CIMAT

Centro de Investigación en Matemáticas, A.C.

ROBUST ESTIMATION OF THE MEAN OF A RANDOM MATRIX: A NON-ASYMPTOTIC STUDY

TESIS

Que para obtener el grado de

Maestro en Ciencias

con Orientación en

Probabilidad y Estadística

Presenta

Roberto Cabal López

Director de Tesis:

Dr. Emilien Joly

Autorización de la versión final

Guanajuato, Gto., 30 de octubre de 2020

Centro de Investigación en Matemáticas, A.C.

CIMAT

ACTA PROVISIONAL

Acta de Examen de Grado

Acta No.:	168
Libro No.:	002
Foja No.:	168

En la Ciudad de Guanajuato, Gto., siendo las 11:00 horas del día 30 de octubre del año 2020, se reunieron los miembros del jurado integrado por los señores:

DR. ROLANDO JOSÉ BISCAY LIRIO	**(CIMAT)**
DR. ROGELIO RAMOS QUIROGA	**(CIMAT)**
DR. MATTHIEU PIERRE LERASLE	**(ENSAE-FRANCIA)**
DR. EMILIEN ANTOINE MARIE JOLY	**(CIMAT)**

Bajo la presidencia del primero y con carácter de secretario el segundo, para proceder a efectuar el examen que para obtener el grado de

MAESTRO EN CIENCIAS
CON ESPECIALIDAD EN PROBABILIDAD Y ESTADÍSTICA

Sustenta

ROBERTO CABAL LÓPEZ

En cumplimiento con lo establecido en los reglamentos y lineamientos de estudios de posgrado del Centro de Investigación en Matemáticas, A.C., mediante la presentación de la tesis

“ROBUST ESTIMATION OF THE MEAN OF A RANDOM MATRIX: A NON-ASYMPTOTIC STUDY”

Los miembros del jurado examinaron alternadamente al (la) sustentante y después de deliberar entre sí resolvieron declararlo (a)

APROBADO

R. Biscay

DR. ROLANDO JOSÉ BISCAY LIRIO
Presidente

Rogelio Ramos Q

DR. ROGELIO RAMOS QUIROGA
Secretario

DR. MATTHIEU PIERRE LERASLE
Vocal

Joly

DR. EMILIEN ANTOINE MARIE JOLY
Vocal

CIMAT DIRECCIÓN GENERAL

Dr. Víctor Manuel Rivero Mercado
Director General

Abstract

This thesis is concerned with the estimation of the mean of a random matrix when there are no assumptions about the tail of the distributions that are related to the matrix. More specifically, the estimation procedure contemplates that the distribution of the elements of the random matrix could be heavy-tailed. For this reason, we develop concentration inequalities for the estimators around the mean matrix in such a way that the theoretical guarantees give us, for example, valuable information about how to choose the hyperparameters related to the estimator. Of particular interest is the robust estimation of the covariance matrix from a random sample, which has numerous applications in statistical science such as Factor Analysis and Principal Components Analysis [37]. Other famous applications of matrix concentration inequalities are in the fields of Matrix Completion and community detection in Random Graphs Theory [47].

Acknowledgements

To CONACyT, for the economic support given within these two years of the master program at CIMAT.

To my teachers at CIMAT, for the dedication and patience in the wonderful courses.

To my friends of the generation 2018-2020, for all the support and the healthy competition trying to be always the best version of us.

To my family, for all the words and advice that helped me live a better life.

To María, for all the love and time invested throughout these years. In the end it was all worth it.

Table of Contents

List of symbols

$\mathcal{M}_{n,p}$ Space of $n \times p$ real matrices

$\mathcal{M}_p$ Space of $p \times p$ real matrices

$\mathcal{S}_p$ Space of $p \times p$ symmetric matrices

$\mathbf{A} \succeq \mathbf{0}$ $\mathbf{A}$ is symmetric non-negative definite

$\mathbf{A} \succeq \mathbf{H}$ $\mathbf{A} - \mathbf{H}$ is non-negative definite; $\mathbf{A}$ and $\mathbf{H}$ are symmetric

$\lambda_j(\mathbf{A})$ j-th greatest eigenvalue of $\mathbf{A} \in \mathcal{M}_p$, i.e., the eigenvalues are ordered as $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_p(\mathbf{A})$

$s_j(\mathbf{B})$ j-th greatest singular value of $\mathbf{B} \in \mathcal{M}_{n,p}$, i.e., the singular values are ordered as $s_1(\mathbf{B}) \geq \cdots \geq s_m(\mathbf{B}) \geq 0$, $m = n \wedge p$

$\boldsymbol{s}(\mathbf{B})$ Vector of singular values of $\mathbf{B} \in \mathcal{M}_{n,p}$

$\|\boldsymbol{x}\|_k$ ℓ_k norm of $\boldsymbol{x} \in \mathbb{R}^p$ equal to $\left(\sum_{j=1}^p |x_j|^k\right)^{1/k}$

$|||\mathbf{B}|||_k$ Shatten k-norm of $\mathbf{B} \in \mathcal{M}_{n,p}$ equal to $\|\boldsymbol{s}(\mathbf{B})\|_k$

$|||\mathbf{B}|||$ Operator norm of $\mathbf{B} \in \mathcal{M}_{n,p}$ equal to $|||\mathbf{B}|||_\infty = s_1(\mathbf{B})$

$|||\mathbf{B}|||_2$ Frobenius norm of $\mathbf{B} \in \mathcal{M}_{n,p}$ equal to $\operatorname{tr}(\mathbf{B}^\intercal\mathbf{B})$ and $\sqrt{s_1(\mathbf{B})^2 + \cdots + s_m(\mathbf{B})^2}$, $m = n \wedge p$

$|||\mathbf{B}|||_1$ Nuclear norm of $\mathbf{B} \in \mathcal{M}_{n,p}$ equal to $\operatorname{tr}(\sqrt{\mathbf{B}^\intercal\mathbf{B}})$ and $s_1(\mathbf{B}) + \cdots + s_m(\mathbf{B})$, $m = n \wedge p$

$|||\mathbf{B}|||_{\max}$ Max-norm of $\mathbf{B} \in \mathcal{M}_{n,p}$ equal to $\max_{i,j} |b_{ij}|$ where b_{ij} is the i,j-th element of $\mathbf{B}$

$\langle \mathbf{B}, \mathbf{B}' \rangle$ Inner product of $\mathbf{B}, \mathbf{B}' \in \mathcal{M}_{n,p}$ defined as $\operatorname{tr}(\mathbf{B}^\intercal\mathbf{B}')$.

$\bar{\boldsymbol{X}}$ Mean vector equal to $\frac{1}{n}\sum_{j=1}^n \boldsymbol{X}_j$ with $\boldsymbol{X}_j \in \mathbb{R}^p$

$\|X\|_{\psi_2}$ sub-Gaussian norm of the random variable X defined as $\inf\{t > 0 : \mathbb{E}\exp(X^2/t^2) \leq 2\}$

$\|X\|_{L^2}$ L^2-norm of the random variable X defined as $(\mathbb{E}|X|^2)^{1/2}$

Chapter 1

Introduction

In this thesis we study the general procedure of Minsker in [32] to obtain *good* estimators of the mean of a random matrix. More precisely, suppose that $\mathbf{X}$ is a $p \times p$ matrix whose entries are real-valued random variables. We call this object a *random matrix*[1]. Suppose that we observe the n iid (independent and identically distributed) copies $\mathbf{X}_1, ..., \mathbf{X}_n$ of $\mathbf{X}$ meaning that the entries of each copy are independent from the other copies and they have the same distribution of the entries of $\mathbf{X}$. We are interested in the estimation of the mean matrix

$$\mathbb{E}\mathbf{X} = \left(\mathbb{E}X_{ij}\right)_{ij}$$

through the matrices $\mathbf{X}_1, ..., \mathbf{X}_n$. One educated guess of a good estimator is the empirical mean $\mathbf{M}$ defined as

$$\mathbf{M} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{X}_j.$$

This is good in the sense that it is an unbiased estimator of $\mathbb{E}\mathbf{X}$, i.e.,

$$\mathbb{E}\mathbf{M} = \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\mathbf{X}_j = \mathbb{E}\mathbf{X}.$$

But how can we quantify the *variability* of this estimator? In the case of scalar random variables it is common to quantify it by the variance, but in this case there is not concrete

[1] A precise definition of random matrix, independence and expectation is given in Chapter 2.

definition of “variance of a random matrix.” Another approach is by defining a metric $d : \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p} \to \mathbb{R}$ and study the quantity

$$d\left(\mathbf{T}, \mathbb{E}\mathbf{X}\right),$$

where $\mathbf{T} = \mathbf{T}(\mathbf{X}_1, ..., \mathbf{X}_n)$ is some estimator of $\mathbb{E}\mathbf{X}$. More specifically, we try to obtain a result of the form

$$\mathbb{P}\left(d\left(\mathbf{T}, \mathbb{E}\mathbf{X}\right) \geq t\right) \leq b(n, p, t), \tag{1.1}$$

where b is non-negative and decreasing on n and t. A natural election of d is a distance induced by some matrix norm $|||\cdot|||$, namely,

$$d(\mathbf{A}, \mathbf{B}) = |||\mathbf{A} - \mathbf{B}|||.$$

Observe that obtaining an inequality like (1.1) is a more powerful approach than study the variability of, for example, the entries of $\mathbf{T}$ since we are bounding the tail distribution of $d\left(\mathbf{T}, \mathbb{E}\mathbf{X}\right)$ and this can gives us valuable information like consistency[2] of $\mathbf{T}$ with respect to the distance d and the values of n and p for which $\mathbf{T}$ is *close* to $\mathbb{E}\mathbf{X}$.

We refer to a result of the form (1.1) as *non-asymptotic* since we are not necessarily studying what happens to $\mathbf{T}$ when $n \to \infty$ or $p \to \infty$. Instead, we are trying to control its behavior around $\mathbb{E}\mathbf{X}$ for fixed values of n and p. Nevertheless, obtaining a result of the form (1.1) can be difficult without making assumptions on the distribution of the entries of $\mathbf{X}$. This is the relevance of work presented in [32] since the author hands over a novel approach to obtain results of the form (1.1) with very few distributional assumptions. The type of estimators that require minimal assumptions are called *robust*.

Incidentally, a non-asymptotic viewpoint leads us to results relevant in a *high-dimensional* setting, i.e., when the matrix dimension p is very large. For example, if for every $t \geq 0$ we have that $b(n, p, t) \to 0$ when $n \to \infty$ even if $p = p(n) \to \infty$, we obtain that $\mathbf{T}$ is consistent. Similarly, even in the case $p \gg n$ we can determine a sample size $n = n(p, t, \delta)$ such that

$$\mathbb{P}\left(d\left(\mathbf{T}, \mathbb{E}\mathbf{X}\right) < t\right) \geq 1 - \delta, \quad \delta \in (0, 1).$$

In what follows we give some examples of random matrices in statistics and we dive more into the concepts of robustness and non-asymptotic results.

[2] We say that $\mathbf{T}$ is a consistent estimator of $\mathbb{E}\mathbf{X}$ with respect to the distance d if $d(\mathbf{T}, \mathbb{E}\mathbf{X}) \to 0$ in probability when $n \to \infty$.

1.1 Random matrices in statistics

In this section we briefly summarize some of the applications of random matrices that are covered in this thesis.

1. **Covariance matrix estimation**. Suppose that $\boldsymbol{X} \in \mathbb{R}^p$ is a random vector with mean vector $\mathbb{E}\boldsymbol{X} = \boldsymbol{\mu}$ and covariance matrix $\text{Cov}\boldsymbol{X} = \boldsymbol{\Sigma}$[3]. If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid copies of $\boldsymbol{X}$ then the empirical estimator of the covariance matrix $\boldsymbol{\Sigma}$ is

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{j=1}^{n}(\boldsymbol{X}_j - \bar{\boldsymbol{X}})(\boldsymbol{X}_j - \bar{\boldsymbol{X}})^\intercal,$$

   where $\bar{\boldsymbol{X}} = n^{-1}\sum_{j=1}^{n}\boldsymbol{X}_j$. Here the random matrix of interest is $\widehat{\boldsymbol{\Sigma}}$ and it can be shown that $\mathbb{E}\widehat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}$. This estimator is studied broadly in Chapters 3 and 5.

2. **Stochastic block model**. Let $G = (V, E)$ be a undirected graph with vertex set $V$and edge set E. Let $\mathbf{A}$ be a matrix such that $a_{ij} = 1$ whenever $(i, j) \in E$. If the set E is random, i.e., two vertices $i, j \in V$ are connected with certain probability, then the matrix $\mathbf{A}$ is random. Additionally, we suppose that there is an structure of groups in G, so vertices that belong to the same group have higher probability of being connected than vertices in different groups. Here we are interested in estimating $\mathbb{E}\mathbf{A}$ through the matrix $\mathbf{A}$ with entries contaminated with noise. This application is studied in Chapter 4 and is related to the problem of community detection.

3. **Matrix completion**. Let $\mathbf{A}$ be a $n \times m$ matrix. We make the following experiment: choose a pair $(i, j) \in \{1, ..., n\} \times \{1, ..., m\}$ uniformly at random and observe the quantity $Y = a_{ij} + \xi$, where ξ is some real-valued random noise. Suppose that we observe a sequence of pairs $(i_1, j_1), ..., (i_n, j_n)$ and iid copies $Y_1, ..., Y_n$ of Y. Our interest is to recovering the matrix $\mathbf{A}$ from this sample. This problem is called matrix completion and is fully covered in Chapter 6. The random matrix underlying this problem is constructed from the sample through a *trace regression model*.

Motivated by the estimation of large covariance matrices, in [21] it is mentioned that most of the work done to this objective hinges on distributional assumptions like Gaussianity or

[3] $\mathbb{E}\boldsymbol{X} = (\mathbb{E}X_1, ..., \mathbb{E}X_p)^\intercal$ and $\text{Cov}\boldsymbol{X} = \mathbb{E}\left[(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})^\intercal\right]$, where the expectation is taken entrywise.

sub-Gaussianity (a concept that is defined in Chapter 3.) Diverse applications of covariance matrix estimation such as functional magnetic resonance imaging (fMRI) data [13], genomics [26] and quantitative finances [10] points out that the usual assumptions are not valid. In particular, in [13] the authors mention that the evidence suggests that the principal cause of the invalid inference done with fMRI data is that the spatial autocorrelation (or covariance) do not follow the Gaussianity assumptions. In this sense, there is a need for robust methods that require few (or none) distributional assumptions for the data to perform covariance matrix estimation, and in general, a robust method for mean matrix estimation.

1.2 Robust estimation

In [19], Huber indicates that an estimator is robust if it is "insensible to small deviations from the assumptions." To be more specific, suppose that we observe the independent random variables $X_1, ..., X_n$ and some fraction $\epsilon \in (0,1)$ of them has a different distribution that is heavy-tailed, i.e., the sample is contaminated by *outliers*. Let $T = T(X_1, ..., X_n)$ be an estimator of, for example, the mean of the $n(1-\epsilon)$ variables of interest. Huber defined loosely in [19] that the *breakdown point* is the smallest ϵ such that T takes arbitrarily large aberrant values, and T can be considered robust whenever ϵ is large. Nevertheless, this is not the exact definition that Huber made, since in [19] he precisely define the concepts of *qualitative* and *quantitative* robustness.

Despite of getting into conflict with the ideas presented in [19], in this thesis we conceive the concept of robusteness more related to *distribution-freeness*, i.e., we consider that the estimator T is robust whenever we can obtain a result of the form (1.1) making no assumptions of the tail-behavior of the distribution of the variables $X_1, ..., X_n$. The reason for this is that, as mentioned in [21], the appearance of outliers in the sample indicates that the phenomenon of interest can be modeled with a heavy-tailed distribution and in order to obtain results of the form (1.1) we need a procedure that can handle arbitrary distributions. Therefore, our concept of robustness can be best understood as *tail-robustness*.

We want to emphasize that the distributional generality of the procedures presented along this thesis makes it reasonable to classify them as robust. This procedures are not influenced and do not require a concrete distributional assumption for the sample. So even if there exist some underlying distributional assumptions that serve a whole pipeline of statistical procedures, this methods serve its purpose by giving estimations that are not affected by outliers.

1.3 A non-asymptotic viewpoint

As mentioned earlier, throughout this thesis our objective will be to attain results of the form (1.1). This type of inequalities are called Concentration Inequalities (or Concentration Bounds). In [4] the authors mention that a concentration inequality is "a way to quantify random fluctuations of functions of independent random variables, typically by bounding the probability that such a function differs from its expected value (or from its median) by more than a certain amount." To fix ideas, suppose that we are given an iid sample $X_1, ..., X_n$ or real-valued random variables and we want to estimate $\mathbb{E}X_1 < \infty$. We want a result of the form (1.1) for the sample mean $\bar{X}_n = n^{-1} \sum_{j=1}^{n} X_j$, namely,

$$\mathbb{P}\left(|\bar{X}_n - \mathbb{E}X_1| \geq t\right) \leq b(n,t).$$

This can be done whenever $\mathrm{Var}X_1 < \infty$. Indeed, by Chebyshev's inequality we obtain

$$\mathbb{P}\left(|\bar{X}_n - \mathbb{E}X_1| \geq t\right) \leq \frac{\mathrm{Var}X_1}{nt^2}.$$

Despite of being the most basic concentration inequality, this result tell us that the probability $\mathbb{P}\left(|\bar{X}_n - \mathbb{E}X_1| \geq t\right)$ decreases to zero at least linearly. But, can we do better? More precisely, can we construct an estimator $T = T(X_1, ..., X_n)$ such that the probability $\mathbb{P}\left(|T - \mathbb{E}X_1| \geq t\right)$ decreases, for example, exponentially to zero? Surprisingly, the answer to this question is affirmative. In order to obtain such estimator T we need more sophisticated concentration inequalities than Chebyshev. We present one famous result called Hoeffding's inequality. The proof of the following Theorem can be found in [4, p. 34].

Theorem 1.3.1 (Hoeffding's inequality)**.** *Let $X_1, ..., X_n$ be independent random variables such that X_i takes values in $[a_i, b_i]$ almos surely for all $i \leq n$. Then, for every $t > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^{n} (X_i - \mathbb{E}X_i) \geq t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

In order to see the power of Theorem 1.3.1 we give the following example that is directed to the estimation of the mean of a random variable. The method presented is called the Median of Means estimator. To see more sophisticated applications of this method one can see the recent work in [30] and [29].

Example 1.3.1 (Median of means estimator)**.** Let X be a random variable with $\mathbb{E}X = \mu$ and $\mathrm{Var}X = \sigma^2 < \infty$. Suppose we observe the iid copies $X_1, X_2, ..., X_n$ of X and that we want to obtain a estimator T of μ such that we guarantee for any $\epsilon > 0$ and $\delta \in (0, 1)$ that with an specific sample size $n = n(\sigma^2, \epsilon, \delta)$ we get that

$$\mathbb{P}\left(|T - \mu| < \epsilon\right) \geq 1 - \delta.$$

To do so, suppose that we can construct a partition of $\{1, 2, ..., n\}$ in k groups where k is an even number. Denote as $C_1, ..., C_k$ this partition with $|C_i| = m > 0$ for all i. For any C_i denote its empirical mean with respect to the sample $X_1, ..., X_n$ as

$$\hat{\mu}_i = \frac{1}{m} \sum_{j \in C_i} X_j.$$

By Chebyshev's inequality,

$$\mathbb{P}\left(|\hat{\mu}_i - \mu| \geq \epsilon\right) \leq \frac{\sigma^2}{m\epsilon^2}.$$

So by taking $m = \sigma^2/(4\epsilon^2)$ we guarantee for any i that

$$\mathbb{P}\left(|\hat{\mu}_i - \mu| < \epsilon\right) \geq \frac{3}{4}.$$

Now, define the estimator T as[4]

$$T = \operatorname{med}\left(\hat{\mu}_1, ..., \hat{\mu}_k\right).$$

Also, define the random variables $Y_i = \mathbf{1}(|\hat{\mu}_i - \mu| \geq \epsilon)$ which are iid Bernoulli random variables with success parameter $q \leq 1/4$ (assuming that $m = \sigma^2/(4\epsilon^2)$.) By definition of median of a finite set, if $|T - \mu| \geq \epsilon$ then at least $k/2$ of the variables $\hat{\mu}_1, ..., \hat{\mu}_k$ have to satisfy that $|\hat{\mu}_i - \mu| \geq \epsilon$. Therefore,

$$\mathbb{P}\left(|T - \mu| \geq \epsilon\right) \leq \mathbb{P}\left(\sum_{i=1}^{k} Y_i \geq \frac{k}{2}\right).$$

On the other hand, as $Y_i \in [0, 1]$ almost surely, we obtain by Hoeffding's inequality of Theorem 1.3.1 that

$$\mathbb{P}\left(\sum_{i=1}^{k} (Y_i - q) \geq t\right) \leq \exp\left(\frac{-2t^2}{k}\right).$$

[4] $\operatorname{med}(x_1, ..., x_N)$ denote the median value of the set $\{x_1, ..., x_N\} \subset \mathbb{R}$.

Even more, since $q \leq 1/4$,

$$\mathbb{P}\left(\sum_{i=1}^{k} Y_i \geq \frac{k}{2}\right) = \mathbb{P}\left(\sum_{i=1}^{k}\left(Y_i - \frac{1}{4}\right) \geq \frac{k}{4}\right) \leq \mathbb{P}\left(\sum_{i=1}^{k}(Y_i - q) \geq \frac{k}{4}\right).$$

Finally we obtain that

$$\mathbb{P}\left(|T-\mu| \geq \epsilon\right) \leq \exp\left(\frac{-k}{8}\right), \tag{1.2}$$

so by choosing[5] $n = 2\sigma^2 \log(\delta^{-1})/\epsilon^2$ we guarantee that

$$\mathbb{P}\left(|T-\mu| < \epsilon\right) \geq 1 - \delta.$$

♣

From (1.2) of Example 1.3.1 we conclude that there exist an estimator T such that[6]

$$\mathbb{P}\left(|T-\mu| \geq t\right) \leq \exp\left(\frac{-t^2 n}{2\sigma^2}\right),$$

which decreases exponentially with n and t. Another remarkable thing of this approach is that we made no distributional assumption over the iid sample $X_1, ..., X_n$ other than $\text{Var} X_1 < \infty$.

Example 1.3.1 encapsulates some of the ideas of the procedures followed in this thesis. However, we'll go after a different methodology called the *Crámer-Chernoff method*, which consists on bounding the moment generating function. This is motivated from the next observation: for iid random variables $X_1, ..., X_n$ and some function $h : \mathbb{R} \to \mathbb{R}$, we get from Markov's inequality that for any $\theta > 0$,

$$\begin{aligned}
\mathbb{P}\left(\sum_{j=1}^{n} h(X_i) - \mathbb{E}X_1 \geq t\right) &= \mathbb{P}\left(e^{\sum_{j=1}^{n} \theta h(X_i)} \geq e^{\theta(\mathbb{E}X_1 + t)}\right) \\
&\leq e^{-\theta(\mathbb{E}X_1+t)} \mathbb{E} e^{\sum_{j=1}^{n} \theta h(X_i)} \\
&= e^{-\theta(\mathbb{E}X_1+t)} \prod_{j=1}^{n} \mathbb{E} e^{\theta h(X_j)} \\
&= e^{-\theta(\mathbb{E}X_1+t)} \left(\mathbb{E} e^{\theta h(X_1)}\right)^n.
\end{aligned}$$

Therefore, by bounding $\mathbb{E}e^{\theta h(X_1)}$ we can get a concentration inequality for the estimator $\sum_{j=1}^{n} h(X_j)$. This approach, applied to random matrices, is explored in Chapter 2.

[5] $n = mk = \frac{\sigma^2}{4\epsilon^2} 8 \log(\delta^{-1})$

[6] Substitute $k = \frac{n}{m} = \frac{4\epsilon^2}{\sigma^2} n$.

1.4 Matrix assumptions and notations

Motivated by the applications of Section 1.1, throughout this thesis we will work with matrices with real entries with special emphasis on real symmetric matrices. The set $\mathcal{M}_{p,r}$ denotes the set of $p \times r$ matrices and $\mathcal{M}_p$ is the of $p \times p$ (squared) matrices. Also, $\mathcal{S}_p$ is the set of $p \times p$ symmetric matrices.

Since any matrix $\mathbf{A} \in \mathcal{S}_p$ has real eigenvalues we denote them as $\lambda_j(\mathbf{A})$, $j = 1, ..., p$ and suppose that they are arranged in decreasing fashion, i.e., $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_p(\mathbf{A})$.

Similarly, for any matrix $\mathbf{B} \in \mathcal{M}_{p,r}$ we denote its singular values as $s_1(\mathbf{B}) \geq \cdots \geq s_m(\mathbf{B}) \geq 0$, $m = p \wedge r$. See Appendix A for a definition of singular values and Singular Value Decomposition (SVD.) Also, its vector of singular values is

$$\boldsymbol{s}(\mathbf{B}) = (s_1(\mathbf{B}), ..., s_m(\mathbf{B}))^\intercal .$$

With this notation we define the Schatten k-norm of $\mathbf{B}$ as

$$|||\mathbf{B}|||_k = \|\boldsymbol{s}(\mathbf{B})\|_k,$$

where $\|\cdot\|_k$ is the usual ℓ_k vector norm. Of particular interest are the *operator norm*

$$|||\mathbf{B}||| = |||\mathbf{B}|||_\infty := \lim_{k\to\infty} |||\mathbf{B}|||_k = \max_{1\leq j\leq m} s_j(\mathbf{B}) = s_1(\mathbf{B}),$$

the *Frobenius norm*

$$|||\mathbf{B}|||_2 = \sqrt{\operatorname{tr}(\mathbf{B}^\intercal\mathbf{B})} = \sqrt{s_1^2(\mathbf{B}) + \cdots + s_m^2(\mathbf{B})},$$

and the *Nuclear norm*

$$|||\mathbf{B}|||_1 = s_1(\mathbf{B}) + \cdots + s_m(\mathbf{B}).$$

See Appendix A for a more precise development of this matrix norms. Note that for any $\mathbf{A} \in \mathcal{S}_p$ we have that

$$|||\mathbf{A}||| = \max\{|\lambda_1(\mathbf{A})|, |\lambda_p(\mathbf{A})|\} = \max\{\lambda_1(\mathbf{A}), -\lambda_p(\mathbf{A})\}.$$

To better work with general matrices, we'll use the symmetric dilation $\mathcal{H} : \mathcal{M}_{p,r} \to \mathcal{S}_{p+r}$ defined as

$$\mathcal{H}(\mathbf{B}) = \begin{pmatrix} \mathbf{0} & \mathbf{B} \\ \mathbf{B}^\intercal & \mathbf{0} \end{pmatrix}.$$

The notation $\mathbf{A} \succeq \mathbf{0}$ indicates that the matrix $\mathbf{A} \in \mathcal{S}_p$ is non-negative definite, i.e., for every $\boldsymbol{x} \in \mathbb{R}^p$ we have $\boldsymbol{x}^\intercal \mathbf{A} \boldsymbol{x} \geq 0$. We use the non-negative order (or semidefinite order) $\mathbf{A} \succeq \mathbf{H}$ as short for $\mathbf{A} - \mathbf{H} \succeq \mathbf{0}$, i.e., the symmetric matrix $\mathbf{A} - \mathbf{H}$ is non-negative definite. The notation $\mathbf{A} \succeq \mathbf{H}$ indicates implicitly that $\mathbf{A}$ and $\mathbf{H}$ are symmetric. The same reasoning goes for $\mathbf{A} \succ \mathbf{0}$ and $\mathbf{A} \succ \mathbf{H}$ which indicates positive definiteness.

Any vector $\boldsymbol{x} \in \mathbb{R}^p$ will be understood as a column vector with p entries. For any random vector we denote $\mathbb{E}\boldsymbol{X}$ as its mean vector and $\mathrm{Cov}\boldsymbol{X}$ as its covariance matrix, i.e.,

$$\mathbb{E}\boldsymbol{X} = \begin{pmatrix} \mathbb{E}X_1 \\ \vdots \\ \mathbb{E}X_p \end{pmatrix}$$

and

$$\begin{aligned} \mathrm{Cov}\boldsymbol{X} &= \mathbb{E}\left[(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})^\intercal\right] \\ &= \begin{pmatrix} \mathbb{E}[(X_1 - \mathbb{E}X_1)^2] & \cdots & \mathbb{E}[(X_1 - \mathbb{E}X_1)(X_p - \mathbb{E}X_p)] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[(X_1 - \mathbb{E}X_1)(X_p - \mathbb{E}X_p)] & \cdots & \mathbb{E}[(X_p - \mathbb{E}X_p)^2] \end{pmatrix}. \end{aligned}$$

If $\boldsymbol{X} \in \mathbb{R}^p$ is a Gaussian random vector with $\mathbb{E} = \boldsymbol{\mu}$ and $\mathrm{Cov}\boldsymbol{X} = \boldsymbol{\Sigma}$ we write $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. See [16, Chapter 5] for a concise review on Gaussian random vectors.

1.5 Chapters description and contributions

The following is a brief description of Chapters 2 through 6.

Chapter 2. This chapter is devoted to the basic techniques of [45] for obtaining concentration inequalities for random matrices. To do so, we first define the function of a matrix and the methods of matrix ordering.

Chapter 3. In order to compare the methods presented in Chapter 2, this chapter is directed to obtaining concentration inequalities for the classical covariance matrix estimation with some of the most known techniques that require distributional assumptions.

Chapter 4. This chapter contains the most important results of the thesis that are taken from [32]. We obtain concentration bounds for symmetric random matrices and generalize it to rectangular matrices.

Chapter 5. Here we also explore the covariance matrix estimator, but unlike Chapter 3, we use the general technique of Chapter 4 by defining a new robust estimator taken from [21]. We complement the analysis with a simulation study.

Chapter 6. To give a different application of the technique of Chapter 4, in this chapter we present the Matrix Completion problem with theoretical guarantees in a robust setting. Additional to the main result, we give a method for calculating the estimator.

The major contribution of this thesis is to give a complete presentation of the theory necessary to understand the work of Minsker in [32]. The incorporation of the appendices A, B, and C makes this thesis a self-contained presentation of the corresponding theory.

Additionally, there are several contributions to some chapters. In Chapter 4 we present two applications of the methodology, namely, PCA and Community Detection. As far as the author is concerned these applications are not presented in the literature. In Chapter 5 we give a theorem that ensures the solution of an equation regarding a method called forth moment estimation for choosing a hyperparameter. Finally, in Chapter 6 we present a method to calculate the robust estimator in the context of Matrix Completion.

Chapter 2

Matrix concentration inequalities

This chapter presents the basic techniques of [43] for obtaining concentration bounds of the form (1.1). For this purpose, in Section 1 we give a brief overview of how to order symmetric matrices and the definition of a function of a symmetric matrix. Then, in Section 2 we give a formal definition of random matrix and clarify what we mean by independent and identically distributed random matrices. Following this discussion, Section 3 presents the definition of moment generating function of a random matrix and in Section 4 we show how to use this concept to develop concentration inequalities.

