*

$$\left\|\hat{\Sigma}^\lambda - \Sigma\right\|_2^2 \leq \inf_{S \in \mathcal{S}_p} \left\{ \|S - \Sigma\|_2^2 + \min\left\{ 2\lambda \|S\|_1, \frac{(1+\sqrt{2})^2}{8} \lambda^2 \operatorname{rank}(S) \right\} \right\}, \tag{3.1}$$

and

$$\left\|\hat{\Sigma}^\lambda - \Sigma\right\|_\infty \leq \lambda. \tag{3.2}$$

As we see in Theorem 1, the regularization parameter λ should be chosen sufficiently large such that the condition $\lambda \geq 2\|\tilde{\Sigma}_n - \Sigma\|_\infty$ holds with probability close to 1. The optimal choice of λ depends on the unknown distribution of the observations. We consider now the case of sub-Gaussian random vector $X \in \mathbb{R}^p$.

Assumption 1 (Sub-Gaussian observations). *The random vector $X \in \mathbb{R}^p$ is sub-Gaussian, that is $\|X\|_{\psi_2} < \infty$. In addition, there exist a numerical constant $c_1 > 0$ such that*

$$\mathbb{E}\big[\langle X, u\rangle^2\big] \geq c_1 \left\|\langle X, u\rangle\right\|_{\psi_2}^2 \qquad \forall u \in \mathbb{R}^p. \tag{3.3}$$

Note that Gaussian distributions satisfy Assumption 1. Under the above condition, we can study the stochastic quantity $\|\tilde{\Sigma}_n - \Sigma\|_\infty$ and thus properly tune the regularization parameter λ.

We have the following result, which requires no condition on the covariance matrix Σ.

Proposition 3. *Let $X_1, \dots, X_n \in \mathbb{R}^p$ be i.i.d. random vectors satisfying Assumption 1. Let $Y_1, \dots, Y_n$ be defined in (1.1) with $\delta \in (0, 1]$. Then, for any $t > 1 \vee \log n$, we have with probability at least $1 - \mathrm{e}^{-t}$*

$$\begin{aligned} \|\tilde{\Sigma}_n - \Sigma\|_\infty \leq C \frac{\|\Sigma\|_\infty}{c_1} \\ \times \max\left\{ \sqrt{\frac{\mathbf{r}(\Sigma)(t + \log(2p))}{\delta^2 n}}, \frac{\mathbf{r}(\Sigma)(t + \log(2p))}{\delta^2 n} (c_1 \delta + t + \log n) \right\}, \end{aligned} \tag{3.4}$$

and

$$\left|\operatorname{tr}(\tilde{\Sigma}_n) - \operatorname{tr}(\Sigma)\right| \leq C \frac{\operatorname{tr}(\Sigma)}{c_1 \delta} \max\left\{ \sqrt{\frac{t}{n}}, \frac{t}{n} \right\}, \tag{3.5}$$

where $C > 0$ is an absolute constant.

1. The natural choice for t is of the order of magnitude $\log(2p)$. Then the conclusions of Proposition 3 hold true with probability at least $1-\frac{1}{2p}$. In addition, if the number of measurements n is sufficiently large

$$n \geq c\frac{\mathbf{r}(\Sigma)}{\delta^2}\log^2\big((2p)\vee n\big), \tag{3.6}$$

where $c>0$ is a sufficiently large numerical constant, then an acceptable choice for the regularization parameter λ is

$$\lambda = C\frac{\|\Sigma\|_\infty}{c_1}\sqrt{\frac{\mathbf{r}(\Sigma)\log(2p)}{\delta^2 n}}, \tag{3.7}$$

where the absolute constant $C>0$ is sufficiently large.

2. Proposition 3 and Equation (3.7) give some insight on the tuning of the regularization parameter:

$$\lambda = C\frac{\sqrt{\mathrm{tr}(\Sigma)\|\Sigma\|_\infty}}{c_1\delta}\sqrt{\frac{\log(2p)}{n}},$$

where $C>0$ is a sufficiently large absolute constant. We see that this choice of λ depends on $\mathrm{tr}(\Sigma)$ and $\|\Sigma\|_\infty$ which are typically unknown. Therefore, we propose to use instead

$$\lambda = C\frac{\sqrt{\mathrm{tr}(\tilde{\Sigma}_n)\|\tilde{\Sigma}_n\|_\infty}}{\delta}\sqrt{\frac{\log 2p}{n}}, \tag{3.8}$$

where $C>0$ is a large enough constant. Note that the above choice of λ does not depend on the unknown quantities $\|\Sigma\|_\infty$ or $\mathrm{tr}(\Sigma)$ and constitutes thus an interesting choice in practice. We prove in the next lemma that $2\|\tilde{\Sigma}_n-\Sigma\|_\infty\leq\lambda$ with probability at least $1-\frac{1}{2p}$.

3. As we claimed in the introduction, Proposition 3 requires no condition on Σ whatsoever. However, for the result to be of any practical interest, we need the bound in (3.4) to be small, which is the case if the condition (3.6) is satisfied. This condition is interesting since it shows that the number of measurements sufficient to guarantee a precise enough estimation of the spectrum of Σ grows with the effective rank $\mathbf{r}(\Sigma)$. In particular, when no observation is missing ($\delta=1$), if Σ is approximately low-rank so that $\mathbf{r}(\Sigma)\ll p$, then only $n=\mathcal{O}(\mathbf{r}(\Sigma)\log^2(2p))$ measurements are sufficient to estimate precisely the spectrum of the $p\times p$ covariance matrix Σ.

4. Note that if we assume that $\|Y\otimes Y\|_\infty=|Y|_2^2\leq U$ a.s. for some constant $U>0$, then we can eliminate the $(c_1\delta+t+\log n)$ factor in (3.4). Consequently, we can replace the condition (3.6) on the number of measurements by the following less restrictive one

$$n\geq c\frac{\mathbf{r}(\Sigma)}{\delta^2}\log(2p)$$

for some absolute constant $c>0$ sufficiently large. When there is no missing observation ($\delta=1$), we obtain the standard condition on the number of measurements (see Remark 5.53

in [33]). When some observations are missing ($\delta < 1$), we have the additional quantity δ^2 in the denominators of (3.4) and (3.6). The bound (3.4) is degraded in the case $\delta < 1$ since we observe less entries per measurement. Consequently, as we can see it in (3.6), if we denote by $N(\varepsilon)$ the number of necessary measurements to estimate Σ with a precision ε when no observation is missing ($\delta = 1$), then we will need at least $\mathcal{O}(N(\varepsilon)/\delta^2)$ measurements in order to estimate Σ with the same precision ε when some observations are missing ($\delta < 1$). In Theorem 2, we prove in particular that the dependence of the bound (3.4) on δ is sharp by establishing a minimax lower bound.

5. In the full observations case ($\delta = 1$) and for sub-Gaussian distributions with low rank covariance matrix Σ, a simple modification of the ε-net argument used in [33] to prove Theorem 5.39 yields an inequality similar to (3.4) with an upper bound of the order $\|\Sigma\|_\infty \sqrt{\frac{\operatorname{rank}(\Sigma)+t}{n}}$ without logarithmic factor $\log 2p$. Note however that this bound is sub-optimal when $\mathbf{r}(\Sigma)\log^2((2p)\vee n) \ll \operatorname{rank}(\Sigma)$ (see the discussion on the intrinsic dimension of a matrix in Section 2.2). In addition, in the missing observations framework $\delta < 1$, the matrix $\Sigma^{(\delta)}$ can have full rank even if the matrix Σ is low rank. Therefore, the ε-net argument will yield an upper bound of the order $\|\Sigma\|_\infty \sqrt{\frac{p+t}{\delta^2 n}}$ which is much larger than the bound derived in (3.4).

Lemma 1. *Let the assumptions of Proposition 3 be satisfied. Assume in addition that (3.6) holds true. Take λ as in (3.8) with $C > 0$ a large enough constant that can depend only on c_1. Then, we have with probability at least $1 - \frac{1}{2p}$ that*

$$2\|\tilde{\Sigma}_n - \Sigma\|_\infty \le \lambda \le C'\|\Sigma\|_\infty \sqrt{\frac{\mathbf{r}(\Sigma)\log(2p)}{\delta^2 n}},$$

where $C' > 0$ can depend only on c_1.

We obtain the following corollary of Theorem 1.

Corollary 1. *Let Assumption 1 and condition (3.6) be satisfied. Consider the estimator (1.5) with the regularization parameter λ satisfying (3.8). Then we have, with probability at least $1 - \frac{1}{2p}$ that*

$$\left\|\hat{\Sigma}^\lambda - \Sigma\right\|_2^2 \le \inf_{S\in\mathcal{S}_p}\left\{\|\Sigma - S\|_2^2 + C_1\|\Sigma\|_\infty^2 \frac{\mathbf{r}(\Sigma)\log 2p}{\delta^2 n}\operatorname{rank}(S)\right\}, \tag{3.9}$$

and

$$\left\|\hat{\Sigma}^\lambda - \Sigma\right\|_\infty \le C_2\|\Sigma\|_\infty \sqrt{\frac{\mathbf{r}(\Sigma)\log 2p}{\delta^2 n}}, \tag{3.10}$$

where $C_1, C_2 > 0$ can depend only on c_1.

The proof of this corollary is immediate by combining Theorem 1 with Proposition 3 and Lemma 1 and up to a rescaling of the constants.

4. Lower bounds

For any integer $1 \le r \le p$, define

$$\mathcal{C}_r = \big\{S \in \mathcal{S}_p : \mathbf{r}(S) \le r\big\}.$$

We also introduce $\mathcal{P}_r$ the class of probability distributions on $\mathbb{R}^p$ with covariance matrix $\Sigma \in \mathcal{C}_r$.

We now establish a minimax lower bound that guarantees the rates we obtained in Corollary 1 are optimal up to a logarithmic factor on the probability distribution class $\mathcal{P}_r$. In particular, the dependence of our rates on δ, $\|\Sigma\|_\infty$ and $\mathbf{r}(\Sigma)$ is sharp.

Theorem 2. *Fix $\delta \in (0,1]$. Let $n, r \ge 1$ be integers such that $n \ge \delta^{-2} r^2$. Let $X_1, \dots, X_n$ be i.i.d. random vectors in $\mathbb{R}^p$ with covariance matrix $\Sigma \in \mathcal{C}_r$. We observe n i.i.d. random vectors $Y_1, \dots, Y_n \in \mathbb{R}^p$ such that*

$$Y_i^j = \delta_{ij} X_i^{(j)}, \qquad 1 \le i \le n, 1 \le j \le p,$$

where $(\delta_{ij})_{1 \le i \le n, 1 \le j \le p}$ is an i.i.d. sequence of Bernoulli $B(\delta)$ random variables independent of $X_1, \dots, X_n$.

