Covariance estimation with missing observation

Karim Lounici   Grégoire Pacreau

March 2022

1 Introduction

Let $X, X_1, \dots, X_n$ be i.i.d. zero mean vectors with unknown covariance matrix $\Sigma = \mathbb{E}\left[X \bigotimes X\right]$. 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 or wrongly 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 with probability $\delta \in (0, 1]$. In the first part of our study, we will assume that the remaining components are not observed, i.e.:

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

where δ_{ij} are independent realisations of a bernoulli random variable of parameter δ. In the second part, we will assume the missing data is replaced by another independent distribution, representing either a poisoning of the data or random mistakes in measurements. The observations are then:

$$Y_i^{(j)} = \delta_{i,j} X_i^{(j)} + (1 - \delta_{ij}) \xi_i^{(j)}, \qquad 1 \le i \le n, 1 \le j \le p \tag{1.2}$$

where $\xi_1, \dots \xi_n$ are erroneous measurements following a i.i.d. subgaussian distribution.

The missing values case corresponds to the Missing Completely At Random (MCAR) case in Rubin (1976), compared with missing at random (MAR), where some dimensions are more prone to not being observed, or not missing at random (NMAR), where the lack observation is a deterministic function of the realisations of the random variables (not missing at random). Unbiased estimator based on maximum likelihood and the expectation maximisation algorithm exist to compute the variance with MCAR data (Jamshidian and Bentler, 1999). We focus instead on the unbiased estimator of Lounici (2014), which avoids the many steps of EM like algorithms while guarantying a known rate of convergence and minimax error lower bound according to the missingness rate δ. In this paper, we improve upon the theoretical results of Lounici (2014).

GP: faut-il expliquer plus pourquoi on se focalise sur ce problème?

The contamination case relates to the Fully Independent Contamination Model (FICM) as described in Alqallaf et al. (2009), which is a direct extension of Huber contamination (Huber, 1964) on samples to a more cell-wise approach. While the latter is much studied in the litterature, with robust algorithms for estimating both means, such as Tukey's median (Tukey, 1978), and covariance, such as the Minimum Covariance Determinant (Hubert et al., 2018) or Tukey's S-estimator (Rousseeuw and Yohai, 1984). On the other hand, the former is a more recent problem with much less litterature, as far as covariance estimation is concerned. Some papers propose to adapt Huber-style procedures to this generalisation (Farcomeni, 2014; Rousseeuw and den Bossche, 2018) or showcase expensive procedures based on expectation maximisation and the Mahalanobis distance (Raymaekers and Rousseeuw, 2021). However, to our knowledge, only the consistency of these methods have been studied, without consideration for the rates of convergence.

GP: Doit-on mentionner l'absence d'étude dans le cas adversariel?

We focus on the high dimensional case, in which most procedures we cited fail due to both computational errors and high time requirements. In particular, we suppose that the true covariance matrix of the X_i has a low rank structure. This makes the Mahalanobis distance based methods inpracticle, since it would require the inversion of a matrix with many clos to zero eigenvalues. Our proposal is base on a correction of

GP: ajouter des exemples pratiques?

the classical covariance estimator on $Y_1, \dots, Y_n$ first introduced in Lounici (2014) for the case with missing values. The procedure is based on the following observation, with Σ^Y the covariance of the data with missing values and Σ the true covariance:

$$\Sigma = \left(\delta^{-1} - \delta^{-2}\right) \operatorname{diag}(\Sigma^Y) + \delta^{-2}\Sigma^Y \tag{1.3}$$

We then extend this correction to the case where the data is contaminated, first by introducing a new term, then by returning to a missing values problem by eliminating outliers using a detection procedure. We argue that the latter technique is more promissing, assuming a sufficiently accurate detection algorithm, since the existence of outliers cause the estimator to bear the full effect of the high dimension p, whereas ideally we would like to be constrained by the rank of the true covariance matrix.

2 Missing Values

We place ourselves in the setting described in equation 1.1. We provide an updated lower and upper bound to the estimator defined in 1.3, which are sharper than those in Lounici (2014).

2.1 Upper bound

Using the correction of equation 1.3 we are able to construct an unbiased estimator of the covariance matrix. In this section, we provide an upper bound of the estimation error in operator norm. This upper bound depends on the effective rank of Σ, the true covariance matrix, which is a measure of the intrinsic dimension of a symmetric matrix. The effective rank is defined as:

$$\boldsymbol{r}(\Sigma) := \frac{\mathbb{E}\left\|X\right\|^2}{\|\Sigma\|} = \frac{\operatorname{tr}(\Sigma)}{\|\Sigma\|} \tag{2.1}$$

We can see that $0 \leq \boldsymbol{r}(\Sigma) \leq \operatorname{rank}(\Sigma)$. Furthermore, $\boldsymbol{r}(\Sigma) \ll \operatorname{rank}(\Sigma)$ for approximately low rank matrices, i.e. matrices with few eigenvalues significantly larger than 0.

GP: Passer en espace de Hilbert

Theorem 1 *Let $X_1, \dots, X_n$ be i.i.d. subgaussian random variables in $\mathbb{R}^p$, with covariance matrix Σ, and let $\delta_i^j, i \in [1,n], j \in [1,p]$ be i.i.d bernoulli random variables with probability of success δ. We write $Y_i = \delta_i \odot X_i$. Let $\hat{\Sigma}$ be the classical covariance estimator applied on Y corrected as described in equation 1.3. There exists an absolute constant C such that, for $t > 0$, with probability at least $1 - e^{-t}$:*

$$\left\|\hat{\Sigma} - \Sigma\right\| \leq C\frac{\|\Sigma\|}{\delta}\left(\sqrt{\frac{\boldsymbol{r}(\Sigma)}{n}} \vee \frac{\boldsymbol{r}(\Sigma)}{n} \vee \sqrt{\frac{t}{n}} \vee \frac{t}{n}\right) \tag{2.2}$$

This fact is the consequence of two lemmas, under the same assumptions:

Lemma 1 *Under the same assumptions, let $\Sigma^Y = \mathbb{E}Y \otimes Y$ and $\hat{\Sigma}^Y = n^{-1}\sum_{i=1}^n Y_i \otimes Y_i$. There exist an absolute constant c_1 such that, for $t > 0$, with probability at least $1 - e^{-t}$:*

$$\left\|\hat{\Sigma}^Y - \Sigma^Y\right\| \leq c_1\delta\left\|\Sigma\right\|\left(\sqrt{\frac{\boldsymbol{r}(\Sigma)}{n}} \vee \frac{\boldsymbol{r}(\Sigma)}{n} \vee \sqrt{\frac{t}{n}} \vee \frac{t}{n}\right) \tag{2.3}$$

and

Lemma 2 *Under the same assumptions and notations, there exist an absolute constant c_2 such that, for $t > 0$, with probability at least $1 - e^{-t}$:*

$$\left\|diag\left(\hat{\Sigma}^Y - \Sigma^Y\right)\right\| \leq c_2 \max_j \Sigma_{jj}\left(\sqrt{\frac{t}{n}} \vee \frac{t}{n}\right) \tag{2.4}$$

We provide proof of these lemma and of the upper bound in section 6.2

2.2 Lower bound

Theorem 2 *Let $\mathcal{S}_r$ the set of all covariance matrices of rank r. Then:*

$$\inf_{\hat{\Sigma}} \max_{\Sigma \in \mathcal{S}_r} \mathbb{P}_\Sigma \left(\left\| \hat{\Sigma} - \Sigma \right\| \geq C \sqrt{\frac{r}{\delta n}} \right) \geq \beta \tag{2.5}$$

for C and β two absolute constants and where $\inf_{\hat{\Sigma}}$ represents the infimum over all estimators of matrixΣ.

We see that the dependency in δ is the same in both upper and lower bounds. Furthermore, for n much larger than $\boldsymbol{r}(\Sigma)$, which in the low rank case is very easy to verify, the square roots in the upper bound simplify, leading to the same structure as our lower bound.