2.1 Symmetric matrix operator and ordering

We now present a definition of a function of a symmetric matrix that will be used extensively throughout this thesis. This definition indicates essentially that a function of a symmetric matrix operates over the spectrum the matrix.

Definition 2.1.1 (Symmetric matrix operator)**.** *Let* $\mathbf{A}$ *be a* $p \times p$ *symmetric matrix with spectral decomposition*

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^\intercal,$$

where $\mathbf{D} = \operatorname{diag}(\lambda_1(\mathbf{A}), ..., \lambda_p(\mathbf{A}))$ *and* $\mathbf{U}$ *is orthogonal. Assume that* $\lambda_j(\mathbf{A}) \in \mathcal{C} \subset \mathbb{R}$, $j = 1, ..., p$*. If* f *is a real-valued function defined on* $\mathcal{C}$ *then the matrix* $f(\mathbf{A})$ *is defined as*

$$f(\mathbf{A}) = \mathbf{U}f(\mathbf{D})\mathbf{U}^\intercal,$$

where

$$f(\mathbf{D}) = \begin{pmatrix} f\left(\lambda_1(\mathbf{A})\right) & 0 & \cdots & 0 \\ 0 & f\left(\lambda_2(\mathbf{A})\right) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f\left(\lambda_p(\mathbf{A})\right) \end{pmatrix}.$$

Notice that the orthogonal matrix $\mathbf{U}$ of the previous definition is not unique, so one can think that $f(\mathbf{A})$ is not well defined. In Appendix A we show that this is not the case since the product $\mathbf{U}f(\mathbf{D})\mathbf{U}^\intercal$ is always the same regardless of the choice of $\mathbf{U}$.

According to the previous definition, for any $\mathbf{A} \in \mathcal{S}_p$ such that $\lambda_j(\mathbf{A}) \in \mathcal{C} \subset \mathbb{R}$ and any function $f : \mathcal{C} \to \mathbb{R}$, the eigenvalues of $f(\mathbf{A})$ are the set

$$\{f(\lambda_1(\mathbf{A})), ..., f(\lambda_1(\mathbf{A}))\}.$$

Note that in general $\lambda_1(f(\mathbf{A})) \geq f(\lambda_j(\mathbf{A}))$ for all j. If f is strictly increasing we have that $\lambda_j(f(\mathbf{A})) = f(\lambda_j(\mathbf{A}))$, $j = 1, ..., p$.

As stated in the definition, the matrix $e^{\mathbf{A}}$ exist for any symmetric matrix $\mathbf{A}$. This is an alternative for the usual definition

$$e^{\mathbf{A}} = \mathbf{I} + \sum_{k=1}^{\infty} \frac{\mathbf{A}^k}{k!}, \quad \mathbf{A} \in \mathcal{M}_p, \tag{2.1}$$

where $\mathbf{I}$ is the $p \times p$ identity matrix and $\mathcal{M}_p$ is the space of $p \times p$ matrices. When $\mathbf{A} \in \mathcal{S}_p$, the matrix exponential definition (2.1) and the one obtained by Definition 2.1.1 are the same.

Example 2.1.1 (Matrix logarithm)**.** According to Definition 2.1.1 the matrix $\log \mathbf{A}$ is well defined only when the eigenvalues of $\mathbf{A}$ are positive, i.e., when $\mathbf{A}$ is positive definite. But note that we can always represent $\mathbf{A}$ as

$$\mathbf{A} = \log e^{\mathbf{A}},$$

since $e^{\mathbf{A}}$ is well defined for any $\mathbf{A} \in \mathcal{S}_p$ and $e^{\mathbf{A}}$ is positive definite. This will be a useful representation when we need to apply Lieb's Theorem and Jensen's inequality with the aid of the concavity of $x \mapsto \log x$. ♣

Example 2.1.2 (Rank one matrices)**.** If $\mathbf{A} \in \mathcal{S}_p$ has rank one then it has one non-zero eigenvalue λ and its spectral decomposition is given by

$$\mathbf{A} = \lambda \boldsymbol{u}\boldsymbol{u}^\intercal,$$

where $\boldsymbol{u}$ is the unitary eigenvector associated to λ. Then, for every function f for which $f(\lambda)$ is well defined, we have that

$$f(\mathbf{A}) = f(\lambda)\boldsymbol{u}\boldsymbol{u}^\intercal.$$

In particular, for any vector $\boldsymbol{x} \in \mathbb{R}^p$, $\boldsymbol{x} \neq \mathbf{0}$, the symmetric matrix $\boldsymbol{x}\boldsymbol{x}^\intercal$ is rank one and has non-zero eigenvalue $\|\boldsymbol{x}\|_2^2$ with associated eigenvector $\boldsymbol{x}/\|\boldsymbol{x}\|_2$. Hence, for any $f : (0, \infty) \to \mathbb{R}$,

$$f(\boldsymbol{x}\boldsymbol{x}^\intercal) = f\left(\|\boldsymbol{x}\|_2^2\right)\left(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}\right)\left(\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}\right)^\intercal.$$

This equality will be useful for computation purposes when we are working with matrix estimators of the form

$$\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^\intercal,$$

where $\boldsymbol{X}_i \in \mathbb{R}^p$, $i = 1, ..., n$. ♣

Example 2.1.3 (Identity addition)**.** Define for a matrix $\mathbf{A} \in \mathcal{S}_p$ a real-valued function g with domain on the spectrum of $\mathbf{A}$. Define the real-valued function f_1 as $f_1(x) = g(x) + 1$. If the spectral decomposition of $\mathbf{A}$ is $\mathbf{U}\mathbf{D}\mathbf{U}^\intercal$, then

$$g(\mathbf{A}) + \mathbf{I} = \mathbf{U}g(\mathbf{D})\mathbf{U}^\intercal + \mathbf{U}\mathbf{I}\mathbf{U}^\intercal = \mathbf{U}\left(g(\mathbf{D}) + \mathbf{I}\right)\mathbf{U}^\intercal,$$

where $g(\mathbf{D}) + \mathbf{I} = \operatorname{diag}\left(g(\lambda_j(\mathbf{A})) + 1\right)$[1]. Then, $f_1(\mathbf{A}) = g(\mathbf{A}) + \mathbf{I}$. Furthermore, define the real-valued function h with domain the image of f_1, and the function $f_2(x) = h(g(x) + 1)$. Since,

$$h\left(g(\mathbf{A}) + \mathbf{I}\right) = h\left(\mathbf{U}\left(g(\mathbf{D}) + \mathbf{I}\right)\mathbf{U}^\intercal\right) = \mathbf{U}h(g(\mathbf{D}) + \mathbf{I})\mathbf{U}^\intercal,$$

where $h(g(\mathbf{D})+\mathbf{I}) = \operatorname{diag}\left[h(g(\lambda_j(\mathbf{A})) + 1)\right]$, we have that $f_2(\mathbf{A}) = h\left(g(\mathbf{A}) + \mathbf{I}\right)$. Therefore, whenever a "+1" term appears in the mapping it's translated as a "+**I**" when the mapping is applied to the matrix. ♣

The next proposition stipulates that the semidefinite order is preserved under matrix addition.

[1] For any collection $x_1, ..., x_n \in \mathbb{R}$, the matrix $\operatorname{diag}(x_1, ..., x_n)$ is the matrix of zeros with $x_1, ..., x_n$ in the diagonal in this specific order. For short we just write $\operatorname{diag}(x_j)$.

Proposition 2.1.2. *Let* $\mathbf{A}_1, ..., \mathbf{A}_n$ *and* $\mathbf{H}_1, ..., \mathbf{H}_n$ *be* $p \times p$ *symmetric matrices such that* $\mathbf{A}_j \succeq \mathbf{H}_j$, $j = 1, ..., n$. *Then* $\sum_{j=1}^n \mathbf{A}_j \succeq \sum_{j=1}^n \mathbf{H}_j$.

Proof. For every $\boldsymbol{x} \in \mathbb{R}^p$ we have that $\boldsymbol{x}^\intercal \mathbf{A}_j \boldsymbol{x} \geq \boldsymbol{x}^\intercal \mathbf{H}_j \boldsymbol{x}$, $j = 1, ..., n$. Summation over j yields $\boldsymbol{x}^\intercal \left(\sum_{j=1}^n \mathbf{A}_j - \sum_{j=1}^n \mathbf{H}_j \right) \boldsymbol{x} \geq 0$ which proves the affirmation. □

Other operation that preserves the semidefinite order is the one presented by Proposition 2.1.3. Its is useful for proving that certain functions are *operator monotone*: we say that a function $f : \mathcal{C} \subset \mathbb{R} \to \mathbb{R}$ is operator monotone if $\mathbf{A} \succeq \mathbf{H}$ implies that $f(\mathbf{A}) \succeq f(\mathbf{H})$.

Proposition 2.1.3. *If* $\mathbf{A}, \mathbf{H} \in \mathcal{S}_p$ *and* $\mathbf{A} \succeq \mathbf{H}$, *then for every* $\mathbf{B} \in \mathcal{M}_p$ *we have that* $\mathbf{B}^\intercal \mathbf{A} \mathbf{B} \succeq \mathbf{B}^\intercal \mathbf{H} \mathbf{B}$.

Proof. Since $\mathbf{A} - \mathbf{H} \succeq 0$, we can define a symmetric matrix $\mathbf{M}$ such that $\mathbf{A} - \mathbf{H} = \mathbf{M}^2$. Indeed, as $\mathbf{A} - \mathbf{H}$ is symmetric with non-negative eigenvalues we can define it's square root $\mathbf{M} = \sqrt{\mathbf{A} - \mathbf{H}}$ according to Definition 2.1.1. Then, for every $\boldsymbol{v} \in \mathbb{R}^p$ we have that

$$\begin{aligned} \boldsymbol{v}^\intercal \left(\mathbf{B}^\intercal \mathbf{A} \mathbf{B} - \mathbf{B}^\intercal \mathbf{H} \mathbf{B} \right) \boldsymbol{v} &= \boldsymbol{v}^\intercal \mathbf{B}^\intercal \mathbf{M}^2 \mathbf{B} \boldsymbol{v} \\ &= \left(\mathbf{M}^\intercal \mathbf{B} \boldsymbol{v} \right)^\intercal \left(\mathbf{M}^\intercal \mathbf{B} \boldsymbol{v} \right) \\ &\geq 0. \end{aligned}$$

Therefore $\mathbf{B}^\intercal \mathbf{A} \mathbf{B} - \mathbf{B}^\intercal \mathbf{H} \mathbf{B} \succeq 0$. □

We know that a matrix $\mathbf{A} \in \mathcal{S}_p$ is non-negative definite if $\lambda_j(\mathbf{A}) \geq 0$ for all j. So the non-negative order $\succeq$ is saying something about the eigenvalues of $\mathbf{A}$. What does it say about the eigenvalues of $\mathbf{A}$ and $\mathbf{H}$ when $\mathbf{A} \succeq \mathbf{H}$? Lemma 2.1.4 gives us the expected answer to this question.

Lemma 2.1.4. *Let* $\mathbf{A}$ *and* $\mathbf{H}$ *be two* $p \times p$ *real symmetric matrices such that* $\mathbf{A} \succeq \mathbf{H}$ *and* $\lambda_j(\mathbf{A}), \lambda_j(\mathbf{H}) \in \mathcal{C}$ *for all* j. *Then,*

(a) $\lambda_j(\mathbf{A}) \geq \lambda_j(\mathbf{B})$ *for all* j.

(b) $\operatorname{tr} f(\mathbf{A}) \geq \operatorname{tr} f(\mathbf{H})$ *for any non-decreasing* $f : \mathcal{C} \to \mathbb{R}$.

Proof. (a) By hypothesis we have that for any $\boldsymbol{x} \in \mathbb{R}^p$,

$$\boldsymbol{x}^\intercal \mathbf{A} \boldsymbol{x} \geq \boldsymbol{x}^\intercal \mathbf{H} \boldsymbol{x}. \tag{2.2}$$

Define the subspace $W \subset \mathbb{R}^p$ as $W = \operatorname{span}\{\boldsymbol{v}_j, ..., \boldsymbol{v}_p\}$. Then, by the Rayleigh quotient (Appendix A) we get that

$$\lambda_j(\mathbf{A}) = \max_{\substack{\boldsymbol{x} \in W \\ \|x\|_2 = 1}} \boldsymbol{x}^\intercal \mathbf{A} \boldsymbol{x}.$$

Then, applying maximum in both sides of (2.2) over the set $\{\boldsymbol{x} \in W : \|\boldsymbol{x}\|_2 = 1\}$ we obtain that

$$\lambda_j(\mathbf{A}) \geq \max_{\substack{\boldsymbol{x} \in W \\ \|x\|_2 = 1}} \boldsymbol{x}^\intercal \mathbf{H} \boldsymbol{x}. \tag{2.3}$$

On the other hand, in virtue of the Fischer-Courant min-max principle (Appendix A) we have that

$$\lambda_j(\mathbf{H}) = \min_{\substack{W \subset \mathbb{R}^p \\ \dim(W) = p-j+1}} \max_{\substack{\boldsymbol{x} \in W \\ \|x\|_2 = 1}} \boldsymbol{x}^\intercal \mathbf{H} \boldsymbol{x}.$$

Therefore, by taking minimum over W in (2.3) we conclude that $\lambda_j(\mathbf{A}) \geq \lambda_j(\mathbf{H})$, for all j.

To prove part (b) note that since the set of eigenvalues of $f(\mathbf{A})$ is $\{f(\lambda_1(\mathbf{A})), ..., f(\lambda_p(\mathbf{A}))\}$ we have that

$$\operatorname{tr} f(\mathbf{A}) = \sum_{j=1}^{p} f(\lambda_j(\mathbf{A})) \geq \sum_{j=1}^{p} f(\lambda_j(\mathbf{H})) = \operatorname{tr} f(\mathbf{H}),$$

where the inequality follows from non-decreasing assumption for f. □

From Lemma 2.1.4 we get that whenever $\mathbf{A} \succeq \mathbf{H}$,

$$\operatorname{tr}(\mathbf{A}) \geq \operatorname{tr}(\mathbf{H}) \quad \text{and} \quad \operatorname{tr} e^{\mathbf{A}} \geq \operatorname{tr} e^{\mathbf{H}},$$

since the functions $x \mapsto x$ and $x \mapsto e^x$ are increasing.

One question that arises immediately is that if $\lambda_j(\mathbf{A}) \geq \lambda_j(\mathbf{H})$ for all j, implies that $\mathbf{A} \succeq \mathbf{H}$. This is not true in general as indicted by the next example.

Example 2.1.4. Let $\mathbf{P} \in \mathcal{S}_p$ be a projection matrix[2] and write $\mathbf{Q} = \mathbf{I} - \mathbf{P}$. Let $C(\mathbf{P})$ be the column space of $\mathbf{P}$, i.e.

$$C(\mathbf{P}) = \{\boldsymbol{y} : \mathbf{P}\boldsymbol{x} = \boldsymbol{y} \text{ for some } \boldsymbol{x} \in \mathbb{R}^p\}.$$

[2] $\mathbf{P} = \mathbf{P}^\intercal$ and $\mathbf{P}^2 = \mathbf{P}$.

Assume that $\operatorname{rank}(\mathbf{P}) = r$, where $p - r \leq r \leq p$. We know that $\lambda_j(\mathbf{P})$ and $\lambda_j(\mathbf{Q})$ are zero or one for each j[3], and that

$$\begin{aligned} r = \operatorname{tr}\mathbf{P} &= \sum_{j=1}^{p} \lambda_j(\mathbf{P}) \\ p - r = \operatorname{tr}\mathbf{Q} &= \sum_{j=1}^{p} \lambda_j(\mathbf{Q}), \end{aligned}$$

i.e., there are more ones in $\{\lambda_j(\mathbf{P})\}$ than in $\{\lambda_j(\mathbf{Q})\}$. Then $\lambda_j(\mathbf{P}) \geq \lambda_j(\mathbf{Q})$ for all j. But for $\boldsymbol{x} \in C^{\perp}(\mathbf{P})$ we have that

$$\boldsymbol{x}^{\intercal}\mathbf{P}\boldsymbol{x} - \boldsymbol{x}^{\intercal}\mathbf{Q}\boldsymbol{x} = 0 - \|\boldsymbol{x}\|_2^2 \leq 0.$$

Therefore $\mathbf{P} \not\succeq \mathbf{Q}$. ♣

Despite of not being true in general, the statement of the previous question is true for diagonal matrices. If $\mathbf{S}$ and $\mathbf{G}$ are $p \times p$ diagonal matrices such that $\lambda_j(\mathbf{S}) \geq \lambda_j(\mathbf{G})$ for each j then for every $\boldsymbol{x} \in \mathbb{R}^p$

$$\boldsymbol{x}^{\intercal}\mathbf{S}\boldsymbol{x} = \sum_{j=1}^{p} \lambda_j(\mathbf{S}) x_j^2 \geq \sum_{j=1}^{p} \lambda_j(\mathbf{G}) x_j^2 = \boldsymbol{x}^{\intercal}\mathbf{G}\boldsymbol{x},$$

so $\mathbf{S} \succeq \mathbf{G}$. We can extend this to any $\mathbf{A} \in \mathcal{S}_p$ that is compared to the identity $\mathbf{I}$. Suppose that $\lambda_j(\mathbf{A}) \geq 1$ for each j an that the spectral decomposition of $\mathbf{A}$ is $\mathbf{U}\mathbf{D}\mathbf{U}^{\intercal}$. Using that

$$\mathbf{A} - \mathbf{I} = \mathbf{U}^{\intercal}\left(\mathbf{D} - \mathbf{I}\right)\mathbf{U}, \quad \mathbf{D} - \mathbf{I} \succeq \mathbf{0}$$

and Proposition 2.1.3 we conclude that $\mathbf{A} \succeq \mathbf{I}$. Also, note that $\lambda_j(\mathbf{A}^{-1}) \leq 1$ for each j, so $\mathbf{A}^{-1} \preceq \mathbf{I}$. Furthermore, assume that $\mathbf{A}$ and $\mathbf{H}$ are any positive definite matrices such that $\mathbf{A} \succeq \mathbf{H}$. Then, applying Proposition 2.1.3 we get that

$$\mathbf{H}^{-1/2}\mathbf{A}\mathbf{H}^{-1/2} \succeq \mathbf{I} \quad \text{and} \quad \mathbf{H}^{1/2}\mathbf{A}^{-1}\mathbf{H}^{1/2} \preceq \mathbf{I}.$$

And utilizing again Proposition 2.1.3 we conclude that $\mathbf{A}^{-1} \preceq \mathbf{H}^{-1}$. Therefore the mapping $x \mapsto x^{-1}$ is order reversing.

Thanks to the previous reasoning we can prove that some functions are operator monotone. This is shown in the next example.

[3] Indeed, if $\mathbf{P}\boldsymbol{v} = \lambda\boldsymbol{v}$ then $\lambda\boldsymbol{v} = \mathbf{P}^2\boldsymbol{v} = \lambda^2\boldsymbol{v}$, so λ is either 1 or 0.

Example 2.1.5. The function $f : (0, \infty) \to \mathbb{R}$ defined by $f(x) = -(x+a)^{-1}$, $a \geq 0$, is operator monotone in the cone of positive definite matrices. To see this, let $\mathbf{A}, \mathbf{H} \in \mathcal{S}_p$ be positive definite matrices and suppose that $\mathbf{A} \succeq \mathbf{H}$. We want to prove that $f(\mathbf{A}) \succeq f(\mathbf{H})$.

First, it is clear that $\mathbf{A} + a\mathbf{I} \succeq \mathbf{H} + a\mathbf{I}$. Then, because the map $x \mapsto x^{-1}$ is order reversing we get that

$$(\mathbf{A} + a\mathbf{I})^{-1} \preceq (\mathbf{H} + a\mathbf{I})^{-1}.$$

Also, the map $x \mapsto -x$ is clearly order reversing, so

$$-(\mathbf{A} + a\mathbf{I})^{-1} \succeq -(\mathbf{H} + a\mathbf{I})^{-1}.$$

Which proofs that f is operator monotone. ♣

The class of operator monotone functions is not that extensive as we may think. For example, we know that the function $t \mapsto t^2$ is monotone on the positive real line, but this is not the case on the set of positive definite matrices. Also, the function $t \mapsto e^t$ is monotone on the real line, but it is not operator monotone on $\mathcal{S}_p$. The counterexamples of this affirmations and a complete development of operator monotone functions can be found in [3, Chapter 5]. Nevertheless, matrix logarithm is operator monotone as stated by the next proposition.

Proposition 2.1.5 (Logarithm is operator monotone)**.** *For matrices* $\mathbf{A}, \mathbf{H} \succ \mathbf{0}$*, we have that*

$$\mathbf{A} \succeq \mathbf{H} \quad \textit{implies} \quad \log \mathbf{A} \succeq \log \mathbf{H}.$$

Proof. First, note that if $\mathbf{T}(y), \mathbf{U}(y) \in \mathcal{S}_p$ are matrices that depend on $y \in \mathbb{R}$, then if $\mathbf{T}(y) \succeq \mathbf{U}(y)$ for every $y \in \mathcal{C}$,

$$\int_{\mathcal{C}} \mathbf{T}(y)\,\mathrm{d}y \succeq \int_{\mathcal{C}} \mathbf{U}(y)\,\mathrm{d}y,$$

where the integral is taken entry-wise. Indeed, just use that fact that

$$\int_{\mathcal{C}} \boldsymbol{x}^\intercal \mathbf{T}(y) \boldsymbol{x}\,\mathrm{d}y = \boldsymbol{x}^\intercal \left(\int_{\mathcal{C}} \mathbf{T}(y)\,\mathrm{d}y \right) \boldsymbol{x} \quad \forall \boldsymbol{x} \in \mathbb{R}^p.$$

Now, from Example 2.1.5, we have that for any $y \geq 0$,

$$(1+y)^{-1}\mathbf{I} - (\mathbf{A} + y\mathbf{I})^{-1} \succeq (1+y)^{-1}\mathbf{I} - (\mathbf{H} + y\mathbf{I})^{-1}.$$

Hence,

$$\int_0^\infty (1+y)^{-1}\mathbf{I} - (\mathbf{A}+y\mathbf{I})^{-1} \,\mathrm{d}y \succeq \int_0^\infty (1+y)^{-1}\mathbf{I} - (\mathbf{H}+y\mathbf{I})^{-1} \,\mathrm{d}y.$$

On the other hand, the integral representation of the logarithm of Appendix C, indicates that

$$\log \mathbf{A} = \int_0^\infty (1+y)^{-1}\mathbf{I} - (\mathbf{A}+y\mathbf{I})^{-1} \,\mathrm{d}y.$$

This ends the proof. □

There are characteristics of a function $f : \mathcal{C} \subseteq \mathbb{R} \to \mathbb{R}$ that are not inherited on $\mathcal{S}_p$. An example of this is the monotonicity, as we have just mentioned. But, what if we have some other function $g : \mathcal{C} \subseteq \mathbb{R} \to \mathbb{R}$ and know about some relationship between f and g? For example, if $f(x) \geq g(x)$ for each x, can we say that $f(\mathbf{A}) \succeq g(\mathbf{A})$ for each $\mathbf{A} \in \mathcal{S}_p$? This is what is stated in the next Lemma.

Lemma 2.1.6. *Let $\mathbf{A} \in \mathcal{S}_p$ and f, g be two real valued functions defined in the subset of $\mathbb{R}$ that contains all the eigenvalues of $\mathbf{A}$. If f, g are such that $f(\lambda_j(\mathbf{A})) \geq g(\lambda_j(\mathbf{A}))$, $j = 1, ..., p$. Then, $f(\mathbf{A}) \succeq g(\mathbf{A})$.*

Proof. Since $f(\mathbf{A}) = \mathbf{U}\mathrm{diag}\left[f(\lambda_j(\mathbf{A}))\right]\mathbf{U}^\intercal$ and $g(\mathbf{A}) = \mathbf{U}\mathrm{diag}\left[g(\lambda_j(\mathbf{A}))\right]\mathbf{U}^\intercal$, then

$$f(\mathbf{A}) - g(\mathbf{A}) = \mathbf{U}\mathrm{diag}\left[f(\lambda_j(\mathbf{A})) - g(\lambda_j(\mathbf{A}))\right]\mathbf{U}^\intercal.$$

Given than $f(\lambda_j(\mathbf{A})) \geq g(\lambda_j(\mathbf{A}))$ we get that $f(\mathbf{A}) - g(\mathbf{A}) \succeq 0$. □

One final result that is used extensively throughout this thesis is Lemma 2.1.7, which is related to what we developed in Lemma 2.1.4.

Lemma 2.1.7. *For any $\mathbf{A} \in \mathcal{S}_p$ such that $\lambda_j(\mathbf{A}) \in \mathcal{C} \subset \mathbb{R}$, $j = 1, ..., p$, and any non-negative function $f : \mathcal{C} \to [0, \infty)$ we have that, for all j,*

$$f\left(\lambda_j(\mathbf{A})\right) \leq \mathrm{tr}\, f\left(\mathbf{A}\right) \leq p f\left(|||\mathbf{A}|||\right).$$

Proof. Observe that the set

$$\{\lambda_1(f(\mathbf{A})), ..., \lambda_p(f(\mathbf{A}))\}$$

of eigenvalues of $f(\mathbf{A})$ is a subset of $[0,\infty)$, i.e., $f(\mathbf{A}) \succeq 0$. Even more, for any j we have that

$$f\left(\lambda_j(\mathbf{A})\right) \in \left\{\lambda_1(f(\mathbf{A})), ..., \lambda_p(f(\mathbf{A}))\right\}.$$

In consequence,

$$f\left(\lambda_j(\mathbf{A})\right) \leq \sum_{k=1}^{p} f\left(\lambda_k(\mathbf{A})\right) = \sum_{k=1}^{p} \lambda_k(f(\mathbf{A})) = \operatorname{tr} f(\mathbf{A}).$$

Finally, since $\lambda_j(\mathbf{A}) \leq \max\{|\lambda_1(\mathbf{A})|, |\lambda_p(\mathbf{A})|\} = |\!|\!|\mathbf{A}|\!|\!|$ for any j, we obtain that

$$\operatorname{tr} f(\mathbf{A}) = \sum_{k=1}^{p} f\left(\lambda_k(\mathbf{A})\right) \leq p f(|\!|\!|\mathbf{A}|\!|\!|),$$

which ends the proof. □

Of particular interest for future development is the implication of Lemma 2.1.7 for the exponential function on the maximum eigenvalue:

$$e^{\lambda_1(\mathbf{A})} \leq \operatorname{tr} e^{\mathbf{A}} \leq p e^{|\!|\!|\mathbf{A}|\!|\!|}.$$

2.2 Random matrices

An intuitive and convenient way of thinking of a random matrix $\mathbf{Z} \in \mathcal{M}_{p,q}$ is as a $p \times q$ matrix whose entries are random variables which can be correlated. Just to keep things formal, we present the following definition.

Definition 2.2.1 (Random matrix)**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space. We say that* $\mathbf{Z} = (Z_{ij}) \in \mathcal{M}_{p,q}$ *is a random matrix in* $(\Omega, \mathcal{F}, \mathbb{P})$ *if* (Z_{ij}) *is collection of random variables in* $(\Omega, \mathcal{F}, \mathbb{P})$*, i.e., for all* i, j*,* $Z_{ij}^{-1}(B) \in \mathcal{F}$ *for each*[4] $B \in \mathcal{B}(\mathbb{R})$*.*

When talking about the *distribution of a random matrix* we'll refer to the distribution of the entries. There is no need to be precise about this concept since in each case we'll define the distribution of the entries if needed. However, we need the concept of equality in distribution.

[4] $Z_{ij}^{-1}(B)$ stand for $\{\omega \in \Omega : Z_{ij}(\omega) \in B\}$ and $\mathcal{B}(\mathbb{R})$ stands for the Borel σ-álgebra in $\mathbb{R}$.

Definition 2.2.2 (Equality in distribution row-wise)**.** *Write $\boldsymbol{Z} \sim \boldsymbol{S}$ whenever two random vectors $\boldsymbol{Z}, \boldsymbol{S} \in \mathbb{R}^q$ have the same distribution. Let $\mathbf{Z}, \mathbf{S} \in \mathcal{M}_{p,q}$ be two random matrices with rows $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_p$ and $\boldsymbol{S}_1, ..., \boldsymbol{S}_p$, respectively. We say that the random matrices $\mathbf{Z}$ and $\mathbf{S}$ have the same distribution row-wise if $\boldsymbol{Z}_i \sim \boldsymbol{S}_i$ for all i, and we write $\mathbf{Z} \sim \mathbf{S}$.*

Note that if $\mathbf{Z} \sim \mathbf{S}$, then $Z_{ij} \sim S_{ij}$ for all i, j, where Z_{ij} and S_{ij} are the entries of $\mathbf{Z}$ and $\mathbf{S}$, respectively.

Definition 2.2.3 (Expectation of a random matrix)**.** *Let $\mathbf{Z} \in \mathcal{M}_{p,q}$ be a random matrix. The expectation (or mean) of $\mathbf{Z}$ is the matrix $\mathbb{E}\mathbf{Z} \in \mathcal{M}_{p,q}$ defined as*

$$(\mathbb{E}\mathbf{Z})_{ij} = \mathbb{E}Z_{ij}, \quad \forall i, j.$$

From this definition it's easy to see that for fixed (non-random) matrices $\mathbf{B}' \in \mathcal{M}_{n,p}$ and $\mathbf{B}^* \in \mathcal{M}_{q,m}$, we have that

$$\mathbb{E}\left[\mathbf{B}'\mathbf{Z}\mathbf{B}^*\right] = \mathbf{B}'\mathbb{E}\left[\mathbf{Z}\right]\mathbf{B}^*.$$

Also, if $\mathbf{X}, \mathbf{Y} \in \mathcal{S}_p$ are symmetric random matrices[5] such that with probability one $\mathbf{X} \preceq \mathbf{Y}$, then

$$\mathbb{E}\mathbf{X} \preceq \mathbb{E}\mathbf{Y},$$

i.e., matrix expectation preserves semidefinite order. Additionally, it's straightforward to note that

$$\operatorname{tr}\mathbb{E}\mathbf{Z} = \mathbb{E}\operatorname{tr}\mathbf{Z}.$$

We define the independence of random matrices in the same entry-wise manner:

Definition 2.2.4 (Independence of random matrices)**.** *Let $\mathbf{Z} \in \mathcal{M}_{p,q}$ and $\mathbf{S} \in \mathcal{M}_{r,t}$ be two random matrices. We say that $\mathbf{Z}$ and $\mathbf{S}$ are independent if the collection of random variables (Z_{ij}) is independent of the collection of random variables (S_{ij}).*

From Definition 2.2.4 we can deduce that if $\mathbf{Z} \in \mathcal{M}_{p,q}$ and $\mathbf{S} \in \mathcal{M}_{q,r}$ are independent, then

$$\mathbb{E}\left[\mathbf{Z}\mathbf{S}\right] = \mathbb{E}\left[\mathbf{Z}\right]\mathbb{E}\left[\mathbf{S}\right].$$

[5] $\mathbf{X} \in \mathcal{S}_p$ means that $\mathbb{P}(X_{ij} = X_{ji}) = 1$ for all i, j.