Then, there exist absolute constants $\beta \in (0,1)$ and $c > 0$ such that

$$\inf_{\hat{\Sigma}} \sup_{\mathbb{P}_\Sigma \in \mathcal{P}_r} \mathbb{P}_\Sigma \left(\|\hat{\Sigma} - \Sigma\|_2^2 > c \|\Sigma\|_\infty^2 \frac{\mathbf{r}(\Sigma)}{\delta^2 n} \operatorname{rank}(\Sigma) \right) \ge \beta, \qquad \textit{below Assumption 1} \tag{4.1}$$

and

$$\inf_{\hat{\Sigma}} \sup_{\mathbb{P}_\Sigma \in \mathcal{P}_r} \mathbb{P}_\Sigma \left(\|\hat{\Sigma} - \Sigma\|_\infty > c \|\Sigma\|_\infty \sqrt{\frac{\mathbf{r}(\Sigma)}{\delta^2 n}} \right) \ge \beta, \tag{4.2}$$

where $\inf_{\hat{\Sigma}}$ denotes the infimum over all possible estimators $\hat{\Sigma}$ of Σ based on $Y_1, \dots, Y_n$.

5. Proofs

5.1. Proof of Theorem 1

The proof of the first inequality adapts to covariance matrix estimation the arguments used in the trace regression problem to prove Theorems 1 and 11 in [23].

Proof of Theorem 1. By definition of $\hat{\Sigma}^\lambda$, we have for any $S \in \mathcal{S}_p$

$$\big\|\hat{\Sigma}^\lambda - \Sigma\big\|_2^2 \le \|S - \Sigma\|_2^2 + \lambda \|S\|_1 + 2\big\langle \Sigma - \tilde{\Sigma}_n, S - \hat{\Sigma}^\lambda \big\rangle - \lambda \big\|\hat{\Sigma}^\lambda\big\|_1.$$

If $\lambda \ge 2\|\tilde{\Sigma}_n - \Sigma\|_\infty$, we deduce from the previous display that

$$\big\|\hat{\Sigma}^\lambda - \Sigma\big\|_2^2 \le \|S - \Sigma\|_2^2 + 2\lambda \|S\|_1 \qquad \forall S \in \mathcal{S}_p.$$

Next, a necessary and sufficient condition of minimum for problem (1.5) implies that there exists $\hat{V} \in \partial\|\hat{\Sigma}^\lambda\|_1$ such that for all $S \in \mathcal{S}_p$

$$-2\langle \tilde{\Sigma}_n - \hat{\Sigma}^\lambda, \hat{\Sigma}^\lambda - S\rangle + \lambda\langle \hat{V}, \hat{\Sigma}^\lambda - S\rangle \le 0. \tag{5.1}$$

For any $S \in \mathcal{S}_p$ of rank r with spectral representation $S = \sum_{j=1}^r \sigma_j u_j \otimes u_j$ and support L. It follows from (5.1) that

$$2\langle \hat{\Sigma}^\lambda - \Sigma, \hat{\Sigma}^\lambda - S\rangle + \lambda\langle \hat{V} - V, \hat{\Sigma}^\lambda - S\rangle \le -\lambda\langle V, \hat{\Sigma}^\lambda - S\rangle + 2\langle \tilde{\Sigma}_n - \Sigma, \hat{\Sigma}^\lambda - S\rangle \tag{5.2}$$

for an arbitrary $V \in \partial\|S\|_1$. Note that $\langle \hat{V} - V, \hat{\Sigma}^\lambda - S\rangle \ge 0$ by monotonicity of subdifferentials of convex functions and that the following representation holds

$$V = \sum_{j=1}^r u_j \otimes u_j + P_L^\perp W P_L^\perp,$$

where W is an arbitrary matrix with $\|W\|_\infty \le 1$. In particular, there exists W with $\|W\|_\infty \le 1$ such that

$$\langle P_L^\perp W P_L^\perp, \hat{\Sigma}^\lambda - S\rangle = \|P_L^\perp \hat{\Sigma}^\lambda P_L^\perp\|_1.$$

For this choice of W, we get from (5.2) that

$$\begin{aligned} &\|\hat{\Sigma}^\lambda - \Sigma\|_2^2 + \|\hat{\Sigma}^\lambda - S\|_2^2 + \lambda\|P_L^\perp \hat{\Sigma}^\lambda P_L^\perp\|_1 \\ &\quad \le \|S - \Sigma\|_2^2 + \lambda\|P_L(\hat{\Sigma}^\lambda - S)P_L\|_1 + 2\langle \tilde{\Sigma}_n - \Sigma, \hat{\Sigma}^\lambda - S\rangle, \end{aligned} \tag{5.3}$$

where we have used the following facts

$$2\langle \hat{\Sigma}^\lambda - \Sigma, \hat{\Sigma}^\lambda - S\rangle = \|\hat{\Sigma}^\lambda - \Sigma\|_2^2 + \|\hat{\Sigma}^\lambda - S\|_2^2 - \|S - \Sigma\|_2^2,$$

and

$$\left\|\sum_{j=1}^r u_j \otimes u_j\right\|_\infty = 1, \qquad \left\langle \sum_{j=1}^r u_j \otimes u_j, \hat{\Sigma}^\lambda - S\right\rangle = \left\langle \sum_{j=1}^r u_j \otimes u_j, P_L(\hat{\Sigma}^\lambda - S)P_L\right\rangle.$$

For any $A \in \mathbb{R}^{p\times p}$ define $\mathcal{P}_L(A) = A - P_L^\perp A P_L^\perp$. Set $\Delta_1 = \tilde{\Sigma}_n - \Sigma$. We have

$$\langle \Delta_1, \hat{\Sigma}^\lambda - S\rangle = \langle \Delta_1, \mathcal{P}_L(\hat{\Sigma}^\lambda - S)\rangle + \langle \Delta_1, P_L^\perp(\hat{\Sigma}^\lambda - S)P_L^\perp\rangle.$$

Using Cauchy–Schwarz's inequality and trace duality, we get

$$\begin{aligned} |\langle \Delta_1, \mathcal{P}_L(\hat{\Sigma}^\lambda - S)\rangle| &\le \sqrt{2\operatorname{rank}(S)}\|\Delta_1\|_\infty \|\hat{\Sigma}^\lambda - S\|_2, \\ \|P_L(\hat{\Sigma}^\lambda - S)P_L\|_1 &\le \sqrt{\operatorname{rank}(S)}\|\hat{\Sigma}^\lambda - S\|_2, \\ |\langle \Delta_1, P_L^\perp(\hat{\Sigma}^\lambda - S)P_L^\perp\rangle| &\le \|\Delta_1\|_\infty \|P_L^\perp \hat{\Sigma}^\lambda P_L^\perp\|_1. \end{aligned}$$

The above display combined with (5.3) give

$$\begin{aligned}&\big\|\hat\Sigma^\lambda-\Sigma\big\|_2^2+\big\|\hat\Sigma^\lambda-S\big\|_2^2+\big(\lambda-2\|\Delta_1\|_\infty\big)\big\|P_L^\perp\hat\Sigma^\lambda P_L^\perp\big\|_1\\&\quad\le\|S-\Sigma\|_2^2+\big(\sqrt{2}\|\Delta_1\|_\infty+\lambda\big)\sqrt{r}\big\|\hat\Sigma^\lambda-S\big\|_2.\end{aligned}$$

A decoupling argument gives

$$\begin{aligned}&\big\|\hat\Sigma^\lambda-\Sigma\big\|_2^2+\big\|\hat\Sigma^\lambda-S\big\|_2^2+\big(\lambda-2\|\Delta_1\|_\infty\big)\big\|P_L^\perp\hat\Sigma^\lambda P_L^\perp\big\|_1\\&\quad\le\|S-\Sigma\|_2^2+\left(\frac{1}{\sqrt{2}}\|\Delta_1\|_\infty+\frac{\lambda}{2}\right)^2 r+\big\|\hat\Sigma^\lambda-S\big\|_2^2.\end{aligned}$$

Finally, we get on the event $\lambda\ge 2\|\Delta_1\|_\infty$ that

$$\big\|\hat\Sigma^\lambda-\Sigma\big\|_2^2\le\|S-\Sigma\|_2^2+\frac{(1+\sqrt{2})^2}{8}\lambda^2\operatorname{rank}(S)\qquad\forall S\in\mathcal{S}_p.$$

We now prove the spectral norm bound. Note first that the solution of (1.5) is given by

$$\hat\Sigma^\lambda=\sum_j\left(\sigma_j(\tilde\Sigma_n)-\frac{\lambda}{2}\right)_+u_j(\tilde\Sigma_n)\otimes u_j(\tilde\Sigma_n),\tag{5.4}$$

where $x_+=\max\{0,x\}$ and $\tilde\Sigma_n$ admits the spectral representation

$$\tilde\Sigma_n=\sum_j\sigma_j(\tilde\Sigma_n)u_j(\tilde\Sigma_n)\otimes u_j(\tilde\Sigma_n),$$

with positive eigenvalues $\sigma_j(\tilde\Sigma_n)\ge 0$ and orthonormal eigenvectors $u_j(\tilde\Sigma_n)$. Indeed, the solution of (1.5) is unique since the functional $S\to F(S)=\|\tilde\Sigma_n-S\|_2^2+\lambda\|S\|_1$ is strictly convex. A sufficient condition of minimum is $\mathbf{0}\in\partial F(\hat\Sigma^\lambda)=-2(\tilde\Sigma_n-\hat\Sigma^\lambda)+\lambda\hat V$ with $\hat V\in\partial\|\hat\Sigma^\lambda\|_1$. We consider the following choice of $\hat V=\sum_{j:\sigma_j(\tilde\Sigma_n)\ge\lambda/2}u_j(\tilde\Sigma_n)\otimes u_j(\tilde\Sigma_n)+W\in\partial\|\hat\Sigma^\lambda\|_1$ with

$$W=\sum_{j:\sigma_j(\tilde\Sigma_n)<\lambda/2}\frac{2\sigma_j(\tilde\Sigma_n)}{\lambda}u_j(\tilde\Sigma_n)\otimes u_j(\tilde\Sigma_n).$$

It is easy to check that $\partial F(\hat\Sigma^\lambda)=-2(\tilde\Sigma_n-\hat\Sigma^\lambda)+\lambda\hat V=\mathbf{0}$.

Next, we have on the event $\lambda\ge 2\|\Delta_1\|_\infty$

$$\big\|\hat\Sigma^\lambda-\Sigma\big\|_\infty\le\big\|\hat\Sigma^\lambda-\tilde\Sigma_n\big\|_\infty+\|\Delta_1\|_\infty\le\lambda.\qquad\square$$

5.2. Proof of Proposition 3

The delicate part of this proof is to obtain the sharp dependence on δ. As a consequence, the proof is significantly more technical as compared to the case of full observations $\delta=1$. To simplify the understanding of this proof, we decomposed it into three lemmas that we prove below.