3 Contaminations

Let us now look at the case where $Y_i^{(j)} = \delta_{ij} X_i^{(j)} + \varepsilon_{ij} \xi_i j$, where $X_1, \dots X_n$ are i.i.d. vectors of lax $\mathcal{N}(0, \Sigma)$ and $\xi_1, \dots \xi_n$ i.i.d. vectors following a subgaussian law of variance Λ, with Λ a diagonal matrix of size p. Let also $\lambda = \max_i \Lambda_i = \|\Lambda\|$. We suppose that the X_i and the ξ_i are mutually independent. However, the boolean random variables δ_{ij} and ϵ_{ij} cannot be both equal to 1 (a component cannot be both correctly observed and contaminated). This is a slight generalisation of the cell-wise contamination of Alqallaf et al. (2009), where $epsilon_{ij} = 1 - \delta_{ij}$. This generalisation allows us to showcase the influence of a accurate filtering of the contaminated data, where in the filtered data we find true values with probability δ (preferably close to 1) and contaminated data with probability ε (preferably close to 0). This means that we observe nothing with probability $1 - \delta - \varepsilon$.

Let $\Sigma^Y = \mathbb{E}\left(Y \otimes Y\right)$ and let $\hat{\Sigma}^Y = n^{-1} \sum_{i=1}^n Y_i \otimes Y_i$ the empirical covariance matrix. Assuming knowledge of Λ, we get the following correction formula (see section 8.2 for the detail):

GP: PEut-être que les notations ε et δ portent à confusion avec les vecteurs $varepsilon_i$ et δ_i?

$$\Sigma = (\delta^{-1} - \delta^{-2}) \mathrm{diag}\left(\Sigma^Y\right) + \delta^{-2} \Sigma^Y - \frac{\varepsilon}{\delta} \Lambda \tag{3.1}$$

3.1 Upper bound

We derive an upper bound through the following triangular inequality: First, let us look at the error of estimation on Σ^Y. Here is a decomposition of this norm:

$$\begin{aligned} \left\| \hat{\Sigma}^Y - \Sigma^Y \right\| &= \left\| \left(\hat{\Sigma}^\delta - \Sigma^\delta \right) + \left(\hat{\Lambda}^\varepsilon - \mathbb{E} \hat{\Lambda}^\varepsilon \right) + \hat{\Sigma}^{X,\xi,\delta,\varepsilon} \right\| \\ &\leq \left\| \hat{\Sigma}^\delta - \Sigma^\delta \right\| + \left\| \hat{\Lambda}^\varepsilon - \mathbb{E} \hat{\Lambda}^\varepsilon \right\| + \left\| \hat{\Sigma}^{X,\xi,\delta,\varepsilon} \right\| \end{aligned} \tag{3.2}$$

where the three empirical matrices are:

1. $\hat{\Sigma}^\delta = n^{-1} \sum_{i=1}^n (\delta_i \otimes \delta_i) \odot (X_i \otimes X_i)$, the empirical covariance matrix of the $\delta_i \otimes X_i$;
2. $\hat{\Lambda}^\varepsilon = n^{-1} \sum_{i=1}^n (\varepsilon_i \otimes \varepsilon_i) \odot (\xi_i \otimes \xi_i)$, the empirical covariance of the $\varepsilon_i \odot \xi_i$;
3. $\hat{\Sigma}^{X,\xi,\delta,\varepsilon} = n^{-1} \sum_{i=1}^n (\delta_i \otimes \varepsilon_i) \odot (X_i \otimes \xi_i) + (\varepsilon_i \otimes \delta_i) \odot (\xi_i \otimes X_i)$, the empirical covariance terms between the $\delta_i \otimes X_i$ and the $(1 - \delta_i) \otimes \xi_i$, that should all convergence towards 0.

When bounding those three terms independently, and then by looking at the correction above, we find the following theorem:

Theorem 3 *For $t > 0$, with probability $1 - e^{-t}$:*

$$\begin{aligned}\delta \left\|\hat{\Sigma} - \Sigma\right\| \lesssim (\|\Sigma\| + \lambda) &\left(\sqrt{\frac{t}{n}} \vee \frac{t}{n}\right) \\ &+ \|\Sigma\| \left(\sqrt{\frac{\boldsymbol{r}(\Sigma)}{n}} \vee \frac{\boldsymbol{r}(\Sigma)}{n} \vee \sqrt{\frac{t}{n}} \vee \frac{t}{n}\right) \\ &+ \varepsilon p \sqrt{\lambda \|\Sigma\|} \left(\sqrt{\frac{t}{n}} \vee \frac{t}{n}\right) \\ &+ \varepsilon \left(\sqrt{\frac{p}{n}} \vee \frac{p}{n} \vee \sqrt{\frac{t}{n}} \vee \frac{t}{n}\right)\end{aligned} \tag{3.3}$$

Observe that, even in the highly unlikely case we know about Λ, we feel the full weight of the dimension p due to the contaminated data being of full rank.

3.2 Lower bound

Just as before, we adapt the lower bound proof in the missing data case to the contaminated case.

Theorem 4

$$\inf_{\hat{\Sigma}} \sup_{\mathbb{P}_\Sigma} \mathbb{P}_\Sigma \left[\left\|\hat{\Sigma} - \Sigma\right\| \geq C \left(\sqrt{\frac{r}{\delta n}} \wedge p \frac{r}{\varepsilon n}\right)\right] \geq \beta \tag{3.4}$$

This lower bound also makes us bear the weight of the full dimensionalaty. Since in the high dimensional setting we seek to avoid this at all costs, we clearly need to artificially reduce the value ε so that the terms in p become negligeable in the upper bound. Regarding the lower bound, for ε close enough to 0, the term in p will simplify if

$$\varepsilon \leq 1 \vee p \sqrt{\frac{\delta r}{n}} \tag{3.5}$$

Paradoxically, the higher the number of samples, the more error is induced by the contamination.

4 Filtering contaminations

We argue that cell-wise contamination can be managed by censuring contaminated data and using the missing value correction. A similar approach can be found in Farcomeni (2014), who develop a clustering method based on expectancy maximisation under the assumption that at least one sample from each cluster has no contamination.

We combined the detection method of Rousseeuw and den Bossche (2018) with the missing value estimator by setting all contaminated values to zero and by counting the number of detected contaminated values to get an estimate of δ. This estimator is called DDCMV in the rest of the paper. Another possible approach is to use the results of section 4.1 to randomly hide values in the dataset, so as to cover with high probability the outliers. This methods only works, however, if the contamination parameter is small. This estimator is called RandomMV in this paper.

4.1 Random covering of the error

In the previous sections, we supposed known the probability with which values are missed. Let us now set up the case with which several are contaminated at random, and we try to correct the estimation by randomly erasing values.

KL: Qu'est-ce qu'il passe quand on fait: 1) une partition multinomiales (K classes) entrees de la matrices. 2) on calcule les estimateurs de la matrice de cov sur chaque partition. 3) on fait une sorte de mediane ou trimmed mean sur les K estimateurs

In the case where the contamination probability ε of a value in the dataset being recorded with error, but where the contaminated values' position in the dataset are unknown, a simple calculus informs us that in order to delete all contaminated values with probability at least $Q \in [0, 1]$ one needs to set

$$\delta \leq \varepsilon^{-1} \left(1 - M^{1/np}\right) \tag{4.1}$$

For too large an ε, the risk is that δ will become too small for any meaningful estimation. However, cell-wise outlier detection algorithms exist, such as the one in Rousseeuw and den Bossche (2018). In section 5, we will compare the performances of a random hiding of data with high probability and the hiding of detected outlier using this method.

4.2 Robust Cell-Wise estimators

Since its introduction by Alqallaf et al. (2009), several estimators have been proposed to estimate the mean and covariance in a robust fashion under this contamination. In particular, the focus has been on detecting the outliers in the data using techniques similar to those developed for Huber contamination, but with extra steps (Farcomeni, 2014; Rousseeuw and den Bossche, 2018). However, most techniques use altered expectation maximisation algorithms and the Mahalanobis distance Raymaekers and Rousseeuw (2020), which is impractical in both high dimensional and low rank settings. Indeed, the Mahalanobis distance requires the inversion of the covariance matrix, which is unstable in a low-rank setting setting.