Indeed, the equality follows since $\left(\mathbb{E}\left[\mathbf{ZS}\right]\right)_{ij} = \mathbb{E}(\mathbf{ZS})_{ij}$ and

$$\begin{aligned}\mathbb{E}(\mathbf{ZS})_{ij} &= \mathbb{E}\sum_{k=1}^{q} Z_{ik}S_{kj} \\ &= \sum_{k=1}^{q} \mathbb{E}[Z_{ik}S_{kj}] \\ &= \sum_{k=1}^{q} \mathbb{E}[Z_{ik}]\mathbb{E}[S_{kj}] \\ &= \left(\mathbb{E}[\mathbf{Z}]\mathbb{E}[\mathbf{S}]\right)_{ij}.\end{aligned}$$

Throughout this thesis we will be working in the space $\mathcal{S}_p$ of symmetric matrices, and of main interest will be to study the maximum eigenvalue $\lambda_1(\cdot)$. To be sure that we have no theoretical burdens, we justify in Appendix C that for any random matrix $\mathbf{X} \in \mathcal{S}_p$, the quantity $\lambda_1(\mathbf{X})$ is measurable.

2.3 Matrix moment generating function

We define the moment generating function of the random matrix $\mathbf{X} \in \mathcal{S}_p$ as

$$\boldsymbol{\Upsilon}_{\mathbf{X}}(\theta) = \mathbb{E}e^{\theta\mathbf{X}},$$

provided that the expectations $(\mathbb{E}e^{\theta\mathbf{X}})_{ij}$ are finite for $|\theta| < \theta_0$, for some $\theta_0 > 0$[6]. The next proposition gives us an idea on how the moment generating function can help us to obtain concentration bounds.

Proposition 2.3.1. *For $\mathbf{X} \in \mathcal{S}_p$ and $t \in \mathbb{R}$*

$$\mathbb{P}(\lambda_1(\mathbf{X}) \geq t) \leq e^{-\theta t}\mathbb{E}\mathrm{tr}\, e^{\theta\mathbf{X}}.$$

Proof. Using Lemma 2.1.7 and Markov inequality we have that for any $\theta > 0$

$$\begin{aligned}\mathbb{P}(\lambda_1(\mathbf{X}) \geq t) &= \mathbb{P}(\theta\lambda_1(\mathbf{X}) \geq t\theta) \\ &= \mathbb{P}(\lambda_1(\theta\mathbf{X}) \geq t\theta)\end{aligned}$$

[6] It is clear that it is well defined on $\theta = 0$ for any $\mathbf{X}$, in which case we have $\boldsymbol{\Upsilon}_{\mathbf{X}}(0) = \mathbf{I}$.

$$
\begin{aligned}
&= \mathbb{P}\left(e^{\lambda_1(\theta\mathbf{X})} \geq e^{t\theta}\right) \\
&\leq e^{-t\theta}\mathbb{E}e^{\lambda_1(\theta\mathbf{X})} \\
&\leq e^{-t\theta}\mathbb{E}\mathrm{tr}\, e^{\theta\mathbf{X}}.
\end{aligned}
$$

□

Since the inequality of the previous proposition works for any $\theta > 0$ we can take the infimum to get

$$\mathbb{P}(\lambda_1(\mathbf{X}) \geq t) \leq \inf_{\theta>0} \left\{e^{-\theta t}\mathbb{E}\mathrm{tr}\, e^{\theta\mathbf{X}}\right\}.$$

Due to the exchangeability between expectation and trace, the inequality of Proposition 2.3.1 could be written as

$$\mathbb{P}(\lambda_1(\mathbf{X}) \geq t) \leq \inf_{\theta>0} \left\{e^{-\theta t}\mathrm{tr}\, \boldsymbol{\Upsilon}_{\mathbf{X}}(\theta)\right\}.$$

Proposition 2.3.2. *For $\mathbf{X} \in \mathcal{S}_p$ and every $t \geq 0$ and $\theta > 0$*

$$
\begin{aligned}
\mathbb{P}(\|\!|\mathbf{X}|\!\| \geq t) &\leq \mathbb{P}(\lambda_1(\mathbf{X}) \geq t) + \mathbb{P}(\lambda_1(-\mathbf{X}) \geq t) \\
&\leq e^{-t\theta}\mathbb{E}\mathrm{tr}\, e^{\theta\mathbf{X}} + e^{-t\theta}\mathbb{E}\mathrm{tr}\, e^{-\theta\mathbf{X}}
\end{aligned}
$$

Proof. Observe that $\|\!|\mathbf{X}|\!\| = \max\{\lambda_1(\mathbf{X}), -\lambda_p(\mathbf{X})\}$. Hence,

$$
\begin{aligned}
\mathbb{P}(\|\!|\mathbf{X}|\!\| \geq t) &= \mathbb{P}([\lambda_1(\mathbf{X}) \geq t] \cup [-\lambda_p(\mathbf{X}) \geq t]) \\
&\leq \mathbb{P}(\lambda_1(\mathbf{X}) \geq t) + \mathbb{P}(-\lambda_p(\mathbf{X}) \geq t).
\end{aligned}
$$

Also, $\lambda_1(-\mathbf{X}) = -\lambda_p(\mathbf{X})$, so applying the same procedure of Proposition 2.3.1 the result follows. □

From Proposition 2.3.2, If $\mathbf{X} \sim -\mathbf{X}$, then

$$\mathbb{P}(\|\!|\mathbf{X}|\!\| \geq t) \leq 2\mathbb{P}(\lambda_1(\mathbf{X}) \geq t) \leq 2e^{-t\theta}\mathbb{E}\mathrm{tr}\, e^{\theta\mathbf{X}}.$$

Similar to Proposition 2.3.1, the inequality of Proposition 2.3.2 can be written in terms of $\boldsymbol{\Upsilon}_{\mathbf{X}}$ as

$$\mathbb{P}(\|\!|\mathbf{X}|\!\| \geq t) \leq e^{-t\theta}\mathrm{tr}\, \left(\boldsymbol{\Upsilon}_{\mathbf{X}}(\theta) + \boldsymbol{\Upsilon}_{\mathbf{X}}(-\theta)\right), \quad \theta > 0.$$

If $\boldsymbol{\Upsilon}_{\mathbf{X}}(\theta)$ exists for all θ in a non trivial interval around $\theta = 0$ and $\mathbf{X} \sim -\mathbf{X}$ we get that $\boldsymbol{\Upsilon}_{\mathbf{X}}(\theta) = \boldsymbol{\Upsilon}_{\mathbf{X}}(-\theta)$ and

$$\mathbb{P}(|||\mathbf{X}||| \geq t) \leq 2 \inf_{\theta \in \mathbb{R}} \left\{ e^{-t\theta} \operatorname{tr} \boldsymbol{\Upsilon}_{\mathbf{X}}(\theta) \right\}.$$

As we can see, knowledge of $\boldsymbol{\Upsilon}_{\mathbf{X}}(\theta)$ could give us informative bounds for the concentration of $|||\mathbf{X}|||$.

Example 2.3.1. Let $g \sim \mathcal{N}(0,1)$ and $\mathbf{A}$ a fix matrix in $\mathcal{S}_p$. Define the random matrix $\mathbf{X} \in \mathcal{S}_p$ as

$$\mathbf{X} = g\mathbf{A}.$$

The moment generating function of a standard Gaussian is the function $M_g(\theta) = e^{\theta^2/2}$, $\theta \in \mathbb{R}$, so the trace of $\boldsymbol{\Upsilon}_{\mathbf{X}}(\theta)$, $\theta \in \mathbb{R}$, is given by

$$\begin{aligned}
\operatorname{tr} \boldsymbol{\Upsilon}_{\mathbf{X}}(\theta) &= \operatorname{tr} \mathbb{E} e^{g\theta\mathbf{A}} \\
&= \sum_{j=1}^{p} \mathbb{E} e^{g\theta\lambda_j(\mathbf{A})} \\
&= \sum_{j=1}^{p} M_g\left(\theta\lambda_j(\mathbf{A})\right) \\
&= \sum_{j=1}^{p} e^{(\theta\lambda_j(\mathbf{A}))^2/2} \\
&\leq p e^{\theta^2 |||\mathbf{A}|||^2/2}.
\end{aligned}$$

Since $M_g(\theta) = M_g(-\theta)$, we get that $\boldsymbol{\Upsilon}_{\mathbf{X}}(\theta) = \boldsymbol{\Upsilon}_{\mathbf{X}}(-\theta)$. Then, by Proposition 2.3.2, for $t \geq 0$,

$$\mathbb{P}\left(|||\mathbf{X}||| \geq t\right) \leq 2pe^{-t\theta + \theta^2 |||\mathbf{A}|||^2/2}, \quad \theta \in \mathbb{R}.$$

Optimizing the last bound over θ, i.e., deriving $-t\theta + \theta^2 |||\mathbf{A}|||^2/2$ with respect to θ, finding the root[7] and substitute it in the bound we arrive at

$$\mathbb{P}\left(|||\mathbf{X}||| \geq t\right) \leq 2pe^{-t^2/(2|||\mathbf{A}|||^2)}.$$

♣

[7] $\theta \mapsto -t\theta + \theta^2 |||\mathbf{A}|||^2/2$ is convex on $\mathbb{R}$, so any root of its derivative is a minimum.

In the rest of the thesis we will be mainly interested in constructing concentration bounds for $\mathbf{X} \in \mathcal{S}_p$ written as the sum

$$\mathbf{X} = \mathbf{X}_1 + \cdots + \mathbf{X}_n,$$

where of course $\mathbf{X}_i \in \mathcal{S}_p$, $i = 1, ..., n$. As the intuition may suggest, in general we do not have that

$$\Upsilon_{\mathbf{X}}(\theta) = \prod_{i=1}^{n} \Upsilon_{\mathbf{X}_i}(\theta).$$

When we are working with scalar random variables the previous equality is satisfied. But for random matrices is not true since for $\mathbf{A}, \mathbf{H} \in \mathcal{S}_p$

$$e^{\mathbf{A}+\mathbf{H}} \neq e^{\mathbf{A}} e^{\mathbf{H}}, \quad \text{unless } \mathbf{A} \text{ and } \mathbf{H} \text{ commute.}$$

One can argue that the situation improves if we introduce the trace. This is actually the case, i.e., it can be proved that

$$\operatorname{tr} e^{\mathbf{A}+\mathbf{H}} \leq \operatorname{tr} \left(e^{\mathbf{A}} e^{\mathbf{H}} \right). \tag{2.4}$$

This is called the Golden-Thompson inequality. The proof of (2.4) is beyond the scope of this thesis. One can see [34] for an sketched proof and [35] for an application of this inequality in the context of random matrices. Unfortunately, the Golden-Thompson inequality fails for more than two matrices.

In the next section we approach a result called Lieb's inequality which gives us a method to obtain concentration inequalities for the sum of independent random matrices, circumventing the problem of non-commutativity. The power of Lieb's inequality in this context is provided mainly in [45].

2.4 Lieb's Theorem and sum of random matrices

The next theorem is proved succinctly in [43]. We strongly recommend Chapter 8 of [45] for a theoretical background on the theory necessary to prove Lieb's inequality. In particular, one need to prove with the aid of several other results that the function $D(\mathbf{A}; \mathbf{H})$, $\mathbf{A}, \mathbf{H} \succ 0$ defined as

$$D(\mathbf{A}; \mathbf{H}) = \operatorname{tr} \left[\mathbf{A}(\log \mathbf{A} - \log \mathbf{H}) - (\mathbf{A} - \mathbf{H}) \right]$$

is convex, meaning that for $\mathbf{A}_i, \mathbf{H}_i \succ 0$, $i = 1, 2$,

$$D(t\mathbf{A}_1 + (1-t)\mathbf{A}_2;\, t\mathbf{H}_1 + (1-t)\mathbf{H}_2) \leq tD(\mathbf{A}_1; \mathbf{H}_1) + (1-t)D(\mathbf{A}_2; \mathbf{H}_2), \quad t \in [0, 1].$$

Theorem 2.4.1 (Lieb's inequality)**.** *Let* $\mathbf{H}$ *be a* $p \times p$ *symmetric matrix. The mapping*

$$\mathbf{A} \mapsto \mathrm{tr}\ \exp\left(\mathbf{H} + \log(\mathbf{A})\right)$$

is concave on $\mathcal{S}_p^+ = \{\mathbf{A} \in \mathcal{S}_p : \lambda_p(\mathbf{A}) > 0\}$.

A simple but powerful consequence of Lieb's Theorem is the next corollary.

Corollary 2.4.2. *Let* $\mathbf{X} \in \mathcal{S}_p$ *be a random matrix and* $\mathbf{H} \in \mathcal{S}_p$ *a fixed matrix. Then,*

$$\mathbb{E}\mathrm{tr}\ \exp\left(\mathbf{X} + \mathbf{H}\right) \leq \mathrm{tr}\ \exp\left(\log \mathbb{E} e^{\mathbf{X}} + \mathbf{H}\right).$$

Proof. Let $\mathbf{Y} = e^{\mathbf{X}}$. Then, by Lieb's Theorem and Jensen's inequality (Lemma C.3.1 of Appendix C),

$$\begin{aligned}\mathbb{E}\mathrm{tr}\ \exp\left(\mathbf{X} + \mathbf{H}\right) &= \mathbb{E}\mathrm{tr}\ \exp\left(\log \mathbf{Y} + \mathbf{H}\right)\\ &\leq \mathrm{tr}\ \exp\left(\log \mathbb{E}\mathbf{Y} + \mathbf{H}\right)\\ &= \mathrm{tr}\ \exp\left(\log \mathbb{E} e^{\mathbf{X}} + \mathbf{H}\right).\end{aligned}$$

□

Corollary 2.4.2 is directly applied to the next lemma in order to obtain a more general bound on the expectation. Lemma 2.4.3 will be subsequently applied to get a concentration bound on the greatest eigenvalue.

Lemma 2.4.3 (Tropp's expectation bound)**.** *Let* $\mathbf{X}_1, ..., \mathbf{X}_n \in \mathcal{S}_p$ *be independent random matrices and* $\mathbf{H} \in \mathcal{S}_p$ *a deterministic matrix. Then,*

$$\mathbb{E}\mathrm{tr}\ \exp\left(\sum_{j=1}^{n} \mathbf{X}_j + \mathbf{H}\right) \leq \mathrm{tr}\ \exp\left(\sum_{j=1}^{n} \log \mathbb{E} e^{\mathbf{X}_j} + \mathbf{H}\right).$$

Proof. Using the notation $\mathbb{E}_1[\cdot] = \mathbb{E}[\cdot|\mathbf{X}_2, ..., \mathbf{X}_n]$, $\mathbb{E}_n[\cdot] = \mathbb{E}[\cdot|\mathbf{X}_1, ..., \mathbf{X}_{n-1}]$ and $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot|\mathbf{X}_1, ..., \mathbf{X}_{k-1}, \mathbf{X}_{k+1}, ..., \mathbf{X}_n]$ for $1 < k < n$, we have that

$$\mathbb{E}\mathrm{tr}\ \exp\left(\sum_{j=1}^{n} \mathbf{X}_j + \mathbf{H}\right) = \mathbb{E}\mathbb{E}_n\mathrm{tr}\ \exp\left(\sum_{j=1}^{n-1} \mathbf{X}_j + \mathbf{H} + \mathbf{X}_n\right). \tag{2.5}$$

Applying Corollary 2.4.2 to the inside expectation on (2.5) we obtain that

$$\begin{aligned}\mathbb{E}\mathrm{tr}\ \exp\left(\sum_{j=1}^{n}\mathbf{X}_j+\mathbf{H}\right)&\leq\mathbb{E}\mathrm{tr}\ \exp\left(\sum_{j=1}^{n-1}\mathbf{X}_j+\mathbf{H}+\log\mathbb{E}_n e^{\mathbf{X}_n}\right)\\&=\mathbb{E}\mathrm{tr}\ \exp\left(\sum_{j=1}^{n-1}\mathbf{X}_j+\mathbf{H}+\log\mathbb{E}e^{\mathbf{X}_n}\right),\end{aligned}$$

where in the last equality the expectation $\mathbb{E}_n$ is replaced by $\mathbb{E}$ due to the independence of $\mathbf{X}_1, ..., \mathbf{X}_n$. Applying the same argument recursively we get

$$\begin{aligned}\mathbb{E}\mathrm{tr}\ \exp\left(\sum_{j=1}^{n}\mathbf{X}_j+\mathbf{H}\right)&\leq\mathbb{E}\mathbb{E}_{n-1}\mathrm{tr}\ \exp\left(\sum_{j=1}^{n-2}\mathbf{X}_j+\mathbf{H}+\log\mathbb{E}e^{\mathbf{X}_n}+\mathbf{X}_{n-1}\right)\\&\leq\mathbb{E}\mathrm{tr}\ \exp\left(\sum_{j=1}^{n-2}\mathbf{X}_j+\mathbf{H}+\log\mathbb{E}e^{\mathbf{X}_{n-1}}+\log\mathbb{E}e^{\mathbf{X}_n}\right)\\&\vdots\\&\leq\mathbb{E}\mathrm{tr}\ \exp\left(\mathbf{X}_1+\mathbf{H}+\sum_{j=2}^{n}\log\mathbb{E}e^{\mathbf{X}_j}\right)\\&=\mathbb{E}\mathbb{E}_1\mathrm{tr}\ \exp\left(\mathbf{H}+\sum_{j=2}^{n}\log\mathbb{E}e^{\mathbf{X}_j}+\mathbf{X}_1\right)\\&\leq\mathbb{E}\mathrm{tr}\ \exp\left(\mathbf{H}+\sum_{j=1}^{n}\log\mathbb{E}e^{\mathbf{X}_j}\right).\end{aligned}$$

□

Theorem 2.4.4 (Tropp's probability bound)**.** *Let* $\mathbf{X}_1, ..., \mathbf{X}_n \in \mathcal{S}_p$ *be independent random matrices and* $\mathbf{H} \in \mathcal{S}_p$ *a deterministic matrix. Then, for every* $\theta > 0$ *and* $t \geq 0$

$$\mathbb{P}\left(\lambda_1\left(\sum_{j=1}^{n}\mathbf{X}_j+\mathbf{H}\right)\geq t\right)\leq e^{-\theta t}\mathrm{tr}\ \exp\left(\sum_{j=1}^{n}\log\mathbb{E}e^{\theta\mathbf{X}_j}+\theta\mathbf{H}\right).$$

Proof. Using Proposition 2.3.1 and Lemma 2.4.3 we get

$$\mathbb{P}\left(\lambda_1\left(\sum_{j=1}^{n}\mathbf{X}_j+\mathbf{H}\right)\geq t\right)=\mathbb{P}\left(\theta\lambda_1\left(\sum_{j=1}^{n}\mathbf{X}_j+\mathbf{H}\right)\geq\theta t\right)$$

$$
\begin{aligned}
&\leq e^{-\theta t}\mathbb{E}\mathrm{tr}\,\exp\left(\sum_{j=1}^{n}\theta\mathbf{X}_j+\theta\mathbf{H}\right)\\
&\leq e^{-\theta t}\mathrm{tr}\,\exp\left(\sum_{j=1}^{n}\log\mathbb{E}e^{\theta\mathbf{X}_j}+\theta\mathbf{H}\right).
\end{aligned}
$$

□

As in Proposition 2.3.1 we can take the infimum over $\theta > 0$ to optimize the bound. Theorem 2.4.4 is presented in [45] as one of the *master bounds* for sum of independent random matrices. The other master bound is on $\mathbb{E}|||\mathbf{X}|||$. In this thesis, we'll not use bounds on the expected value. Instead, we'll be interested in classical concentration inequalities of the form of Theorem 2.4.4 and Examples 2.3.1 and 2.4.1. Besides, once we have obtained a bound on $\mathbb{P}(|||\mathbf{X}||| \geq t)$, we can get bounds on $\mathbb{E}|||\mathbf{X}|||$ with the formula

$$\mathbb{E}|||\mathbf{X}||| = \int_0^\infty \mathbb{P}\left(|||\mathbf{X}||| \geq t\right)\, dt.$$

Example 2.4.1. Like in Example 2.3.1, here we obtain a bound on the matrix $\mathbf{X} \in \mathcal{S}_p$ defined as

$$\mathbf{X} = \sum_{j=1}^{n} g_j\mathbf{A}_j,$$

where $g_1, ..., g_n$ are iid $\mathcal{N}(0,1)$ and $\mathbf{A}_1, ..., \mathbf{A}_n$ are fixed $p \times p$ symmetric matrices. Let $g \sim \mathcal{N}(0,1)$ and $\mathbf{A} \in \mathcal{S}_p$ fixed. Now, instead of bounding $\mathbb{E}\mathrm{tr}\, e^{\theta g\mathbf{A}}$, $\theta > 0$, we'll obtain an explicit expression for $\mathbb{E}e^{\theta g\mathbf{A}}$. It is well known that

$$\mathbb{E}g^k = \begin{cases} 0, & k \text{ odd},\\ \dfrac{k!}{2^{k/2}(k/2)!} & k \text{ even}.\end{cases}$$

Then by definition of matrix exponential (2.1) we get that

$$
\begin{aligned}
\mathbb{E}e^{\theta g\mathbf{A}} &= \mathbf{I} + \sum_{k=1}^{\infty}\frac{(\theta\mathbf{A})^k\mathbb{E}g^k}{k!}\\
&= \mathbf{I} + \sum_{k=1}^{\infty}\frac{(\theta\mathbf{A})^{2k}}{(2k)!}\frac{(2k)!}{2^k k!}
\end{aligned}
$$

$$\begin{aligned} &= \mathbf{I} + \sum_{k=1}^{\infty} \frac{(\theta^2 \mathbf{A}^2/2)^k}{k!} \\ &= e^{\theta^2 \mathbf{A}^2/2}. \end{aligned}$$

Then, $\log \mathbb{E} e^{\theta g \mathbf{A}} = \theta^2 \mathbf{A}^2/2$ and

$$\sum_{j=1}^{n} \log \mathbb{E} e^{\theta g_j \mathbf{A}_j} = \frac{\theta^2}{2} \sum_{j=1}^{n} \mathbf{A}_j^2.$$

Therefore, by Theorem 2.4.4 (with $\mathbf{H} = \mathbf{0}$) we conclude that

$$\begin{aligned} \mathbb{P}\left(\lambda_1(\mathbf{X}) \geq t\right) &\leq e^{-t\theta} \mathrm{tr}\ \exp\left(\frac{\theta^2}{2} \sum_{j=1}^{n} \mathbf{A}_j^2 \right) \\ &\leq e^{-t\theta} p \exp\left(\frac{\theta^2}{2} \lambda_1 \left(\sum_{j=1}^{n} \mathbf{A}_j^2 \right) \right) \\ &\leq p e^{-t\theta} \exp\left(\frac{\theta^2}{2} \left\vert\left\vert\left\vert \sum_{j=1}^{n} \mathbf{A}_j^2 \right\vert\right\vert\right\vert \right) \end{aligned}$$

If we define $\sigma^2 = \vert\vert\vert \mathbb{E} \mathbf{X}^2 \vert\vert\vert$, then, by independence of $g_1, ..., g_n$,

$$\sigma^2 = \left\vert\left\vert\left\vert \sum_{j=1}^{n} \mathbb{E} g_j^2 \mathbf{A}_j^2 \right\vert\right\vert\right\vert = \left\vert\left\vert\left\vert \sum_{j=1}^{n} \mathbf{A}_j^2 \right\vert\right\vert\right\vert.$$

Thus,

$$\mathbb{P}\left(\lambda_1(\mathbf{X}) \geq t\right) \leq p \exp\left(-t\theta + \frac{\theta^2}{2} \sigma^2 \right).$$

And optimizing on $\theta > 0$ we finally get that

$$\mathbb{P}\left(\lambda_1(\mathbf{X}) \geq t\right) \leq p e^{-\frac{t^2}{2\sigma^2}}.$$

To get a bound on $\vert\vert\vert \mathbf{X} \vert\vert\vert$, note that $\mathbf{X} \sim -\mathbf{X}$, and by Proposition 2.3.2,

$$\mathbb{P}\left(\vert\vert\vert \mathbf{X} \vert\vert\vert \geq t\right) \leq 2p e^{-\frac{t^2}{2\sigma^2}}. \tag{2.6}$$

The bound (2.6) is useful, i.e., is less that or equal to one, when $t \geq \sigma \sqrt{2 \log(2p)}$, so we can rewrite it in the following way: for any $\delta \geq 0$ we have that

$$\vert\vert\vert \mathbf{X} \vert\vert\vert \geq \sigma \sqrt{2(1+\delta) \log(2p)},$$

with probability at most $(2p)^{-\delta}$. This is obtained taking $t = \sigma \sqrt{2(1+\delta) \log(2p)}$. ♣

2.5 Summary and observations

In this chapter we saw how to define the function $f(\mathbf{A})$ of a symmetric matrix $\mathbf{A} \in \mathcal{S}_p$ and how to order matrices in $\mathcal{S}_p$ through the semidefinite order $\preceq$. Then, we presented a more formal definition of a random matrix and the moment generating function of a random symmetric matrix. Finally, we showed how to apply the Lieb's inequality in order to obtain general concentration inequalities for the sum of iid symmetric random matrices in terms of the spectral norm.

Examples 2.3.1 and 2.4.1 illustrate most of the technique used in this thesis:

1. Define a symmetric random matrix $\mathbf{X} = \sum_{j=1}^{n} \mathbf{X}_j$. We didn't make use of it in example 2.4.1, but usually we also define a deterministic matrix $\mathbf{H}$ which represents the matrix we want to estimate.
2. Find a upper bound for $\mathbb{E}e^{\theta \mathbf{X}_j}$ in terms of semidefinite order. In Example 2.4.1 we were able to find the expectation explicitly, but this would not be the case in the future. In fact this will be the goal of the robust methodology.
3. Use the stated results to find a bound for $\mathbb{P}(\lambda_1(\mathbf{X}) \geq t)$.
4. Verify that $\mathbb{P}(\lambda_1(\mathbf{X}) \geq t) = \mathbb{P}(\lambda_1(-\mathbf{X}) \geq t)$ to obtain a bound for $\mathbb{P}(\|\|\mathbf{X}\|\| \geq t)$.

Applications of the methodology presented in this chapter can be found in [45], which ranges from covariance matrix estimation, randomized numerical linear algebra (see also [31]), data matrix sparsification, random feature maps, and random graphs.

In the next chapter we deviate from the previous development in order to emphasize the simplicity and generality of the results obtained. Chapter 3 will be devoted to the applications of the theory of metric entropy and concentration bounds for Gaussian and sub-Gaussian processes.

Chapter 3

Covariance matrix estimation

This chapter presents some of the basic known techniques to obtain concentration inequalities in the context of covariance matrix estimation. The objective of this chapter is to contrast the development of Chapter 2 with the classical results in one of the most known applications of random matrix theory. The first techniques will deviate broadly from the ones developed in Chapter 2 since we'll make strong distributional assumptions. The main sources of the theory presented are [4], [47] and [48].

First, we'll present the basic sub-Gaussian assumption for random variables and how can be used to obtain concentration bounds. Then, the following sections are dedicated to obtaining theoretical guarantees under Gaussian or sub-Gaussian assumptions. To do this, we present some known results which are proved in Appendix B. The last section is dedicated to obtaining a first robust estimation with the techniques of Chapter 2. Before presenting all the results, let us state the context of the estimation problem.

Suppose that $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are n iid copies of the random vector $\boldsymbol{X} \in \mathbb{R}^p$ with $\mathbb{E}\boldsymbol{X} = \boldsymbol{\mu}$ and $\mathrm{Cov}\boldsymbol{X} = \boldsymbol{\Sigma}$. We want to estimate $\boldsymbol{\Sigma}$ through $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$. Usually we will do it with the empirical covariance matrix $\widehat{\boldsymbol{\Sigma}}$ defined as

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{X}_j - \bar{\boldsymbol{X}})(\boldsymbol{X}_j - \bar{\boldsymbol{X}})^{\intercal},$$

where $\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i$. Is not difficult to show that $\frac{n}{n-1}\widehat{\boldsymbol{\Sigma}}$ is and unbiased estimator[1] of $\boldsymbol{\Sigma}$ (see Appendix C.) A major difficulty in obtaining concentration bounds for $\widehat{\boldsymbol{\Sigma}}$ is that the

[1] This implies that $\mathbb{E}\widehat{\boldsymbol{\Sigma}} = \frac{n-1}{n}\boldsymbol{\Sigma} \to \boldsymbol{\Sigma}$, $n \to \infty$, entry-wise.

addends are not independent, since each one depends on $\bar{\boldsymbol{X}}$. To avoid this burden, we'll assume throughout this chapter that $\mathbb{E}\boldsymbol{X} = \mathbf{0}$. In Chapter 5 we'll obtain concentration inequalities for $\widehat{\mathbf{\Sigma}}$ without this assumption. In the case of this chapter, we can use

$$\widehat{\mathbf{\Sigma}}_0 = \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{X}_j \boldsymbol{X}_j^\intercal$$

as an unbiased estimator of $\mathbf{\Sigma}$, since the mean of this random matrix is

$$\mathbb{E}\widehat{\mathbf{\Sigma}}_0 = \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}\boldsymbol{X}_j \boldsymbol{X}_j^\intercal = \mathbf{\Sigma}.$$

Define the $n \times p$ random matrix $\mathbf{X}$ as the *ensemble*

$$\mathbf{X} = \begin{pmatrix} \boldsymbol{X}_1^\intercal \\ \vdots \\ \boldsymbol{X}_n^\intercal \end{pmatrix}.$$

Then, $\widehat{\mathbf{\Sigma}}_0$ can be expressed as

$$\widehat{\mathbf{\Sigma}}_0 = \frac{1}{n} \mathbf{X}^\intercal \mathbf{X}.$$

The asymptotic distribution of the matrix $\widehat{\mathbf{\Sigma}}_0$ generated by the ensemble $\mathbf{X}$ is studied broadly in random matrix theory (see for example [52]). Instead, as it is done in all of this thesis, we'll do a non-asymptotic study of this estimator giving concentration bounds for the random variable

$$|||\widehat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}|||.$$

To do so, we need to impose some assumptions on the underlying distribution of $\mathbf{X}$, i..e., on the distribution of the rows $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$. We'll present precise definitions of this in the following sections.

3.1 Sub-Gaussian variables and Lipschitz functions

One class of random variables which automatically gives us straightforward concentration bounds are sub-Gaussian random variables, which are defined as follows.

Definition 3.1.1 (Sub-Gaussian random variables)**.** *Let X be a real-valued random variable with $\mathbb{E}X = \mu$. We say that X is sub-Gaussian with parameter $\sigma > 0$ if*

$$\log \mathbb{E} e^{\theta(X-\mu)} \le \frac{\theta^2\sigma^2}{2}, \quad \forall \theta \in \mathbb{R},$$

and we write $X \sim \mathcal{SG}(\sigma)$.