Proof of Proposition 3. We start with (3.5). For any $t > 0$, Lemma 2 gives, with probability at least $1 - \mathrm{e}^{-t}$ that

$$\left|\operatorname{tr}\left(\Sigma_n^{(\delta)}\right) - \delta \operatorname{tr}(\Sigma)\right| \leq C c_1^{-1} \operatorname{tr}(\Sigma) \max\left\{\sqrt{\frac{t}{n}}, \frac{t}{n}\right\} \tag{5.5}$$

for some numerical constant $C > 0$. Noting that $\operatorname{tr}(\tilde{\Sigma}_n) = \delta^{-1} \operatorname{tr}(\Sigma_n^{(\delta)})$, we can deduce (3.5) immediately from (5.5).

We now treat (3.4). For any $t \geq 0$, Lemma 3 gives with probability at least $1 - \mathrm{e}^{-t}$ that

$$\left\|\operatorname{diag}\left(\Sigma_n^{(\delta)} - \Sigma^{(\delta)}\right)\right\|_\infty \leq t_1, \tag{5.6}$$

with

$$t_1 = C c_1^{-1} \max_{1 \leq j \leq p} (\Sigma_{jj}) \max\left\{\sqrt{\frac{t}{n}}, \frac{t}{n}\right\},$$

where $C > 0$ is some absolute constant.

Define

$$A_n^{(\delta)} = \Sigma_n^{(\delta)} - \operatorname{diag}\left(\Sigma_n^{(\delta)}\right), \qquad A^{(\delta)} = \Sigma^{(\delta)} - \operatorname{diag}\left(\Sigma^{(\delta)}\right).$$

For any $t \geq 1 \vee \log n$, Lemma 4 gives with probability at least $1 - \mathrm{e}^{-t}$ that

$$\left\|A_n^{(\delta)} - A^{(\delta)}\right\|_\infty \leq t_2, \tag{5.7}$$

with

$$t_2 = \frac{C}{c_1} \delta \|\Sigma\|_\infty \max\left\{\sqrt{\frac{\mathbf{r}(\Sigma)(t + \log(2p))}{n}}, \mathbf{r}(\Sigma)(c_1\delta + t + \log n)\frac{t + \log(2p)}{\delta n}\right\},$$

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

Set now

$$\bar{t} = \frac{C'}{c_1} \|\Sigma\|_\infty \max\left\{\sqrt{\frac{\mathbf{r}(\Sigma)(t + \log(2p))}{\delta^2 n}}, \mathbf{r}(\Sigma)(c_1\delta + t + \log n)\frac{t + \log(2p)}{\delta^2 n}\right\},$$

where $C' > 0$ is a large enough numerical constant such that $\bar{t} \geq \frac{t_1}{\delta} + \frac{t_2}{\delta^2}$, where we have used that $\max_{1 \leq j \leq p}(\Sigma_{jj}) \leq \sqrt{\operatorname{tr}(\Sigma)\|\Sigma\|_\infty} \leq \operatorname{tr}(\Sigma)$.

Combining (5.6) and (5.7) with a union bound argument, we get for any $t \geq 1 \vee \log n$, with probability at least $1 - 2\mathrm{e}^{-t}$ that

$$\begin{aligned} \|\tilde{\Sigma}_n - \Sigma\|_\infty &\leq \delta^{-1}\left\|\operatorname{diag}\left(\Sigma_n^{(\delta)} - \Sigma^{(\delta)}\right)\right\|_\infty + \delta^{-2}\left\|A_n^{(\delta)} - A^{(\delta)}\right\|_\infty \\ &\leq \frac{t_1}{\delta} + \frac{t_2}{\delta^2} \leq \bar{t}. \end{aligned} \tag{5.8}$$

A union bound argument again gives that (5.5) and (5.8) hold valid simultaneously with probability at least $1 - 3\mathrm{e}^{-t}$ for any $t \geq 1 \vee \log n$. Up to a rescaling of the constants, we can assume that (5.5) and (5.8) hold valid with probability at least $1 - \mathrm{e}^{-t}$. □

Lemma 2. *Under the assumptions of Proposition 3, we have with probability at least* $1-\mathrm{e}^{-t}$ *that*

$$\bigl|\operatorname{tr}\bigl(\Sigma_n^{(\delta)}\bigr)-\delta\operatorname{tr}(\Sigma)\bigr|\le Cc_1^{-1}\operatorname{tr}(\Sigma)\max\biggl\{\sqrt{\frac{t}{n}},\frac{t}{n}\biggr\},\tag{5.9}$$

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

Proof. In view of Assumption 1, we have for any $1\le j\le p$ that $\|(Y^{(j)})^2\|_{\psi_1}\le\|(X^{(j)})^2\|_{\psi_1}\le 2\|X^{(j)}\|_{\psi_2}^2\le 2c_1^{-1}\Sigma_{jj}$ and

$$\bigl\||Y|_2^2\bigr\|_{\psi_1}\le\sum_{j=1}^p\bigl\|\bigl(Y^{(j)}\bigr)^2\bigr\|_{\psi_1}\le\bigl\|\bigl(X^{(j)}\bigr)^2\bigr\|_{\psi_1}\le 2\sum_{j=1}^p\bigl\|X^{(j)}\bigr\|_{\psi_2}^2\le 2c_1^{-1}\operatorname{tr}(\Sigma).$$

Next, we have

$$\begin{aligned}\operatorname{tr}\bigl(\Sigma_n^{(\delta)}\bigr)-\delta\operatorname{tr}(\Sigma)&=\operatorname{tr}\bigl(\Sigma_n^{(\delta)}-\Sigma^{(\delta)}\bigr)\\&=\operatorname{tr}\biggl(\frac{1}{n}\sum_{i=1}^n Y_i\otimes Y_i-\mathbb{E}[Y\otimes Y]\biggr)\\&=\frac{1}{n}\sum_{i=1}^n\operatorname{tr}(Y_i\otimes Y_i)-\mathbb{E}\bigl[\operatorname{tr}(Y\otimes Y)\bigr]\\&=\frac{1}{n}\sum_{i=1}^n|Y_i|_2^2-\mathbb{E}\bigl[|Y|_2^2\bigr].\end{aligned}$$

Next, we have

$$\bigl\||Y_i|_2^2-\mathbb{E}\bigl[|Y|_2^2\bigr]\bigr\|_{\psi_1}\le c\bigl\||Y_i|_2^2\bigr\|_{\psi_1}\le 2\frac{c}{c_1}\operatorname{tr}(\Sigma)$$

for some numerical constant $c>0$. Then, we can apply Proposition 1 to get the result. □

Lemma 3. *Under the assumptions of Proposition 3, we have with probability at least* $1-\mathrm{e}^{-t}$ *that*

$$\bigl\|\operatorname{diag}\bigl(\Sigma_n^{(\delta)}-\Sigma^{(\delta)}\bigr)\bigr\|_\infty\le Cc_1^{-1}\max_{1\le j\le p}(\Sigma_{jj})\max\biggl\{\sqrt{\frac{t+\log p}{n}},\frac{t+\log p}{n}\biggr\},\tag{5.10}$$

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

Proof. We have

$$\bigl\|\operatorname{diag}\bigl(\Sigma_n^{(\delta)}-\Sigma^{(\delta)}\bigr)\bigr\|_\infty=\max_{1\le j\le p}\biggl|\frac{1}{n}\sum_{i=1}^n\delta_{i,j}^2\bigl(X_i^{(j)}\bigr)^2-\delta\Sigma_{jj}\biggr|.$$

Next, since the random variables $X_i^{(j)}$ are sub-Gaussian for any i, j, we have

$$\left\|\left(\delta_{i,j}X_i^{(j)}\right)^2\right\|_{\psi_1} \le 2\left\|\delta_{i,j}X_i^{(j)}\right\|_{\psi_2}^2 \le 2\left\|X_i^{(j)}\right\|_{\psi_2}^2 \le 2c_1^{-1}\Sigma_{jj},$$

where we have used Assumption 1 in the last inequality. We can apply Proposition 1 to get for any $1 \le j \le p$ with probability at least $1 - \mathrm{e}^{-t'}$ that

$$\left|\frac{1}{n}\sum_{i=1}^n \delta_{i,j}^2\left(X_i^{(j)}\right)^2 - \delta\Sigma_{jj}\right| \le Cc_1^{-1}\Sigma_{jj}\max\left\{\sqrt{\frac{t'}{n}}, \frac{t'}{n}\right\},$$

where $C > 0$ is an absolute constant. Next, taking $t' = t + \log p$ combined with a union bound argument we get the result. □

Lemma 4. *Under the assumptions of Proposition 3, we have for any $t \ge 1 \vee \log n$ with probability at least $1 - \mathrm{e}^{-t}$ that*

$$\begin{aligned}&\left\|A_n^{(\delta)} - A^{(\delta)}\right\|_\infty \\ &\quad\le \frac{C}{c_1}\delta\|\Sigma\|_\infty \max\left\{\sqrt{\frac{\mathbf{r}(\Sigma)(t+\log(2p))}{n}}, \mathbf{r}(\Sigma)(c_1\delta + t + \log n)\frac{t+\log(2p)}{\delta n}\right\},\end{aligned} \tag{5.11}$$

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

Proof. We have

$$A_n^{(\delta)} - A^{(\delta)} = \frac{1}{n}\sum_{i=1}^n Z_i - \mathbb{E}[Z_i],$$

where

$$Z_i = Y_i \otimes Y_i - \operatorname{diag}(Y_i \otimes Y_i), \qquad 1 \le i \le n.$$

Define $Y = (\delta_1 X^{(1)}, \dots, \delta_p X^{(p)})^\top$ where $\delta_1, \dots, \delta_p$ are i.i.d. Bernoulli random variables with parameter δ independent from X and $Z = Y \otimes Y - \operatorname{diag}(Y \otimes Y)$.

Fact 1. *We have that*

$$\mathbb{P}\left(\bigcap_{i=1}^n \left\{|Y_i|_2^2 \le U\right\}\right) \ge 1 - \mathrm{e}^{-t} \qquad \forall t \ge 1, \tag{5.12}$$

where $U = \frac{C}{c_1}\operatorname{tr}(\Sigma)(c_1\delta + t + \log n)$ and $C > 0$ is some numerical constant.