4.3 Adapting the Missing Values estimator

5 Experiments

In this section, we test the missing value correction based estimators on two types of cell-wise contamination: one based on the bernoulli setup described in section **??** and another adversarial perturbation of the data. We compare the performance of the estimators with that of state of the art Huber robust estimator TSGS (Agostinelli et al., 2014) and cell-wise robust estimator DI (Raymaekers and Rousseeuw, 2020). To provide an idea of the perturbation caused by the contaminations, we also provide the monte-carlo estimated bias of the classical covariance estimator.

5.1 Contamination models

In our experiments, we test our methods on three contamination models: two FICM contaminations with the outliers following a isotropic gaussian or a uniform distribution, and an adversarial perturbation of the data.

The adversarial model aims at disturbing the first eigenspace of the matrix. We find θ^{adv} a sparse projector of dimension $\epsilon * p$, such that $\theta^{\text{adv}}\theta^1 = 0$, ie the adversarial subspace and the first eigenspace are orthogonal.

6 Proof of upper bounds

6.1 Tools and definitions

6.1.1 Sub-exponential random vectors

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

Definition 1 *For any $\alpha \geq 1$, the ψ_α-norms of a real-valued random variable V are defined as:*

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

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.

Here are two well-known properties of sub-exponential random variables:

1. For any real-valued variable V such that $\|V\|_\alpha < \infty$ for some $\alpha > 1$, we have
$$\mathbb{E}\left[|V|^m\right] \leq 2\frac{m}{\alpha}\Gamma\left(\frac{m}{\alpha}\right)\|V\|_{\psi_\alpha}^m \qquad \forall m \geq 1,$$
where $\Gamma(\dot{})$ is the Gamma function.

2. If a real-valued random variable V is sub-Gaussian, then V^2 is sub-exponential. Indeed, we have:
$$\left\|V^2\right\|_{\psi_1} \leq 2\left\|V\right\|_{\psi_2}^2$$

Definition 2 *A random vector $X \in \mathbb{R}^p$ is sub-exponential if $\langle X, x\rangle$ are sub-exponential random variables for $x \in \mathbb{R}^p$. The ψ_α-norms of a random vector X are defined as:*
$$\|X\|_{\psi_\alpha} = \sup_{x\in\mathbb{R}^p, |x|_2=1} \|\langle X, x\rangle\|_{\psi_\alpha}, \qquad \alpha \geq 1$$

Bernstein's inequality can be adapted to the matrix setup as follows (see corollary 5.17 in Vershynin (2011)):

Proposition 1 *Let $X_1, \dots X_n$ be sub-exponential random variables and $K = \max_i \|X_i\|_{\psi_1}$, then, for $t > 0$, with probability at least $1 - e^{-t}$:*
$$\left|n^{-1}\sum_{i=1}^n Y_i\right| \leq CK\left(\sqrt{\frac{t}{n}} \vee \frac{t}{n}\right) \tag{6.1}$$
where C is an absolute constant.

6.1.2 Talagrand's chaining and covariance concentration inequalities

Talagrand's work on generic chaining complexities for empirical processes allows for sharper upper bounds on covariance matrix estimation (Koltchinskii and Lounici, 2014). In this section we introduce a form of Talagrand's theorem adapted to empirical processes.

Let (T, d) be a metric space and let $N_n := 2^{2^n}, n \geq 1$ and $N_0 := 1$. We define an increasing sequence Δ_n of partitions of T as admissible if, and only if, $\mathrm{card}(\Delta_n) \leq N_n$. Given such a sequence, for $t \in T$, let $\Delta_n(t)$ be the unique set of a Δ_n containing t. Let us finally define:
$$\gamma_2(T, d) = \inf\sup_{t\in T}\sum_{n=0}^{\infty} 2^{n/2} D\left(\Delta_n(t)\right), \tag{6.2}$$
where D denotes the diameter of a subset of T and the infimum is taken over all admissible sequences. Talagrand's result can then be stated as follows:

Proposition 2 *Let $X(t), t \in T$ be a centered Gaussian process and let, for $t, s \in T$,*
$$d(t, s) := \mathbb{E}^{1/2}\left(X(t) - X(s)\right)^2 \tag{6.3}$$
Then there exists an absolute constant $K > 0$ such that
$$\mathbb{E}\sup_{t\in T} X(t) \geq K^{-1}\gamma_2(T, d)$$

One can apply this fact to the case where $T = \mathcal{F}$ is a class of functions on a probability space $(S, \mathcal{A}, P)$ to provide upper bounds on the error when estimating covariance matrices (theorem 8, Koltchinskii and Lounici (2014)):

Proposition 3 *Let $X, X_1, \ldots, X_n$ be i.i.d. random variables in S with distribution P. Let $\mathcal{F}$ be a class of measurable functions on $(S, \mathcal{A})$ such that $f \in \mathcal{F}$ implies $-f \in \mathcal{F}$ and $\mathbb{E}f(X) = 0$. Then there exist a constant C such that, for $t > 0$, with probability at least $1 - e^{-t}$:*

$$\sup_{f\in\mathcal{F}} \left| n^{-1} \sum_{i=1}^{n} f^2(X_i) - \mathbb{E}f^2(X) \right| \leq C \max\left\{ \sup_{f\in\mathcal{F}} \|f\|_{\psi_1} \frac{\gamma_2(\mathcal{F}, \psi_2)}{\sqrt{n}}, \frac{\gamma_2^2(\mathcal{F}, \psi_2)}{n}, \sup_{f\in\mathcal{F}} \|f\|_{\psi_1}^2 \sqrt{\frac{t}{n}}, \sup_{f\in\mathcal{F}} \|f\|_{\psi_1}^2 \frac{t}{n} \right\} \tag{6.4}$$

6.2 Proof of theorem 1

Let $X_1, \ldots, X_n$ be i.i.d. zero mean vectors following a $\mathcal{N}(0, \Sigma)$ law, with Σ an unknown positive definite Hermitian matrix of size $p \times p$. Let for $1 \leq i \leq n$ and $1 \leq j \leq p$, δ_{ij} follows an Bernoulli lax $\mathcal{B}(\delta)$, with $\delta \in [0, 1]$, such that δ_{ij} is independent both from $X_i^{(j)}$, that is the jth component of X_i, and of any other Bernoulli random variable. Let finally $Y_i^{(h)} = \delta_{ij} X_i^{(j)}$ the observed random variable with missing values. We will denote by $\lesssim$ the fact that the left side term is dominated by the right side term.

6.2.1 Proof of lemma 1

By definition of the operator norm, we can express this error in terms of Rayleigh's quotient:

$$\begin{aligned} \left\| n^{-1} \sum_i Y_i \otimes Y_i - \mathbb{E}\left[Y \otimes Y\right] \right\|_2 &= \max_{\|u\|=1} u \left(n^{-1} \sum_i Y_i \otimes Y_i - \mathbb{E}\left[Y \otimes Y\right] \right) u^\top \\ &= \max_{\|u\|=1} n^{-1} \sum_i u(Y_i \otimes Y_i) u^\top - u\mathbb{E}\left[Y \otimes Y\right] u^\top \\ &= \max_{\|u\|=1} n^{-1} \sum_i \langle \delta_i \odot X_i, u\rangle^2 - u^\top \mathbb{E}\left[Y \otimes Y\right] u \end{aligned} \tag{6.5}$$

Let X and $\tilde{\delta}$ be two random variables of same distribution to, respectively, the X_i and δ_i. We can rewrite the expectation as:

$$\begin{aligned} u^\top \mathbb{E}\left[Y \otimes Y\right] u &= \mathbb{E} u^\top (\tilde{\delta} \odot X) \otimes (\tilde{\delta} \odot X) u \\ &= \mathbb{E}\langle \tilde{\delta} \odot X, u\rangle^2 \end{aligned} \tag{6.6}$$