Note that the parameter σ is not unique since for every $v \ge \sigma$, $X \sim \mathcal{SG}(\sigma)$ implies $X \sim \mathcal{SG}(v)$.

Recall that one way to measure the *magnitude* of a random variable is through its L^2-norm defined as

$$\|X\|_{L^2} = \left(\mathbb{E}|X|^2\right)^{1/2}.$$

In a similar way, the sub-Gaussian norm of a random variable is defined as

$$\|X\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}\exp(X^2/t^2) \le 2\right\}.$$

See Appendix B for more details on the sub-Gaussian norm. This quantity is called a norm because it is a norm over the set of real-valued random variables. Even more, it can be proved that whenever $\|X\|_{\psi_2} < \infty$, then $X \sim \mathcal{SG}(\sigma)$ for some $\sigma > 0$, and in fact $\|X\|_{\psi_2}$ is proportional to σ (see Proposition 2.5.2 of [47]). We'll come back to this quantity when we work with sub-Gaussian random vectors, a concept defined letter on.

There is an easy method for obtaining concentration inequalities for sub-Gaussian random variables, in fact, this method motivates its definition. If $X \sim \mathcal{SG}(\sigma)$, from Markov's inequality we deduce that for $t \in \mathbb{R}$ and any $\theta > 0$,

$$\begin{aligned}\mathbb{P}\left(X - \mu \ge t\right) &= \mathbb{P}\left(e^{\theta(X-\mu)} \ge e^{\theta t}\right)\\ &\le e^{-\theta t}\mathbb{E}e^{\theta(X-\mu)}\\ &\le e^{-\theta t + \frac{\theta^2\sigma^2}{2}}.\end{aligned}$$

From the last inequality we can see why sub-Gaussian random variables are so useful: if one can bound the moment generating function of $X - \mu$, then we obtain in immediate bound for the tail of the distribution of $X - \mu$. As was done in the Examples 2.3.1 and 2.4.1 of the previous chapter, optimizing over $\theta \in \mathbb{R}$ we turn the previous bound into

$$\mathbb{P}\left(X \ge \mu + t\right) \le \inf_{\theta > 0} e^{-\theta t + \frac{\theta^2\sigma^2}{2}} = e^{-\frac{t^2}{2\sigma^2}}.$$

As a matter of fact, the last inequality characterize sub-Gaussian random variables. This is stated in the next proposition proved in [47, p. 22].

Proposition 3.1.2. *Let X be a real-valued random variable with $\mathbb{E}X = \mu$. The next statements are equivalent.*

(a) $X \sim \mathcal{SG}(\sigma)$.

(b) For some $K > 0$,

$$\mathbb{P}\left(|X - \mu| \geq t\right) \leq 2\exp\left(-t^2/K^2\right), \quad \forall t \geq 0.$$

The constants σ and K are proportional.

Up until this moment we haven't shown if there exist random variables that belong to this sub-Gaussian class. Indeed, it's not difficult to see that at least Gaussian and bounded random variables (like Rademacher) are sub-Gaussian (see for example Chapter 2 of [48].) For Gaussian variables this is obvious since its moment generating function is given by

$$t \mapsto \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right), \quad \mu \in \mathbb{R}, \sigma > 0.$$

The next interesting result tells us that we can obtain sub-Gaussian random variables from Gaussian random vectors. This is done with the aid of Lipschitz functions, which are defined in the following way.

Definition 3.1.3 (Lipschitz function)**.** *Let $\mathcal{Y}$ be some normed space. We say that a function $f : \mathcal{Y} \to \mathbb{R}$ is L-Lipschitz with respect to the norm N if*

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq LN(\boldsymbol{x} - \boldsymbol{y}), \quad \forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathcal{Y}.$$

The next theorem is proved in [4, p. 125]. The authors do this with the aid of a machinery called *logarithmic Sobolev inequalities*, which provides a differential equation inequality that consequently gives a bound on the moment generating function of $f(\boldsymbol{X})$.

Theorem 3.1.4. *Let $\boldsymbol{X} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I})$ and $f : \mathbb{R}^p \to \mathbb{R}$ be a L-Lipschitz function with respect to the ℓ_2 norm. Then,*

$$f(\boldsymbol{X}) - \mathbb{E}f(\boldsymbol{X}) \sim \mathcal{SG}(L).$$

An immediate consequence of Theorem 3.1.4 is that

$$\mathbb{P}\left(f(\boldsymbol{X}) \geq \mathbb{E}f(\boldsymbol{X}) + t\right) \leq e^{-\frac{t^2}{2L^2}}.$$

As well, it is clear that f is L-Lipschitz if and only if $-f$ is L-Lipschitz, so

$$\mathbb{P}\left(f(\boldsymbol{X}) \leq \mathbb{E}f(\boldsymbol{X}) - t\right) = \mathbb{P}\left(-f(\boldsymbol{X}) \geq -\mathbb{E}f(\boldsymbol{X}) + t\right) \leq e^{-\frac{t^2}{2L^2}},$$

or

$$\mathbb{P}\left(f(\boldsymbol{X}) \geq \mathbb{E}f(\boldsymbol{X}) - t\right) \geq 1 - e^{-\frac{t^2}{2L^2}}. \tag{3.1}$$

In the context of random matrices, we are interested in the function $s_1 : \mathcal{M}_{n,p} \to [0, \infty)$ which gives us the maximum singular value of a matrix in $\mathcal{M}_{n,p}$ (recall that $s_1(\mathbf{F}) = |||\mathbf{F}|||$). From a consequence of Weyl's Theorem (see Appendix A) or by the triangle inequality[2] we obtain that

$$|s_1(\mathbf{F}) - s_1(\mathbf{G})| \leq |||\mathbf{F} - \mathbf{G}|||, \quad \mathbf{F}, \mathbf{G} \in \mathcal{M}_{n,p}.$$

The general proof of the previous inequality (i.e. for any singular value s_i) is presented in Proposition A.5.1. Observe that by monotonicity of the norm[3],

$$|s_1(\mathbf{F}) - s_1(\mathbf{G})| \leq |||\mathbf{F} - \mathbf{G}|||_2$$

From the previous result, we can see s_1 as a function from $\mathbb{R}^{np}$ to $[0, \infty)$ and deduce that it is 1-Lipschitz with respect to the ℓ_2 norm. More precisely, define the mapping $\gamma : \mathcal{M}_{n,p} \to \mathbb{R}^{np}$ as

$$\gamma(\mathbf{F}) = (f_{11}, ..., f_{n1}, ..., f_{1p}, ..., f_{np})^{\intercal},$$

i.e., $\gamma(\mathbf{F})$ is obtained by stacking the columns of $\mathbf{F}$ in one big column. Then, γ is an isometry from $(\mathcal{M}_{n,p}, |||\cdot|||_2)$ to $(\mathbb{R}^{np}, \|\cdot\|_2)$. If we define the function $\tilde{s}_1 : \mathbb{R}^{np} \to [0, \infty)$ as

$$\tilde{s}_1(\boldsymbol{x}) = s_1(\gamma^{-1}(\boldsymbol{x})),$$

then, $\tilde{s}_1$ is 1-Lipschitz since

$$|\tilde{s}_1(\boldsymbol{x}) - \tilde{s}_1(\boldsymbol{y})| = |s_1(\gamma^{-1}(\boldsymbol{x})) - s_1(\gamma^{-1}(\boldsymbol{y}))| \leq |||\gamma^{-1}(\boldsymbol{x}) - \gamma^{-1}(\boldsymbol{y})|||_2 = \|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

[2] Indeed, one just need to see that $|||\mathbf{F}||| \leq |||\mathbf{F} - \mathbf{G}||| + |||\mathbf{G}|||$ and $|||\mathbf{G}||| \leq |||\mathbf{F} - \mathbf{G}||| + |||\mathbf{F}|||$.

[3] For $1 \leq k \leq q \leq \infty$, $|||\mathbf{F}|||_k \geq |||\mathbf{F}|||_q$, see Proposition A.4.3.

For the last equality, note that the Frobenius norm can be calculated by summation of all the squared terms of a matrix (see Proposition A.4.3). This observations stress that for a random matrix $\mathbf{X}$, working with $|||\mathbf{X}||| = s_1(\mathbf{X})$ is equivalent to working with $\tilde{s}_1(\gamma(\mathbf{X}))$. So establishing that a real-valued matrix function is Lipschitz enables us to use Theorem 3.1.4 whenever $\gamma(\mathbf{X})$ is a standard Gaussian random vector. This is what we'll do in the next section.

3.2 Gaussian ensembles

As it is always the case in Statistics, we'll begin the journey of obtaining good estimators starting from Gaussian samples. In this case we need to define what is a *Gaussian matrix.* Instead of doing it entry-wise, we adopt a more general definition and do it row-wise as an ensemble. Recall that the matrix $\mathbf{X}$ is an ensemble of the vectors $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ if

$$\mathbf{X} = \begin{pmatrix} \boldsymbol{X}_1^\intercal \\ \vdots \\ \boldsymbol{X}_n^\intercal \end{pmatrix}.$$

Observe that, as we mentioned in the beginning of the chapter, all the vectors considered will be centered, i.e., $\mathbb{E}\boldsymbol{X}_i = \mathbf{0}$.

Definition 3.2.1 (Gaussian ensemble)**.** *If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid random vectors in $\mathbb{R}^p$ with distribution $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, then the $n \times p$ ensemble $\mathbf{X}$ of the vectors $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ is said to be a $\boldsymbol{\Sigma}$-Gaussian ensemble and we write $\mathbf{X} \sim \mathcal{N}_{n\times p}(\boldsymbol{\Sigma})$.*

In rest of this section we'll use the following observation extensively.

Remark 3.2.1. Define $\mathbf{W} \sim \mathcal{N}_{n\times p}(\mathbf{I})$, so the entries of $\mathbf{W}$ are iid $\mathcal{N}(0,1)$. If $\boldsymbol{X}_i$ and $\boldsymbol{W}_i$ are the Gaussian vectors that conforms the ensembles $\mathbf{X}$ and $\mathbf{W}$ respectively, then $\boldsymbol{X}_i \sim \sqrt{\boldsymbol{\Sigma}}\boldsymbol{W}_i$, so

$$\mathbf{X} \sim \begin{pmatrix} \boldsymbol{W}_1^\intercal\sqrt{\boldsymbol{\Sigma}} \\ \vdots \\ \boldsymbol{W}_n^\intercal\sqrt{\boldsymbol{\Sigma}} \end{pmatrix} = \mathbf{W}\sqrt{\boldsymbol{\Sigma}}.$$

♠

The next theorem is a consequence of Theorem 3.1.3, and tells us how much the norm of a Gaussian ensemble deviates from its mean value.

Theorem 3.2.2. *Let $\mathbf{X} \sim \mathcal{N}_{n\times p}(\mathbf{\Sigma})$. Then, for any $t \geq 0$*

$$\mathbb{P}\left(|||\mathbf{X}||| \geq \mathbb{E}|||\mathbf{X}||| + t\right) \leq \exp\left(\frac{-t^2}{2\lambda_1(\mathbf{\Sigma})}\right).$$

Proof. Define $\mathbf{W} \sim \mathcal{N}_{n\times p}(\mathbf{I})$ and the mapping $z : \mathcal{M}_{n,p} \to [0,\infty)$ as

$$z(\mathbf{F}) = s_1(\mathbf{F}\sqrt{\mathbf{\Sigma}}).$$

Due to the fact that s_1 is 1-Lipschitz, we get by submultiplicity of the norm[4] that

$$|z(\mathbf{F}) - z(\mathbf{G})| \leq |||(\mathbf{F}-\mathbf{G})\sqrt{\mathbf{\Sigma}}||| \leq |||\mathbf{F}-\mathbf{G}||| |||\sqrt{\mathbf{\Sigma}}||| \leq |||\mathbf{F}-\mathbf{G}|||_2 \lambda_1(\sqrt{\mathbf{\Sigma}}),$$

so z is $\lambda_1(\sqrt{\mathbf{\Sigma}})$-Lipschitz with respect to the Frobenius norm. Then, since $\gamma(\mathbf{W}) \sim \mathcal{N}_{np}(\mathbf{0}, \mathbf{I})$, Theorem 3.1.4 implies that

$$\begin{aligned}\mathbb{P}\left(|||\mathbf{X}||| \geq \mathbb{E}|||\mathbf{X}||| + t\right) &= \mathbb{P}\left(z(\mathbf{W}) \geq \mathbb{E}z(\mathbf{W}) + t\right)\\ &\leq \exp\left(-\frac{t^2}{2\lambda_1(\sqrt{\mathbf{\Sigma}})^2}\right).\end{aligned}$$

Noting that $\lambda_1(\sqrt{\mathbf{\Sigma}})^2 = \lambda_1(\mathbf{\Sigma})$ ends the proof. □

Theorem 3.2.2 is interesting in its own sake. But in the case of covariance estimation we'll like to obtain concentration results for matrices of the form $\mathbf{X}^\intercal\mathbf{X}$, which are not Gaussian if $\mathbf{X}$ is a Gaussian ensemble. To see this one can note that the elements of its diagonal are always positive, so we can not impose a Gaussian distribution in all the entries of $\mathbf{X}^\intercal\mathbf{X}$.

Instead of using directly Theorem 3.2.2 to obtain concentrations guarantees for the covariance estimation, we'll apply some heavy machinery from the theory of concentration for Gaussian processes (see for example [24, Chapter 3] and [47, Chapter 7]). The next important result will rely strongly on the following lemma, which proof can be found in Appendix B.

Lemma 3.2.3. *Let $\mathbf{X} \sim \mathcal{N}_{n\times p}(\mathbf{\Sigma})$ with $\mathbf{\Sigma} \succ \mathbf{0}$. Then,*

[4] For $p \geq 1$, $|||\mathbf{FG}|||_p \leq |||\mathbf{F}|||_p |||\mathbf{G}|||_p$, see Proposition A.4.3.

(a)

$$\mathbb{E}s_1(\mathbf{X}) \leq \sqrt{n}\lambda_1(\sqrt{\mathbf{\Sigma}}) + \sqrt{\operatorname{tr}\mathbf{\Sigma}}.$$

(b) If $n \geq p$,

$$\mathbb{E}\left[\min_{\boldsymbol{v}\in V(R)} \frac{\|\mathbf{X}\boldsymbol{v}\|_2}{\sqrt{n}}\right] \geq 1 - R\sqrt{\frac{\operatorname{tr}\mathbf{\Sigma}}{n}},$$

where $R = 1/\lambda_p(\sqrt{\mathbf{\Sigma}})$ and $V(R) = \{\boldsymbol{v} \in \mathbb{R}^p : \|\sqrt{\mathbf{\Sigma}}\boldsymbol{v}\|_2 = 1, \|\boldsymbol{v}\|_2 \leq R\}$.

The proof of Lemma 3.2.3 relies strongly in the fact that

$$s_1(\mathbf{X}) = \max_{\substack{\|\boldsymbol{u}\|_2=1\\ \|\boldsymbol{v}\|_2=1}} \boldsymbol{u}^\intercal\mathbf{X}\boldsymbol{v} \quad \text{and} \quad s_p(\mathbf{X}) = \min_{\substack{\|\boldsymbol{u}\|_2=1\\ \|\boldsymbol{v}\|_2=1}} \boldsymbol{u}^\intercal\mathbf{X}\boldsymbol{v},$$

(see Theorems A.5.4 and A.5.5), and that $\boldsymbol{u}^\intercal\mathbf{W}\boldsymbol{v}$, where $\mathbf{W} \sim \mathcal{N}_{n\times p}(\mathbf{0}, \mathbf{I})$, is a Gaussian process in $\mathbb{R}^n \times \mathbb{R}^p$. To obtain the inequalities one uses to powerful bounds on Gaussian processes called Sudakov-Fernique and Gordon. See Appendix B for more details.

The next theorem is our first concentration result on covariance matrix estimation.

Theorem 3.2.4. *Let $\mathbf{X} \sim \mathcal{N}_{n\times p}(\mathbf{\Sigma})$ with $n \geq p$ and $\mathbf{\Sigma} \succ \mathbf{0}$. Then, for any $\delta > 0$,*

$$\mathbb{P}\left(\frac{|||\widehat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}|||}{|||\mathbf{\Sigma}|||} \geq 2\epsilon + \epsilon^2\right) \leq 2e^{-n\delta^2/2},$$

where $\widehat{\mathbf{\Sigma}}_0 = n^{-1}\mathbf{X}^\intercal\mathbf{X}$ and $\epsilon = \delta + \sqrt{p/n}$.

Proof. First, we'll obtain concentration bounds for $s_1(\mathbf{X})$ and $s_p(\mathbf{X})$ and then we proceed to bound $|||\widehat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}|||$. Throughout the proof we denote $\mathbf{W} \sim \mathcal{N}_{n\times p}(\mathbf{I})$. Also, notice that $\lambda_i(\mathbf{\Sigma}) = s_i(\mathbf{\Sigma})$ for all i since $\mathbf{\Sigma}$ is positive definite.

Bounds on singular values. Define $z_1 = \lambda_1(\sqrt{\mathbf{\Sigma}})$. From Theorem 3.2.2 we get that, for $\delta \geq 0$,

$$\mathbb{P}\left(s_1(\mathbf{X}) \geq \mathbb{E}s_1(\mathbf{X}) + \sqrt{n}z_1\delta\right) \leq e^{-n\delta^2/2}.$$

From Lemma 3.2.3 (a) we know that

$$\mathbb{E}s_1(\mathbf{X}) \leq \sqrt{n}z_1 + \sqrt{\operatorname{tr}\mathbf{\Sigma}},$$

so

$$\mathbb{P}\left(\frac{s_1(\mathbf{X})}{\sqrt{n}} \geq z_1(1+\delta) + \sqrt{\frac{\operatorname{tr}\mathbf{\Sigma}}{n}}\right) \leq e^{-n\delta^2/2}. \tag{3.2}$$

Now to obtain a similar bound for $s_p(\mathbf{X})$, define $z_p = \lambda_p(\sqrt{\mathbf{\Sigma}})$, $R = 1/z_p$ and $V(R) = \{\boldsymbol{v} \in \mathbb{R}^p : \|\sqrt{\mathbf{\Sigma}}\boldsymbol{v}\|_2 = 1, \|\boldsymbol{v}\|_2 \leq R\}$. Note that by definition[5]

$$z_p = \min_{\|\boldsymbol{w}\|_2=1} \|\sqrt{\mathbf{\Sigma}}\boldsymbol{w}\|_2.$$

Suppose that the following inequality holds:

$$\min_{\boldsymbol{v}\in V(R)} \frac{\|\mathbf{X}\boldsymbol{v}\|_2}{\sqrt{n}} \geq 1 - \delta - R\sqrt{\frac{\operatorname{tr}\mathbf{\Sigma}}{n}}. \tag{3.3}$$

Then, for any $\boldsymbol{w}$ such that $\|\boldsymbol{w}\|_2 = 1$ define $\boldsymbol{v} = \boldsymbol{w}/\|\sqrt{\mathbf{\Sigma}}\boldsymbol{w}\|_2$. By construction we have

$$\|\sqrt{\mathbf{\Sigma}}\boldsymbol{v}\|_2 = 1 \quad \text{and} \quad \|\boldsymbol{v}\|_2 = \frac{1}{\|\sqrt{\mathbf{\Sigma}}\boldsymbol{w}\|_2} \leq \frac{1}{z_p} = R,$$

so $\boldsymbol{v} \in V(R)$. Observe that

$$\frac{\|\mathbf{X}\boldsymbol{w}\|_2}{\sqrt{n}} = \|\sqrt{\mathbf{\Sigma}}\boldsymbol{w}\|_2 \frac{\|\mathbf{X}\boldsymbol{v}\|_2}{\sqrt{n}} \geq z_p \min_{\boldsymbol{v}\in V(R)} \frac{\|\mathbf{X}\boldsymbol{v}\|_2}{\sqrt{n}}.$$

Therefore, by assumption (3.3),

$$\frac{s_p(\mathbf{X})}{\sqrt{n}} = \min_{\|\boldsymbol{w}\|_2=1} \frac{\|\mathbf{X}\boldsymbol{w}\|_2}{\sqrt{n}} \geq (1-\delta)z_p - \sqrt{\frac{\operatorname{tr}\mathbf{\Sigma}}{n}}. \tag{3.4}$$

Now, we want to obtain a probability bound for (3.3). As was done in Theorem 3.2.2, we can easily see that the mapping

$$\mathbf{F} \mapsto \min_{\boldsymbol{v}\in V(R)} \frac{\|\mathbf{F}\sqrt{\mathbf{\Sigma}}\boldsymbol{v}\|_2}{\sqrt{n}},$$

is $(1/\sqrt{n})$-Lipschitz[6]. Then, because $\mathbf{X} \sim \mathbf{W}\sqrt{\mathbf{\Sigma}}$, we obtain from Theorem 3.1.4 and (3.1) that

$$\mathbb{P}\left(\min_{\boldsymbol{v}\in V(R)} \frac{\|\mathbf{X}\boldsymbol{v}\|_2}{\sqrt{n}} \geq \mathbb{E}\left[\min_{\boldsymbol{v}\in V(R)} \frac{\|\mathbf{X}\boldsymbol{v}\|_2}{\sqrt{n}}\right] - \delta\right) \geq 1 - e^{-n\delta^2/2}.$$

[5] See the representation of s_p of Theorem A.5.5.

[6] One just need to see that $\|\mathbf{F}\sqrt{\mathbf{\Sigma}}\boldsymbol{v}\|_2 \leq \|\mathbf{G}\sqrt{\mathbf{\Sigma}}\boldsymbol{v}\|_2 + \|(\mathbf{F}-\mathbf{G})\sqrt{\mathbf{\Sigma}}\boldsymbol{v}\|_2 \leq \|\mathbf{G}\sqrt{\mathbf{\Sigma}}\boldsymbol{v}\|_2 + \|\mathbf{F}-\mathbf{G}\|$, because $\boldsymbol{v} \in V(R)$.

Hence, by Lemma 3.2.3 and because (3.3) implies (3.4), we have that

$$\mathbb{P}\left(\frac{s_p(\mathbf{X})}{\sqrt{n}} \geq (1-\delta)z_p - \sqrt{\frac{\operatorname{tr}\boldsymbol{\Sigma}}{n}}\right) \geq 1 - e^{-n\delta^2/2}. \tag{3.5}$$

Bound on the spectral norm. Suppose for the moment that $\boldsymbol{\Sigma} = \mathbf{I}$, so $\mathbf{X} \sim \mathbf{W}$ and $\widehat{\boldsymbol{\Sigma}}_0 = n^{-1}\mathbf{W}^\intercal\mathbf{W}$. Then,

$$\begin{aligned} |||\widehat{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}||| &= \left|\left|\left|\frac{1}{n}\mathbf{W}^\intercal\mathbf{W} - \mathbf{I}\right|\right|\right| \\ &\leq \max\left\{\frac{[s_1(\mathbf{W})]^2}{n} - 1, \left|\frac{[s_p(\mathbf{W})]^2}{n} - 1\right|\right\}, \end{aligned} \tag{3.6}$$

where we used the fact that $s_j(\mathbf{W}^\intercal\mathbf{W}) = \lambda_j(\mathbf{W}^\intercal\mathbf{W}) = [s_j(\mathbf{W})]^2$ and that for any $p \times p$ symmetric matrices $\mathbf{A}, \mathbf{H}$,

$$|||\mathbf{A} + \mathbf{H}||| = \max\left\{\lambda_1(\mathbf{A}+\mathbf{H}), |\lambda_p(\mathbf{A}+\mathbf{H})|\right\} \leq \max\left\{\lambda_1(\mathbf{A}) + \lambda_1(\mathbf{H}), |\lambda_p(\mathbf{A}) + \lambda_p(\mathbf{H})|\right\}.$$

Note that $\lambda_1(\mathbf{I}^{1/2}) = \lambda_p(\mathbf{I}^{1/2}) = 1$ and $\operatorname{tr}\mathbf{I} = p$. Define $\epsilon = \delta + \sqrt{p/n}$, $\delta > 0$, then from (3.2) and (3.5) of the first part, the event

$$\left[\frac{s_1(\mathbf{W})}{\sqrt{n}} \leq 1 + \epsilon\right] \cap \left[\frac{s_p(\mathbf{W})}{\sqrt{n}} \geq 1 - \epsilon\right] \tag{3.7}$$

occurs with probability at least $1 - 2e^{-n\delta^2/2}$. Furthermore, the event (3.7) implies that

$$(1-\epsilon)^2 - 1 \leq \frac{[s_p(\mathbf{W})]^2}{n} - 1 \leq \frac{[s_1(\mathbf{W})]^2}{n} - 1 \leq (1+\epsilon)^2 - 1,$$

and by (3.6),

$$\begin{aligned} |||\widehat{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}||| &\leq \max\left\{(1+\epsilon)^2 - 1, |(1-\epsilon)^2 - 1|\right\} \\ &= \max\left\{2\epsilon + \epsilon^2, |-2\epsilon + \epsilon^2|\right\} \\ &= 2\epsilon + \epsilon^2 \end{aligned}$$

with probability at least $1 - 2e^{-n\delta^2/2}$.

Now suppose that $\boldsymbol{\Sigma} \neq \mathbf{I}$. Using that $\mathbf{X} \sim \mathbf{W}\sqrt{\boldsymbol{\Sigma}}$and the submultiplicity of the spectral norm we obtain that

$$|||\widehat{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}||| = \left|\left|\left|\frac{1}{n}\mathbf{X}^\intercal\mathbf{X} - \boldsymbol{\Sigma}\right|\right|\right|$$

$$
\begin{aligned}
&\sim \left|\!\left|\!\left| \mathbf{\Sigma}^{1/2}\left(\frac{1}{n}\mathbf{W}^{\intercal}\mathbf{W}-\boldsymbol{I}\right)\mathbf{\Sigma}^{1/2} \right|\!\right|\!\right| \\
&\leq |||\mathbf{\Sigma}||| \left|\!\left|\!\left| \frac{1}{n}\mathbf{W}^{\intercal}\mathbf{W}-\boldsymbol{I} \right|\!\right|\!\right|.
\end{aligned}
$$

By the previous formulation done with $\mathbf{W}$ we can deduce that

$$
\frac{|||\widehat{\mathbf{\Sigma}}_0-\mathbf{\Sigma}|||}{|||\mathbf{\Sigma}|||}\leq 2\epsilon+\epsilon^2,
$$

with probability at least $1-2e^{-n\delta^2/2}$, where $\epsilon=\delta+\sqrt{p/n}$, $\delta>0$. This ends the proof. □

From Theorem 3.2.4, we can deduce that with a sample size of

$$
n\geq\frac{2}{\delta^2}\log\left(\frac{2}{\alpha}\right),\quad \delta>0,\ \alpha\in(0,1),
$$

we guarantee that

$$
|||\widehat{\mathbf{\Sigma}}_0-\mathbf{\Sigma}|||\leq\left[2\left(\delta+\sqrt{\frac{p}{n}}\right)+\left(\delta+\sqrt{\frac{p}{n}}\right)^2\right]|||\mathbf{\Sigma}|||
$$

with probability at least $1-\alpha$. Even more[7], by Proposition C.5.1, as long as $p/n\to 0$ when $n\to\infty$,

$$
|||\widehat{\mathbf{\Sigma}}_0-\mathbf{\Sigma}|||\xrightarrow{\mathbb{P}}0,\ n\to\infty,
$$

i.e., $\widehat{\mathbf{\Sigma}}_0$ is *consistent estimator* of $\mathbf{\Sigma}$. In the next section we'll obtain a concentration bound relaxing the Gaussian assumption over $\mathbf{X}$ and instead we'll work with sub-Gaussian random variables directly.

3.3 Sub-Gaussian ensembles

At the beginning of the chapter we presented sub-Gaussian random variables. Now, we define sub-Gaussian random vectors in the following way.

[7] We write $X_n\xrightarrow{\mathbb{P}}0$ if for any $t\geq 0$, $P(|X_n|\geq t)\to 0$, $n\to\infty$.

Definition 3.3.1 (Sub-Gaussian random vectors)**.** *A random vector $\boldsymbol{X} \in \mathbb{R}^p$ is called sub-Gaussian if the one dimensional marginals $\langle \boldsymbol{X}, \boldsymbol{x} \rangle$ are sub-Gaussian random variables for all $\boldsymbol{x} \in \mathbb{R}^p$.*

Just as we did with in the unidemensional case, we define the sub-Gaussian norm of a vector as

$$\|\boldsymbol{X}\|_{\psi_2} = \sup_{\|x\|_2=1} \|\langle \boldsymbol{X}, \boldsymbol{x} \rangle\|_{\psi_2}.$$

Again, it is clear that if $\|\boldsymbol{X}\|_{\psi_2} < \infty$, then $\boldsymbol{X}$ is sub-Gaussian. Unlike random variables we are not really interest in parametrize the distribution of a sub-Guassian random vector (e.g. $\boldsymbol{X} \sim \mathcal{SG}(\sigma)$) because we'll obtain concentration bounds that depends on the norm directly.

As the previous section, starting from an ensemble $\mathbf{X}$, we want to obtain concentration guarantees for the matrix $\mathbf{X}^\intercal\mathbf{X}$, but instead of assuming $\boldsymbol{X}_i$ to be Gaussian we'll assume that it is sub-Gaussian. The next theorem is proved in Appendix B and gives us our first concentration bound for sub-Gaussian ensembles. We'll assume first that the vectors $\boldsymbol{X}_i$ of the ensemble are *isotropic*, meaning that

$$\mathbb{E}\boldsymbol{X}_i\boldsymbol{X}_i^\intercal = \mathbf{I}.$$

Since we are working with centered vectors, this means that $\mathrm{Cov}\boldsymbol{X}_i = \mathbf{I}$.

Theorem 3.3.2. *Let $\mathbf{Z}$ be an $n \times p$ ensemble of the independent random vectors $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n \in \mathbb{R}^p$, such that $\boldsymbol{Z}_i$ is sub-Gaussian, isotropic and centered. Then, for any $t \geq 0$,*

$$\mathbb{P}\left(\left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \frac{1}{n}\mathbf{Z}^\intercal\mathbf{Z} - \mathbf{I} \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert \geq LK^2(\delta \vee \delta^2) \right) \leq 2e^{-t^2},$$

where

$$K = \max_i \|\boldsymbol{Z}_i\|_{\psi_2}, \quad \delta = C\left(\sqrt{\frac{p}{n}} + \frac{t}{\sqrt{n}}\right),$$

and $L, C > 0$ are absolute constants.

The proof of Theorem 3.3.2 relies on the fact that for any symmetric matrix $\mathbf{A} \in \mathcal{S}_p$, we can bound $|||\mathbf{A}|||$ by the maximum of $|\langle \mathbf{A}\boldsymbol{x}, \boldsymbol{x} \rangle|$ over all $\boldsymbol{x} \in N$, where N is a finite subset of the sphere in $\mathbb{R}^p$. Such subset N is called an ϵ-net. One can see this fact and definition in Appendix B. The key observation is that bounding the norm with the maximum over a finite set enables us to use the union bound to obtain concentration results.

A consequence of Theorem 3.3.2 is the next concentration bound for the covariance estimator under sub-gaussianity.