The proof of Fact 1 can be found in Section 5.3 below. Define now the truncated random matrices

$$\tilde{Z}_i = Z_i \mathbb{1}_{|Y_i|_2^2 \le U}, \qquad 1 \le i \le n,$$

where $U > 0$ is given in Fact 1. We have, on the event $\bigcap_{i=1}^n \{|Y_i|_2^2 \le U\}$, that

$$\begin{aligned} A_n^{(\delta)} - A^{(\delta)} &= \frac{1}{n}\sum_{i=1}^n \tilde{Z}_i - \mathbb{E}[Z_i] \\ &= \frac{1}{n}\sum_{i=1}^n \big(\tilde{Z}_i - \mathbb{E}[\tilde{Z}_i]\big) + \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Z_i \mathbb{1}_{|Y_i|_2^2 > U}] \\ &= \frac{1}{n}\sum_{i=1}^n \big(\tilde{Z}_i - \mathbb{E}[\tilde{Z}_i]\big) + \mathbb{E}[Z \mathbb{1}_{|Y|_2^2 > U}]. \end{aligned}$$

Thus we get, on the event $\bigcap_{i=1}^n \{|Y_i|_2^2 \le U\}$, that

$$\big\|A_n^{(\delta)} - A^{(\delta)}\big\|_\infty \le \left\|\frac{1}{n}\sum_{i=1}^n \big(\tilde{Z}_i - \mathbb{E}[\tilde{Z}_i]\big)\right\|_\infty + \big\|\mathbb{E}[Z \mathbb{1}_{|Y|_2^2 > U}]\big\|_\infty. \tag{5.13}$$

We study now $\|\mathbb{E}[Z \mathbb{1}_{|Y|_2^2 > U}]\|_\infty$. Set $\mathbf{S}^{p-1} = \{\theta \in \mathbb{R}^p : |\theta|_2 = 1\}$. We have

$$\begin{aligned} \big\|\mathbb{E}[Z \mathbb{1}_{|Y|_2^2 > U}]\big\|_\infty &= \max_{\theta \in \mathbf{S}^{p-1}} \big\{\mathbb{E}\big[\theta^\top Z \theta \mathbb{1}_{|Y|_2^2 > U}\big]\big\} \\ &\le \sqrt{\max_{\theta \in \mathbf{S}^{p-1}} \mathbb{E}\big[\big(\theta^\top Z \theta\big)^2\big]}\sqrt{\mathbb{P}\big(|Y|_2^2 > U\big)}. \end{aligned}$$

Fact 2. *We have*

$$\max_{\theta \in \mathbf{S}^{p-1}} \mathbb{E}\big[\big(\theta^\top Z \theta\big)^2\big] \le \frac{16}{c_1^2}\delta^2 \|\Sigma\|_\infty^2.$$

The proof of Fact 2 can be found in Section 5.4 below. Next, we note that as a by product of the proof of Fact 1, we also have that

$$\mathbb{P}\big(|Y|_2^2 > U\big) \le \mathrm{e}^{-t}.$$

Combining the last three displays, we get that

$$\big\|\mathbb{E}[Z \mathbb{1}_{|Y|_2^2 > U}]\big\|_\infty \le \frac{4}{c_1}\delta \|\Sigma\|_\infty \mathrm{e}^{-t/2}. \tag{5.14}$$

We now want to apply the non-commutative Bernstein inequality to $\|\frac{1}{n}\sum_{i=1}^n (\tilde{Z}_i - \mathbb{E}[\tilde{Z}_i])\|_\infty$. To this end, we need to study the quantities $\|\mathbb{E}(\tilde{Z} - \mathbb{E}[\tilde{Z}])^2\|_\infty$ and $\|\tilde{Z} - \mathbb{E}[\tilde{Z}]\|_\infty$.

Fact 3. *We have*

$$\big\|\mathbb{E}\big(\tilde{Z}-\mathbb{E}[\tilde{Z}]\big)^2\big\|_\infty \le \frac{C}{c_1^2}\delta^2 \operatorname{tr}(\Sigma)\|\Sigma\|_\infty, \quad \textit{and} \quad \big\|\tilde{Z}-\mathbb{E}[\tilde{Z}]\big\|_\infty \le 2U, \tag{5.15}$$

where $C>0$ is an absolute constant.

The proof of Fact 3 can be found in Section 5.5 below. Combining (5.15) with Proposition 2, we get for any $t>0$

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n\big(\tilde{Z}_i-\mathbb{E}[\tilde{Z}_i]\big)\right\|_\infty \ge t_1\right) \le \mathrm{e}^{-t}, \tag{5.16}$$

where

$$\begin{aligned} t_1 &= \frac{C}{c_1}\max\left\{\delta\sqrt{\|\Sigma\|_\infty \operatorname{tr}(\Sigma)}\sqrt{\frac{t+\log(2p)}{n}}, \operatorname{tr}(\Sigma)(c_1\delta+t+\log n)\frac{t+\log(2p)}{n}\right\} \\ &= \frac{C}{c_1}\delta\|\Sigma\|_\infty \max\left\{\sqrt{\mathbf{r}(\Sigma)\frac{t+\log(2p)}{n}}, \mathbf{r}(\Sigma)(c_1\delta+t+\log n)\frac{t+\log(2p)}{\delta n}\right\} \end{aligned}$$

for some numerical constant $C>0$.

For any $t>1$, we set

$$\eta = \frac{4}{c_1}\delta\|\Sigma\|_\infty \mathrm{e}^{-t/2}, \quad \text{and} \quad t_2 = t_1+\eta.$$

We have for any $t>0$ that

$$\begin{aligned} &\mathbb{P}\big(\big\|A_n^{(\delta)}-A^{(\delta)}\big\|_\infty \ge t_2\big) \\ &\le \mathbb{P}\left(\big\{\big\|A_n^{(\delta)}-A^{(\delta)}\big\|_\infty \ge t_2\big\} \cap \bigcap_{i=1}^n\big\{|Y_i|_2^2 \le U\big\}\right) + \mathbb{P}\left(\bigcup_{i=1}^n\big\{|Y_i|_2^2 > U\big\}\right) \\ &\le \mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n\big(\tilde{Z}_i-\mathbb{E}[\tilde{Z}_i]\big)\right\|_\infty \ge t_1 + \big(\eta - \big\|\mathbb{E}[Z\mathbb{1}_{|Y|_2^2>U}]\big\|_\infty\big) \,\Big|\, \bigcap_{i=1}^n\big\{|Y_i|_2^2 \le U\big\}\right) \\ &\quad + \mathbb{P}\left(\bigcup_{i=1}^n\big\{|Y_i|_2^2 > U\big\}\right) \\ &\le \mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^n\big(\tilde{Z}_i-\mathbb{E}[\tilde{Z}_i]\big)\right\|_\infty \ge t_1\right) + \mathbb{P}\left(\bigcup_{i=1}^n\big\{|Y_i|_2^2 > U\big\}\right) \\ &\le 2\mathrm{e}^{-t}, \end{aligned}$$

where we have used (5.14), Fact 1 and (5.16).

Next, we have for any $t \geq \log n$ that

$$\eta = \frac{4}{c_1}\delta\|\Sigma\|_\infty \mathrm{e}^{-t/2} \leq \frac{4}{c_1}\delta\|\Sigma\|_\infty \frac{1}{\sqrt{n}} \leq \frac{4}{c_1}\delta\|\Sigma\|_\infty \sqrt{\mathbf{r}(\Sigma)\frac{(t+\log(2p))}{n}},$$

since $\mathbf{r}(\Sigma) = \mathrm{tr}(\Sigma)/\|\Sigma\|_\infty \geq 1$. Thus, we have for any $t \geq 1 \vee \log n$ that $t_2 \leq t_3$ with

$$t_3 = \frac{C'}{c_1}\delta\|\Sigma\|_\infty \max\left\{\sqrt{\mathbf{r}(\Sigma)\frac{t+\log(2p)}{n}}, \mathbf{r}(\Sigma)(c_1\delta + t + \log n)\frac{t+\log(2p)}{\delta n}\right\}$$

for some sufficiently large numerical constant $C' > 0$. Up to a rescaling of the constants, we get the result with probability at least $1 - \mathrm{e}^{-t}$. □

5.3. Proof of Fact 1

In view of Assumption 1, we have

$$\big\||Y|_2^2\big\|_{\psi_1} \leq \sum_{j=1}^p \big\|\delta_j\big(X^{(j)}\big)^2\big\|_{\psi_1} \leq 2\sum_{j=1}^p \big\|X^{(j)}\big\|_{\psi_2}^2 \leq 2c_1^{-1}\,\mathrm{tr}(\Sigma).$$

Note also that $\mathbb{E}[|Y|_2^2] = \delta\,\mathrm{tr}(\Sigma)$. We apply Proposition 1 to get for any $1 \leq i \leq n$ and $t' > 0$ that

$$\mathbb{P}\left(|Y_i|_2^2 > \mathrm{tr}(\Sigma)\left(\delta + \frac{C}{c_1}\max\big(\sqrt{t'}, t'\big)\right)\right) \leq \mathrm{e}^{-t'},$$

where $C > 0$ is a numerical constant. Next, combining the above display with a union bound argument, we get for any $t' \geq 1$

$$\mathbb{P}\left(\max_{1\leq i\leq n}|Y_i|_2^2 > \mathrm{tr}(\Sigma)\left(\delta + \frac{C}{c_1}t'\right)\right) \leq n\mathrm{e}^{-t'}.$$

Replacing now $t' = t + \log n$, we get for any $t \geq 1$

$$\mathbb{P}\left(\max_{1\leq i\leq n}|Y_i|_2^2 \leq \mathrm{tr}(\Sigma)\left(\delta + \frac{C}{c_1}(t+\log n)\right)\right) \geq 1 - \mathrm{e}^{-t},$$

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

5.4. Bounding of the moment $\mathbb{E}[(\theta^\top Z\theta)^2]$

Set $\bar{Z} = XX^\top - \mathrm{diag}(XX^\top)$. For any $\theta = (\theta^{(1)}, \ldots, \theta^{(p)})^\top \in \mathbb{R}^p$ and $\delta_1, \ldots, \delta_p \in \{0,1\}^p$, we set $\theta_\delta = (\delta_1\theta^{(1)}, \ldots, \delta_p\theta^{(p)})^\top$. Note that

$$\theta^\top Z\theta = \theta^\top\big[YY^\top - \mathrm{diag}\big(YY^\top\big)\big]\theta = \theta_\delta^\top \bar{Z}\theta_\delta.$$

Recall that $\mathbf{S}^{p-1} = \{\theta \in \mathbb{R}^p : |\theta|_2 = 1\}$. We have for any $\theta \in \mathbf{S}^{p-1}$ that

$$\begin{aligned}
\big(\theta_\delta^\top \bar{Z}\theta_\delta\big)^2 &= \Bigg(\sum_{j\neq k} \delta_j\delta_k\theta^{(j)}\theta^{(k)}X^{(j)}X^{(k)}\Bigg)^2 \\
&= \sum_{j_1,j_2:j_1\neq j_2} \delta_{j_1}\delta_{j_2}\big(\theta^{(j_1)}\big)^2\big(\theta^{(j_2)}\big)^2\big(X^{(j_1)}\big)^2\big(X^{(j_2)}\big)^2 \\
&\quad + \sum_{j_1,j_2,j_3 \text{ distinct}} \delta_{j_1}\delta_{j_2}\delta_{j_3}\big(\theta^{(j_1)}\big)^2\theta^{(j_2)}\theta^{(j_3)}\big(X^{(j_1)}\big)^2X^{(j_2)}X^{(j_3)} \\
&\quad + \sum_{j_1,j_2,j_3,j_4 \text{ distinct}} \delta_{j_1}\delta_{j_2}\delta_{j_3}\delta_{j_4}\theta^{(j_1)}\theta^{(j_2)}\theta^{(j_3)}\theta^{(j_4)}X^{(j_1)}X^{(j_2)}X^{(j_3)}X^{(j_4)}.
\end{aligned}$$