Let $\mathcal{F} = \{\langle \cdot, u\rangle, \|u\| \leq 1\}$. Since X is subgaussian, $\tilde{\delta} \odot X$ is too. This means that the ψ_1 and ψ_2 norms of linear functionals $\langle \tilde{\delta} \odot X, u\rangle$ are both equivalent to the L_2-norm. Thus:

$$\sup_{f\in\mathcal{F}} \|f\|_{\psi_1} \lesssim \sup_{\|u\|\leq 1} \mathbb{E}^{1/2} \langle \tilde{\delta} \odot X, u\rangle^2 \leq \mathbb{E}^{1/2} \left\| \tilde{\delta} \odot X \right\|^2 = \mathbb{E}^{1/2} \sum_{i=1}^{p} \tilde{\delta}_i^2 X_i^2 \tag{6.7}$$

Since $\tilde{\delta}$ is a Boolean vector, $\forall i, \tilde{\delta}_i^2 = \tilde{\delta}_i$. Thus, by the tower property:

$$\begin{aligned} \sup_{f\in\mathcal{F}} \|f\|_{\psi_1} &\lesssim \mathbb{E}^{1/2} \mathbb{E}_{\tilde{\delta}} \sum_{i=1}^{p} \tilde{\delta}_i X_i^2 \\ &= \mathbb{E}^{1/2} \delta \left\|X\right\|^2 \\ &\leq \sqrt{\delta \left\|\Sigma\right\|} \end{aligned} \tag{6.8}$$

Now let us focus on $\gamma_2(\mathcal{F}, \psi_2)$. The norm equivalence and Talagrand's theorem (property 2) gives:

$$\gamma_2(\mathcal{F}, \psi_2) \lesssim \gamma_2(\mathcal{F}, L_2) \lesssim \mathbb{E} \sup_{\|u\| \leq 1} \langle \tilde{\delta} \odot X, u \rangle \leq \sqrt{\delta} \mathbb{E} \|X\| \tag{6.9}$$

Thus, by proposition 6.4, there exist an absolute constant c_1 such that, for $t > 0$, with probability at least $1 - e^{-t}$:

$$\begin{aligned} \mathbb{E} \left\| \hat{\Sigma}^Y - \Sigma^Y \right\| &\lesssim \max \left\{ \sqrt{\delta} \|\Sigma\|^{1/2} \frac{\sqrt{\delta} \mathbb{E} \|X\|}{\sqrt{n}}, \frac{\delta \mathbb{E} \|X\|^2}{n}, \delta \|\Sigma\| \sqrt{\frac{t}{n}}, \delta \|\Sigma\| \frac{t}{n} \right\} \\ &= \delta \|\Sigma\| \left(\sqrt{\frac{\boldsymbol{r}(\Sigma)}{n}} \vee \frac{\boldsymbol{r}(\Sigma)}{n} \vee \sqrt{\frac{t}{n}} \vee \frac{t}{n} \right) \end{aligned} \tag{6.10}$$

6.2.2 Proof of lemma 2

Since taking the operator norm of a diagonal matrix is the same as taking the operator norm of the vector containing the diagonal values, we get:

$$\left\| \operatorname{diag} \left(n^{-1} \sum_{1}^{n} Y_i \otimes Y_i \right) - \operatorname{diag} \left(\mathbb{E} Y \otimes Y \right) \right\| = \max_{j = \{1, \ldots p\}} \left| n^{-1} \sum_{i=1}^{n} \left(\delta_i^{(j)} X_i^{(j)} \right)^2 - \delta \Sigma_{jj} \right| \tag{6.11}$$

Since, for a any given $j \in \{1 \ldots\}$, $X^{(j)}$ is sub-gaussian, we have:

$$\left\| \left(\delta_i^{(j)} X_i^{(j)} \right)^2 \right\|_{\psi_1} \leq 2 \left\| \delta_i^{(j)} X_i^{(j)} \right\|_{\psi_2}^2 \leq 2 \left\| X_i^{(j)} \right\|_{\psi_2}^2 \leq 2 c_2^{-1} \Sigma_{jj} \tag{6.12}$$

for c_2 a constant. Bernstein's inequality as introduced in proposition 1 tells us that, for $t > 0$, with probability at least $1 - e^{-t}$, there exist an absolute constant c_2 such that:

$$\left\| \operatorname{diag} \left(\hat{\Sigma}^Y - \Sigma^Y \right) \right\| \leq c_2 \max_j \Sigma_{jj} \left(\sqrt{\frac{t}{n}} \vee \frac{t}{n} \right) \tag{6.13}$$

Furthermore, note that $\max_j \Sigma_{jj} \leq \|\Sigma\|$.

6.2.3 Proof of theorem 1

Now that we have proven lemmas 1 and 2, we can combine them to obtain the final upper bound.

We are looking for an upper bound on:

$$\begin{aligned} \left\| \hat{\Sigma} - \Sigma \right\| &= \left\| (\delta^{-1} - \delta{-}2) \operatorname{diag} \left(\hat{\Sigma}^Y - \Sigma^Y \right) + \delta^{-2} \left(\hat{\Sigma}^Y - \Sigma^Y \right) \right\| \\ &\leq (\delta^{-1} - \delta^{-2}) \left\| \operatorname{diag} \left(\hat{\Sigma}^Y - \Sigma^Y \right) \right\| + \delta^{-2} \left\| \hat{\Sigma}^Y - \Sigma^Y \right\| \\ &\leq \delta^{-1} \left\| \operatorname{diag} \left(\hat{\Sigma}^Y - \Sigma^Y \right) \right\| + \delta^{-2} \left\| \hat{\Sigma}^Y - \Sigma^Y \right\| \end{aligned} \tag{6.14}$$

Combining lemmas 1 and 2 with a union bound argument, and by reajusting the constants, we get that, for $t > 0$, with probability at least $1 - e^{-t}$:

$$\left\| \hat{\Sigma} - \Sigma \right\| \leq C \frac{\|\Sigma\|}{\delta} \left(\sqrt{\frac{\boldsymbol{r}(\Sigma)}{n}} \vee \frac{\boldsymbol{r}(\Sigma)}{n} \vee \sqrt{\frac{t}{n}} \vee \frac{t}{n} \right) \tag{6.15}$$

with $C > (c1 \vee c2)$ an absolute constant.

6.3 Proof of the upper bound in the contaminated case

6.3.1 Bounding the error on the full matrix

Using the previous result, we know that, with probability at least $1 - e^{-t}$ and for an absolute constant C:

$$\left\|\hat{\Sigma}^{\delta} - \Sigma^{\delta}\right\| \leq \delta C \left\|\Sigma\right\| \left(\sqrt{\frac{\boldsymbol{r}(\Sigma)}{n}} \vee \frac{\boldsymbol{r}(\Sigma)}{n} \vee \sqrt{\frac{t}{n}} \vee \frac{t}{n}\right) \tag{6.16}$$

and

$$\left\|\hat{\Lambda}^{\varepsilon} - \mathbb{E}\hat{\Lambda}^{\varepsilon}\right\| \leq \varepsilon C \left(\sqrt{\frac{p}{n}} \vee \frac{p}{n} \vee \sqrt{\frac{t}{n}} \vee \frac{t}{n}\right) \tag{6.17}$$

Now we need to control the norm of $\hat{\Sigma}^{X,\xi,\delta,\varepsilon}$. Notice that this matrix has no diagonal terms, since $\delta_i^{(j)}$ and $\varepsilon_i^{(j)}$ cannot be both equal to 1. For $l, k \in \{1, \ldots, p\}$, the distribution of the $d_i^l(1 - d_i^k)X_i^{(l)}\xi_i^{(k)}$ is sub-exponential, since, as shown in appendix 8.1,

$$\left\|\delta_i^l \varepsilon_i^k X_i^{(l)} \xi_i^{(k)}\right\|_{\psi_1} \leq \delta\varepsilon \left\|X_i^{(l)}\right\|_{\psi_2} \left\|\xi_i^{(k)}\right\|_{\psi_2} < \infty \tag{6.18}$$

which comes from the fact both the X_i and ξ are sub-gaussian. Bernstein's inequality then gives:

$$\left\|\hat{\Sigma}^{X,\xi,\delta,\varepsilon}\right\|_{\max} \leq \max_{1 \leq j \leq p} c_3 \delta\varepsilon \sqrt{\lambda \Sigma_{jj}} \left(\sqrt{\frac{t}{n}} \vee \frac{t}{n}\right) \tag{6.19}$$