Theorem 3.3.3. *Let $\boldsymbol{X}$ be a mean-zero sub-Gaussian random vector in $\mathbb{R}^p$ with covariance matrix $\boldsymbol{\Sigma} \succ \mathbf{0}$. More precisely, assume that there exist $K \geq 1$ such that*

$$\|\langle \boldsymbol{X}, \boldsymbol{x} \rangle\|_{\psi_2} \leq K \|\langle \boldsymbol{X}, \boldsymbol{x} \rangle\|_{L^2}, \quad \forall \boldsymbol{x} \in \mathbb{R}^p.$$

Let $\mathbf{X}$ be the $n \times p$ ensemble of $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$, a set of iid copies of $\boldsymbol{X}$, and define $\widehat{\boldsymbol{\Sigma}}_0 = n^{-1}\mathbf{X}^\intercal \mathbf{X}$. Then, for every $u \geq 0$,

$$\frac{|||\widehat{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}|||}{|||\boldsymbol{\Sigma}|||} \leq JK^2 \left(\sqrt{\frac{p}{n} + u} + \frac{p}{n} + u \right),$$

with probability at least $1 - 2e^{-un}$, for some constant $J > 0$.

Proof. Define the random vector $\boldsymbol{Z} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}$. Then

$$\mathbb{E}\boldsymbol{Z}\boldsymbol{Z}^\intercal = \boldsymbol{\Sigma}^{-1/2} \left(\mathbb{E}\boldsymbol{X}\boldsymbol{X}^\intercal \right) \boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1/2} = \mathbf{I},$$

so $\boldsymbol{Z}$ is isotropic. Even more, because

$$\|\langle \boldsymbol{X}, \boldsymbol{x} \rangle\|_{L^2}^2 = \mathbb{E}(\boldsymbol{x}^\intercal \boldsymbol{X})^2 = \mathbb{E}\left[\boldsymbol{x}^\intercal \boldsymbol{X}\boldsymbol{X}^\intercal \boldsymbol{x} \right] = \|\sqrt{\boldsymbol{\Sigma}}\boldsymbol{x}\|_2^2,$$

we have, by hypothesis, that

$$K \geq \frac{\|\langle \boldsymbol{X}, \boldsymbol{x} \rangle\|_{\psi_2}}{\|\langle \boldsymbol{X}, \boldsymbol{x} \rangle\|_{L^2}} = \frac{\|\langle \boldsymbol{Z}, \boldsymbol{y} \rangle\|_{\psi_2}}{\|\boldsymbol{y}\|_2},$$

where $\boldsymbol{y} = \sqrt{\boldsymbol{\Sigma}}\boldsymbol{x}$. This implies that

$$\|\boldsymbol{Z}\|_{\psi_2} = \sup_{\|\boldsymbol{y}\|_2 = 1} \|\langle \boldsymbol{Z}, \boldsymbol{y} \rangle\|_{\psi_2} \leq K.$$

Let $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n$ be iid copies of Z and define the random matrix $\mathbf{Y} \in \mathcal{S}_p$ as

$$\mathbf{Y} = \frac{1}{n} \sum_{j=1}^{n} (\boldsymbol{Z}_j \boldsymbol{Z}_j^\intercal - \mathbf{I}).$$

From Theorem 3.3.2 we know that for $t \geq 0$

$$\mathbb{P}\left(\mathbf{Y} \leq LK^2(\delta \vee \delta^2)\right) \geq 1 - 2e^{-t^2}, \quad \delta = C\left(\sqrt{\frac{p}{n}} + \sqrt{\frac{t^2}{n}}\right).$$

On the other hand, by submultiplicity of the norm,

$$|||\widehat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}||| = |||\mathbf{\Sigma}^{-1/2}\mathbf{Y}\mathbf{\Sigma}^{-1/2}||| \leq |||\mathbf{Y}|||\,|||\mathbf{\Sigma}|||.$$

Then, take $u = t^2$ and observe that

$$\delta \vee \delta^2 \leq \delta + \delta^2 = C\sqrt{\frac{p+u}{n}} + C^2\frac{p+u}{n},$$

Define a large enough constant $C' > 0$ such that[8]

$$C\sqrt{\frac{p+u}{n}} + C^2\frac{p+u}{n} \leq C'\left(\sqrt{\frac{p+u}{n}} + \frac{p+u}{n}\right), \quad \forall u \geq 0.$$

Therefore, we get that

$$|||\widehat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}||| \leq JK^2\left(\sqrt{\frac{p+u}{n}} + \frac{p+u}{n}\right)|||\mathbf{\Sigma}|||,$$

with probability at least $1 - e^{-u}$, where $J = LC'$. Taking the mapping $u \mapsto u/n$ ends the proof. □

As in the previous section, we can conclude from Theorem 3.3.3 that with a sample size of

$$n \geq \frac{1}{u}\log\left(\frac{2}{\alpha}\right), \quad u > 0,\ \alpha \in (0,1),$$

we guarantee that

$$|||\widehat{\mathbf{\Sigma}}_0 - \mathbf{\Sigma}||| \leq JK^2\left(\sqrt{\frac{pu}{\log(2/\alpha)} + u} + \frac{pu}{\log(2/\alpha)} + u\right)|||\mathbf{\Sigma}|||$$

with probability at least $1 - \alpha$, and $\widehat{\mathbf{\Sigma}}_0$ is a consistent estimator of $\mathbf{\Sigma}$.

To relax the Gaussian and sub-Gaussian assumptions, in the next section we'll work with almost arbitrarily distributions.

[8] For example, one can take $C' = \sup_{u\geq 0} g(u)$, where $g(u) = \frac{C\sqrt{(p+u)/n}+C^2(p+u)/n}{\sqrt{(p+u)/n}+(p+u)/n}$. Observe that g is a bounded and monotone function.

3.4 Matrix Bernstein inequality: first robust estimation

In this section we'll show the power of the methods developed in Chapter 2 when additional assumptions are added to the ensemble $\mathbf{X}$. More specifically, we'll assume that the vectors $\boldsymbol{X}_i$ have bounded norm. Despite that this assumption can be viewed as very restrictive since heavy-tailed distributions are not bounded, we can think of it as first approximation to arbitrary distributions. Indeed, if the density function of $\boldsymbol{X}_i$ has heavy tails or is not a *bell curve* like in the Gaussian case, one can assume that $\|\boldsymbol{X}_i\| \leq B$, for some $B < \infty$, in order to obtain a concentration inequality, leaving the form of the density to be arbitrary inside the ball $\{\boldsymbol{x} \in \mathbb{R}^p : \|\boldsymbol{x}\|_2 \leq B\}$.

Before stating the first important result of this section, the Matrix Bernstein bound, we need to establish the next technical lemma.

Lemma 3.4.1. *Let $\mathbf{X} \in \mathcal{S}_p$ be a random matrix such that $\mathbb{E}\mathbf{X} = \mathbf{0}$ and $|||\mathbf{X}||| \leq K$. Define $h : \mathbb{R} \to \mathbb{R}$ as*

$$h(\theta) = \frac{\theta^2/2}{1 - |\theta|K/3}.$$

Then,

(a) for any $|\theta| < 3/K$,

$$\mathbb{E}e^{\theta\mathbf{X}} \preceq \exp\left(h(\theta)\mathbb{E}\mathbf{X}^2\right).$$

(b) Fix $t \geq 0$, $\sigma^2 > 0$. The value $\theta_0 = t/(\sigma^2 + Kt/3)$ is in the range $(-3/K, 3/K)$ and the function $\theta \mapsto -t\theta + h(\theta)\sigma^2$ evaluated at θ_0 is $-t^2/(2(\sigma^2 + Kt/3))$.

Proof. (a) Take $x, \theta > 0$ such that $x\theta < 3$. Then, by Taylor expansion,

$$\begin{aligned} e^{x\theta} &= 1 + x\theta + \sum_{j=2}^{\infty} \frac{(x\theta)^j}{j!} \\ &= 1 + x\theta + \frac{(x\theta)^2}{2} \sum_{j=2}^{\infty} \frac{(x\theta)^{j-2}2}{j!} \\ &\leq 1 + x\theta + \frac{(x\theta)^2}{2} \sum_{j=2}^{\infty} \frac{(x\theta)^{j-2}}{3^{j-2}}, \quad (j! \geq 2(3^{j-2}),\ j \geq 2) \end{aligned}$$

$$= 1 + x\theta + \frac{\theta^2/2}{1 - x\theta/3}x^2, \quad (x\theta < 3).$$

Therefore, by taking any $x \in [-K, K]$ and any $\theta \in [-3/K, 3/K]$, we have that $1 - x\theta/3 \geq 1 - |\theta|K/3$ and

$$e^{x\theta} \leq 1 + x\theta + \frac{\theta^2/2}{1 - |\theta|K/3}x^2 = \underbrace{1 + x\theta + h(\theta)x^2}_{=\varphi(x)}.$$

Since $|||\mathbf{X}||| \leq K$ and $|||\mathbf{X}||| = \max\{|\lambda_1(\mathbf{X})|, ..., |\lambda_p(\mathbf{X})|\}$, all the eigenvalues of $\mathbf{X}$ are inside the interval $[-K, K]$ and we arrive at the bound

$$e^{\mathbf{X}\theta} \preceq \varphi(\mathbf{X}) = \mathbf{I} + \mathbf{X}\theta + h(\theta)\mathbf{X}^2. \tag{3.8}$$

Taking expectation on both sides of (3.8) and using that $\mathbb{E}\mathbf{X} = \mathbf{0}$ we get

$$\mathbb{E}e^{\theta\mathbf{X}} \preceq \mathbf{I} + h(\theta)\mathbb{E}\mathbf{X}^2,$$

and because $1 + x \leq e^x$ for all $x \in \mathbb{R}$,

$$\mathbb{E}e^{\theta\mathbf{X}} \preceq \exp\left(h(\theta)\mathbb{E}\mathbf{X}^2\right).$$

(b) Note that $3\sigma^2/K > 0$, so $t > 3\sigma^2/K + 3(Kt/3)/K$ and $\theta_0 = t/(\sigma^2 + Kt/3)$ is in $(-3/K, 3/K)$. Now, we just need to perform a simple evaluation:

$$\begin{aligned}
-t\theta_0 + h(\theta_0)\sigma^2 &= -\frac{t^2}{\sigma^2 + Kt/3} + \sigma^2 \frac{\dfrac{t^2}{2(\sigma^2 + Kt/3)^2}}{1 - \dfrac{Kt/3}{\sigma^2 + Kt/3}} \\
&= -\frac{t^2}{\sigma^2 + Kt/3} + \sigma^2 \frac{t^2(\sigma^2 + Kt/3)/2}{(\sigma^2 + Kt/3 - Kt/3)(\sigma^2 + Kt/3)^2} \\
&= -\frac{t^2}{\sigma^2 + Kt/3} + \frac{t^2/2}{\sigma^2 + Kt/3} \\
&= -\frac{t^2/2}{\sigma^2 + Kt/3}
\end{aligned}$$

□

The next theorem relies on Theorem 2.4.4 since it establishes a bound on the sum of independent and symmetric random matrices, with the additional assumption that they have zero mean and have bounded spectral norm. It is called Matrix Bernstein because the hypothesis and the conclusion resembles the ones of the classic Bernstein inequality for random variables (see for example [4, p. 36]). This result was obtained from [45, Chapter 6].

Theorem 3.4.2 (Matrix Bernstein). *Let $\mathbf{X}_1, ..., \mathbf{X}_n \in \mathcal{S}_p$ be independent random matrices such that $\mathbb{E}\mathbf{X}_j = \mathbf{0}$ and $|||\mathbf{X}_j||| \le K$, $j = 1, ..., n$. Then, for any $t \ge 0$,*

$$\mathbb{P}\left(\left|\left|\left| \sum_{j=1}^{n} \mathbf{X}_j \right|\right|\right| \ge t \right) \le 2p \exp\left(\frac{-t^2/2}{\sigma^2 + Kt/3} \right),$$

where $\sigma^2 = ||| \sum_{j=1}^{n} \mathbb{E}\mathbf{X}_j^2 |||$.

Proof. Let $\mathbf{X} = \sum_{j=1}^{n} \mathbf{X}_j$. Because $\mathbb{E}\mathbf{X}_j = \mathbf{0}$ and $|||\mathbf{X}_j||| \le K$, $j = 1, ..., n$, we get from Lemma 3.4.1 (a) that

$$\mathbb{E}e^{\theta X_j} \preceq \exp\left(h(\theta)\mathbb{E}\mathbf{X}_j^2 \right), \quad |\theta| < 3/K, \, j = 1, ..., n. \tag{3.9}$$

Furthermore, because logarithm is operator monotone (Proposition 2.1.5) we get that

$$\log \mathbb{E}e^{\theta X_j} \preceq h(\theta)\mathbb{E}\mathbf{X}_j^2, \quad |\theta| < 3/K, \, j = 1, ..., n.$$

Then, Theorem 2.4.4 implies that

$$\begin{aligned} \mathbb{P}\left(\lambda_1\left(\mathbf{X}\right) \ge t\right) &\le e^{-\theta t} \mathrm{tr} \exp\left(h(\theta) \sum_{j=1}^{n} \mathbb{E}\mathbf{X}_j^2 \right) \\ &\le p e^{-\theta t} \exp\left(h(\theta) \left|\left|\left| \sum_{j=1}^{n} \mathbb{E}\mathbf{X}_j^2 \right|\right|\right| \right) \\ &= p \exp\left(-\theta t + h(\theta)\sigma^2 \right). \end{aligned}$$

Finally, by Lemma 3.4.1 (b), we choose the value $\theta = t/(\sigma^2 + Kt/3)$ to get

$$\mathbb{P}\left(\lambda_1\left(\mathbf{X}\right) \ge t\right) \le p \exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3} \right).$$

On the other hand, because the bound (3.9) is valid for $\theta \in (-3/K, 3/K)$ we get that

$$\mathbb{E}e^{-\theta X_j} \preceq \exp\left(h(-\theta)\mathbb{E}\mathbf{X}_j^2 \right), \quad |\theta| < 3/K, \, j = 1, ..., n.$$

Observing that $h(\theta) = h(-\theta)$ and applying the same reasoning we obtain that

$$\begin{aligned}\mathbb{P}\left(-\lambda_p\left(\mathbf{X}\right) \geq t\right) &= \mathbb{P}\left(\lambda_1\left(-\mathbf{X}\right) \geq t\right) \\ &\leq p \exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right).\end{aligned}$$

Therefore, the result follows by Proposition 2.3.2. □

In the upcoming theorem we derive a concentration bound on the matrix estimator under the bounded norm assumption using Theorem 3.4.2. This result was obtained from [45, Chapter 1].

Theorem 3.4.3. *Let $\boldsymbol{X} \in \mathbb{R}^p$ be a random vector such that $\mathbb{E}\boldsymbol{X} = \mathbf{0}$,* Cov$\boldsymbol{X} = \boldsymbol{\Sigma}$ *and $\|\boldsymbol{X}\|_2^2 \leq B$. Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ iid copies of $\boldsymbol{X}$ from which we construct $\widehat{\boldsymbol{\Sigma}}_0$. Then, for every $t \geq 0$,*

$$\mathbb{P}\left(|||\widehat{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}||| \geq t\right) \leq 2p \exp\left(\frac{-nt^2/2}{B|||\boldsymbol{\Sigma}||| + 2Bt/3}\right).$$

Proof. Define the matrices $\mathbf{Y}_1, ..., \mathbf{Y}_n \in \mathcal{S}_p$ as

$$\mathbf{Y}_j = \frac{1}{n}\left(\boldsymbol{X}_j \boldsymbol{X}_j^\intercal - \boldsymbol{\Sigma}\right), \quad j = 1, ..., n.$$

Also, define $\mathbf{Y} = \sum_{j=1}^n \mathbf{Y}_j$, so

$$\mathbf{Y} = \widehat{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}.$$

Note that $\mathbb{E}\mathbf{Y}_j = \mathbf{0}$ since $\mathbb{E}\boldsymbol{X}\boldsymbol{X}^\intercal = \boldsymbol{\Sigma}$, and

$$\begin{aligned}|||\mathbf{Y}_j||| &= \frac{1}{n}|||\boldsymbol{X}_j\boldsymbol{X}_j^\intercal - \boldsymbol{\Sigma}||| \\ &\leq \frac{1}{n}\left(|||\boldsymbol{X}_j\boldsymbol{X}_j^\intercal||| + |||\boldsymbol{\Sigma}|||\right), \quad \text{(by triangle inequality)} \\ &\leq \frac{1}{n}\left(B + |||\mathbb{E}\boldsymbol{X}\boldsymbol{X}^\intercal|||\right), \quad \text{(by Example 2.1.2)} \\ &\leq \frac{1}{n}\left(B + \mathbb{E}|||\boldsymbol{X}\boldsymbol{X}^\intercal|||\right), \quad \text{(by Jensen's inequality)} \\ &\leq \frac{2B}{n}.\end{aligned}$$

Also, by definition of $\mathbf{Y}_j$,

$$\begin{aligned}\mathbb{E}\mathbf{Y}_j^2 &= \frac{1}{n^2}\mathbb{E}\left[\|\boldsymbol{X}_j\|_2^2\boldsymbol{X}_j\boldsymbol{X}_j^\intercal - \boldsymbol{X}_j\boldsymbol{X}_j^\intercal\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{X}_j\boldsymbol{X}_j^\intercal + \boldsymbol{\Sigma}^2\right] \\ &\preceq \frac{1}{n^2}\left(B\mathbb{E}\boldsymbol{X}_j\boldsymbol{X}_j^\intercal - \boldsymbol{\Sigma}^2 - \boldsymbol{\Sigma}^2 + \boldsymbol{\Sigma}^2\right) \\ &\preceq \frac{B\boldsymbol{\Sigma}}{n^2}.\end{aligned}$$

Then,

$$\sum_{j=1}^n \mathbb{E}\mathbf{Y}_j^2 \preceq \frac{B\boldsymbol{\Sigma}}{n} \quad \text{and} \quad \left|\!\left|\!\left|\sum_{j=1}^n \mathbb{E}\mathbf{Y}_j^2\right|\!\right|\!\right| \leq \frac{B}{n}|\!|\!|\boldsymbol{\Sigma}|\!|\!|.$$

Therefore, by direct application of Theorem 3.4.2, we get that

$$\begin{aligned}\mathbb{P}\left(|\!|\!|\widehat{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}|\!|\!| \geq t\right) &= \mathbb{P}\left(|\!|\!|\mathbf{Y}|\!|\!| \geq t\right) \\ &= 2p\exp\left(-\frac{t^2/2}{\frac{1}{n}B|\!|\!|\boldsymbol{\Sigma}|\!|\!| + \frac{1}{n}2Bt/3}\right).\end{aligned}$$

This ends the proof. □

As in the two previous sections, we conclude from Theorem 3.4.3 that

$$\mathbb{P}\left(|\!|\!|\widehat{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}|\!|\!| \leq \sqrt{2t(B|\!|\!|\boldsymbol{\Sigma}|\!|\!| + 2Bt/3)}\right) \geq 1 - 2pe^{-nt},$$

so with a sample size of

$$n \geq \frac{1}{t}\log\left(\frac{2p}{\alpha}\right), \quad t > 0,\, \alpha \in (0,1),$$

we guarantee that

$$|\!|\!|\widehat{\boldsymbol{\Sigma}}_0 - \boldsymbol{\Sigma}|\!|\!| \leq \sqrt{2t(B|\!|\!|\boldsymbol{\Sigma}|\!|\!| + 2Bt/3)},$$

with probability at least $1-\alpha$. And of course $\widehat{\boldsymbol{\Sigma}}_0$ is a consistent estimator of $\boldsymbol{\Sigma}$.

Summary and observations

In this chapter we presented concentration inequalities for the random matrix $\widehat{\mathbf{\Sigma}}_0$ around its mean $\mathbf{\Sigma}$, were closeness is measured in terms of the spectral norm. To do so we had to make the general assumption that $\mathbb{E}\boldsymbol{X} = \mathbf{0}$. Additionally, we assumed that $\boldsymbol{X}$ had a Gaussian or sub-Gaussian distribution, and in the final section we supposed that $\|\boldsymbol{X}\|_2$ is bounded.

We can observe that the concentration bounds under Gaussian or sub-Gaussian assumptions are similar. In particular the Gaussian bound is sharper since it doesn't depend on any unknown constant. On the other hand, the bond obtained by the Matrix Bernstein inequality differ from the previous two in the incorporation of the norm bound and that in the probability bound there is an explicit dependence on the dimension. This last difference is due to the fact that the technique used are distinct. One important observation is that the assumption $\|\|\mathbf{X}\|\| \leq K$ of the Matrix Bernstein inequality do not imply immediately a Hoeffding-type result like in Theorem 1.3.1 of Chapter 1. For that type of results one uses the stronger assumption $\mathbf{X}^2 \preceq \mathbf{K}^2$ [44]. As far as the author is concerned, there is no result that requires only the hypothesis $\|\|\mathbf{X}\|\| \leq K$ in order to obtain Hoeffding-type bounds.

We want to stress that Theorem 3.4.3 depends on the more succinct results of Chapter 2, but to make those results useful we needed to add the bounded condition. In the following chapter we continue exploiting the results of Chapter 2, but we'll do it in a way that is more general in the sense that very few distributional assumptions will be made.

Chapter 4

Robust estimation of a matrix

In this chapter we present the main results studied in this thesis. We describe the procedure of Minsker in [32] for the estimation of the mean of a random matrix with heavy tail entries. The originality of this approach is the application of Catoni's influence function, presented in [9], to the matrix concentration techniques provided by Tropp in [45].

As it is indicated in the title of this thesis, we focus on the estimation of the mean of a random matrix, since it is common to rely on matrix estimators that are unbiased. This was the case for the covariance estimator presented in the previous chapter. More specifically, suppose that $\mathbf{Y}_1, ..., \mathbf{Y}_n \in \mathcal{S}_p$ are independent random matrices. We want to obtain a good estimator of $\mathbb{E}[\sum_{j=1}^n \mathbf{Y}_j]$, i.e., an estimator $\mathbf{X} = \mathbf{X}(\mathbf{Y}_1, ..., \mathbf{Y}_n)$ that guarantees that, for $t \geq 0$,

$$\left\|\left|\mathbf{X} - \sum_{j=1}^{n} \mathbb{E}\mathbf{Y}_j\right|\right\| \leq t$$

with high probability. In Chapter 3, for the case of the covariance estimator, we emphasized that we needed distributional assumptions to do so. In this chapter we apply the aforementioned techniques to obtain theoretical guarantees without strong distributional assumptions.

In Section 1 we give a theorem that motivates the application of Catoni's influence function. In Section 2 we present the Catoni's influence function and how it's used to obtain concentration inequalities. Section 3 shows the main result of [32] regarding symmetric random matrices, and an additional generalization for rectangular matrices is given in Section 4. Section 5 gives a data-dependent method to choose the hyperparameters of the estimator with theoretical guarantees. And finally, Section 6 shows two application of the main result: PCA and community detection.

4.1 Motivation for the influence function

As mentioned in Chapter 2, in order to make useful the Tropp's probability bound on symmetric matrices of Theorem 2.4.4, we need to be able to found upper bounds for

$$\mathbb{E}e^{\theta \mathbf{X}_j}, \quad j = 1, ..., n. \tag{4.1}$$

One such bound was found in Section 3.4 to obtain the Matrix Bernstein concentration inequality (Theorem 3.4.2.) But this was done at the cost of assuming that the random matrices were bounded. In this section we'll suppose that we have been able to find bounds for (4.1) to obtain generic concentration inequalities. This results will motivate the use of Catoni's influence function.

Our first lemma is a consequence of Lemma 2.4.3 and gives us a generic bound on the expectation of the trace of the exponential function.

Lemma 4.1.1. *Let $\mathbf{X}_1, ..., \mathbf{X}_n \in \mathcal{S}_p$ be independent random matrices and $\mathbf{H} \in \mathcal{S}_p$ a deterministic matrix. If $\mathbf{W}_1, ..., \mathbf{W}_n \in \mathcal{S}_p$ are independent random matrices such that $\mathbf{X}_j \preceq \mathbf{W}_j$ for any j, then*

$$\mathbb{E}\mathrm{tr}\ \exp\left(\sum_{j=1}^{n} \mathbf{X}_j + \mathbf{H}\right) \leq \mathrm{tr}\ \exp\left(\sum_{j=1}^{n} \log \mathbb{E}e^{\mathbf{W}_j} + \mathbf{H}\right). \tag{4.2}$$

Proof. Since $\mathbf{X}_j \preceq \mathbf{W}_j$, $j = 1, ..., n$, we have that

$$\sum_{j=1}^{n} \mathbf{X}_j + \mathbf{H} \preceq \sum_{j=1}^{n} \mathbf{W}_j + \mathbf{H},$$

and because $\mathbf{A}_1 \preceq \mathbf{A}_2$ imply $\mathrm{tr}\, e^{\mathbf{A}_1} \leq \mathrm{tr}\, e^{\mathbf{A}_2}$ (Lemma 2.1.4) we get

$$\mathrm{tr}\ \exp\left(\sum_{j=1}^{n} \mathbf{X}_j + \mathbf{H}\right) \leq \mathrm{tr}\ \exp\left(\sum_{j=1}^{n} \mathbf{W}_j + \mathbf{H}\right).$$

Taking expectations and applying Tropp's expectation bound of Lemma 2.4.3 to the right expression we arrive at

$$\begin{aligned}\mathbb{E}\mathrm{tr}\ \exp\left(\sum_{j=1}^{n}\mathbf{X}_j+\mathbf{H}\right) &\leq \mathbb{E}\mathrm{tr}\ \exp\left(\sum_{j=1}^{n}\mathbf{W}_j+\mathbf{H}\right)\\ &\leq \mathrm{tr}\ \exp\left(\sum_{j=1}^{n}\log\mathbb{E}e^{\mathbf{W}_j}+\mathbf{H}\right).\end{aligned}$$

□

Note that the matrices $\mathbf{W}_1,...,\mathbf{W}_n$ of the previous lemma are independent but not necessarily independent from $\mathbf{X}_1,...,\mathbf{X}_n$. This observation allows us to define the matrices $\mathbf{W}_i$ that depend directly on $\mathbf{X}_i$, for example. This will also be the case in Theorem 4.1.2.

From the previous lemma and Tropp's probability bound on symmetric matrices, we deduce the next generic concentration inequality for the eigenvalues of the sum of independent symmetric matrices.

Theorem 4.1.2. *Let $\mathbf{X}_1,...,\mathbf{X}_n\in\mathcal{S}_p$ be independent random matrices and $\mathbf{H}=\sum_{j=1}^{n}\mathbf{H}_j$ with $\mathbf{H}_j\in\mathcal{S}_p$ deterministic matrices. If $\mathbf{W}_1(\theta),...,\mathbf{W}_n(\theta)\in\mathcal{S}_p$ are independent random matrices such that $\theta\mathbf{X}_j\preceq\mathbf{W}_j(\theta)$, for any j and $\theta>0$, and $\mathbf{M}_1(\theta),...,\mathbf{M}_n(\theta)\in\mathcal{S}_p$ are deterministic matrices such that*

$$\log\mathbb{E}e^{\mathbf{W}_j(\theta)}+\theta\mathbf{H}_j\preceq\mathbf{M}_j(\theta),\quad\forall j,\ \theta>0.$$

then

$$\mathbb{P}\left(\lambda_1\left(\sum_{j=1}^{n}\mathbf{X}_j+\mathbf{H}\right)\geq t\right)\leq p\exp\left(-t\theta+\left|\left|\left|\sum_{j=1}^{n}\mathbf{M}_j(\theta)\right|\right|\right|\right)$$

Proof. By hypothesis and Lemma 2.1.2 we have that

$$\sum_{j=1}^{n}\log\mathbb{E}e^{\mathbf{W}_j(\theta)}+\theta\mathbf{H}\preceq\sum_{j=1}^{n}\mathbf{M}_j(\theta),$$

and by Lemma 2.1.4

$$\mathrm{tr}\ \exp\left(\sum_{j=1}^{n}\log\mathbb{E}e^{\mathbf{W}_j(\theta)}+\theta\mathbf{H}\right)\leq\mathrm{tr}\ \exp\left(\sum_{j=1}^{n}\mathbf{M}_j(\theta)\right).\tag{4.3}$$

On the other hand, applying the same procedure of Tropp's probability bound of Theorem 2.4.4 we have that

$$\mathbb{P}\left(\lambda_1\left(\sum_{j=1}^n \mathbf{X}_j + \mathbf{H}\right) \geq t\right) \leq e^{-\theta t}\mathbb{E}\mathrm{tr}\ \exp\left(\sum_{j=1}^n \theta\mathbf{X}_j + \theta\mathbf{H}\right).$$

Now, from the previous Lemma 4.1.1 and inequality (4.3),

$$\begin{aligned}\mathbb{E}\mathrm{tr}\ \exp\left(\sum_{j=1}^n \theta\mathbf{X}_j + \theta\mathbf{H}\right) &\leq \mathrm{tr}\ \exp\left(\sum_{j=1}^n \log\mathbb{E}e^{\mathbf{W}_j(\theta)} + \theta\mathbf{H}\right)\\ &\leq \mathrm{tr}\ \exp\left(\sum_{j=1}^n \mathbf{M}_j(\theta)\right).\end{aligned}$$

Note that for any $\mathbf{A} \in \mathcal{S}_p$ it is true that $\mathrm{tr}\, e^{\mathbf{A}} \leq pe^{\lambda_1(\mathbf{A})} \leq pe^{\|\!|\mathbf{A}|\!\|}$ (Lemma 2.1.7), so

$$e^{-\theta t}\mathrm{tr}\ \exp\left(\sum_{j=1}^n \mathbf{M}_j(\theta)\right) \leq p\exp\left(-\theta t + \left|\!\left|\!\left|\sum_{j=1}^n \mathbf{M}_j(\theta)\right|\!\right|\!\right|\right).$$

This ends the proof. □

Observe that the matrices $\mathbf{W}_j(\theta)$ and $\mathbf{M}_j(\theta)$ of Theorem 4.1.2 depend on $\theta > 0$, so at the end one can optimize over θ to get a better bound.

Now that we have obtained the generic bound of Theorem 4.1.2, we need to address the question of why is it useful. To see this, consider a sample of iid symmetric random matrices $\mathbf{Y}_1, ..., \mathbf{Y}_n$ from which we want to estimate $\mathbb{E}\mathbf{Y}_1$. Certainly, it can be difficult to come up with matrices $\mathbf{W}_j(\theta)$ and $\mathbf{M}_j(\theta)$ to get the desired bounds of Theorem 4.1.2. But we can think of functions $f, g : \mathbb{R} \to \mathbb{R}$ such that

$$f(x) \leq g(x), \quad \forall x \in \mathbb{R},$$

and define $\mathbf{X}_j = f(\mathbf{Y}_j)$ and $\mathbf{W}_j = g(\mathbf{Y}_j)$, so by Lemma 2.1.6, $\mathbf{X}_j \preceq \mathbf{W}_j$. This reasoning motivates the idea that with a clever definition of f and g and by defining $\mathbf{H}_j = \mathbb{E}\mathbf{Y}_1$ in Theorem 4.1.2, one can obtain good concentration inequalities for the estimator $\sum_j f(\mathbf{Y}_j)$ of the matrix $\mathbb{E}\mathbf{Y}_1$. This formulation is explored in the next section.