Taking the expectation w.r.t. to $\delta_1, \ldots, \delta_p$, we get

$$\begin{aligned}
\mathbb{E}_\delta\big[\big(\theta_\delta^\top \bar{Z}\theta_\delta\big)^2\big] &= \delta^2 \sum_{j_1,j_2:j_1\neq j_2} \big(\theta^{(j_1)}\big)^2\big(\theta^{(j_2)}\big)^2\big(X^{(j_1)}\big)^2\big(X^{(j_2)}\big)^2 \\
&\quad + \delta^3 \sum_{j_1,j_2,j_3 \text{ distinct}} \big(\theta^{(j_1)}\big)^2\theta^{(j_2)}\theta^{(j_3)}\big(X^{(j_1)}\big)^2X^{(j_2)}X^{(j_3)} \\
&\quad + \delta^4 \sum_{j_1,j_2,j_3,j_4 \text{ distinct}} \theta^{(j_1)}\theta^{(j_2)}\theta^{(j_3)}\theta^{(j_4)}X^{(j_1)}X^{(j_2)}X^{(j_3)}X^{(j_4)}.
\end{aligned}$$

Set

$$\begin{aligned}
A &= \sum_{j_1,j_2:j_1\neq j_2} \big(\theta^{(j_1)}\big)^2\big(\theta^{(j_2)}\big)^2\big(X^{(j_1)}\big)^2\big(X^{(j_2)}\big)^2, \\
B &= \sum_{j_1,j_2,j_3 \text{ distinct}} \big(\theta^{(j_1)}\big)^2\theta^{(j_2)}\theta^{(j_3)}\big(X^{(j_1)}\big)^2X^{(j_2)}X^{(j_3)}, \\
C &= \sum_{j_1,j_2,j_3,j_4 \text{ distinct}} \theta^{(j_1)}\theta^{(j_2)}\theta^{(j_3)}\theta^{(j_4)}X^{(j_1)}X^{(j_2)}X^{(j_3)}X^{(j_4)}.
\end{aligned}$$

We have

$$\begin{aligned}
\mathbb{E}_\delta\big[\big(\theta_\delta^\top \bar{Z}\theta_\delta\big)^2\big] &= \delta^2A + \delta^3B + \delta^4C \\
&= \big(\delta^2 - \delta^4\big)A + \big(\delta^3 - \delta^4\big)B + \delta^4(A + B + C) \\
&= \big[(\delta^2 - \delta^4 - \big(\delta^3 - \delta^4\big)\big]A + \big(\delta^3 - \delta^4\big)(A + B) + \delta^4(A + B + C) \\
&= \delta^2(1 - \delta)A + \delta^3(1 - \delta)(A + B) + \delta^4(A + B + C).
\end{aligned} \tag{5.17}$$

Next, we note that

$$A+B+C=\big(\theta^\top \bar{Z}\theta\big)^2=\bigg(\big(\theta^\top X\big)^2-\sum_j\big(\theta^{(j)}X^{(j)}\big)^2\bigg)^2$$
$$\le 2\big(\theta^\top X\big)^4+2\bigg(\sum_j\big(\theta^{(j)}\big)^2\big(X^{(j)}\big)^2\bigg)^2$$
$$\le 2\big(\theta^\top X\big)^4+2\sum_j\big(\theta^{(j)}\big)^2\big(X^{(j)}\big)^4,$$

where the last inequality comes from convexity of $x\to x^2$.

Taking now the expectation w.r.t. X, we get for any $\theta\in\mathbf{S}^{p-1}$ that

$$\begin{aligned}\mathbb{E}_X[A+B+C]&\le 2\mathbb{E}_X\big[\big(\theta^\top X\big)^4\big]+2\sum_j\big(\theta^{(j)}\big)^2\mathbb{E}_X\big[\big(X^{(j)}\big)^4\big]\\&\le 8\big\|\theta^\top X\big\|_{\psi_2}^4+8\sum_j\big(\theta^{(j)}\big)^2\big\|X^{(j)}\big\|_{\psi_2}^4\\&\le\frac{8}{c_1^2}\big(\mathbb{E}_X\big[\big(\theta^\top X\big)^2\big]\big)^2+\frac{8}{c_1^2}\sum_j\big(\theta^{(j)}\big)^2\big(\mathbb{E}_X\big[\big(X^{(j)}\big)^2\big]\big)^2\\&\le\frac{16}{c_1^2}\|\Sigma\|_\infty^2,\end{aligned}\tag{5.18}$$

where we have used (2.1) and Assumption 1.

We now treat $A+B$ similarly. We have

$$\begin{aligned}A+B&=\sum_{j_1}\big(\theta^{(j_1)}\big)^2\big(X^{(j_1)}\big)^2\bigg(\sum_{j_2,j_3:j_2\ne j_1,j_3\ne j_1}\theta^{(j_2)}\theta^{(j_3)}X^{(j_2)}X^{(j_3)}\bigg)\\&=\sum_{j_1}\big(\theta^{(j_1)}\big)^2\big(X^{(j_1)}\big)^2\big(\theta^\top X-\theta^{(j_1)}X^{(j_1)}\big)^2\\&\le 2\sum_{j_1}\big(\theta^{(j_1)}\big)^2\big(X^{(j_1)}\big)^2\big(\theta^\top X\big)^2+2\sum_{j_1}\big(\theta^{(j_1)}\big)^4\big(X^{(j_1)}\big)^4.\end{aligned}$$

Next, we note that

$$\mathbb{E}_X\big[\big(X^{(j_1)}\big)^2\big(\theta^\top X\big)^2\big]\le\sqrt{\mathbb{E}_X\big[\big(X^{(j_1)}\big)^4\big]}\sqrt{\mathbb{E}_X\big[\big(\theta^\top X\big)^4\big]}\le\frac{4}{c_1^2}\|\Sigma\|_\infty^2.$$

Combining the last two displays, we get

$$\mathbb{E}_X[A+B]\le\frac{16}{c_1^2}\|\Sigma\|_\infty^2.\tag{5.19}$$

We now deal with A. We have

$$A \le \left(\sum_{j_1} \big(\theta^{(j_1)}\big)^2 \big(X^{(j_1)}\big)^2\right)^2 - \sum_{j_1} \big(\theta^{(j_1)}\big)^4 \big(X^{(j_1)}\big)^4 \le \sum_{j_1} \big(\theta^{(j_1)}\big)^2 \big(X^{(j_1)}\big)^4.$$

Taking the expectation w.r.t. X, we get

$$\mathbb{E}_X[A] \le \mathbb{E}_X\left(\sum_{j_1} \big(\theta^{(j_1)}\big)^2 \big(X^{(j_1)}\big)^4\right) \le \frac{4}{c_1^2}\|\Sigma\|_\infty^2, \tag{5.20}$$

by convexity of $x \to x^2$.

Combining (5.17)–(5.20), we get

$$\max_{\theta\in\mathbf{S}^{p-1}} \mathbb{E}\big[\big(\theta_\delta^\top \bar{Z}\theta_\delta\big)^2\big] \le \frac{16}{c_1^2}\delta^2\|\Sigma\|_\infty^2\big(1-\delta+\delta(1-\delta)+\delta^2\big) = \frac{16}{c_1^2}\delta^2\|\Sigma\|_\infty^2.$$

5.5. Proof of Fact 3

We note first that

$$\big\|\mathbb{E}\big(\tilde{Z}-\mathbb{E}[\tilde{Z}]\big)^2\big\|_\infty = \big\|\mathbb{E}\big[\tilde{Z}^2\big] - \big(\mathbb{E}[\tilde{Z}]\big)^2\big\|_\infty \le \max\big\{\big\|\mathbb{E}\big[\tilde{Z}^2\big]\big\|_\infty, \big\|\mathbb{E}[\tilde{Z}]\big\|_\infty^2\big\}. \tag{5.21}$$

Next, we have

$$\begin{aligned}
\big\|\mathbb{E}[\tilde{Z}]\big\|_\infty^2 &= \big\|\mathbb{E}[Z\mathbb{1}_{|Y|_2^2\le U}]\big\|_\infty^2 \\
&= \max_{\theta\in\mathbf{S}^{p-1}} \big\{\mathbb{E}\big[\theta^\top Z\theta \mathbb{1}_{|Y|_2^2\le U}\big]\big\}^2 \\
&\le \max_{\theta\in\mathbf{S}^{p-1}} \mathbb{E}\big[\big(\theta^\top Z\theta\big)^2\big]\mathbb{P}\big(|Y|_2^2 \le U\big) \\
&\le \max_{\theta\in\mathbf{S}^{p-1}} \mathbb{E}\big[\big(\theta^\top Z\theta\big)^2\big] \le \frac{16}{c_1^2}\delta^2\|\Sigma\|_\infty^2,
\end{aligned} \tag{5.22}$$

in view of Fact 2.

We now treat $\|\mathbb{E}[\tilde{Z}^2]\|_\infty$. Note first that $\|\mathbb{E}[\tilde{Z}^2]\|_\infty = \max_{\theta\in\mathbf{S}^{p-1}} \mathbb{E}[\theta^\top Z^2\theta \mathbb{1}_{|Y|_2^2\le U}] \le \|\mathbb{E}[Z^2]\|_\infty$. Next, we set $V = Z + \delta\,\mathrm{diag}(X\otimes X)$ and $W = \delta\,\mathrm{diag}(X\otimes X)$. Some easy algebra yields that

$$\begin{aligned}
\big\|\mathbb{E}\big[Z^2\big]\big\|_\infty &= \big\|\mathbb{E}\big[V^2\big] + \mathbb{E}\big[W^2\big] - \mathbb{E}[VW] - \mathbb{E}[WV]\big\|_\infty \\
&= \max_{\theta\in\mathbf{S}^{p-1}} \big\{\mathbb{E}\big[\theta^\top V^2\theta\big] + \mathbb{E}\big[\theta^\top W^2\theta\big] + 2\mathbb{E}\big[\theta^\top VW\theta\big]\big\}.
\end{aligned}$$

Next, we have for any $\theta \in \mathbf{S}^{p-1}$ that

$$\mathbb{E}\big[\theta^\top VW\theta\big] \le \mathbb{E}\big[\sqrt{\theta^\top V^2\theta}\sqrt{\theta^\top W^2\theta}\big] \le \sqrt{\mathbb{E}\big[\theta^\top V^2\theta\big]}\sqrt{\mathbb{E}\big[\theta^\top W^2\theta\big]},$$

where we have used Cauchy–Schwarz's inequality twice w.r.t. to the scalar product in $\mathbb{R}^p$ first and then w.r.t. $\mathbb{E}$. Combining the last two displays, we get that

$$\begin{aligned}\big\|\mathbb{E}\big[Z^2\big]\big\|_\infty &\le \max_{\theta\in\mathbf{S}^{p-1}}\big\{\mathbb{E}\big[\theta^\top V^2\theta\big]+\mathbb{E}\big[\theta^\top W^2\theta\big]+2\sqrt{\mathbb{E}\big[\theta^\top V^2\theta\big]}\sqrt{\mathbb{E}\big[\theta^\top W^2\theta\big]}\big\}\\ &\le \max_{\theta\in\mathbf{S}^{p-1}}\Big(\sqrt{\mathbb{E}\big[\theta^\top V^2\theta\big]}+\sqrt{\mathbb{E}\big[\theta^\top W^2\theta\big]}\Big)^2\\ &\le \Big(\sqrt{\max_{\theta\in\mathbf{S}^{p-1}}\mathbb{E}\big[\theta^\top V^2\theta\big]}+\sqrt{\max_{\theta\in\mathbf{S}^{p-1}}\mathbb{E}\big[\theta^\top W^2\theta\big]}\Big)^2\\ &\le \Big(\sqrt{\big\|\mathbb{E}V^2\big\|_\infty}+\sqrt{\big\|\mathbb{E}W^2\big\|_\infty}\Big)^2.\end{aligned}\tag{5.23}$$

We now treat $\|\mathbb{E}[V^2]\|_\infty$ and $\|\mathbb{E}[W^2]\|_\infty$ separately. Denote by $\mathbb{E}_\delta$ and $\mathbb{E}_X$ the expectations w.r.t. $(\delta_1,\dots,\delta_p)$ and X respectively. We have $\mathbb{E}[V^2]=\mathbb{E}_X\mathbb{E}_\delta[V^2]$.