6.3.2 Bounding the diagonal terms

The diagonal term can be bounded using Bernstein's inequality. Since, for any given j, $X^{(j)} + \xi^{(j)}$ is sub-gaussian:

$$\begin{aligned}\left\|\left(\delta_i^{(j)} \odot X_i^{(j)} + \varepsilon_i \odot \xi_i^{(j)}\right)^2\right\|_{\psi_1} &\leq 2\left\|\delta_i^{(j)} X_i^{(j)} + \varepsilon_i \odot \xi_i^{(j)}\right\|_{\psi_2}^2 \leq \left(\left\|\delta_i^{(j)} X_i^{(j)}\right\|_{\psi_2} + \left\|\varepsilon_i^{(j)} \xi_i^{(j)}\right\|_{\psi_2}\right)^2 \\ &\leq 2\left(\left\|X_i^{(j)}\right\|_{\psi_2} + \left\|\xi_i^{(j)}\right\|_{\psi_2}\right)^2 \leq 2c_2^{-1}(\sqrt{\Sigma_{jj}} + \sqrt{\lambda}))^2\end{aligned} \tag{6.20}$$

we get, using Bernstein's inequality:

$$\left\|\operatorname{diag}\left(\hat{\Sigma}^{Y} - \Sigma^{Y}\right)\right\| \leq c_2 \left(\max_j \sqrt{\Sigma_{jj}} + \sqrt{\lambda}\right)^2 \left(\sqrt{\frac{t}{n}} \vee \frac{t}{n}\right) \tag{6.21}$$

Finally, similarly to the previous case and since the terms in I_p simplify, we have that:

$$\left\|\hat{\Sigma} - \Sigma\right\| = \left\|(\delta^{-1} - \delta^{-2})\operatorname{diag}\left(\hat{\Sigma}^{Y} - \Sigma^{Y}\right) + \delta^{-2}\left(\hat{\Sigma}^{Y} - \Sigma^{Y}\right)\right\| \tag{6.22}$$

7 Proof of theorem 2

This demonstration takes inspiration to the lower bound proof of Lounici (2014), which we improved by changing the set of hypotheses with the ideas of Koltchinskii et al. (2015).

7.1 Hypothesis construction in a Grassmannian manifold

Let $p \geq 2$ be the dimension of our observations and let $1 \leq r \leq p$ be the intrinsic dimension of Σ. Although the problem at hand is p-dimensional, we are most interested in correctly estimating the r eigenspaces related to the largest eigenvalues. We will thus look at p dimensional matrices that are projection in $\mathbb{R}^p$ of r dimensional kernels.

We set γ to be a constant larger than 0. Let H be a $p \times r$ matrix with orthonormal rows. Each matrix H describes a subspace U_H of $\mathbb{R}^p$, where $\dim(U_H) = r$ and $H^\top H$ is its projector in $\mathbb{R}^p$. The set of all U_H is the Grassmannian manifold $G_r(\mathbb{R}^p)$, which is the set of all r-dimensional subspaces of $\mathbb{R}^p$. The Grassmannian manifold is a smooth manifold of dimension $d = r(p-r)$, where one can define a metric for all subspaces $U, \bar{U} \in G_r(\mathbb{R}^p)$:

$$d(U, \bar{u}) = \|P_U - P_{\bar{U}}\|_F = \left\|H^\top H - \bar{H}^\top \bar{H}\right\| \tag{7.1}$$

where P_U and $P_{\bar{U}}$ are the projectors to the subspaces U and $\bar{U}$ respectively and H and $\bar{H}$ are the $r \times p$ matrix with orthonormal rows associated with U and $\bar{U}$ respectively. In the remainder of the proof, we will identify the projectors to the subspaces. A result on the entropy of Grassmanian manifolds (Pajor, 1998) shows that:

Proposition 4 *For all $\varepsilon > 0$, there exists a family of orthonormal projectors $\mathcal{U} \subset G_r(\mathbb{R}^p)$ such that:*

$$|\mathcal{U}| \geq \left\lfloor \frac{\bar{c}}{\varepsilon} \right\rfloor^d \tag{7.2}$$

and, $\forall P, Q \in G_r(\mathbb{R}^p), P \neq Q$,

$$\bar{c}\varepsilon\sqrt{r} \leq \|P - Q\|_F \leq \frac{\varepsilon\sqrt{r}}{\bar{c}} \tag{7.3}$$

for some small enough absolute constant $\bar{c}$, where $|\mathcal{U}|$ is the cardinal of set $\mathcal{U}$.

Without loss of generality, we can suppose that the block matrix $P_1 = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}$ belongs to the set. Indeed, change of basis doesn't impact the Frobenius norm.

Let us then build such a set $\mathcal{U}$ of hypotheses. Let $\gamma = a\sqrt{p/\delta n}$, for $a > 0$ a constant. We set $N = |\mathcal{U}|$ and $\mathcal{U} = \{P_1, \dots, P_N\}$. Remember that P_1 is the diagonal matrix with r diagonal values equal to 1 and the rest to 0. Let us define the family $\Sigma_1, \dots, \Sigma_N$ of symmetric $p \times p$ approximately low rank covariances matrices such that, $\forall j \in \{1, N\}, \Sigma_j = \sigma I_p + \gamma P_j$, where I_p is the p dimensional identity matrix. These matrices are the superposition of two domains, one of low eigenvalues akin to a isotropic noise, and a one with large eigenvalues which acts as the signal. To ensure that the effective rank of those matrices is controlled by r, one must choose σ sufficiently small.

Then, we can see that, for $i, j \in \{1, \dots N\}$, by setting $\varepsilon = 1/2$:

$$\|\Sigma_i - \Sigma_j\|_F^2 = \gamma^2 \|P_i - P_j\|_F^2 > a^2 \bar{c}^2 \frac{pr}{2\delta n} \tag{7.4}$$

7.2 KL-divergence of hypotheses

Now that we have our candidate covariances $\Sigma_0, \dots, \Sigma_N$, let us define the associated distributions. For $j \in \{0, N\}$, let $X_1, \dots X_n$ be i.i.d. random variables following a gaussian $\mathcal{N}(0, \Sigma_j)$ law. Let $\delta_1, \dots \delta_n$ be each vectors of p i.i.d bernoulli random variables of probability of success δ, and let $Y_1, \dots Y_n$ be random variables such that, $\forall i \in \{1, n\}, Y_i = \delta_i \odot X_i$, with $\odot$ the Hadamard or term-by-term product. Let us also define as $\mathbb{P}_j$ the distribution of $Y_1, \dots Y_n$ and $\mathbb{P}_j^{(\delta)}$ the conditional distribution of the $Y_1, \dots Y_n$ knowing $\delta_1, \dots \delta_n$. Finally, let $\mathbb{E}_j$ be the expectation given the distribution associated with the j-th projector and $\mathbb{E}_\delta$ the expectation given $\delta_1, \dots \delta_n$.

For $j \in \{2, \dots, N\}$, let us compute the Kullback-Leibler divergence from $\mathbb{P}_1$ to $\mathbb{P}_j$.