4.2 Catoni's robust method

Let Y be a real-valued random variable with $\mathbb{E}Y = \mu$ and $\mathbb{E}Y^2 < \infty$. One of the main objectives of Catoni in [9] is to estimate the mean μ trough the iid copies $Y_1, ..., Y_n$ of Y. Since the distribution of Y can be heavy-tailed, the usual empirical estimator $n^{-1}\sum_{i=1}^{n} Y_i$ can deviate broadly from μ for finite samples. The novel approach[1] presented in [9] to overcome this issue relies on the definition of the *influence function* $\psi : \mathbb{R} \to \mathbb{R}$ as a non-decreasing function such that

$$-\log(1 - x + x^2/2) \leq \psi(x) \leq \log(1 + x + x^2/2). \tag{4.4}$$

See Appendix C to verify that the bounds of ψ are well defined. As pointed out by Catoni, there exists at least two monotone functions that satisfies (4.4): the *widest* possible choice ψ^W and the *narrowest* choice ψ^N defined as

$$\psi^W(x) = \begin{cases} \log(1 + x + x^2/2), & x \geq 0, \\ -\log(1 - x + x^2/2), & x < 0, \end{cases}$$

and

$$\psi^N(x) = \begin{cases} \log(2), & x \geq 1, \\ -\log(1 - x + x^2/2), & 0 \leq x < 1, \\ \log(1 + x + x^2/2), & -1 \leq x < 0, \\ -\log(2), & x < -1. \end{cases}$$

Figure 4.1 shows the functions ψ^W and ψ^N. We can observe that around zero they are almost identical to the identity function, but outside this range they tent to depart from it. The two functions give less weight to great values that the identity, this is why it is called the influence function, since values very far away from zero in the sample, i.e. influential values, are downsized to get a more stable estimation.

We can increase or decrease the range of values in which the function ψ is similar to the identity by considering the function $x \mapsto \psi(\theta x)/\theta$, for some fixed $\theta > 0$. This is shown in Figure 4.2. Here we graph the function $x \mapsto \psi^N(\theta x)/\theta$ and observe that whenever θ is big, the range of values in which the function is close to the identity is smaller and vice versa.

[1] It is important to mention that the general idea of *influence function* or *truncation function* in the context of robust estimation is related originally to the work of Huber [19]

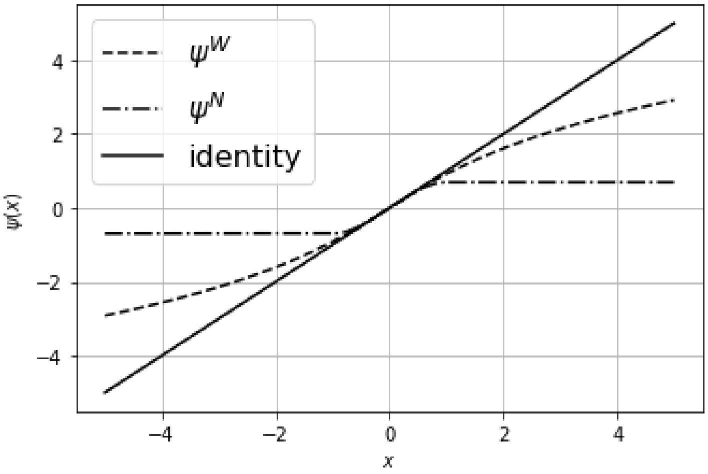


Figure 4.1: Functions ψ^W and ψ^N compared to the identity.

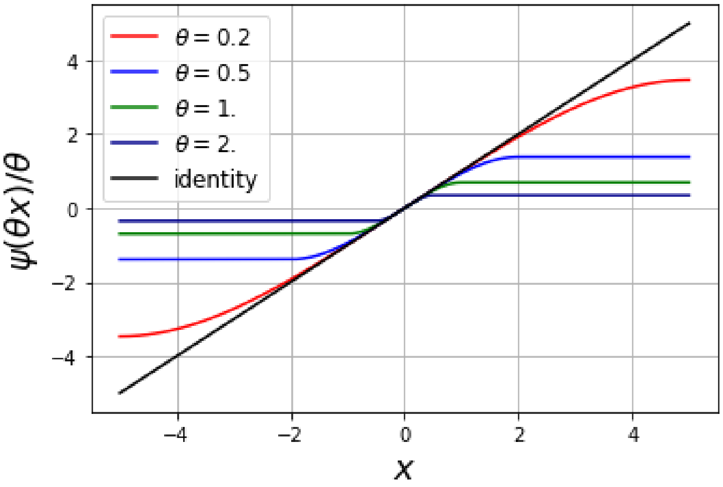


Figure 4.2: Functions $x \mapsto \psi^N(\theta x)/\theta$ for different values of θ.

Similar to the influence function ψ, in [21] the authors define the truncation operator ψ_τ as

$$\psi_\tau(x) = (|x| \wedge \tau)\,\mathsf{sign}(x), \quad \tau > 0.$$

Note that $\psi_\tau(x) = \tau\psi_1(x/\tau)$. The function ψ_τ doesn't satisfies (4.4), but ψ_1 satisfies the weaker inequality (see Appendix C)

$$-\log(1 - x + x^2) \leq \psi_1(x) \leq \log(1 + x + x^2),$$

therefore,

$$-\tau\log(1 - x\tau^{-1} + x^2\tau^{-2}) \leq \psi_\tau(x) \leq \tau\log(1 + x\tau^{-1} + x^2\tau^{-2}).$$

This will be the function used in Chapter 5 to obtain concentration inequalities for covariance estimation. As pointed out by Minsker in [32] the truncation operator ψ_τ makes the concentration results for the estimations valid but with worst constant factors.

The work of Catoni in [9] has a M-estimation perspective. More precisely, he propose the estimator $\hat{\mu}_\alpha$ of the mean μ that satisfies

$$\sum_{j=1}^{n} \psi\left(\alpha(Y_j - \hat{\mu}_\alpha)\right) = 0, \quad \text{for some } \alpha > 0.$$

Nevertheless, the introduction of this kind of influence functions rises a variety of interesting results per se. To see this, suppose that we want to estimate μ from $Y_1, ..., Y_n$. For some $\theta > 0$, define the estimator

$$T = T(\theta) = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{\theta}\psi(\theta Y_j).$$

We'll focus on the properties of the estimator T instead of $\hat{\mu}_\alpha$. By definition of ψ we obtain the next bound:

$$\begin{aligned}
\mathbb{E}e^{n\theta T} &= \prod_{j=1}^{n} \mathbb{E}e^{\psi(\theta Y_j)} \\
&\leq \prod_{j=1}^{n} \left(1 + \theta\mu + \frac{\theta^2}{2}\mathbb{E}Y^2\right) \\
&= \left(1 + \theta\mu + \frac{\theta^2}{2}\mathbb{E}Y^2\right)^n.
\end{aligned}$$

Then, since $\log(1+x) \leq x$ for $x > -1$,

$$\begin{aligned}\mathbb{E}e^{n\theta T - n\theta\mu} &= \exp\left(\log \mathbb{E}e^{n\theta T} - n\theta\mu\right)\\ &\leq \exp\left(n\log\left(1+\theta\mu+\frac{\theta^2}{2}\mathbb{E}Y^2\right) - n\theta\mu\right)\\ &\leq \exp\left(n\left(\theta\mu+\frac{\theta^2}{2}\mathbb{E}Y^2\right) - n\theta\mu\right)\\ &= \exp\left(n\frac{\theta^2}{2}\mathbb{E}Y^2\right).\end{aligned}$$

Similarly, using the lower bound for ψ, we obtain that

$$\mathbb{E}e^{-(n\theta T - n\theta\mu)} \leq \exp\left(n\frac{\theta^2}{2}\mathbb{E}Y^2\right).$$

Therefore, by Markov's inequality, for $t \geq 0$,

$$\begin{aligned}\mathbb{P}\left(|T-\mu| \geq t\right) &= \mathbb{P}\left(n\theta(T-\mu) \geq n\theta t\right) + \mathbb{P}\left(-n\theta(T-\mu) \geq n\theta t\right)\\ &\leq 2\exp\left(-n\theta t + n\frac{\theta^2}{2}\mathbb{E}Y^2\right).\end{aligned}$$

Optimizing over $\theta > 0$ we get that the last probability bound is minimized at $\tilde{\theta} = t/\mathbb{E}Y^2$ and consequently

$$\mathbb{P}\left(|T(\tilde{\theta}) - \mu| \geq t\right) \leq 2\exp\left(\frac{-nt^2}{2\mathbb{E}Y^2}\right).$$

This indicates that $T(\tilde{\theta}) - \mu \sim \mathcal{SG}(\sigma)$ where σ^2 is proportional to $2n^{-1}\mathbb{E}Y^2$. This fact follows from Proposition 3.1.2. In [9] it is shown that a similar result for $\hat{\mu}_\alpha$ holds, but the probability bound depends on $\mathrm{Var}Y$ instead of $\mathbb{E}Y^2$ and the value of t depends on the sample size n. This is a trade-off since in general $\mathrm{Var}Y \leq \mathbb{E}Y^2$, but a result in which t doesn't depend on n could lead to better bounds for the absolute value of the deference.

From the previous development we conclude that, just requiring that $\mathbb{E}Y^2 < \infty$, we've been able to find an estimator T of μ such that $|T(\tilde{\theta}) - \mu| \leq t$ with high probability for any $t \geq 0$. This is a remarkable observation since we don't need any strong distributional assumption for Y. The only drawback is that $\mathbb{E}Y^2$ is unknown and the optimal choice $\tilde{\theta}$ depends on it. Of course we can estimate it by $\sum_{j=1}^n Y_j^2$, but if the data is heavy-tailed this will be a poor estimation. We'll overcome this burden in Section 4.5 with the Lepski's method.

Almost immediately we can take a further step and apply this methodology to the matrix case, i.e., to the problem of estimating $\mathbb{E}\mathbf{Z}$ from an iid sample $\mathbf{Z}_1, ..., \mathbf{Z}_n$ of the random matrix $\mathbf{Z} \in \mathcal{M}_{p,r}$. To do so, in the next section we develop the same methodology for symmetric random matrices, since from Chapter 2 we know how to apply a function to this kind of matrices and how to obtain concentration bounds using Lieb's inequality.

4.3 Concentration of the symmetric robust estimator

The first step in applying the buildout of the previous section is to see how the non-decreasing function ψ defined by inequalities (4.4) behaves in the matrix case. This is done by the next straightforward lemma.

Lemma 4.3.1. *For* $\mathbf{A} \in \mathcal{S}_p$ *we have that,*

(a) if $\mathbf{I} + \mathbf{A} \succ \mathbf{0}$ *then* $\log(\mathbf{I} + \mathbf{A}) \preceq \mathbf{A}$*;*

(b) $\mathbf{I} + \mathbf{A} + \frac{1}{2}\mathbf{A}^2 \succ 0$*;*

(c) for every $\psi : \mathbb{R} \to \mathbb{R}$ *such that* $-\log(1 - x + x^2/2) \leq \psi(x) \leq \log(1 + x + x^2/2)$*,*

$$-\log\left(\mathbf{I} - \mathbf{A} + \frac{1}{2}\mathbf{A}^2\right) \preceq \psi(\mathbf{A}) \preceq \log\left(\mathbf{I} + \mathbf{A} + \frac{1}{2}\mathbf{A}^2\right).$$

Proof. (a) This is an immediate consequence of the inequality $\log(1 + x) \leq x$ for $x > -1$ and Lemma 2.1.6. The condition $\mathbf{I} + \mathbf{A} \succ \mathbf{0}$ ensures that $\log(\mathbf{I} + \mathbf{A})$ is well defined.

(b) The function $f(x) = 1 + x + x^2/2$ is positive in $\mathbb{R}$ so the eigenvalues of $f(\mathbf{A})$ are positive for any $\mathbf{A} \in \mathcal{S}_p$.

(c) This is clear from (b), the definition of ψ and Lemma 2.1.6. The matrix $-\log\left(\mathbf{I} - \mathbf{A} + \frac{1}{2}\mathbf{A}^2\right)$ is well defined for an argument similar to the one made in (b). □

Property (c) of Lemma 4.3.1 ensures that we can bound the random matrices of the iid sample $\mathbf{Y}_1, ..., \mathbf{Y}_n$ of $\mathbf{Y} \in \mathcal{S}_p$. Even more, later we will be able to bound

$$\mathbb{E}\mathrm{tr}\, e^{\psi(\theta \mathbf{Y}_j)}, \quad j = 1, ..., n,\ \theta > 0.$$

Remark 4.3.1. As mentioned earlier, the function ψ_τ satisfies

$$-\tau \log(1 - \tau^{-1}x + \tau^{-2}x^2) \leq \tau\psi_1(x/\tau) \leq \tau \log(1 + \tau^{-1}x + \tau^{-2}x^2).$$

Then, for any matrix $\mathbf{A} \in \mathcal{S}_p$,

$$-\log\left(\mathbf{I} - \tau^{-1}\mathbf{A} + \tau^{-2}\mathbf{A}^2\right) \preceq \psi_1(\tau^{-1}\mathbf{A}) \preceq \log\left(\mathbf{I} + \tau^{-1}\mathbf{A} + \tau^{-2}\mathbf{A}^2\right).$$

♠

Recall that $\mathbf{Y}_1, ..., \mathbf{Y}_n \in \mathcal{S}_p$ are independent random matrices. For some $\theta > 0$ define the random matrix $\mathbf{T} \in \mathcal{S}_p$ as

$$\mathbf{T} = \mathbf{T}(\theta) = \frac{1}{n\theta} \sum_{j=1}^{n} \psi\left(\theta \mathbf{Y}_j\right). \tag{4.5}$$

The next theorem will be a consequence of Theorem 4.1.2 and is the main result studied in this thesis. It was first formulated in [32].

Theorem 4.3.2 (Minsker's concentration inequality)**.** *Let* $\mathbf{Y}_1, ..., \mathbf{Y}_n$ *be* $p \times p$ *independent and symmetric random matrices and* $\sigma_n^2 \geq \left\lVert\!\left| \sum_{j=1}^{n} \mathbb{E}\mathbf{Y}_j^2 \right|\!\right\rVert$. *Then, for* $\mathbf{T}$ *defined in* (4.5) *with* $\theta > 0$ *and* $t \geq 0$,

$$\left\lVert\!\left| \mathbf{T} - \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\mathbf{Y}_j \right|\!\right\rVert \geq \frac{t}{\sqrt{n}}$$

with probability at most

$$2p \exp\left(-\theta t\sqrt{n} + \frac{\theta^2 \sigma_n^2}{2}\right). \tag{4.6}$$

Before we prove Theorem 4.3.2, let us illustrate the forms that the later bound can take. It's straightforward to see that the value of θ that optimizes (4.6) is

$$\tilde{\theta} = \frac{t\sqrt{n}}{\sigma_n^2},$$

so we get for any $t > 0$ that the probability bound (4.6) is

$$2p \exp\left(-\tilde{\theta} t\sqrt{n} + \frac{\tilde{\theta}^2 \sigma_n^2}{2}\right) = 2p \exp\left(\frac{-nt^2}{2\sigma_n^2}\right), \tag{4.7}$$

which, except for the dimension factor p, is a *sub-Gaussian bound* in the sense that it resembles the one obtained for sub-Gaussian random variables (see Section 3.1). In this case, when choosing $\tilde{\theta}$, Theorem 4.3.2 will be informative (i.e. the bound (4.7) will be ≤ 1) only for values of t such that $t \geq n^{-1/2}\sigma_n\sqrt{2\log(2p)}$.

Remark 4.3.2. One form that will be useful for the Lepski's method, is the next one. Take $z > 0$, transform $t \mapsto z\sqrt{2t}$, and choose $\theta = \sqrt{\frac{2t}{n}\frac{1}{z}}$. Define the random matrix $\mathbf{Y}$ as

$$\mathbf{Y} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{Y}_j.$$

Then,

$$\mathbb{P}\left(\|\|\mathbf{T} - \mathbb{E}\mathbf{Y}\|\| \geq z\sqrt{\frac{2t}{n}}\right) \leq 2p\exp\left(-2t + \frac{t\sigma_n^2}{nz^2}\right).$$

If z is such that $z > \sigma_n/\sqrt{n}$, then

$$\mathbb{P}\left(\|\|\mathbf{T} - \mathbb{E}\mathbf{Y}\|\| \geq z\sqrt{\frac{2t}{n}}\right) \leq 2p\exp\left(-2t + t\right) \leq 2pe^{-t}.$$

♠

Proof of Theorem 4.3.2. Following Proposition 2.3.2 we will show that the two probabilities

$$\mathbb{P}\left(\lambda_1\left(\sum_{j=1}^{n}\left(\frac{1}{\theta}\psi(\theta\mathbf{Y}_j) - \mathbb{E}\mathbf{Y}_j\right)\right) \geq s\right)$$

and

$$\mathbb{P}\left(-\lambda_p\left(\sum_{j=1}^{n}\left(\frac{1}{\theta}\psi(\theta\mathbf{Y}_j) - \mathbb{E}\mathbf{Y}_j\right)\right) \geq s\right) \tag{4.8}$$

are both bounden by

$$p\exp\left(-\theta s + \frac{\theta^2}{2}\left\|\left\|\sum_{j=1}^{n}\mathbf{Y}_j^2\right\|\right\|\right).$$

Having done this, we can conclude that

$$\begin{aligned}\mathbb{P}\left(\left|\!\left|\!\left|\mathbf{T}-\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\mathbf{Y}_j\right|\!\right|\!\right|\geq\frac{s}{n}\right)&=\mathbb{P}\left(\left|\!\left|\!\left|\sum_{j=1}^{n}\left(\frac{1}{\theta}\psi(\theta\mathbf{Y}_j)-\mathbb{E}\mathbf{Y}_j\right)\right|\!\right|\!\right|\geq s\right)\\&\leq 2p\exp\left(-\theta s+\frac{\theta^2}{2}\left|\!\left|\!\left|\sum_{j=1}^{n}\mathbf{Y}_j^2\right|\!\right|\!\right|\right)\\&\leq 2p\exp\left(-\theta s+\frac{\theta^2\sigma_n^2}{2}\right),\end{aligned}$$

and taking $s=t\sqrt{n}$ will yield the result.

Using the notation of Theorem 4.1.2 set

$$\begin{gathered}\mathbf{X}_j=\frac{1}{\theta}\psi(\theta\mathbf{Y}_j),\quad\mathbf{H}_j=-\mathbb{E}\mathbf{Y}_j,\quad\mathbf{H}=\sum_{j=1}^{n}\mathbf{H}_j\\\mathbf{W}_j(\theta)=\log\left(\mathbf{I}+\theta\mathbf{Y}_j+\frac{\theta^2}{2}\mathbf{Y}_j^2\right),\quad\mathbf{M}_j(\theta)=\frac{\theta^2}{2}\mathbb{E}\mathbf{Y}_j^2.\end{gathered}$$

Because of Lemma 4.3.1 (c) it is clear that $\theta\mathbf{X}_j\preceq\mathbf{W}_j$ for any j. Also,

$$\log\mathbb{E}e^{\mathbf{W}_j(\theta)}+\theta\mathbf{H}_j=\log\left(\mathbf{I}+\theta\mathbb{E}\mathbf{Y}_j+\frac{\theta^2}{2}\mathbb{E}\mathbf{Y}_j^2\right)-\theta\mathbb{E}\mathbf{Y}_j,$$

and by Lemma 4.3.1 (a), taking $\mathbf{A}=\theta\mathbb{E}\mathbf{Y}_j+\frac{\theta^2}{2}\mathbb{E}\mathbf{Y}_j^2$ and rearranging terms, we have that $\log\mathbb{E}e^{\mathbf{W}_j(\theta)}+\theta\mathbf{H}_j\preceq\mathbf{M}_j(\theta)$ for all j. So Theorem 4.1.2 implies that

$$\mathbb{P}\left(\lambda_1\left(\sum_{j=1}^{n}\left(\frac{1}{\theta}\psi(\theta\mathbf{Y}_j)-\mathbb{E}\mathbf{Y}_j\right)\right)\geq s\right)\leq p\exp\left(-\theta s+\frac{\theta^2}{2}\left|\!\left|\!\left|\sum_{j=1}^{n}\mathbf{Y}_j^2\right|\!\right|\!\right|\right).$$

The bound for (4.8) is obtained in the same way but redefining

$$\begin{gathered}\mathbf{X}_j=-\frac{1}{\theta}\psi(\theta\mathbf{Y}_j),\quad\mathbf{H}_j=\mathbb{E}\mathbf{Y}_j,\quad\mathbf{H}=\sum_{j=1}^{n}\mathbf{H}_j\\\mathbf{W}_j(\theta)=\log\left(\mathbf{I}-\theta\mathbf{Y}_j+\frac{\theta^2}{2}\mathbf{Y}_j^2\right),\quad\mathbf{M}_j(\theta)=\frac{\theta^2}{2}\mathbb{E}\mathbf{Y}_j^2.\end{gathered}$$

and noticing by Lemma 4.3.1 (c) that $\theta\mathbf{X}_j=-\psi(\theta\mathbf{Y}_j)\preceq\log\left(\mathbf{I}-\theta\mathbf{Y}_j+\theta^2\mathbf{Y}_j^2/2\right)=\mathbf{W}_j$ for any j. This ends the proof. □

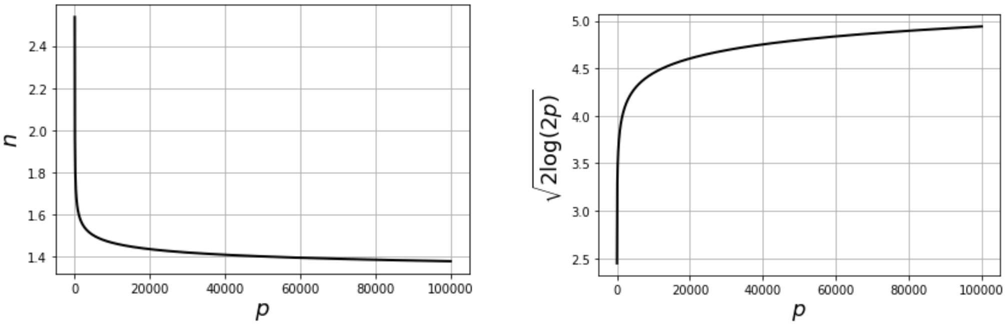


Figure 4.3: Effect of p over the probability bound. We fixed $\alpha = 0.01$ and $\sigma = 1$, and the range of values of p is from 100 to 100,000. Left: p v.s. n. Right: p v.s. $\sqrt{2\log(2p)}$

Suppose that we choose the optimal value $\tilde{\theta}$ for θ. If $\mathbf{Y}_1, ..., \mathbf{Y}_n$ are iid, we define $\sigma_n^2 = n\sigma^2$ for some $\sigma > 0$ such that $\sigma^2 \geq |||\mathbb{E}\mathbf{Y}_1^2|||$, and take $t = \sqrt{n}\sigma\sqrt{2\log(2p)}$ to get that

$$\mathbb{P}\left(|||\mathbf{T} - \mathbb{E}\mathbf{Y}_1||| \geq \sigma\sqrt{2\log(2p)}\right) \leq \frac{1}{(2p)^{n-1}}. \tag{4.9}$$

By this choice, we made the concentration bound to depend only on p, and the probability bound is less than the unity. Therefore, for any $\alpha \in (0, 1)$, if

$$n \geq \frac{\log(2p/\alpha)}{\log(2p)},$$

we guarantee that

$$\mathbb{P}\left(|||\mathbf{T} - \mathbb{E}\mathbf{Y}_1||| < \sigma\sqrt{2\log(2p)}\right) \geq 1 - \alpha$$

In Figure 4.3 we show the effect of the dimension p over this choice of n and over the bound $\sigma\sqrt{2\log(2p)}$ fixing $\alpha = 0.01$ and $\sigma = 1$. We observe that, as p grows, the number of samples needed decreases very quickly. This seems counter intuitive, but it is a consequence that in inequality (4.9) the probability bound decreases both with p and n. What is interesting is that the upper bound $\sigma\sqrt{2\log(2p)}$ increases very slowly with p, and only the term σ can affect this behavior. This indicates that if p is very large, we do not need a very large sample n to guarantee a reasonable deviation for the estimator $\mathbf{T}$ from $\mathbb{E}\mathbf{Y}_1$. With this observation we can conclude that the performance of the estimator $\mathbf{T}$ doesn't seem to be much affected by the dimension p.

Additional to the previous analysis, we observe that in the iid case when we choose $\theta = \tilde{\theta}$ and the mapping $t \mapsto \sqrt{n}t$, we have that

$$\mathbb{P}\left(\|\|\mathbf{T} - \mathbb{E}\mathbf{Y}\|\| \geq t \right) \leq 2p \exp\left(\frac{-nt^2}{2\sigma^2} \right) = 2n \exp\left(\frac{-nt^2}{2\sigma^2} \right) \frac{p}{n}.$$

So as long as $p/n \to 0$ when $n \to \infty$ we get that

$$\|\|\mathbf{T} - \mathbb{E}\mathbf{Y}\|\| \xrightarrow{\mathbb{P}} 0, \quad n \to \infty,$$

i.e., $\mathbf{T}$ is a consistent estimator of $\mathbb{E}\mathbf{Y}$ in terms of the operator norm.

4.4 Bounds for rectangular matrices

In this section we show how the estimator $\mathbf{T}$ is generalized to rectangular random matrices. The procedure is analogous, but we need to make some modifications in order to work with symmetric matrices. Recall that the symmetric dilation of a matrix $\mathbf{B} \in \mathcal{M}_{p_1,p_2}$ is defined as

$$\mathcal{H}(\mathbf{B}) = \begin{pmatrix} \mathbf{0} & \mathbf{B} \\ \mathbf{B}^{\intercal} & \mathbf{0} \end{pmatrix} \in \mathcal{S}_{p_1+p_2}.$$

Since we defined the matrix operator only for symmetric matrices, the concept of dilation enables us to work with matrix operators of rectangular matrices, i.e., instead of defining $f(\mathbf{B})$, we'll work with $f(\mathcal{H}(\mathbf{B}))$. One may argue that we can define $f(\mathbf{B})$ by applying the function f to the singular values of $\mathbf{B}$ in a similar way that we did with the eigenvalues of symmetric matrices. But this formulation wont help since the main results rely on the Lieb's Theorem 2.4.1, and this result assumes that the matrices are symmetric. As far as the author is concerned, there is no generalization of Lieb's Theorem for more general matrices.

Let $\mathbf{Z}_1, ...\mathbf{Z}_n \in \mathcal{M}_{p_1,p_2}$ be independent random matrices and define σ_n^2 such that

$$\sigma_n^2 \geq \max\left\{ \left\|\left\| \sum_{j=1}^{n} \mathbb{E}\mathbf{Z}_j\mathbf{Z}_j^{\intercal} \right\|\right\|, \left\|\left\| \sum_{j=1}^{n} \mathbb{E}\mathbf{Z}_j^{\intercal}\mathbf{Z}_j \right\|\right\| \right\}. \tag{4.10}$$

Define the random matrix $\mathbf{T} \in \mathcal{S}_{p_1+p_2}$ as

$$\mathbf{T} = \mathbf{T}(\theta) = \frac{1}{n\theta} \sum_{j=1}^{n} \psi\left(\theta \mathcal{H}(\mathbf{Z}_j) \right). \tag{4.11}$$

Consider the partition

$$\mathbf{T} = \begin{pmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{12}^{\intercal} & \mathbf{T}_{22} \end{pmatrix}, \tag{4.12}$$

where $\mathbf{T}_{11} \in \mathcal{M}_{p_1}$, $\mathbf{T}_{22} \in \mathcal{M}_{p_2}$ and $\mathbf{T}_{12} \in \mathcal{M}_{p_1,p_2}$. Since

$$\mathbb{E}\left[\sum_{j=1}^{n} \mathbf{Z}_j\right] = \begin{pmatrix} \mathbf{0} & \mathbb{E}\left[\sum_{j=1}^{n} \mathbf{Z}_j\right] \\ \mathbb{E}\left[\sum_{j=1}^{n} \mathbf{Z}_j^{\intercal}\right] & \mathbf{0} \end{pmatrix},$$

it is natural to think that $\mathbf{T}_{12}$ is a good estimator of $\frac{1}{n}\mathbb{E}\left[\sum_{j=1}^{n} \mathbf{Z}_j\right]$. This intuition is confirmed in Theorem 4.4.1. To state this result we'll use that for any matrices $\mathbf{A} \in \mathcal{S}_{p_1}$ and $\mathbf{H} \in \mathcal{S}_{p_2}$,

$$\left\lVert\!\left\lvert \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^{\intercal} & \mathbf{H} \end{pmatrix} \right\rvert\!\right\rVert \geq \left\lVert\!\left\lvert \begin{pmatrix} \mathbf{0} & \mathbf{B} \\ \mathbf{B}^{\intercal} & \mathbf{0} \end{pmatrix} \right\rvert\!\right\rVert,$$

and that $|||\mathcal{H}(\mathbf{B})||| = |||\mathbf{B}|||$. This two results are proved in Lemma A.5.6 and Corollary A.5.3, respectively. The next result was taken from [32].