For any $1\le k\le p$, we have

$$\big(V^2\big)_{k,k}=\delta^2\big(X^{(k)}\big)^4+\sum_{i=1:i\ne k}^{p}\delta_i\delta_k\big(X^{(i)}\big)^2\big(X^{(k)}\big)^2.$$

Taking the expectation w.r.t. $\mathbb{E}_\delta$, we get

$$\begin{aligned}\mathbb{E}_\delta\big[\big(V^2\big)_{k,k}\big]&=\delta^2\big(X^{(k)}\big)^4+\delta^2\sum_{i=1:i\ne k}^{p}\big(X^{(i)}\big)^2\big(X^{(k)}\big)^2\\ &=\delta^2\big(X^{(k)}\big)^2|X|_2^2.\end{aligned}$$

Now for any $1\le k,l\le p$ with $k\ne l$, we have

$$\big(V^2\big)_{k,l}=\delta\delta_k\delta_l\big(X^{(k)}\big)^3X^{(l)}+\delta\delta_k\delta_lX^{(k)}\big(X^{(l)}\big)^3+\sum_{j=1:j\ne k,j\ne l}^{p}\delta_j\delta_k\delta_l\big(X^{(j)}\big)^2X^{(k)}X^{(l)}.$$

Taking the expectation w.r.t. $\mathbb{E}_\delta$, we get

$$\begin{aligned}\mathbb{E}_\delta\big[\big(V^2\big)_{k,l}\big]&=\delta^3\big[\big(X^{(k)}\big)^3X^{(l)}+X^{(k)}\big(X^{(l)}\big)^3\big]+\delta^3\sum_{j=1:j\ne k,j\ne l}^{p}\big(X^{(j)}\big)^2X^{(k)}X^{(l)}\\ &=\delta^3X^{(k)}X^{(l)}|X|_2^2.\end{aligned}$$

Thus, we get

$$\big(\mathbb{E}_\delta V^2\big)_{k,l}=\begin{cases}\delta^2\big(X^{(k)}\big)^2|X|_2^2 & \text{if } k=l,\\ \delta^3X^{(k)}X^{(l)}|X|_2^2 & \text{otherwise.}\end{cases}$$

Consequently, we get for any $\theta = (\theta^{(1)}, \dots, \theta^{(p)})^\top \in \mathbf{S}^{p-1}$ that

$$\begin{aligned}
\mathbb{E}\big[\theta^\top V^2 \theta\big] &= \delta^2 \mathbb{E}_X \Bigg[\sum_{k=1}^p \big(\theta^{(k)}\big)^2 \big(X^{(k)}\big)^2 |X|_2^2 + \delta \sum_{k,l=1:k\neq l}^p \theta^{(k)} \theta^{(l)} X^{(k)} X^{(l)} |X|_2^2 \Bigg] \\
&= \delta^2 \mathbb{E}_X \Bigg[(1-\delta) \sum_{k=1}^p \big(\theta^{(k)}\big)^2 \big(X^{(k)}\big)^2 |X|_2^2 + \delta \sum_{k,l=1}^p \theta^{(k)} \theta^{(l)} X^{(k)} X^{(l)} |X|_2^2 \Bigg] \\
&= \delta^2 \Bigg((1-\delta) \sum_{k=1}^p \big(\theta^{(k)}\big)^2 \mathbb{E}_X \big[|X|_2^2 \big(X^{(k)}\big)^2 \big] + \delta \mathbb{E}_X \big[|X|_2^2 \big(X^\top \theta\big)^2 \big] \Bigg) \\
&\leq \delta^2 \sqrt{\mathbb{E}_X |X|_2^4} \Bigg((1-\delta) \sum_{k=1}^p \big(\theta^{(k)}\big)^2 \sqrt{\mathbb{E}_X \big[\big(X^{(k)}\big)^4 \big]} + \delta \sqrt{\mathbb{E}_X \big[\big(X^\top \theta\big)^4 \big]} \Bigg),
\end{aligned} \tag{5.24}$$

where we have applied Cauchy–Schwarz's inequality.

We have again by Cauchy–Schwarz's inequality and Assumption 1 that

$$\begin{aligned}
\mathbb{E}_X \big[|X|_2^4 \big] &= \sum_{j=1}^p \mathbb{E}\big[\big(X^{(j)}\big)^4 \big] + \sum_{j,k=1:j\neq k}^p \mathbb{E}\big[\big(X^{(j)}\big)^2 \big(X^{(k)}\big)^2 \big] \\
&\leq \sum_{j=1}^p \mathbb{E}\big[\big(X^{(j)}\big)^4 \big] + \sum_{j,k=1:j\neq k}^p \sqrt{\mathbb{E}\big[\big(X^{(j)}\big)^4 \big]} \sqrt{\mathbb{E}\big[\big(X^{(k)}\big)^4 \big]} \\
&\leq \Bigg(\sum_{j=1}^p \sqrt{\mathbb{E}\big[\big(X^{(j)}\big)^4 \big]} \Bigg)^2 \\
&\leq 4 \Bigg(\sum_{j=1}^p \big\| X^{(j)} \big\|_{\psi_2}^2 \Bigg)^2 \\
&\leq 4 c_1^{-2} \big(\operatorname{tr}(\Sigma) \big)^2,
\end{aligned}$$

where we have used (2.1).

The same argument gives

$$\sqrt{\mathbb{E}_X \big[\big(X^\top \theta\big)^4 \big]} \leq 2 \big\| \langle X, \theta \rangle \big\|_{\psi_2}^2 \leq 2 c_1^{-1} \|\Sigma\|_\infty \qquad \forall \theta \in \mathbf{S}^{p-1},$$

and

$$\sqrt{\mathbb{E}_X \big[\big(X^{(k)}\big)^4 \big]} \leq 2 \big\| X^{(k)} \big\|_{\psi_2}^2 \leq 2 c_1^{-1} \Sigma_{k,k}, \qquad 1 \leq k \leq p.$$

Combining the three above displays with (5.24), we get

$$\begin{aligned}\big\|\mathbb{E}\big[V^2\big]\big\|_\infty &\le Cc_1^{-1}\delta^2\operatorname{tr}(\Sigma)\Big[(1-\delta)\max_{1\le k\le p}(\Sigma_{k,k})+\delta\|\Sigma\|_\infty\Big]\\ &\le 4c_1^{-2}\delta^2\operatorname{tr}(\Sigma)\|\Sigma\|_\infty,\end{aligned} \tag{5.25}$$

and

$$\big\|\mathbb{E}\big[W^2\big]\big\|_\infty=\delta^2\max_{1\le k\le p}\mathbb{E}_X\big[\big(X^{(k)}\big)^4\big]\le 4c_1^{-2}\delta^2\max_{1\le k\le p}\big(\Sigma_{k,k}^2\big)\le 4c_1^{-2}\delta^2\operatorname{tr}(\Sigma)\|\Sigma\|_\infty.$$

Combining the two above displays with (5.23), we get

$$\big\|\mathbb{E}\big[Z^2\big]\big\|_\infty\le 16c_1^{-2}\delta^2\operatorname{tr}(\Sigma)\|\Sigma\|_\infty.$$

Combining the above display with (5.21) and (5.22), we get that

$$\mathbb{E}\big[\big(\tilde{Z}-\mathbb{E}[\tilde{Z}]\big)^2\big]\le\frac{16}{c_1^2}\delta^2\|\Sigma\|_\infty\operatorname{tr}(\Sigma).$$

Next, we treat $\|\tilde{Z}-\mathbb{E}[\tilde{Z}]\|_\infty$. We have

$$\begin{aligned}\big\|\tilde{Z}-\mathbb{E}[\tilde{Z}]\big\|_\infty&\le\|\tilde{Z}\|_\infty+\big\|\mathbb{E}[\tilde{Z}]\big\|_\infty\le\|\tilde{Z}\|_\infty+\mathbb{E}\big[\|\tilde{Z}\|_\infty\big]\\ &\le|Y|_2^2\mathbb{1}_{|Y|_2^2\le U}+\mathbb{E}\big[|Y|_2^2\mathbb{1}_{|Y|_2^2\le U}\big]\le 2U,\end{aligned}$$

where we have used that $\|Z\|_\infty\le\max\{\|Y\otimes Y\|_\infty,\|\operatorname{diag}(Y\otimes Y)\|_\infty\}\le|Y|_2^2$. □

5.6. Proof of Lemma 1

In view of Proposition 3, we have on an event $\mathcal{A}$ of probability at least $1-\frac{1}{2p}$ that

$$\|\tilde{\Sigma}_n-\Sigma\|_\infty\le C\frac{\|\Sigma\|_\infty}{c_1}\sqrt{\frac{\mathbf{r}(\Sigma)\log 2p}{\delta^2 n}}. \tag{5.26}$$

We assume further that (3.6) is satisfied with a sufficiently large constant c so that we have, in view of (3.4) and (3.5), on the same event $\mathcal{A}$ that

$$\|\tilde{\Sigma}_n-\Sigma\|_\infty\le\frac{\|\Sigma\|_\infty}{2}$$

and

$$\big|\operatorname{tr}(\tilde{\Sigma}_n)-\operatorname{tr}(\Sigma)\big|\le\frac{\operatorname{tr}(\Sigma)}{2}.$$

We immediately get on the event $\mathcal{A}$ that

$$\tfrac{1}{2}\|\Sigma\|_\infty\le\|\tilde{\Sigma}_n\|_\infty\le\tfrac{3}{2}\|\Sigma\|_\infty,$$

and

$$\tfrac{1}{2}\operatorname{tr}(\Sigma) \le \operatorname{tr}(\tilde{\Sigma}_n) \le \tfrac{3}{2}\operatorname{tr}(\Sigma).$$

Combining these simple facts with (5.26), we get the result.