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_1, \mathbb{P}_j) &= \mathbb{E}_1 \log\left(\frac{d\mathbb{P}_1}{d\mathbb{P}_j}\right) = \mathbb{E}_1 \log\left(\frac{d\mathbb{P}_\delta \otimes \mathbb{P}_1^{(\delta)}}{d\mathbb{P}_\delta \otimes \mathbb{P}_j^{(\delta)}}\right) \\ &= \mathbb{E}_\delta \mathrm{KL}(\mathbb{P}_1^{(\delta)}, \mathbb{P}_j^{(\delta)}) = \sum_{i=1}^{n} \mathbb{E}_\delta \mathrm{KL}(\mathbb{P}_1^{(\delta_i)}, \mathbb{P}_j^{(\delta_i)})\end{aligned} \tag{7.5}$$

Since $\forall i \in \{1, \dots, n\}$, $Y_i | \delta_i \sim \mathcal{N}(0, (\delta_i \otimes \delta_i) \odot \Sigma)$, for all $j \in \{1, \dots N\}$ and for each realisation $\delta(\omega) \in \{0,1\}^p$, $\mathbb{P}_j \gg \mathbb{P}_1$, thus $\mathrm{KL}(\mathbb{P}_1, \mathbb{P}_j) < \infty$.

Define $J_i = \{j : \delta_{i,k} = 1, 1 \leq k \leq p\}$ the set of indices kept by vector δ_i and $d_i = \sum_{k=1}^{p} \delta_i^k \sim \mathcal{B}(p, \delta)$. Then, define the mapping $Q_i : \mathbb{R}^p \to \mathbb{R}^{d_i}$ such that $Q_i(x) = x_{J_i}$, such that x_{J_i} is a d_i dimensional vector containing the components of x whose index are in J_i. Let $Q_i^* : \mathbb{R}^{d_i} \to \mathbb{R}^p$ the right inverse of Q_i.

Note that $\forall j \in \{1, N-1\}$, $\Sigma_j = (\sigma + \gamma) P_j + \sigma P_j^\perp$, with $P_j^\perp$ the projector to the subspace of $\mathbb{R}^p$ orthogonal to the one described by P_j. Let us define $\Sigma_j^{(\delta_i)} = Q_i \Sigma_j Q_i^*$. Then, observe that $\Sigma_1^{(\delta_i)}$ is invertible, with inverse $Q_i \left(\frac{1}{\gamma+\sigma} P_1 + \frac{1}{\sigma} P_1^\perp\right) Q_i^*$ since P_1 and $P_1^\perp$ are diagonal matrices. We thus get, for $i \in \{1, \dots n\}$:

GP: Est-ce qu'on peut poser $\sigma = 1$ dans le reste de la preuve? ça simplifirait la vie pour le controle du log det dans Pinsker

$$\mathrm{KL}(\mathbb{P}_1^{(\delta_i)}, \mathbb{P}_j^{(\delta_i)}) = \frac{1}{2}\left(\mathrm{tr}\left({\Sigma_1^{(\delta_i)}}^{-1} \Sigma_j^{(\delta_i)}\right) - d_i - \log(\det({\Sigma_1^{(\delta_i)}}^{-1} \Sigma_j^{(\delta_i)}))\right) \tag{7.6}$$

First, using a result of linear algebra described in section 8.3, we show that:

GP: A revoir, des erreurs de raisonements dans la preuve

$$\begin{aligned}-\mathbb{E}_\delta \log(\det({\Sigma_1^{(\delta_i)}}^{-1} \Sigma_j^{(\delta_i)})) &\leq (p - r)(\delta - 1) \log \gamma + r\delta \log \sigma \\ &\leq r\delta \log \gamma \\ &\lesssim \frac{rp}{n}\end{aligned} \tag{7.7}$$

Next, let us focus on bounding $\frac{1}{2}\mathrm{tr}\left({\Sigma_1^{(\delta_i)}}^{-1}(\Sigma_j^{(\delta_i)} - \Sigma_1^{(\delta_i)})\right)$. Using the fact that $\Sigma_1^{-1} = \frac{1}{\sigma+\gamma} P_1 + \frac{1}{\sigma} P_1^\perp$, we get:

$$\begin{aligned}\mathrm{tr}\left({\Sigma_1^{(\delta_i)}}^{-1}(\Sigma_j^{(\delta_i)} - \Sigma_1^{(\delta_i)})\right) &= \frac{\gamma}{\sigma+\gamma}\mathrm{tr}\left(Q_i P_1 (P_j - P_1) Q_i^*\right) + \frac{\gamma}{\sigma}\mathrm{tr}\left(Q_i P_1^\perp (P_j - P_1) Q_i^*\right) \\ &= \frac{\gamma}{\sigma+\gamma}\left(\mathrm{tr}\left(Q_i P_1 P_j Q_i^*\right) - \mathrm{tr}\left(Q_i P_1 Q_i^*\right)\right) + \frac{\gamma}{\sigma}\mathrm{tr}\left(Q_i \left(I_p - P_1\right) P_j Q_i^*\right) \\ &= \left(\frac{\gamma}{\sigma+\gamma} - \frac{\gamma}{\sigma}\right)\left(\mathrm{tr}\left(Q_i P_1 P_j Q_i^*\right) - d_i\right) \\ &= \frac{\gamma^2}{2(\sigma+\gamma)\sigma} \left\|Q_i (P_j - P_1) Q_i^*\right\|_F^2\end{aligned} \tag{7.8}$$

Finally, using the fact demonstrated in appendix 8.4 and the upper bound of proposition 4, we get that:

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_1, \mathbb{P}_j) &\leq \sum_{i=1}^{n} \mathbb{E}_\delta \frac{\gamma^2}{2(\sigma+\gamma)\sigma} \left\|Q_i (P_j - P_1) Q_i^*\right\|_F^2 \\ &\leq \sum_{i=1}^{n} \frac{\gamma^2 \delta}{2(\sigma+\gamma)\sigma} \left\|P_j - P_1\right\|_F^2 \\ &\leq \sum_{i=1}^{n} \frac{\gamma^2 \delta r}{8\bar{c}^2(\sigma+\gamma)\sigma} \\ &= \frac{a^2 rp}{8\bar{c}^2(\sigma+\gamma)\sigma}\end{aligned} \tag{7.9}$$

Thus, since $N \geq \lfloor 2\bar{c} \rfloor^{r(p-r)}$, and as we can set, w.l.o.g. $p > 2r$:

$$\mathrm{KL}(\mathbb{P}_1, \mathbb{P}_j) \leq \alpha \log(N) \tag{7.10}$$

for any α so long as $a > 0$ is chosen small enough. Along with equation 7.4, according to theorem 2.5 of Tsybakov (2009), we get that:

$$\inf_{\hat{\Sigma}} \sup_{\mathbb{P}_\Sigma} \mathbb{P}_\Sigma \left(\left\| \hat{\Sigma} - \Sigma \right\|_F^2 \geq C \frac{r}{\delta n} p \right) \geq \beta \tag{7.11}$$

with C and β two absolute constants. This fact, in turn, implies the lower bound of theorem 2, since, for all Σ_1, Σ_2 matrices of our hypothesis set:

$$\|\Sigma_1 - \Sigma_2\|^2 \geq C \frac{r}{\delta n} \tag{7.12}$$

Indeed, otherwise, we would get:

$$\|\Sigma_1 - \Sigma_2\|_F^2 < p \|\Sigma_1 - \Sigma_2\|^2 < C \frac{r}{\delta n} p \tag{7.13}$$

which contradicts equation 7.4.

7.3 Proof of theorem ??

This proof is based on that of theorem 2 in Lounici (2014), where we change slightly the construction of the distributions of which we bound the KL-divergence.

Let $X_1, \dots X_n \in \mathbb{R}^p$ be i.i.d. samples of $\mathcal{N}(0, \Sigma_j)$, with $\Sigma_j \in \mathcal{A}^0$. Let $\varepsilon_1, \dots \varepsilon_0$ be random vectors with i.i.d. entries ε_i, j following $B(\varepsilon)$. Finally, let $\delta_i = 1 - \varepsilon_i$, and :

$$Y_i = \delta_i \odot X_i + \varepsilon_i \odot \xi \tag{7.14}$$

with $\xi_1, \dots \xi_n$ be i.i.d. samples of a distribution with variance the diagonal matrix V.