Theorem 4.4.1. *For independent random matrices $\mathbf{Z}_1, ..., \mathbf{Z}_n \in \mathcal{M}_{p_1,p_2}$, consider the definitions of σ_n^2 in (4.10), and $\mathbf{T}$ in (4.11) and (4.12). Then,*

$$\left\lVert\!\left\lvert \mathbf{T}_{12} - \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\mathbf{Z}_j \right\rvert\!\right\rVert \geq \frac{t}{\sqrt{n}}$$

with probability at most

$$2(p_1 + p_2)\exp\left(-\theta t\sqrt{n} + \frac{\theta^2\sigma_n^2}{2}\right).$$

Proof. First note that

$$\mathcal{H}(\mathbf{Z}_j)^2 = \begin{pmatrix} \mathbf{Z}_j\mathbf{Z}_j^{\intercal} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_j^{\intercal}\mathbf{Z}_j \end{pmatrix}$$

and that

$$\sum_{j=1}^{n} \mathbb{E}\left[\mathcal{H}(\mathbf{Z}_j)^2\right] = \begin{pmatrix} \sum_{j=1}^{n} \mathbb{E}\mathbf{Z}_j\mathbf{Z}_j^{\intercal} & \mathbf{0} \\ \mathbf{0} & \sum_{j=1}^{n} \mathbb{E}\mathbf{Z}_j^{\intercal}\mathbf{Z}_j \end{pmatrix}.$$

Then, since

$$\left|\left|\left|\begin{pmatrix}\mathbf{A}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{A}_2\end{pmatrix}\right|\right|\right| = \max\left\{\|\mathbf{A}_1\|, \|\mathbf{A}_2\|\right\},$$

(Lemma A.5.2) we have that

$$\left|\left|\left|\sum_{j=1}^{n}\mathbb{E}\left[\mathcal{H}(\mathbf{Z}_j)^2\right]\right|\right|\right| = \max\left\{\left|\left|\left|\sum_{j=1}^{n}\mathbb{E}\mathbf{Z}_j\mathbf{Z}_j^\intercal\right|\right|\right|, \left|\left|\left|\sum_{j=1}^{n}\mathbb{E}\mathbf{Z}_j^\intercal\mathbf{Z}_j\right|\right|\right|\right\} \leq \sigma_n^2.$$

Then, by Minsker's concentration inequality of Theorem 4.3.2, since $\mathbf{T}$ and $\mathcal{H}(\mathbf{Z}_j)$, $j = 1, ..., n$, are symmetric and $\mathbb{E}\mathcal{H}(\mathbf{Z}_j) = \mathcal{H}(\mathbb{E}\mathbf{Z}_j)$, we get that

$$\left|\left|\left|\mathbf{T} - \frac{1}{n}\sum_{j=1}^{n}\mathcal{H}(\mathbb{E}\mathbf{Z}_j)\right|\right|\right| < \frac{t}{\sqrt{n}}, \quad t \geq 0,$$

with probability at least

$$1 - 2(p_1 + p_2)\exp\left(-\theta t\sqrt{n} + \frac{\theta^2\sigma_n^2}{2}\right). \tag{4.13}$$

Now, by Lemma A.5.6 and Corollary A.5.3,

$$\begin{aligned}\left|\left|\left|\mathbf{T} - \frac{1}{n}\sum_{j=1}^{n}\mathcal{H}(\mathbb{E}\mathbf{Z}_j)\right|\right|\right| &= \left|\left|\left|\begin{pmatrix}\mathbf{T}_{11} & \mathbf{T}_{12} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\mathbf{Z}_j\\ \mathbf{T}_{12}^\intercal - \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\mathbf{Z}_j^\intercal & \mathbf{T}_{12}\end{pmatrix}\right|\right|\right|\\ &\geq \left|\left|\left|\begin{pmatrix}\mathbf{0} & \mathbf{T}_{12} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\mathbf{Z}_j\\ \mathbf{T}_{12}^\intercal - \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\mathbf{Z}_j^\intercal & \mathbf{0}\end{pmatrix}\right|\right|\right|\\ &= \left|\left|\left|\mathcal{H}\left(\mathbf{T}_{12} - \sum_{j=1}^{n}\mathbb{E}\mathbf{Z}_j\right)\right|\right|\right|\\ &= \left|\left|\left|\mathbf{T}_{12} - \sum_{j=1}^{n}\mathbb{E}\mathbf{Z}_j\right|\right|\right|.\end{aligned}$$

This provokes that

$$\left|\left|\left|\mathbf{T}_{12} - \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\mathbf{Z}_j\right|\right|\right| < \frac{t}{\sqrt{n}},$$

with probability at least (4.13). □

All the analysis made for symmetric matrices in the previous section can be applied for matrices in $\mathcal{M}_{p_1,p_2}$ with the only difference being the dimension factors change from p to p_1+p_2. In particular, for matrices in $\mathcal{M}_p$ the probability bound is

$$4p\exp\left(-\theta t\sqrt{n}+\frac{\theta^2\sigma_n^2}{2}\right),$$

i.e., the double of what was obtained for matrices in $\mathcal{S}_p$.

Despite that we found the optimal value of the hyperparameter θ is $\tilde{\theta} = t\sqrt{n}/\sigma_n^2$, the ignorance of the parameter σ_n^2 avoid that we can fully calculate $\mathbf{T}$. In the next section we present a method that tries to overcome this difficulty.

4.5 Lepskii's method of calculation

As before, $\mathbf{Y}_1, ..., \mathbf{Y}_n \in \mathcal{S}_p$ are iid random matrices and we want to estimate $\frac{1}{n}\sum_{j=1}^n \mathbb{E}\mathbf{Y}_j$. Recall that we do not know the parameter $\sigma_n^2 \geq \|\!|\sum_{j=1}^n \mathbb{E}\mathbf{Y}_j^2|\!\|$. The adaptive method of Lepskii for estimation [25] consists in defining an upper and lower bound for σ_n^2 and to define sub-intervals whose union is the interval defined by the upper and lower bound. In each interval we define a θ_j and choose the first index j such that the estimation $\mathbf{T}(\theta_j)$ is closed to the others $\mathbf{T}(\theta_k)$, $k > j$. More precisely, consider a sequence $z_0, z_1, ...$ of positive numbers and define for $j = 0, 1, ...$

$$\mathbf{T}_j = \frac{1}{n\theta_j}\sum_{i=1}^n \psi\left(\theta_j \mathbf{Y}_i\right), \quad \text{where } \theta_j = \sqrt{\frac{2t}{n}\frac{1}{z_j}}.$$

Suppose we know some positive bounds $\sigma_{\max}$ and $\sigma_{\min}$ such that

$$\sigma_{\min} \leq \frac{\sigma_n}{\sqrt{n}} \leq \sigma_{\max},$$

where $\sigma_n^2 \geq \|\!|\sum_{j=1}^n \mathbb{E}\mathbf{Y}_j^2|\!\|$. We define the sequence $(z_j)_{j\geq 0}$ as $z_j = \gamma^j\sigma_{\min}$, where $1 < \gamma < 2$ is arbitrary. The index set $\mathcal{J}$ is defined as

$$\mathcal{J} = \left\{j \geq 0 \,:\, \sigma_{\min} \leq z_j < \gamma\sigma_{\max}\right\}.$$

The cardinality of $|\mathcal{J}|$ is at most $1 + \log(\sigma_{\max}/\sigma_{\min})/\log(\gamma)$. To see this, let N be the power of γ such that $\sigma_{\min}\gamma^N = \gamma\sigma_{\max}$. Then, $(N-1)\log(\gamma) = \log(\sigma_{\max}/\sigma_{\min})$, and solve for N to get the bound.

Input: sample $\mathbf{Y}_1, ..., \mathbf{Y}_n$ and parameters $\sigma_{\max}, \sigma_{\min}, \gamma, t$.
Output: robust estimator $\mathbf{T}^*$.

1. Calculate the cardinality $c = \lfloor 1 + \log(\sigma_{\max}/\sigma_{\min})/\log(\gamma) \rfloor$ of the set $|\mathcal{J}|$ and the vector $\boldsymbol{z} = (z_j)_{j=0,...,c-1}$, $z_j = \gamma^j \sigma_{\min}$.

2. For $j = 0, ..., c-1$, choose the first j such that
$$|||\mathbf{T}_k - \mathbf{T}_j||| \leq 2z_k\sqrt{\frac{2t}{n}}$$
for all $k = j+1, ..., c-1$. Define this minimum j as j_*. If the minimum doesn't exist, define $j_* = j_{c-1}$.

3. Return the matrix $\mathbf{T}^* = \mathbf{T}_{j_*}$.

Figure 4.4: Lepskii's algorithm to obtain the robust estimator $\mathbf{T}^*$.

Define the special random index

$$j_* = \min\left\{ j \in \mathcal{J} \; : \; |||\mathbf{T}_k - \mathbf{T}_j||| \leq 2z_k\sqrt{\frac{2t}{n}} \; \forall k > j, k \in \mathcal{J} \right\}, \tag{4.14}$$

where $\min \emptyset := \infty$ and $z_\infty := z_{|\mathcal{J}|-1}$. The robust estimator is defined as $\mathbf{T}^* = \mathbf{T}_{j_*}$. A resume of the Lepskii's procedure is presented in Figure 4.4.

To prove the next result, we also define the auxiliary (non-random) index j_0 as

$$j_0 = \min\left\{ j \in \mathcal{J} \; : \; z_j \geq \frac{\sigma_n}{\sqrt{n}} \right\}. \tag{4.15}$$

Note that $z_{j_0} \leq \gamma\sigma_n/\sqrt{n}$, because on the contrary we will get $\gamma^{j_0}\sigma_{\min} > \gamma\sigma_n/\sqrt{n}$ and $z_{j_0-1} > \sigma_n/\sqrt{n}$, which contradicts the definition of j_0. The next result is an improvement of the one presented in [32].

Theorem 4.5.1. *Define* $\mathbf{Y} = \frac{1}{n}\sum_{i=1}^n \mathbf{Y}_i$. *Then, for any* $\epsilon > 0$*,*

$$|||\mathbf{T}^* - \mathbb{E}\mathbf{Y}||| \geq (3+\epsilon)\frac{\sigma_n}{\sqrt{n}}\sqrt{\frac{2t}{n}}$$

with probability at most

$$2p\frac{\log(\gamma\sigma_{\max}/\sigma_{\min})}{\log(\gamma)}e^{-t}.$$

Proof. For ease of notation, let the norm an probability bounds be

$$x = (3+\epsilon)\frac{\sigma_n}{\sqrt{n}}\sqrt{\frac{2t}{n}} \quad \text{and} \quad q = 2p\frac{\log(\gamma\sigma_{\max}/\sigma_{\min})}{\log(\gamma)}e^{-t},$$

respectively. We'll proceed as follows. First we'll define two sets $\mathcal{B}$ and $\mathcal{E}$ such that

$$\mathcal{B} \subset \mathcal{E} \subset \{|||\mathbf{T}^* - \mathbb{E}\mathbf{Y}||| \leq x\},$$

and then we'll prove that $\mathbb{P}(\mathcal{B}) \geq 1-q$.

<u>Definition of sets</u>: Define

$$\mathcal{B} = \bigcap_{k\in\mathcal{J},\, k\geq j_0} \left\{|||\mathbf{T}_k - \mathbb{E}\mathbf{Y}||| \leq z_k\sqrt{\frac{2t}{n}}\right\} \quad \text{and} \quad \mathcal{E} = \{j_* \leq j_0\}.$$

To see that $\mathcal{B} \subset \mathcal{E}$, note that

$$|||\mathbf{T}_k - \mathbf{T}_{j_0}||| \leq |||\mathbf{T}_k - \mathbb{E}\mathbf{Y}||| + |||\mathbf{T}_{j_0} - \mathbb{E}\mathbf{Y}|||.$$

So, if $\mathcal{B}$ is true, we have that

$$\begin{aligned} |||\mathbf{T}_{j_0} - \mathbb{E}\mathbf{Y}||| &\leq z_{j_0}\sqrt{\frac{2t}{n}} \leq z_k\sqrt{\frac{2t}{n}}, \quad k > j_0, \\ |||\mathbf{T}_k - \mathbb{E}\mathbf{Y}||| &\leq z_k\sqrt{\frac{2t}{n}}, \quad k > j_0, \end{aligned}$$

since $z_k > z_{j_0}$ for every $k > j_0$. Then, by definition of j_*

$$\mathcal{B} \subset \bigcap_{k\in\mathcal{J},\, k\geq j_0} \left\{|||\mathbf{T}_k - \mathbf{T}_{j_0}||| \leq 2z_k\sqrt{\frac{2t}{n}}\right\} \subset \mathcal{E}.$$

Now, if $\mathcal{B}$ is true,

$$|||\mathbf{T}^* - \mathbf{T}_{j_0}||| \leq 2z_{j_0}\sqrt{\frac{2t}{n}}, \quad (\text{because } \mathcal{B} \subset \mathcal{E})$$

$$\|\|\mathbf{T}_{j_0} - \mathbb{E}\mathbf{Y}\|\| \leq z_{j_0}\sqrt{\frac{2t}{n}}, \quad \text{(because } \mathcal{B} \text{ is true)}.$$

And using again the triangle inequality we get

$$\|\|\mathbf{T}^* - \mathbb{E}\mathbf{Y}\|\| \leq \|\|\mathbf{T}^* - \mathbf{T}_{j_0}\|\| + \|\|\mathbf{T}_{j_0} - \mathbb{E}\mathbf{Y}\|\| \leq 3z_{j_0}\sqrt{\frac{2t}{n}}.$$

Then, by definition of j_0 in (4.15) we have that

$$\|\|\mathbf{T}^* - \mathbb{E}\mathbf{Y}\|\| \leq 3\gamma\frac{\sigma_n}{\sqrt{n}}\sqrt{\frac{2t}{n}}.$$

Fix $\epsilon > 0$. By taking $\gamma = 1 + \epsilon/3$ we get that if $\mathcal{B}$ is true, then $\|\|\mathbf{T}^* - \mathbb{E}\mathbf{Y}\|\| \leq x$.

Bound for $\mathbb{P}(\mathcal{B})$: The complement of the set $\mathcal{B}$ is

$$\mathcal{B}^c = \bigcup_{k\in\mathcal{J},\, k\geq j_0} \left\{\|\|\mathbf{T}_k - \mathbb{E}\mathbf{Y}\|\| > z_k\sqrt{\frac{2t}{n}}\right\}.$$

Recall that, by definition, $z_0 \geq z_1 \geq \cdots$, therefore

$$\begin{aligned}
\mathbb{P}\left(\mathcal{B}^c\right) &\leq \sum_{k\in\mathcal{J},\, k\geq j_0} \mathbb{P}\left(\|\|\mathbf{T}_k - \mathbb{E}\mathbf{Y}\|\| > z_k\sqrt{\frac{2t}{n}}\right) \\
&\leq \left(1 + \frac{\log(\sigma_{\max}/\sigma_{\min})}{\log(\gamma)}\right)\mathbb{P}\left(\|\|\mathbf{T}_k - \mathbb{E}\mathbf{Y}\|\| > z_{j_0}\sqrt{\frac{2t}{n}}\right) \\
&= \frac{\log(\gamma\sigma_{\max}/\sigma_{\min})}{\log(\gamma)}\mathbb{P}\left(\|\|\mathbf{T}_k - \mathbb{E}\mathbf{Y}\|\| > z_{j_0}\sqrt{\frac{2t}{n}}\right),
\end{aligned}$$

where in the second inequality we used that $|\mathcal{J}| \leq 1 + \log(\gamma\sigma_{\max}/\sigma_{\min})/\log(\gamma)$. Finally, by Remark 4.3.2 of Minsker's concentration inequality, we bound the probability in the last equality to get that

$$\begin{aligned}
\mathbb{P}\left(\mathcal{B}^c\right) &\leq 2p\frac{\log(\gamma\sigma_{\max}/\sigma_{\min})}{\log(\gamma)}\exp\left(-2t + \frac{t\sigma_n^2}{nz_{j_0}^2}\right) \\
&\leq 2p\frac{\log(\gamma\sigma_{\max}/\sigma_{\min})}{\log(\gamma)}e^{-t}, \quad \text{(because } z_{j_0} \geq \sigma_n/\sqrt{n}\text{)} \\
&= q.
\end{aligned}$$

Then, $\mathbb{P}\left(\mathcal{B}\right) \geq 1 - q$ which finally proves that $\mathbb{P}\left(\|\|\mathbf{T}^* - \mathbb{E}\mathbf{Y}\|\| \leq x\right) \geq 1 - q$. □

To apply the Lepskii's procedure, one need to establish the parameters $\sigma_{\max}$, $\sigma_{\min}$, γ and t. The values of $\sigma_{\max}, \sigma_{\min}$ can be obtained from the sample through a rough estimation. The value of γ is reasonable to set it as the middle in $(1, 2)$, i.e., $\gamma = 3/2$. From Theorem 4.5.1, we can establish that the values of t such that the probability bound is ≤ 1 are

$$t \geq \log(2p\kappa), \quad \kappa = \frac{\log(\gamma\sigma_{\max}/\sigma_{\min})}{\log(\gamma)}. \tag{4.16}$$

Since greater values of t imply greater values of θ_j, this choice of t can provoke too much regularization of the influence function ψ, i.e., only small values are not reduced. Therefore, depending on the value of p and the other parameters of the process we can choose t as in (4.16) or a much small value to adjust for less regularization. An implementation of this procedure is explored in Chapter 5.

Remark 4.5.1. One major observation of this procedure is that the result of Theorem 4.5.1 is bases in two things: 1) the definition of the algorithm 4.4 and 2) the probability bound of Remark 4.3.2, namely,

$$\mathbb{P}\left(|\!|\!| \mathbf{T} - \mathbb{E}\mathbf{Y} |\!|\!| \geq z\sqrt{\frac{2t}{n}} \right) \leq 2pe^{-t}.$$

Therefore, the Lepskii's procedure can be applied for any matrix estimator $\mathbf{T}$ which guarantees a concentration bound of the previous form. This observation will be useful in Chapter 5 for the case of covariance matrix estimation. ♠

4.6 Applications

In this section we present applications of the methodology of this chapter in order to obtain theoretical guarantees in two common statistical procedures: Principal Components Analysis (PCA) and Community Detection. In order to do it, we need to present the Davis-Kahan Theorem which is the subject of the next subsection.

4.6.1 The Davis-Kahan Theorem

The following theorem was first published in [11]. It gives a way to bound the angle between eigenvalues of two symmetric matrices. We present a simpler version, the proof of which can be found in [53].

Theorem 4.6.1 (Davis-Kahan)**.** *Define* $\mathbf{A}, \mathbf{H} \in \mathcal{S}_p$ *with eigenvectors* $\boldsymbol{v}_i$ *and* $\boldsymbol{w}_i$ *such that* $\mathbf{A}\boldsymbol{v}_i = \lambda_i(\mathbf{A})\boldsymbol{v}_i$ *and* $\mathbf{H}\boldsymbol{w}_i = \lambda_i(\mathbf{H})\boldsymbol{w}_i$*. Fix* $i \in \{1, ..., p\}$ *and assume that*

$$\delta_i := \min\{\lambda_{i-1}(\mathbf{A}) - \lambda_i(\mathbf{A}),\ \lambda_i(\mathbf{A}) - \lambda_{i+1}(\mathbf{A})\} > 0,$$

where $\lambda_0(\mathbf{A}) := \infty$*,* $\lambda_{p+1}(\mathbf{A}) := -\infty$*. Then,*

$$\sin\left(\angle(\boldsymbol{w}_i, \boldsymbol{v}_i)\right) \leq \frac{2\|\mathbf{H} - \mathbf{A}\|}{\delta_i}.$$

As pointed out in [47, p. 89], for unitary eigenvectors $\boldsymbol{w}_i$ and $\boldsymbol{v}_i$, the Davis-Kahan Theorem implies that

$$\exists \rho \in \{-1, 1\} : \quad \|\rho\boldsymbol{w}_i - \boldsymbol{v}_i\|_2 \leq \frac{2^{3/2}\|\mathbf{H} - \mathbf{A}\|}{\delta_i}. \tag{4.17}$$

Indeed, recall that for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$,

$$\sin(\angle(\boldsymbol{x}, \boldsymbol{y})) = \sqrt{1 - \frac{(\boldsymbol{x}^\intercal \boldsymbol{y})^2}{\|\boldsymbol{x}\|_2^2 \|\boldsymbol{y}\|_2^2}}.$$

Then, if $\boldsymbol{w}_i^\intercal \boldsymbol{v}_i \geq 0$,

$$\begin{aligned}
\|\boldsymbol{w}_i - \boldsymbol{v}_i\|_2 &= \sqrt{\|\boldsymbol{w}_i\|_2^2 + \|\boldsymbol{v}_i\|_2^2 - 2\boldsymbol{w}_i^\intercal \boldsymbol{v}_i} \\
&= 2^{1/2}\sqrt{1 - \boldsymbol{w}_i^\intercal \boldsymbol{v}_i} \\
&\leq 2^{1/2}\sqrt{1 - (\boldsymbol{w}_i^\intercal \boldsymbol{v}_i)^2} \\
&= 2^{1/2}\sin(\angle(\boldsymbol{w}_i, \boldsymbol{v}_i)),
\end{aligned}$$

because $\|\boldsymbol{w}_i\|_2 = \|\boldsymbol{w}_i\|_2 = 1$. On the other hand if $\boldsymbol{w}_i^\intercal \boldsymbol{v}_i \leq 0$ then $-\boldsymbol{w}_i^\intercal \boldsymbol{v}_i \geq 0$, and since $-\boldsymbol{w}_i$ is a unitary eigenvector associated to $\lambda_i(\mathbf{H})$ we can perform the same calculations to get

$$\|-\boldsymbol{w}_i - \boldsymbol{v}_i\|_2 \leq 2^{1/2}\sin(\angle(\boldsymbol{w}_i, \boldsymbol{v}_i)).$$

The previous inequality (4.17) will be particularly useful for Community Detection.

4.6.2 Robust PCA

PCA is a method to project data points to a lower dimension, and with that make statistical analysis easier. Suppose $\boldsymbol{X} \in \mathbb{R}^p$ is a random vector with $\mathbb{E}\boldsymbol{X} = \boldsymbol{\mu}$ and $\mathrm{Cov}\boldsymbol{X} = \boldsymbol{\Sigma}$. The PCA procedure is based on finding the unitary vector $\boldsymbol{\delta}$ that maximizes the variance of $\boldsymbol{\delta}^\intercal \boldsymbol{X}$, i.e.,

$$\boldsymbol{\delta} = \underset{\|\boldsymbol{x}\|_2=1}{\arg\max}\, \mathrm{Var}(\boldsymbol{x}^\intercal \boldsymbol{X}) = \underset{\|\boldsymbol{x}\|_2=1}{\arg\max}\, \boldsymbol{x}^\intercal \boldsymbol{\Sigma} \boldsymbol{x}.$$

The Rayleigh quotient of Theorem A.2.4 implies that $\boldsymbol{\delta} = \boldsymbol{v}_1$, where $\boldsymbol{v}_1$ is the unitary eigenvector associated to $\lambda_1(\boldsymbol{\Sigma})$. This says that if we want to project a vector $\boldsymbol{X}$ into a lower dimension that guarantees maximum variance, we need to project it in the direction of $\boldsymbol{v}_1$, the eigenvector of the maximum eigenvalue[2]. More generally, to obtain a lower dimensional representation

$$\begin{pmatrix} \boldsymbol{\delta}_1^\intercal \boldsymbol{X} \\ \vdots \\ \boldsymbol{\delta}_k^\intercal \boldsymbol{X} \end{pmatrix},$$

where $k < p$ and $\boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_k$ are orthonormal, we choose the vector $\boldsymbol{\delta}_i$ to be the unitary eigenvector associated to $\lambda_i(\boldsymbol{\Sigma})$. For more details on the PCA procedure one can see [1, Chapter 11] or any other source on Multivariate Statistical Analysis. In essence, what PCA is doing is to project the data so that we guarantee that the points are *statistically separated*, i.e., we project in the direction of maximum variance.

Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n \in \mathbb{R}^p$ be iid copies of $\boldsymbol{X}$. We don't know $\boldsymbol{\Sigma}$ a priori so we have to estimate it through $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$. As $\boldsymbol{X}$ can be heavy-tailed, we want to formulate a robust estimator of $\boldsymbol{\Sigma}$. For simplicity, we assume that $\mathbb{E}\boldsymbol{X} = \boldsymbol{0}$. Just as we did in the simulation study of the Lepskii's method in the previous section, since $\mathbb{E}\boldsymbol{X}\boldsymbol{X}^\intercal = \boldsymbol{\Sigma}$, we define the robust estimator

$$\widetilde{\boldsymbol{\Sigma}} = \widetilde{\boldsymbol{\Sigma}}(\theta) = \frac{1}{n\theta} \sum_{j=1}^{n} \psi\left(\theta \boldsymbol{X}_j \boldsymbol{X}_j^\intercal\right), \quad \theta > 0.$$

Let $\sigma > 0$ be such that $\sigma^2 \geq |||\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^\intercal)^2||| = |||\mathbb{E}[\|\boldsymbol{X}\|_2^2 \boldsymbol{X}\boldsymbol{X}^\intercal]|||$. Taking $\theta = \sqrt{\frac{t}{n}\frac{1}{\sigma}}$ From Minsker's concentration inequality of Theorem 4.3.2 we can conclude that, for any $t \geq 0$,

$$\mathbb{P}\left(|||\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||| \leq \sigma\sqrt{\frac{t}{n}} \right) \geq 1 - 2pe^{-t}. \tag{4.18}$$

[2] We specifically defined a linear combination of $\boldsymbol{X}$, namely, $\boldsymbol{\delta}^\intercal \boldsymbol{X}$. This is just a model to project $\boldsymbol{X}$, so there could be more sophisticated formulations.

This concentration bound indicates that $\widetilde{\boldsymbol{\Sigma}}$ concentrates well around $\boldsymbol{\Sigma}$, particularly when the sample size n increases. Note that we can establish the concentration bound of Theorem 4.5.1, but for this section we focus on the simpler result (4.18). Since we are interested on how well we estimate the eigenvectors $\boldsymbol{v}_i$ of $\boldsymbol{\Sigma}$, can we derive concentration bounds for the eigevectors $\tilde{\boldsymbol{v}}_i$ of $\widetilde{\boldsymbol{\Sigma}}$ around $\boldsymbol{v}_i$? We can do it almost directly with the Davis-Kahan Theorem as we present bellow.

Let $\tilde{\boldsymbol{v}}_i$ and $\boldsymbol{v}_i$ be unitary vectors satisfying

$$\widetilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{v}}_i = \lambda_i(\widetilde{\boldsymbol{\Sigma}})\tilde{\boldsymbol{v}}_i \quad \text{and} \quad \boldsymbol{\Sigma}\boldsymbol{v}_i = \lambda_i(\boldsymbol{\Sigma})\boldsymbol{v}_i.$$

Using the notation of Davis-Kahan Theorem, we'll assume for $\boldsymbol{\Sigma}$ that for all $i = 1, ..., p$

$$\delta_i = \min\{\lambda_{i-1}(\boldsymbol{\Sigma}) - \lambda_i(\boldsymbol{\Sigma}),\, \lambda_i(\boldsymbol{\Sigma}) - \lambda_{i+1}(\boldsymbol{\Sigma})\} > 0.$$

The next theorem guarantees a concentration bound for the eigenvectors $\tilde{\boldsymbol{v}}_i$.

Theorem 4.6.2. *With the assumptions made above we have that*

$$\mathbb{P}\left(\min_{\rho\in\{-1,1\}} \|\rho\tilde{\boldsymbol{v}}_i - \boldsymbol{v}_i\|_2 \geq \frac{2^{3/2}\sigma}{\delta_i}\sqrt{\frac{t}{n}}\right) \leq 2pe^{-t}.$$

Proof. Observe that for each i,

$$\begin{aligned}\min_{\rho\in\{-1,1\}} \|\rho\tilde{\boldsymbol{v}}_i - \boldsymbol{v}_i\|_2^2 &= \min\{2 - 2\tilde{\boldsymbol{v}}_i^\intercal \boldsymbol{v}_i, 2 + 2\tilde{\boldsymbol{v}}_i^\intercal \boldsymbol{v}_i\} \\ &= 2 - 2|\tilde{\boldsymbol{v}}_i^\intercal \boldsymbol{v}_i| \\ &\leq 2 - 2(\tilde{\boldsymbol{v}}_i^\intercal \boldsymbol{v}_i)^2 \\ &= 2\sin^2(\angle(\tilde{\boldsymbol{v}}_i, \boldsymbol{v}_i)),\end{aligned}$$

so

$$\min_{\rho\in\{-1,1\}} \|\rho\tilde{\boldsymbol{v}}_i - \boldsymbol{v}_i\|_2 \leq 2^{1/2}\sin(\angle(\tilde{\boldsymbol{v}}_i, \boldsymbol{v}_i)).$$

Therefore, from (4.18) and the Davis-Kahan Theorem we get that

$$\mathbb{P}\left(\min_{\rho\in\{-1,1\}} \|\rho\tilde{\boldsymbol{v}}_i - \boldsymbol{v}_i\|_2 \leq \frac{2^{3/2}\sigma}{\delta_i}\sqrt{\frac{t}{n}}\right) \geq \mathbb{P}\left(\|\!|\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}|\!\|\leq \sigma\sqrt{\frac{t}{n}}\right) \geq 1 - 2pe^{-t}.$$

□

One particular model is well suited for this result. The *spiked covariance model* presented in [20] establishes that $\mathbf{\Sigma} = \mu \boldsymbol{u}\boldsymbol{u}^\intercal + \mathbf{I}$, where $\boldsymbol{u} \in \mathbb{R}^p$ is unitary and $\mu > 0$. Then, it is immediate to see that

$$\lambda_1(\mathbf{\Sigma}) = \mu + 1 \quad \text{and} \quad \lambda_i(\mathbf{\Sigma}) = 1,\ i > 1.$$

This is a model used when $p \gg n$, since, according to [20], in this context there is an eigenvalue that is much greater that the others, i.e., a plot of the eigenvalues presents a *spike*. Then, from the notation of Davis-Kahan Theorem, $\delta_1 = \min\{\infty, \mu\} = \mu$, so

$$\mathbb{P}\left(\min_{\rho \in \{-1,1\}} \|\rho \tilde{\boldsymbol{v}}_1 - \boldsymbol{v}_1\|_2 \geq \frac{2^{3/2}\sigma}{\mu}\sqrt{\frac{t}{n}}\right) \leq 2pe^{-t}.$$

4.6.3 Community detection with noise

Consider a undirected graph G with vertex set $V \subset \mathbb{Z}$ and edge set E. We say that G is a *random graph* if the edge set E is random, i.e., two vertices i and j are connected with certain probability. The objective of community detection on graphs is to discern which vertices in V belongs to a certain community. More precisely, we assume that we have the partition $V = V_1 \cup \cdots \cup V_K$, where V_ℓ is called a community, and we say that i belongs to community ℓ if $i \in V_\ell$. We focus on the case $k = 2$ and the model called *Stochastic block model*.

Definition 4.6.3 (Stochastic block model)**.** *Let G be a random graph with $2p$ vertices divided by two set of size p each. Connect any pair of vertices independently with probability δ if they belong to the same community and probability ε, where $\varepsilon < \delta$, if they belong to different communities. This distribution on graphs is called the stochastic block model and is denoted $\mathcal{G}(2p, \delta, \varepsilon)$.*

The random matrix associated to a random graph is called adjacency matrix and is defined as follows.

Definition 4.6.4 (Adjacency matrix)**.** *Let $G \sim \mathcal{G}(2p, \delta, \varepsilon)$. The adjacency matrix $\mathbf{A} \in \mathcal{S}_{2p}$ of G has entries (a_{ij}) with distribution a_{ij} ~Bernoulli(δ) if i, j belongs to the same community and a_{ij} ~Bernoulli(ε) otherwise. We write $\mathbf{A} \sim \mathcal{G}(2p, \delta, \varepsilon)$ to indicate that $\mathbf{A}$ is the adjacency matrix of the graph G.*

The entry a_{ij} of the adjacency matrix $\mathbf{A}$ indicates if the vertices i and j are connected. To identify which community each vertex belongs, we need the matrix $\mathbb{E}\mathbf{A}$, since it has entries

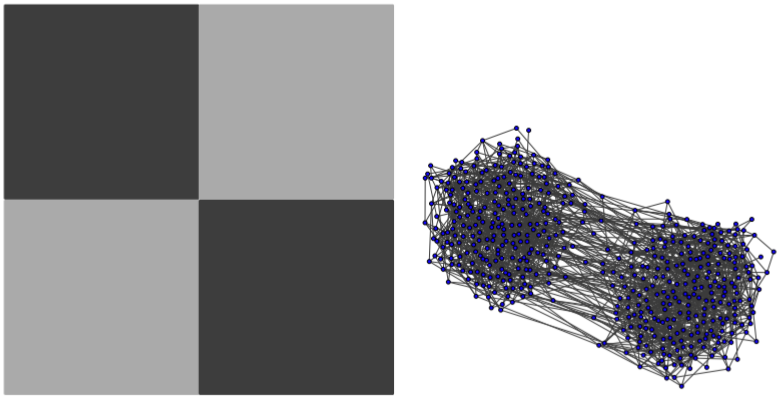


Figure 4.5: Stochastic block model. Image taken from [14].

either δ or ε, and according to our model this identifies the communities. A realization of an stochastic block model is presented in Figure 4.5, where the square represents the matrix $\mathbb{E}\mathbf{A}$ with the rows ordered in communities, i.e., the darker square represents a community different that the lighter square.