5.7. Proof of Theorem 2

This proof uses standard tools of the minimax theory (cf., e.g., [32]). However, as for Proposition 3, the proof with missing observations ($\delta < 1$) is significantly more technical as compared to case of full observations ($\delta = 1$). In particular, the control of the Kullback–Leibler divergence requires a precise description of the conditional distributions of the random variables $Y_1, \dots, Y_n$ given the masked variables $\delta_1, \dots, \delta_n$. To our knowledge, there exists no minimax lower bound result for statistical problem with missing observations in the literature.

Proof of Theorem 2. Set $\gamma = a/\sqrt{\delta^2 n}$ where $a > 0$ is a sufficiently small absolute constant. Define also the set of $p \times p$ matrices $\{E_{k,l} = e_k \otimes e_l, 1 \le k, l \le p\}$ where $e_1, \dots, e_p$ is the canonical basis of $\mathbb{R}^p$.

We consider first the case $r \ge 2$. Define

$$\mathcal{N} = \{E_{k,l} + E_{l,k}, 1 \le k \le r-1, k+1 \le l \le r\}.$$

Set $B_{k,l} = E_{k,l} + E_{l,k}$ for any $1 \le k \le r-1, k+1 \le l \le r$. Consider the associated set of symmetric matrices

$$\mathcal{B}(\mathcal{N}) = \left\{ \Sigma_\varepsilon = \left(\begin{array}{c|c} I_r + \gamma \sum_{k=1}^{r-1} \sum_{l=k+1}^{r} \varepsilon_{k,l} B_{k,l} & O \\ \hline O & O \end{array} \right), \varepsilon = (\varepsilon_{k,l})_{k,l} \in \{0,1\}^{r(r-1)/2} \right\}.$$

Note that any matrix $\Sigma_\varepsilon \in \mathcal{B}(\mathcal{N})$ is positive-semidefinite if $0 < a < 1$ since we have by assumption

$$\left\| \gamma \sum_{k=1}^{r-1} \sum_{l=k+1}^{r} \varepsilon_{k,l} B_{k,l} \right\|_\infty \le \gamma r = a \sqrt{\frac{r^2}{\delta^2 n}} \le a.$$

By construction, any element of $\mathcal{B}(\mathcal{N})$ as well as the difference of any two elements of $\mathcal{B}(\mathcal{N})$ is of rank exactly r. Consequently, $\mathcal{B}(\mathcal{N}) \subset \mathcal{C}_r$ since $\mathbf{r}(\Sigma_\varepsilon) \le \operatorname{rank}(\Sigma_\varepsilon) \le r$ for any $\Sigma_\varepsilon \in \mathcal{B}(\mathcal{N})$. Note also that for any $\Sigma_\varepsilon \in \mathcal{B}(\mathcal{N})$, we have $\operatorname{tr}(\Sigma_\varepsilon) = r$ and $0 < 1 - a \le \|\Sigma_\varepsilon\|_\infty \le 1 + a$ provided that $0 < a < 1$ and consequently $r/(1+a) \le \mathbf{r}(\Sigma_\varepsilon) \le r/(1-a)$. Indeed, we have

$$\|\Sigma_\varepsilon\|_\infty \le 1 + \gamma \left\| \sum_{k=1}^{r-1} \sum_{l=k+1}^{r} \varepsilon_{k,l} B_{k,l} \right\|_\infty \le 1 + \gamma r \le 1 + a,$$

in view of the condition $n \ge \delta^{-2} r^2$. A similar reasoning gives the lower bound.

Denote by A_0 the $p \times p$ block matrix with first block equal to I_r. Varshamov–Gilbert's bound (cf. Lemma 2.9 in [32]) guarantees the existence of a subset $\mathcal{A}^0 \subset \mathcal{B}(\mathcal{N})$ with cardinality $\mathrm{Card}(\mathcal{A}^0) \geq 2^{r(r-1)/16} + 1$ containing A_0 and such that, for any two distinct elements Σ_ε and $\Sigma_{\varepsilon'}$ of $\mathcal{A}^0$, we have

$$\begin{aligned} \|\Sigma_\varepsilon - \Sigma_{\varepsilon'}\|_2^2 &\geq \gamma^2 \frac{r(r-1)}{8} \geq \gamma^2 \frac{r^2}{16} = \frac{a^2}{16} \frac{r^2}{\delta^2 n} \\ &\geq \frac{(1-a)a^2}{16(1+a)^2} \|\Sigma_{\tilde{\varepsilon}}\|_\infty^2 \frac{\mathbf{r}(\Sigma_{\tilde{\varepsilon}})}{\delta^2 n} \operatorname{rank}(\Sigma_{\tilde{\varepsilon}}) \qquad \forall \Sigma_{\tilde{\varepsilon}} \in \mathcal{A}^0. \end{aligned} \tag{5.27}$$

Let $X_1, \ldots, X_n \in \mathbb{R}^p$ be i.i.d. $N(0, \Sigma_\varepsilon)$ with $\Sigma_\varepsilon \in \mathcal{A}^0$. For the sake of brevity, we set $\Sigma = \Sigma_\varepsilon$. Recall that $\delta_1, \ldots, \delta_n$ are random vectors in $\mathbb{R}^p$ whose entries $\delta_{i,j}$ are i.i.d. Bernoulli entries with parameter δ independent from $(X_1, \ldots, X_n)$ and that the observations $Y_1, \ldots, Y_n$ satisfy $Y_i^{(j)} = \delta_{ij} X_i^{(j)}$. Denote by $\mathbb{P}_\Sigma$ the distribution of $(Y_1, \ldots, Y_n)$ and by $\mathbb{P}_\Sigma^{(\delta)}$ the conditional distribution of $(Y_1, \ldots, Y_n)$ given $(\delta_1, \ldots, \delta_n)$. Next, we note that, for any $1 \leq i \leq n$, the conditional random variables $Y_i \mid \delta_i$ are independent Gaussian vectors $N(0, \Sigma^{(\delta_i)})$ where

$$\left(\Sigma^{(\delta_i)}\right)_{j,k} = \begin{cases} \delta_{i,j}\delta_{i,k}\Sigma_{j,k} & \text{if } j \neq k, \\ \delta_{i,j}\Sigma_{j,j} & \text{otherwise.} \end{cases} \tag{5.28}$$

Thus, we have $\mathbb{P}_\Sigma^{(\delta)} = \bigotimes_{i=1}^n \mathbb{P}_{\Sigma^{(\delta_i)}}$. Denote respectively by $\mathbb{P}_\delta$ and $\mathbb{E}_\delta$ the probability distribution of $(\delta_1, \ldots, \delta_n)$ and the associated expectation, and by $\mathbb{E}_{\delta_i}$ the expectation w.r.t. δ_i for any $1 \leq i \leq n$. We also denote by $\mathbb{E}_\Sigma$ and $\mathbb{E}_\Sigma^{(\delta)}$ the expectation and conditional expectation associated respectively, with $\mathbb{P}_\Sigma$ and $\mathbb{P}_\Sigma^{(\delta)}$.

Next, the Kullback–Leibler divergences $K(\mathbb{P}_{A_0}, \mathbb{P}_\Sigma)$ between $\mathbb{P}_{A_0}$ and $\mathbb{P}_\Sigma$ satisfies

$$\begin{aligned} K(\mathbb{P}_{A_0}, \mathbb{P}_\Sigma) &= \mathbb{E}_{A_0} \log\left(\frac{\mathrm{d}\mathbb{P}_{A_0}}{\mathrm{d}\mathbb{P}_\Sigma}\right) = \mathbb{E}_{A_0} \log\left(\frac{\mathrm{d}(\mathbb{P}_\delta \otimes \mathbb{P}_{A_0}^{(\delta)})}{\mathrm{d}(\mathbb{P}_\delta \otimes \mathbb{P}_\Sigma^{(\delta)})}\right) = \mathbb{E}_\delta \mathbb{E}_{A_0}^{(\delta)} \log\left(\frac{\mathrm{d}\mathbb{P}_{A_0}^{(\delta)}}{\mathrm{d}\mathbb{P}_\Sigma^{(\delta)}}\right) \\ &= \mathbb{E}_\delta K\left(\mathbb{P}_{A_0}^{(\delta)}, \mathbb{P}_\Sigma^{(\delta)}\right) = \sum_{i=1}^n \mathbb{E}_{\delta_i} K(\mathbb{P}_{A_0^{(\delta_i)}}, \mathbb{P}_{\Sigma^{(\delta_i)}}). \end{aligned} \tag{5.29}$$

Using that $Y_i \mid \delta_i \sim N(0, \Sigma^{(\delta_i)})$ with $\Sigma^{(\delta_i)}$ defined in (5.28), we get for any $1 \leq i \leq n$, any $\Sigma \in \mathcal{A}^0$ and any realization $\delta_i(\omega) \in \{0,1\}^p$ that

1. $\mathbb{P}_{\Sigma^{(\delta_i(\omega))}} \ll \mathbb{P}_{A_0^{(\delta_i(\omega))}}$ and hence $K(\mathbb{P}_{A_0^{(\delta_i(\omega))}}, \mathbb{P}_{\Sigma^{(\delta_i(\omega))}}) < \infty$.
2. $\mathbb{P}_{\Sigma^{(\delta_i(\omega))}}$ and $\mathbb{P}_{A_0^{(\delta_i(\omega))}}$ are supported on a $d_i(\omega)$-dimensional subspace of $\mathbb{R}^p$ where $d_i = \sum_{j=1}^r \delta_{i,j} \sim \mathrm{Bin}(r, \delta)$.

Define $J_i = \{j : \delta_{i,j} = 1, 1 \leq j \leq r\}$. Define the mapping $P_i : \mathbb{R}^p \to \mathbb{R}^{d_i}$ as follows $P_i(x) = x_{J_i}$ where for any $x = (x^{(1)}, \ldots, x^{(p)})^\top \in \mathbb{R}^p$, $x_{J_i} \in \mathbb{R}^{d_i}$ is obtained by keeping only the components $x^{(k)}$ with their index $k \in J_i$. We denote by $P_i^* : \mathbb{R}^{d_i} \to \mathbb{R}^p$ the right inverse of P_i.