We can show that the empirical variance $\Sigma_1^{(\delta_i)}$ is a diagonal matrix, with d_i diagonal terms with value γ, d_i' values equal to 1 and $p - d_i - d_i'$ values equal to the variance of the noise term, where $d_i \sim \mathcal{B}(r, \delta)$ and $d_i' \sim \mathcal{B}(p-r, \delta)$ representing the values that the mask δ_i has kept in the subspace P_1 and $P_1^\top$ respectively, the rest being filled with the contamination. This shows that this matrix is invertible and, as done in the previous section, we get:

$$\mathrm{KL}\left(\mathbb{P}_1^{(\delta_i)}, \mathbb{P}_j^{(\delta_i)}\right) = \frac{1}{2} \mathrm{tr}\left(\Sigma_1^{(\delta_i)^{-1}} (\Sigma_j^{(\delta_i)} - \Sigma_1^{(\delta_i)}) \right) \tag{7.15}$$

However, here we have that $\Sigma_1^- 1 = (\delta_i \otimes \delta_i) \odot \left((1+\gamma)^{-1} P_1 + P_1^\top\right) + (\varepsilon_i \otimes \varepsilon_i) \odot V^{-1}$. Thus:

$$\begin{aligned} \mathrm{tr}\left(\Sigma_1^{(\delta_i)^{-1}} (\Sigma_j^{(\delta_i)} - \Sigma_1^{(\delta_i)}) \right) = &\left(\frac{\gamma}{1+\gamma} - \gamma \right) \mathrm{tr}\left((\delta_i \otimes \delta_i) \odot (P_1 P_j - I_p) \right) \\ &+ \gamma \mathrm{tr}\left((\varepsilon_i \otimes \varepsilon_i) \odot \left(V^{-1} \left(P_j - P_1 \right) \right) \right) \end{aligned} \tag{7.16}$$

We already know that

$$\mathbb{E}_\delta \mathrm{tr}\left((\delta_i \otimes \delta_i) \odot (P_1 P_j - I_p) \right) = \mathbb{E}_\delta \|Q_i (P_j - P_1) Q_i^*\|_F^2 \leq \frac{(1-\varepsilon) r}{4\bar{c}^2} \tag{7.17}$$

Furthermore, due to the same reasoning as in appendix 8.4 and using the fact that V is diagonal:

$$\begin{aligned} \mathbb{E}_\varepsilon \mathrm{tr}\left((\varepsilon_i \otimes \varepsilon_i) \odot \left(V^{-1} \left(P_j - P_1 \right) \right) \right) &\leq \varepsilon \mathrm{tr}\left(V^{-1} \left(P_j - P_1 \right) \right) \\ &\leq \varepsilon r \left(\max_i V_i^{-1} - \min_i V_i^{-1} \right) \end{aligned} \tag{7.18}$$

which gives by property 4:

$$\mathbb{E}_\varepsilon \mathrm{tr}\left((\varepsilon_i \otimes \varepsilon_i) \odot \left(V^{-1}\left(P_j - P_1\right)\right)\right) \leq \left(\max_i V_i^{-1} - \min_i V_i^{-1}\right) \frac{\varepsilon\sqrt{r}}{2\bar{c}} \tag{7.19}$$

Remember that $\delta = 1-\varepsilon$. Let us also call $b = \left(\max_i V_i^{-1} - \min_i V_i^{-1}\right)$, which is a constant. Thus, by setting $\gamma = a\left(\sqrt{\frac{p}{\delta n}} \wedge \frac{p\sqrt{r}}{\varepsilon n}\right)$, we get:

$$\mathbb{E}_\delta \mathrm{KL}\left(\mathbb{P}_1^{(\delta_i)}, \mathbb{P}_j^{(\delta_i)}\right) \leq \frac{rp}{n}\left[\frac{a^2}{4\bar{c}^2(1+\gamma)} + \frac{ab}{2\bar{c}}\right] \tag{7.20}$$

For a sufficiently small we verify the upper bound condition of theorem 2.5 of Tsybakov (2009). Given this expression of γ, we find the lower bound condition: $\forall i, j \in \{1, N\}$

$$\|\Sigma_i - \Sigma_j\|_F^2 \leq \gamma^2 \|P_i - P_j\|_F^2 \leq a^2 rp\left(\frac{1}{(1-\varepsilon)n} \wedge \frac{pr}{\varepsilon^2 n^2}\right) \tag{7.21}$$

Thus, we have that:

$$\inf_{\hat{\Sigma}} \sup_{\mathbb{P}_\Sigma} \mathbb{P}_\Sigma\left[\left\|\hat{\Sigma} - \Sigma\right\|_F^2 \geq C\left(\frac{rp}{(1-\varepsilon)n} \wedge \left(\frac{rp}{\varepsilon n}\right)^2\right)\right] \geq \beta \tag{7.22}$$

and

$$\inf_{\hat{\Sigma}} \sup_{\mathbb{P}_\Sigma} \mathbb{P}_\Sigma\left[\left\|\hat{\Sigma} - \Sigma\right\|^2 \geq C\left(\frac{r}{(1-\varepsilon)n} \wedge p\left(\frac{r}{\varepsilon n}\right)^2\right)\right] \geq \beta \tag{7.23}$$

8 Other proofs

8.1 Orlicz 1-norm of the components of $\Sigma^{X,\xi,\delta}$

Let X and ξ be two one dimensional random variables following a sub-gaussian distribution, and let d be a bernoulli random variable of mean δ. The Orlicz ψ_1 norm of $d(1-d)X\xi$ is:

$$\begin{aligned} \|d(1-d)X\xi\|_{\psi_1} &= \inf\{u > 0, \mathbb{E}\exp\left(|d(1-d)X\xi|/u\right) \leq 2\} \\ &= \inf\{u > 0, \mathbb{E}\exp\left(d(1-d)|X\xi|/u\right) \leq 2\} \end{aligned} \tag{8.1}$$

Since the bernoulli variables are binary. By the tower property and Jensen equality, we have that, $\forall u$ such that the expectation is well defined :

$$\begin{aligned} \mathbb{E}\exp\left(d(1-d)|X\xi|/u\right) &= \mathbb{E}\mathbb{E}_\delta \exp\left(|d(1-d)X\xi|/u\right) \\ &\geq \mathbb{E}\exp\left(\delta(1-\delta)|X\xi|/u\right) \end{aligned} \tag{8.2}$$

which implies that

$$\{u > 0, \mathbb{E}\exp\left(\delta(1-\delta)|X\xi|/u\right) \leq 2\} \subset \{u > 0, \mathbb{E}\exp\left(d(1-d)|X\xi|/u\right) \leq 2\} \tag{8.3}$$

With a simple change of variable, one can see that:

$$\begin{aligned} \inf\{u > 0, \mathbb{E}\exp\left(\delta(1-\delta)|X\xi|/u\right) \leq 2\} &= \delta(1-\delta)\inf\{u > 0, \mathbb{E}\exp\left(|X\xi|/u\right) \leq 2\} \\ &= \delta(1-\delta)\|X\xi\|_{\psi_1} \\ &\leq \delta(1-\delta)\|X\|_{\psi_2}\|\xi\|_{\psi_2} \end{aligned} \tag{8.4}$$

Hence:

$$\inf\{u > 0, \mathbb{E}\exp\left(d(1-d)|X\xi|/u\right) \leq 2\} \leq \delta(1-\delta)\|X\|_{\psi_2}\|\xi\|_{\psi_2} \tag{8.5}$$

8.2 Proof of the correction formula of equation ??

Let $Y = (\delta_1 \odot X^{(1)} + (1-\delta_1) \odot \xi^{(1)}, \dots, \delta_n \odot X^{(n)} + (1-\delta_n) \odot \xi^{(n)})$ with X and ξ in $\mathbb{R}^{n \times p}$ and δ some p dimensional binary vector.