We note that $P_i A_0^{(\delta_i)} P_i^* = I_{d_i}$ and

$$\begin{aligned} P_i \Sigma^{(\delta_i)} P_i^* &= I_{d_i} + \gamma \sum_{k=1}^{r-1} \sum_{l=k+1}^{r} \varepsilon_{k,l} P_i B_{k,l} P_i^* \mathbb{1}_{k \in J_i} \mathbb{1}_{l \in J_i} \\ &= I_{d_i} + W_i . \end{aligned}$$

Thus, we get that

$$\begin{aligned} K(\mathbb{P}_{A_0^{(\delta_i)}}, \mathbb{P}_{\Sigma^{(\delta_i)}}) &= K(\mathbb{P}_{I_{d_i}}, \mathbb{P}_{I_{d_i}+W_i}) \\ &= \frac{1}{2} \operatorname{tr}(I_{d_i} + W_i) - \frac{1}{2} \log\bigl(\det(I_{d_i} + W_i)\bigr) - \frac{d_i}{2}. \end{aligned}$$

Denote by $\lambda_1, \dots, \lambda_{d_i}$ the eigenvalues of W_i. Note that $\sum_{j=1}^{d_i} \lambda_j = \operatorname{tr}(W_i) = 0$. We get, using the inequality $\log(1+x) \geq x - x^2/2$ for any $x > 0$, that

$$\begin{aligned} K(\mathbb{P}_{A_0^{(\delta_i)}}, \mathbb{P}_{\Sigma^{(\delta_i)}}) &\leq \frac{1}{4} \sum_{j=1}^{d_i} \lambda_j^2 \\ &\leq \frac{1}{4} \| W_i \|_2^2 \\ &\leq \frac{1}{4} \gamma^2 \sum_{k=1}^{r-1} \sum_{l=k+1}^{r} \| B_{k,l} \|_2^2 \mathbb{1}_{k \in J_i} \mathbb{1}_{l \in J_i} \\ &\leq \frac{1}{2} \gamma^2 \bigl(d_i^2 - d_i\bigr). \end{aligned} \tag{5.30}$$

Taking the expectation w.r.t. to δ_i in the above display, we get for any $1 \leq i \leq n$ that

$$\mathbb{E}_{\delta_i} K(\mathbb{P}_{A_0^{(\delta_i)}}, \mathbb{P}_{\Sigma^{(\delta_i)}}) \leq \tfrac{1}{2} \gamma^2 \mathbb{E}_{\delta_i} \bigl(d_i^2 - d_i\bigr) \leq \tfrac{1}{2} \gamma^2 \delta^2 r(r-1),$$

since $d_i \sim \operatorname{Bin}(r, \delta)$. Combining the above display with (5.29), we get

$$K(\mathbb{P}_{A_0}, \mathbb{P}_\Sigma) \leq \frac{n}{2} \gamma^2 \delta^2 r(r-1) = \frac{n}{2} a^2 \frac{1}{\delta^2 n} \delta^2 r(r-1) \leq \frac{a^2}{2} r(r-1).$$

Thus, we deduce from the above display that the condition

$$\frac{1}{\operatorname{Card}(\mathcal{A}^0) - 1} \sum_{\Sigma \in \mathcal{A}^0 \setminus \{A_0\}} K(\mathbb{P}_{A_0}, \mathbb{P}_\Sigma) \leq \alpha \log\bigl(\operatorname{Card}\bigl(\mathcal{A}^0\bigr) - 1\bigr) \tag{5.31}$$

is satisfied for any $\alpha > 0$ if $a > 0$ is chosen as a sufficiently small numerical constant depending on α. In view of (5.27) and (5.31), (4.1) now follows by application of Theorem 2.5 in [32].

The lower bound (4.2) follows from (4.1) by the following simple argument. Consider the set of matrices $\mathcal{A}^0$. For any two distinct matrices Σ_1, Σ_2 of $\mathcal{A}^0$, we have

$$\|\Sigma_1 - \Sigma_2\|_\infty \geq \sqrt{\frac{(1-a)a^2}{16(1+a)^2}}\|\Sigma_{\tilde{\varepsilon}}\|_\infty \sqrt{\frac{\mathbf{r}(\Sigma_{\tilde{\varepsilon}})}{\delta^2 n}} \qquad \forall \Sigma_{\tilde{\varepsilon}} \in \mathcal{A}^0. \tag{5.32}$$

Indeed, if (5.32) does not hold, we get

$$\|\Sigma_1 - \Sigma_2\|_2^2 < \frac{(1-a)a^2}{16(1+a)^2}\|\Sigma_{\tilde{\varepsilon}}\|_\infty^2 \frac{\mathbf{r}(\Sigma_{\tilde{\varepsilon}})}{\delta^2 n}\operatorname{rank}(\Sigma_{\tilde{\varepsilon}}) \qquad \forall \Sigma_{\tilde{\varepsilon}} \in \mathcal{A}^0,$$

since $\operatorname{rank}(\Sigma_1 - \Sigma_2) \leq r$ by construction of $\mathcal{A}^0$. This contradicts (5.27).

Next, (5.31) is satisfied for any $\alpha > 0$ if $a > 0$ is chosen as a sufficiently small numerical constant depending on α.

Combining (5.32) with (5.31) and Theorem 2.5 in [32] gives the result.

The case $r = 1$ can be treated similarly and is actually easier. Indeed if $\mathbf{r}(\Sigma) = 1$, then we have $\operatorname{tr}(\Sigma) = \|\Sigma\|_\infty$ and $\operatorname{rank}(\Sigma) = 1$. Consequently, we can derive the lower bound by testing between the two hypothesis

$$\Sigma_0 = \left(\begin{array}{c|c} 1 & O \\ \hline O & O \end{array}\right) \quad \text{and} \quad \Sigma_1 = \left(\begin{array}{c|c} 1+\gamma & O \\ \hline O & O \end{array}\right),$$

where Σ_0 and Σ_1 are $p \times p$ covariance matrices with only one non-zero component on the first diagonal entry. For these covariance matrices, we have $\operatorname{tr}(\Sigma_0) = \|\Sigma_0\|_\infty = 1$ and $\operatorname{tr}(\Sigma_1) = \|\Sigma_1\|_\infty = 1 + \gamma \leq 2$. Thus, we have

$$\|\Sigma_0 - \Sigma_1\|_\infty^2 = \|\Sigma_0 - \Sigma_1\|_2^2 \geq \frac{a^2}{\delta^2 n} \geq c\|\Sigma_i\|_\infty^2 \frac{\mathbf{r}(\Sigma_i)}{\delta^2 n}, \qquad i = 0, 1$$

for some absolute constant $c > 0$. The rest of the proof is identical to the case $r \geq 2$. □

5.8. Computation of $\mathbb{E}[\Sigma_n^{(\delta)}]$ and (1.3)

Recall that $Y = (\delta_1 X^{(1)}, \ldots, \delta_p X^{(p)})^\top$. We have

$$(Y \otimes Y)_{j,k} = \begin{cases} \delta_j \big(X^{(j)}\big)^2 & \text{if } j = k, \\ \delta_j \delta_k X^{(j)} X^{(k)} & \text{otherwise.} \end{cases}$$

Set $\Sigma^{(\delta)} = \mathbb{E}[\Sigma_n^{(\delta)}] = \mathbb{E}[Y \otimes Y]$. We have

$$\Sigma_{j,k}^{(\delta)} = \begin{cases} \delta \Sigma_{j,j} & \text{if } j = k, \\ \delta^2 \Sigma_{j,k} & \text{otherwise.} \end{cases}$$

Next, we have

$$\frac{1}{\delta}\operatorname{diag}\bigl(\Sigma^{(\delta)}\bigr)+\frac{1}{\delta^2}\bigl(\Sigma^{(\delta)}-\operatorname{diag}\bigl(\Sigma^{(\delta)}\bigr)\bigr)=\frac{1}{\delta}\delta\operatorname{diag}(\Sigma)+\frac{1}{\delta^2}\delta^2\bigl(\Sigma-\operatorname{diag}(\Sigma)\bigr)$$
$$=\operatorname{diag}(\Sigma)+\bigl(\Sigma-\operatorname{diag}(\Sigma)\bigr)=\Sigma.$$

Acknowledgement

Supported in part by NSF Grant DMS-11-06644 and Simons Foundation Grant 209842. I wish to thank Professor Vladimir Koltchinskii for suggesting this problem and the observation (1.3).

References

[1] Ahlswede, R. and Winter, A. (2002). Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory* **48** 569–579. MR1889969

[2] Banerjee, O., El Ghaoui, L. and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9** 485–516. MR2417243

[3] Bickel, P.J. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

[4] Bickel, P.J., Ritov, Y. and Tsybakov, A.B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

[5] 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

[6] Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106** 672–684. MR2847949

[7] Cai, T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

[8] Cai, T.T., Zhang, C.-H. and Zhou, H.H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144. MR2676885

[9] Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

[10] Candès, E.J. and Plan, Y. (2010). Matrix completion with noise. In *Proceedings of the IEEE, Vol.* 98 925–936.

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

[12] Donoho, D.L. and Tanner, J. (2005). Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proc. Natl. Acad. Sci. USA* **102** 9446–9451 (electronic). MR2168715

[13] El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. MR2485011

[14] El Karoui, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.* **36** 2757–2790. MR2485012

[15] Gross, D. (2009) Recovering low-rank matrices from few coefficients in any basis. Available at arXiv:0910.1879.

[16] Johnstone, I.M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961
[17] Johnstone, I.M. and Ma, Z. (2013). Fast approach to the Tracy-Widom law at the edge of GOE and GUE. *Ann. Appl. Probab.* **22** 1962–1988. MR3025686
[18] Keshavan, R.H., Montanari, A. and Oh, S. (2010). Matrix completion from noisy entries. *J. Mach. Learn. Res.* **11** 2057–2078. MR2678022
[19] Klopp, O. (2011). Rank penalized estimators for high-dimensional matrices. *Electron. J. Stat.* **5** 1161–1183. MR2842903
[20] Koltchinskii, V. (2009). The Dantzig selector and sparsity oracle inequalities. *Bernoulli* **15** 799–828. MR2555200
[21] Koltchinskii, V. (2010) Von Neumann entropy penalization and low rank matrix approximation. Available at arXiv:1009.2439.
[22] Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. *Lecture Notes in Math.* **2033**. Heidelberg: Springer. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour [Saint-Flour Probability Summer School]. MR2829871
[23] 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
[24] Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.* **2** 90–102. MR2386087
[25] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363
[26] Recht, B., Fazel, M. and Parrilo, P.A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.* **52** 471–501. MR2680543
[27] Rohde, A. and Tsybakov, A.B. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.* **39** 887–930. MR2816342
[28] Rothman, A.J., Bickel, P.J., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. MR2417391
[29] Rothman, A.J., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* **104** 177–186. MR2504372
[30] Schneider, T. (2001). Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *J. Climate* **14** 853–871.
[31] Tropp, J.A. (2010) User-friendly tail bounds for sums of random matrices. Available at arXiv:1004.4389.
[32] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. New York: Springer. Revised and extended from the 2004 French original. Translated by Vladimir Zaiats. MR2724359
[33] Vershynin, R. (2011) Introduction to the non-asymptotic analysis of random matrices. Available at arXiv:1011.3027v7.
[34] Watson, G.A. (1992). Characterization of the subdifferential of some matrix norms. *Linear Algebra Appl.* **170** 33–45. MR1160950
[35] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824

Received April 2012 and revised September 2012