Thus,

$$(Y \otimes Y)_{jk} = \begin{cases} \left(X^{(j)}\right)^2 & \text{if } j = k \text{ and } \delta_j = 1 \\ \left(\xi^{(j)}\right)^2 & \text{if } j = k \text{ and } \delta_j = 0 \\ \delta_j \delta_k X^{(j)} X^{(k)} & \text{otherwise} \end{cases} \tag{8.6}$$

This means that we have:

$$\Sigma^Y_{jk} = \mathbb{E}\left(Y \otimes Y\right)_{jk} = \begin{cases} \delta \Sigma_{jj} + (1-\delta) V_j & \text{if } j = k \\ \delta^2 \Sigma_{jk} & \text{otherwise} \end{cases} \tag{8.7}$$

Thus:

$$\Sigma_{jk} = \begin{cases} \delta^{-1}\left(\Sigma^Y_{jj} - (1-\delta)V_j\right) & \text{if } j = k \\ \delta^{-2}\Sigma^Y_{jk} & \text{otherwise} \end{cases} \tag{8.8}$$

Which in turn means that:

$$\Sigma = (\delta^{-1} - \delta^{-2})\mathrm{diag}(\Sigma^Y) + \delta^{-2}\Sigma^Y + (1-\delta^{-1})V \tag{8.9}$$

Which gives the general correction formula with independent contamination. For the missing values correction, simply set $V = 0_P$ the $p \times p$ zero matrix.

8.3 bounds on the determinant of in equation 7.7

Theorem 13 of Thompson (1966) states that, for any matrix A of size p with eigenvalues $\lambda_1, \dots \lambda_s$, each with multiplicity $\mu_1, \dots \mu_s$ such that $\sum_{i=1}^s \mu_i = p$, then any principal submatrix $A(j|j)$, that is, a matrix created by removing line j and column j from A, has eigenvalues λ_i with multiplicity $\max(0, \mu_i - 1)$. The other eigenvalues have values between $\min_i \lambda_i$ and $\max_i \lambda_i$.

In our case, the matrix Σ_j has only two eigenvalues: γ and σ, with multiplicity r and $p - r$ respectively. One will easily find by recurrence on the number of deleted dimensions that, with $d_i = \sum_{j=1}^p \delta_i^{(j)}$

$$\det \Sigma_j^{(\delta_i)} = \gamma^{\max(0, r-d_i)} \sigma^{\max(0, p-r-d_i)} \prod_{k=1}^{d_i} \mu_k \tag{8.10}$$

where $\forall k \in \{1, d_i\}$, $\sigma \leq \mu_k \leq \gamma$.

This means, in particular, that:

$$\gamma^{\max(0,r-d_i)} \sigma^{\max(p-r,p-d_i)} \leq \det \Sigma_j^{(\delta_i)} \leq \gamma^{\max(r,p-d_i)} \sigma^{\max(0,p-r-d_i)} \tag{8.11}$$

Now, let us demonstrate the statement in equation 7.7. We have Σ_1 and Σ_j having the same eigenvalues γ and σ with multiplicity respectively r and $p - r$. Let $d_i = \sum_{k=1}^p \delta_i^k$ be the number of deleted components after applying the boolean filter δ_i. Since Σ_1 is diagonal, we know that $\Sigma_1^{(\delta_i)}$ will also have eigenvalues γ and σ, with multiplicity a_i and b_i respectively, where $a_i \sim \mathcal{B}(r, \delta)$ and $b_i \sim \mathcal{B}(p-r, \delta)$ where $\mathcal{B}$ is the binomial distribution.

Then, using the lower bound we just demonstrated, we get that:

$$\begin{aligned} -\mathbb{E}_\delta \log\left(\det\left(\Sigma_1^{(\delta_i)-1}\Sigma_j^{(\delta_i)}\right)\right) &= \mathbb{E}_\delta a_i \log(\gamma) + b_i \log(\sigma) - \log\left(\det\left(\Sigma_j^{(\delta_i)}\right)\right) \\ &\leq \mathbb{E}_\delta a_i \log(\gamma) + b_i \log(\sigma) - \max(0, r-d_i)\log(\gamma) - \max(p-r, p-d_i)\log(\sigma) \\ &\leq (r\delta + \min(0, -r))\log(\gamma) + ((p-r)\delta + \min(r-p, -p))\log(\sigma) \\ &\leq r\delta \log(\gamma) + (p-r)(\delta - 1)\log(\sigma) \\ &\leq r\delta \log(\gamma) \end{aligned} \tag{8.12}$$

since $\delta - 1 \leq 0$. In particular, one can easily see that $\log(x) \leq \frac{1}{2}x^2$ for all positive x, thus

$$\mathbb{E}_\delta \log\left(\det\left(\Sigma_1^{(\delta_i)-1}\Sigma_j^{(\delta_i)}\right)\right) \leq r\delta\gamma^2 \tag{8.13}$$

hence the result.

8.4 Proof of the upper bound of the frobenius norm with missing values

Let $P \in \mathbb{R}^{p\times p}$ be any matrix, then, using the fact that the δ_i are boolean vectors:

$$\begin{aligned}\mathbb{E}_\delta \left\|(\delta_i \otimes \delta_i) \odot P\right\|_F^2 &= \mathbb{E}_\delta \mathrm{tr}\left(((\delta_i \otimes \delta_i) \odot P)^\top ((\delta_i \otimes \delta_i) \odot P)\right) \\ &= \mathbb{E}_\delta \sum_{k=1}^p \sum_{l=1}^p \delta_i^k \delta_i^l P_{kl}^2 \\ &= \sum_{k=1}^p \left(\delta P_{kk} + \sum_{\substack{l=1 \\ l\neq k}}^p \delta^2 P_{kl}^2\right) \\ &\leq \delta \left\|P\right\|_F^2\end{aligned} \tag{8.14}$$

References

Agostinelli, C., Leung, A., Yohai, V. J., and Zamar, R. H. (2014). Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination.

Alqallaf, F., Aelst, S. V., Yohai, V. J., and Zamar, R. H. (2009). Propagation of outliers in multivariate data. *The Annals of Statistics*, 37(1):311–331.

Farcomeni, A. (2014). Robust Constrained Clustering in Presence of Entry-Wise Outliers. *Technometrics*, 56.

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Hubert, M., Debruyne, M., and Rousseeuw, P. J. (2018). Minimum Covariance Determinant and Extensions. *WIREs Computational Statistics*, 10(3).

Jamshidian, M. and Bentler, P. M. (1999). ML Estimation of Mean and Covariance Structures with Missing Data Using Complete Data Routines. *Journal of Educational and Behavioral Statistics*, 24(1):21–41.

Koltchinskii, V. and Lounici, K. (2014). Concentration Inequalities and Moment Bounds for Sample Covariance Operators.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2015). Estimation of Low-Rank Covariance Function.

Lounici, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058.

Pajor, A. (1998). Metric Entropy of the Grassmann Manifold. *Complex Geometry Analysis*, 34:181–188.

Raymaekers, J. and Rousseeuw, P. J. (2020). Handling cellwise outliers by sparse regression and robust covariance. *arXiv:1912.12446 [stat].*

Raymaekers, J. and Rousseeuw, P. J. (2021). Fast robust correlation for high-dimensional data. *Technometrics*, 63(2):184–198.

Rousseeuw, P. and Yohai, V. (1984). Robust Regression by Means of S-Estimators. In Franke, J., Härdle, W., and Martin, D., editors, *Robust and Nonlinear Time Series Analysis*, Lecture Notes in Statistics, pages 256–272, New York, NY. Springer US.

Rousseeuw, P. J. and den Bossche, W. V. (2018). Detecting deviating data cells. *Technometrics*, 60(2):135–145.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63(3):581–592.

Thompson, R. C. (1966). Principal submatrices of normal and Hermitian matrices. *Illinois Journal of Mathematics*, 10(2):296–308.

Tsybakov, A. B. (2009). Nonparametric estimators. In Tsybakov, A. B., editor, *Introduction to Nonparametric Estimation*, Springer Series in Statistics, pages 1–76. Springer, New York, NY.

Tukey, J. W. (1978). The Ninther, a Technique for Low-Effort Robust (Resistant) Location in Large Samples. In David, H. A., editor, *Contributions to Survey Sampling and Applied Statistics*, pages 251–257. Academic Press.

Vershynin, R. (2011). Introduction to the non-asymptotic analysis of random matrices.