As pointed out in [47, p. 88], the second eigenvector of $\mathbb{E}\mathbf{A}$ identifies the communities correctly. This is established in the next proposition.

Proposition 4.6.5. *Let $\mathbf{A} \sim \mathcal{G}(2p, \delta, \varepsilon)$. Define $\boldsymbol{u}_i(\mathbb{E}\mathbf{A})$ as the (non-unitary) eigenvector of $\mathbb{E}\mathbf{A}$ associated with the eigenvalue $\lambda_i(\mathbb{E}\mathbf{A})$. Then, $\mathbb{E}\mathbf{A}$ has rank 2 and non-zero eigenvalues*

$$\lambda_1(\mathbb{E}\mathbf{A}) = \left(\frac{\delta + \varepsilon}{2}\right) p, \quad \lambda_2(\mathbb{E}\mathbf{A}) = \left(\frac{\delta - \varepsilon}{2}\right) p.$$

Furthermore, the entries of $\boldsymbol{u}_2(\mathbb{E}\mathbf{A})$ belongs to $\{-1, 1\}$ and $\boldsymbol{u}_2(\mathbb{E}\mathbf{A})_j = 1$ if an only if the vertex j belongs to the first community.

Proposition 4.6.5 indicates that in order to detect the communities we need to calculate the second eigenvector of $\mathbb{E}\mathbf{A}$ and classify the vertices according to the signs of its entries. Unfortunately, we don't have access to $\mathbb{E}\mathbf{A}$. Instead, we observe a sampled adjacency matrix $\mathbf{A}$.

We are interested in the case where we observe the adjacency matrix plus noise. Let $\mathbf{A} \in \mathcal{S}_{2p}$ be distributed according to $\mathcal{G}(2p, \delta, \varepsilon)$ and define $\mathbf{M} = \mathbb{E}\mathbf{A}$. We observe the noisy random adjacency matrix

$$\mathbf{X} = \mathbf{A} + \mathbf{R}, \tag{4.19}$$

Input: an observed noisy adjacency matrix $\mathbf{X}$ and parameters t, σ^2.
Output: a partition of the rows (or columns) of $\mathbf{X}$ into two communities.

1. Compute the estimator $\widehat{\mathbf{M}} = \theta^{-1}\psi(\theta\mathbf{X})$ where $\theta = t/\sigma^2$.
2. Compute the eigenvector $\boldsymbol{v} = \boldsymbol{v}_2(\mathbf{M})$ corresponding to the eigenvalue $\lambda_2(\mathbf{M})$.
3. Partition the vertices in two communities on the basis of the sign of the coefficients of $\boldsymbol{v}_2(\mathbf{M})$, i.e., if $v_j > 0$ then put vertex j into the first community, otherwise into the second.

Figure 4.6: Spectral clustering algorithm for the stochastic block model.

where $\mathbf{R} \in \mathcal{S}_{2p}$ is a random matrix with $\mathbb{E}\mathbf{R} = \mathbf{0}$. The matrix $\mathbf{R}$ represents the noise of the observations. Note that the distribution of the entries of $\mathbf{R}$ can be heavy-tailed as long as they are centered. Define the robust estimator

$$\widehat{\mathbf{M}} = \frac{\psi(\theta\mathbf{X})}{\theta}, \quad \theta = \frac{t}{\sigma^2},$$

where $\sigma^2 \geq |||\mathbb{E}\mathbf{X}^2|||$. Note that we want to estimate $\mathbf{M} = \mathbb{E}\mathbf{X}$. The algorithm to identify the communities is presented in Figure 4.6 and is refereed to *spectral clustering*.

The next theorem is an adaptation of Theorem 4.5.6 of [47] to the case of general noisy observations. It's important to note that here the sample size can be understood as $n = 1$ or as the number of observed vertices, i.e., $2p$. The concentration bound in the proof of Theorem 4.6.6 is derived from the general result of Theorem 4.1.2 for the case $n = 1$, but taking special parameters we get a bound that depends on p, the number of members in each community.

Theorem 4.6.6 (Spectral clustering for the stochastic block model)**.** *Let $G \sim \mathcal{G}(2p, \delta, \varepsilon)$ with $\min\{\delta, (\delta - \varepsilon)/2\} = \mu > 0$ and adjacency matrix $\mathbf{A}$. Define $\mathbf{X}$ as (4.19). Then, with probability at least $1 - (2p)^{-1}$, the spectral clustering algorithm identifies the communities of G correctly upto $32\sigma^2(2 - 1/p)/\mu^2$ missclassified vertices.*

Proof. Recall the definition of $\boldsymbol{u}_2(\cdot)$ from Proposition 4.6.5 as the (non-unitary) eigenvector associated with $\lambda_2(\cdot)$. Define the set

$$\mathcal{K} = \{j : \operatorname{sign}(u_2(\mathbf{M})_j) \neq \operatorname{sign}(u_2(\widehat{\mathbf{M}})_j)\}$$

and its cardinality $K = |\mathcal{K}|$. We want to prove that

$$\mathbb{P}\left(K \leq \frac{32\sigma^2}{\mu^2}\left(2 - \frac{1}{p}\right)\right) \geq 1 - \frac{1}{2p}. \tag{4.20}$$

To do so we'll make use of the Davis-Kahan Theorem of the previous section.

First, by Theorem 4.1.2, taking $t = 2\sigma\sqrt{\log(2p)}$ and $\theta = t/\sigma^2$, we get the following concentration inequality

$$\mathbb{P}\left(|||\widehat{\mathbf{M}} - \mathbf{M}||| \leq 2\sigma\sqrt{\log(2p)}\right) \geq 1 - \frac{1}{2p}.$$

Note that this is true due to the fact that $\mathbb{E}\mathbf{X} = \mathbf{M}$.

On the other hand,

$$\min\left\{\lambda_1(\mathbf{M}) - \lambda_2(\mathbf{M}), \lambda_2(\mathbf{M})\right\} = \min\left\{\frac{\delta - \varepsilon}{2}, \varepsilon\right\} p = \mu p.$$

Now, using inequality (4.17) of the Davis-Kahan Theorem (Theorem 4.6.1) we get that with probability at least $1 - 1/2p$ there exists a $\rho \in \{-1, 1\}$ such that

$$\|\boldsymbol{v}_2(\mathbf{M}) - \rho\boldsymbol{v}_2(\widehat{\mathbf{M}})\|_2 \leq \frac{2^{5/2}\sigma\sqrt{\log(2p)}}{\mu p},$$

where $\boldsymbol{v}_2(\cdot)$ is the unitary eigenvector associated to $\lambda_2(\cdot)$. Multiplying both sides of the norm inequality by $\sqrt{p}$ we get that with probability at least $1 - 1/2p$ there exists a $\rho \in \{-1, 1\}$ such that

$$\|\boldsymbol{u}_2(\mathbf{M}) - \rho\boldsymbol{u}_2(\widehat{\mathbf{M}})\|_2 \leq \frac{2^{5/2}\sigma}{\mu}\sqrt{\frac{\log(2p)}{p}} \leq \frac{2^{5/2}\sigma}{\mu}\sqrt{2 - \frac{1}{p}}. \tag{4.21}$$

Squaring both sides of (4.21) we get

$$\sum_{j=1}^{p}\left(u_2(\mathbf{M})_j - \rho u_2(\widehat{\mathbf{M}})_j\right)^2 \leq \frac{32\sigma^2}{\mu^2}\left(2 - \frac{1}{p}\right). \tag{4.22}$$

Since $u_2(\mathbf{M})_j \in \{-1, 1\}$ from Proposition 4.6.5 and $\rho \in \{-1, 1\}$, every coefficient j for which the signs $u_2(\mathbf{M})_j$ and $\rho u_2(\widehat{\mathbf{M}})_j$ disagree contributes at least 1 to the left sum in

(4.22). Therefore, the number of disagreeing signs in the left sum of (4.22) must be bounded by

$$\frac{32\sigma^2}{\mu^2}\left(2-\frac{1}{p}\right).$$

More precisely, define the set

$$\mathcal{K}' = \{j : \exists\rho \in \{-1,1\} \text{ such that } \operatorname{sign}(u_2(\mathbf{M})_j) \neq \operatorname{sign}(u_2(\widehat{\mathbf{M}})_j)\}$$

and its cardinality $K' = |\mathcal{K}'|$. Then, whenever there exists a $\rho \in \{-1,1\}$ such that (4.22) is true, we get that

$$\begin{aligned}K' &\leq \sum_{j\in\mathcal{K}'}\left(u_2(\mathbf{M})_j - \rho u_2(\widehat{\mathbf{M}})_j\right)^2\\ &\leq \sum_{j\in\mathcal{K}'}\left(u_2(\mathbf{M})_j - \rho u_2(\widehat{\mathbf{M}})_j\right)^2 + \sum_{j\in\mathcal{K}'^c}\left(u_2(\mathbf{M})_j - \rho u_2(\widehat{\mathbf{M}})_j\right)^2\\ &\leq \frac{32\sigma^2}{\mu^2}\left(2-\frac{1}{p}\right).\end{aligned}$$

Therefore, we just proved that

$$\mathbb{P}\left(K' \leq \frac{32\sigma^2}{\mu^2}\left(2-\frac{1}{p}\right)\right) \geq 1-\frac{1}{2p}.$$

Since $\mathcal{K} \subset \mathcal{K}'$, we get that $K \leq K'$ and inequality (4.20) is proved. □

Summary and observations

In this chapter we presented the main results of the thesis. Specifically, The concentration result of Theorem 4.1.2 guarantee that the general estimator $\mathbf{T} = (n\theta)^{-1}\sum_{j=1}^n \psi(\theta\mathbf{Y}_j)$ is a good robust estimator of $\mathbb{E}\sum_{j=1}^n \mathbf{Y}_j$, in the sense that we assured consistency of $\mathbf{T}$ and only required that

$$\left|\left|\left|\sum_{j=1}^n \mathbb{E}\mathbf{Y}_j^2\right|\right|\right| < \infty, \tag{4.23}$$

without any distributional assumption. Theorem 4.1.2 was shown to be versatile and applicable to different scenarios: we could extended it to the case of arbitrary real matrices in $\mathcal{M}_{p,r}$ and we showed two different applications, namely, PCA and community detection. Also, we presented the Lepskii's method to calculate $\mathbf{T}$ by giving a rough estimation of (4.23). In the following chapters we dive into two more different applications of this robust estimation procedure.

Chapter 5

Robust covariance matrix estimation

Let $\boldsymbol{X} \in \mathbb{R}^p$ be a random vector with $\mathbb{E}\boldsymbol{X} = \boldsymbol{\mu}$ and $\mathrm{Cov}\boldsymbol{X} = \boldsymbol{\Sigma}$. As in Chapter 3, we want to estimate $\boldsymbol{\Sigma}$ through the iid copies $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ of $\boldsymbol{X}$. One way is to use the estimator $\widehat{\boldsymbol{\Sigma}}$. We saw in Chapter 3 that in order to obtain good concentration results for $\widehat{\boldsymbol{\Sigma}}$ we needed to assume that it was calculated from a Gaussian or sub-Gaussian ensemble, or that $\|\boldsymbol{X}\|$ has some upper bound. Instead, in this chapter we will use techniques similar to the ones developed in the previous chapter.

In Chapter 4 we defined the truncation operator

$$\psi_\tau(x) = (|x| \wedge \tau)\,\mathrm{sign}(x), \quad \tau > 0,$$

which satisfies the inequality

$$-\tau \log\left(1 - \tau^{-1}x + \tau^{-2}x^2\right) \le \psi_\tau(x) \le \tau \log\left(1 + \tau^{-1}x + \tau^{-2}x^2\right). \tag{5.1}$$

Figure 5.1 shows the function ψ_τ for different values of τ. We can see that the function truncates values of x that are away from the origin at the value τ or $-\tau$. The bigger the value of τ, the less truncation the function performs. For this reason we call ψ_τ the *truncation operator* and τ the *robustification parameter*. Due to the ease of interpretation of ψ_τ we will use this function throughout this chapter.

Suppose for the moment that $\boldsymbol{\mu} = \boldsymbol{0}$, so a natural estimator of $\boldsymbol{\Sigma}$ is the sum of iid symmetric matrices

$$\widehat{\boldsymbol{\Sigma}}_0 = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}\boldsymbol{X}_i^\intercal.$$

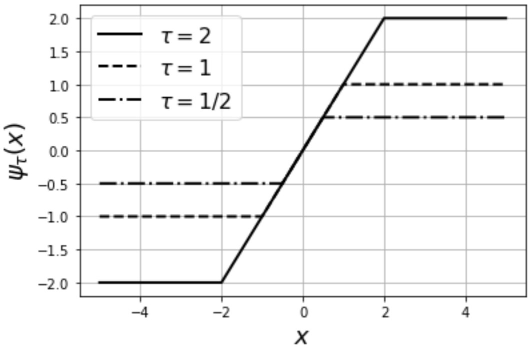


Figure 5.1: Function ψ_τ for different values of τ.

Then, as we did in Chapter 4, to get a robust estimation of $\mathbf{\Sigma}$ it is intuitive to think in the *truncated* version

$$\widehat{\mathbf{\Sigma}}_0^\tau = \frac{1}{n}\sum_{i=1}^{n} \psi_\tau \left(\boldsymbol{X}\boldsymbol{X}_i^\intercal \right).$$

By inequalities (5.1) we can use the same techniques of Chapter 4 to obtain a concentration inequality for $\widehat{\mathbf{\Sigma}}_0^\tau$. More specifically, by Minsker's concentration inequality of Theorem 4.3.2 one can show that for $t \geq 0$

$$\mathbb{P}\left(\left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \widehat{\mathbf{\Sigma}}_0^\tau - \mathbf{\Sigma} \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert \geq t \right) \leq 2p \exp\left(\frac{-nt}{\sigma^2} \right),$$

where $\sigma^2 \geq \left\vert\kern-0.25ex\left\vert\kern-0.25ex\left\vert \mathbb{E}\left[\|\boldsymbol{X}\|_2^2 \boldsymbol{X}\boldsymbol{X}^\intercal \right] \right\vert\kern-0.25ex\right\vert\kern-0.25ex\right\vert$. Note that the factor $1/2$ inside the exponential is missing do to the weaker inequalities that ψ_τ satisfies.

In order to obtain a more general result, suppose that $\boldsymbol{\mu} \neq \mathbf{0}$. Generally, $\widehat{\mathbf{\Sigma}}_0^\tau$ can be thought as a robust estimator of the Gram matrix $\mathbb{E}\boldsymbol{X}\boldsymbol{X}^\intercal$, and with an additional robust estimator of $\boldsymbol{\mu}$ (see for example [28]) we can obtain a robust estimator of $\mathbf{\Sigma}$ since $\mathbf{\Sigma} = \mathbb{E}\boldsymbol{X}\boldsymbol{X}^\intercal - \boldsymbol{\mu}\boldsymbol{\mu}^\intercal$. We want to emphasize that the estimator

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\intercal$$

has no independent addends, so we can not use directly the methods presented in Chapter 4. In this sense, instead of proposing a robust estimator of $\boldsymbol{\mu}$, we'll follow the procedure of [21] to circumvent this complication.

Let $N = \binom{n}{2} = n(n-1)/2$ and define the identically distributed paired data $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_N$ as

$$\begin{aligned} \boldsymbol{Y}_1 &= \boldsymbol{X}_1 - \boldsymbol{X}_2, \\ \boldsymbol{Y}_2 &= \boldsymbol{X}_1 - \boldsymbol{X}_3, \\ &\vdots \\ \boldsymbol{Y}_N &= \boldsymbol{X}_{n-1} - \boldsymbol{X}_n. \end{aligned}$$

Since $\mathbb{E}\boldsymbol{Y}_1 = \boldsymbol{0}$ and $\mathbb{E}\boldsymbol{X}\boldsymbol{X}^\intercal = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\intercal$, we have that

$$\begin{aligned} \mathrm{Cov}\boldsymbol{Y}_1 = \mathbb{E}\left[(\boldsymbol{X}_1 - \boldsymbol{X}_2)(\boldsymbol{X}_1 - \boldsymbol{X}_2)^\intercal\right] &= \mathbb{E}\left[\boldsymbol{X}_1\boldsymbol{X}_1^\intercal - \boldsymbol{X}_1\boldsymbol{X}_2^\intercal - \boldsymbol{X}_2\boldsymbol{X}_1^\intercal + \boldsymbol{X}_2\boldsymbol{X}_2^\intercal\right] \\ &= \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\intercal - 2\boldsymbol{\mu}\boldsymbol{\mu}^\intercal + \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\intercal \\ &= 2\boldsymbol{\Sigma}. \end{aligned} \tag{5.2}$$

So the random vectors $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_N$ are identically distributed with mean $\boldsymbol{0}$ and covariance matrix $2\boldsymbol{\Sigma}$, but they are not independent. It is a classical method in statistics to replace $\widehat{\boldsymbol{\Sigma}}$, which is not biased[1], for the unbiased version or *U-statistic* version[2]

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2}\boldsymbol{Y}_i\boldsymbol{Y}_i^\intercal.$$

See [39, Chapter 5] for more information on U-statistics. It seems that we win very little with this construction because the addends are still dependent. But we'll see later that this representation can be further modified to have a sum of sums of independent random matrices. Additionally, we wont need to estimate directly $\boldsymbol{\mu}$, so any manipulation of $\widehat{\boldsymbol{\Sigma}}$ will be straightforward.

The *spectrum-wise* truncated estimator of $\boldsymbol{\Sigma}$ proposed in [21] is

$$\widehat{\boldsymbol{\Sigma}}^\tau = \frac{1}{N}\sum_{i=1}^{N} \psi_\tau\left(\boldsymbol{Y}_i\boldsymbol{Y}_i^\intercal/2\right). \tag{5.3}$$

Note that $\boldsymbol{Y}_i\boldsymbol{Y}_i^\intercal/2$ its symmetric and has rank one, and that

$$\left(\frac{1}{2}\boldsymbol{Y}_i\boldsymbol{Y}_i^\intercal\right)\frac{\boldsymbol{Y}_i}{\|\boldsymbol{Y}_i\|_2} = \frac{\|\boldsymbol{Y}_i\|_2^2}{2}\frac{\boldsymbol{Y}_i}{\|\boldsymbol{Y}_i\|_2}.$$

[1] One can verify in Appendix C that $(n/(n-1))\widehat{\boldsymbol{\Sigma}}$ is unbiased.

[2] It is more common to define $\boldsymbol{Y}_{ij} = \boldsymbol{X}_i - \boldsymbol{X}_j$ for $i \neq j$ and write $\hat{\Sigma} = \frac{1}{2N}\sum_{i\neq j}\boldsymbol{Y}_i\boldsymbol{Y}_j^\intercal$. But we'll maintain the stated definition for notational convenience.

So it has eigenvalue $\|\boldsymbol{Y}_i\|_2^2/2$ with unitary eigenvector $\boldsymbol{Y}_i/\|\boldsymbol{Y}_i\|_2$. Then, by Example 2.1.2, we can write

$$\widehat{\boldsymbol{\Sigma}}^{\tau} = \frac{1}{N}\sum_{i=1}^{N}\psi_\tau\left(\frac{\|\boldsymbol{Y}_i\|_2^2}{2}\right)\frac{\boldsymbol{Y}_i\boldsymbol{Y}_i^\intercal}{\|\boldsymbol{Y}_i\|_2^2},$$

so the estimator $\widehat{\boldsymbol{\Sigma}}^{\tau}$ is very easy to calculate from the sample since we don't need to spend computational power in any spectral decomposition.

In what follows, we'll develop the theory necessary to obtain concentration inequalities for $\widehat{\boldsymbol{\Sigma}}^{\tau}$. Section 5.1 is devoted to a general framework for concentration results of sum of dependent matrices. Section 5.2 derives the main theorem of this chapter for the estimation of the covariance matrix. Section 5.3 is dedicated to present data dependent methods to determine τ. Finally, Section 5.4 shows a simulation study of the robust procedure from this chapter.

5.1 Concentration for the sum of dependent matrices

Following the procedure of [17] we define the random matrix $\mathbf{X} \in \mathcal{S}_p$ as the convex combination

$$\mathbf{X} = q_1\mathbf{X}_1 + \cdots + q_M\mathbf{X}_M, \tag{5.4}$$

where $\sum_{i=1}^{M} q_i = 1$, $q_i \geq 0$, and $\mathbf{X}_1, ..., \mathbf{X}_M$ are each one a sum of independent $p \times p$ symmetric random matrices, i.e., for $i = 1, ..., M$ define

$$\mathbf{X}_i = \mathbf{X}_{i,1}^* + \cdots + \mathbf{X}_{i,m}^*,$$

where $\mathbf{X}_{i,1}^*, ..., \mathbf{X}_{i,m}^* \in \mathcal{S}_p$ are independent random matrices. We insist that the definition of $\mathbf{X}$ do not requires $\mathbf{X}_1, ..., \mathbf{X}_M$ to be independent, but it does asks for $\mathbf{X}_{i,1}^*, ..., \mathbf{X}_{i,m}^*$ to be independent for each $i = 1, ..., n$. In particular, we allow for $\mathbf{X}_{i,k}^*$ and $\mathbf{X}_{j,\ell}^*$, $i \neq j$, to be dependent, which provokes a possible dependence structure for the matrices $\mathbf{X}_1, ..., \mathbf{X}_M$. For example, if $\boldsymbol{S}_1, \boldsymbol{S}_2 \in \mathbb{R}^p$ are independent random vectors, we can define $\mathbf{X}$ to be

$$\begin{aligned}\mathbf{X} &= \boldsymbol{S}_1\boldsymbol{S}_1^\intercal + \boldsymbol{S}_2\boldsymbol{S}_2^\intercal \\ &= \frac{1}{2}\left(\boldsymbol{S}_1\boldsymbol{S}_1^\intercal + \boldsymbol{S}_2\boldsymbol{S}_2^\intercal\right) + \frac{1}{2}\left(\boldsymbol{S}_1\boldsymbol{S}_1^\intercal + \boldsymbol{S}_2\boldsymbol{S}_2^\intercal\right) \\ &=: \frac{1}{2}\mathbf{X}_1 + \frac{1}{2}\mathbf{X}_2.\end{aligned}$$

Clearly, the matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ are dependent, but the addends of each one are independent.

The next theorem is analogous to Theorem 4.1.2. The difference is that the sum is not of iid matrices.

Theorem 5.1.1. *Let* $\mathbf{X}$ *be defined as* (5.4) *and* $\mathbf{H} \in \mathcal{S}_p$ *a fixed matrix. Also, for each* $i = 1, ..., M$ *let* $\mathbf{W}_{i,1}(\theta), ..., \mathbf{W}_{i,m}(\theta) \in \mathcal{S}_p$ *be independent random matrices such that* $\theta \mathbf{X}^*_{i,j} \preceq \mathbf{W}_{i,j}(\theta)$*, for all* i, j *and* $\theta > 0$*. Then, for every* $\theta > 0$ *and* $t \geq 0$*,*

$$\mathbb{P}\left(\lambda_1\left(\mathbf{X}+\mathbf{H}\right) \geq t\right) \leq e^{-\theta t} \sum_{i=1}^{M} q_i \mathrm{tr}\, \exp\left(\sum_{j=1}^{m} \log \mathbb{E} e^{\mathbf{W}_{i,j}} + \theta \mathbf{H}\right). \tag{5.5}$$

Furthermore, if $\mathbf{H} = \sum_{j=1}^{m} \mathbf{H}_j$ *with* $\mathbf{H}_j \in \mathcal{S}_p$ *fixed and* $\mathbf{M}_{i,j}(\theta) \in \mathcal{S}_p$*,* $i = 1, ..., M$*,* $j = 1, ..., m$*, are fixed matrices such that*

$$\log \mathbb{E} e^{\mathbf{W}_{i,j}(\theta)} + \theta \mathbf{H}_j \preceq \mathbf{M}_{i,j}(\theta), \quad \forall i, j,\, \theta > 0,$$

and the constant $\nu > 0$ *satisfies* $\nu^2 \geq \left\| \sum_{j=1}^{m} \mathbf{M}_{i,j}(\theta) \right\|$ *for all* i *and* $\theta > 0$*, then*

$$\mathbb{P}\left(\lambda_1\left(\mathbf{X}+\mathbf{H}\right) \geq t\right) \leq p \exp\left(-t\theta + \nu^2\right).$$

Proof. Using the convexity of the two mappings $\mathbf{A} \mapsto \lambda_1(\mathbf{A})$, $\mathbf{A} \in \mathcal{S}_p$ (Lemma A.2.8) and $x \mapsto e^x$, $x \in \mathbb{R}$, we obtain that

$$\begin{aligned}
\exp\left(\lambda_1(\mathbf{X}+\mathbf{H})\right) &= \exp\left(\lambda_1\left(\sum_{i=1}^{M} q_i(\mathbf{X}_i + \mathbf{H})\right)\right) \\
&\leq \exp\left(\sum_{i=1}^{M} q_i \lambda_1(\mathbf{X}_i + \mathbf{H})\right) \\
&\leq \sum_{i=1}^{M} q_i \exp\left(\lambda_1(\mathbf{X}_i + \mathbf{H})\right).
\end{aligned} \tag{5.6}$$

Now by Markov inequality, the previous inequality (5.6) and Lemma 2.1.7, we get that

$$\begin{aligned}
\mathbb{P}(\lambda_1(\mathbf{X}+\mathbf{H}) \geq t) &\leq e^{-\theta t} \mathbb{E} \exp\left(\lambda_1(\theta \mathbf{X} + \theta \mathbf{H})\right) \\
&\leq e^{-\theta t} \sum_{i=1}^{M} q_i \mathbb{E} \exp\left(\lambda_1(\theta \mathbf{X}_i + \theta \mathbf{H})\right)
\end{aligned}$$

$$\leq e^{-\theta t} \sum_{i=1}^{M} q_i \mathbb{E}\mathrm{tr}\, \exp\left(\theta \mathbf{X}_i + \theta \mathbf{H}\right).$$

Then, by Lemma 4.1.1 we have that

$$e^{-\theta t} \sum_{i=1}^{M} q_i \mathbb{E}\mathrm{tr}\, \exp\left\{\theta \mathbf{X}_i + \theta \mathbf{H}\right\} \leq e^{-\theta t} \sum_{i=1}^{M} q_i \mathrm{tr}\, \exp\left\{\sum_{j=1}^{m} \log \mathbb{E} e^{\mathbf{W}_{i,j}(\theta)} + \theta \mathbf{H}\right\}.$$

This proves the first inequality (5.5). Finally, the hypothesis $\log \mathbb{E} e^{\mathbf{W}_{i,j}(\theta)} + \theta \mathbf{H}_j \preceq \mathbf{M}_{i,j}(\theta)$ implies that for every $i = 1, ..., M$

$$\sum_{j=1}^{m} \log \mathbb{E} e^{\mathbf{W}_{i,j}(\theta)} + \theta \mathbf{H} \preceq \sum_{j=1}^{m} \mathbf{M}_{i,j}(\theta)$$

and

$$\mathrm{tr}\, \exp\left\{\sum_{j=1}^{m} \log \mathbb{E} e^{\mathbf{W}_{i,j}(\theta)} + \theta \mathbf{H}\right\} \leq \mathrm{tr}\, \exp\left\{\sum_{j=1}^{m} \mathbf{M}_{i,j}(\theta)\right\}.$$

Using the inequality $\mathrm{tr}\, e^{\mathbf{A}} \leq p e^{\|\mathbf{A}\|}$, $\mathbf{A} \in \mathcal{S}_p$, of Lemma 2.1.7, we arrive at

$$\begin{aligned}
\mathbb{P}(\lambda_1(\mathbf{X} + \mathbf{H}) \geq t) &\leq e^{-\theta t} \sum_{i=1}^{M} q_i \mathrm{tr}\, \exp\left\{\sum_{j=1}^{m} \log \mathbb{E} e^{\mathbf{W}_{i,j}(\theta)} + \theta \mathbf{H}\right\} \\
&\leq e^{-\theta t} \sum_{i=1}^{M} q_i \mathrm{tr}\, \exp\left\{\sum_{j=1}^{m} \mathbf{M}_{i,j}(\theta)\right\} \\
&\leq e^{-\theta t} \sum_{i=1}^{M} q_i p \exp\left\{\left\|\sum_{j=1}^{m} \mathbf{M}_{i,j}(\theta)\right\|\right\} \\
&\leq e^{-\theta t} \sum_{i=1}^{M} q_i p \exp\left\{\nu^2\right\} \\
&\leq p \exp\left\{-\theta t + \nu^2\right\},
\end{aligned}$$

which ends the proof. □

In the next section we'll see how to decompose the estimator $\widehat{\mathbf{\Sigma}}^{\tau}$ defined in (5.3) in a sum of the form (5.4). This representation, with the aid of the truncation operator ψ_τ, will give us the desired concentration guarantee for $\widehat{\mathbf{\Sigma}}^{\tau}$.

5.2 The Hoeffding decomposition

Remember that $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid copies of $\boldsymbol{X} \in \mathbb{R}^p$ with $\mathbb{E}\boldsymbol{X} = \boldsymbol{\mu}$ and $\mathrm{Cov}\boldsymbol{X} = \boldsymbol{\Sigma}$. The sample $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_N$, $N = n(n-1)/2$, is defined as

$$\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y}_N\} = \{\boldsymbol{X}_1 - \boldsymbol{X}_2, \boldsymbol{X}_1 - \boldsymbol{X}_3, ..., \boldsymbol{X}_n - \boldsymbol{X}_{n-1}\} \tag{5.7}$$

In this section we want to see that the estimator

$$\widehat{\boldsymbol{\Sigma}}^{\tau} = \frac{1}{N} \sum_{i=1}^{N} \psi_\tau(\boldsymbol{Y}_i \boldsymbol{Y}_i^\intercal / 2)$$

can be represented as in (5.4). To see this define $h : \mathbb{R}^p \times \mathbb{R}^p \to \mathcal{M}_p$ as

$$h(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}(\boldsymbol{x} - \boldsymbol{y})(\boldsymbol{x} - \boldsymbol{y})^\intercal$$

Note that $\mathbb{E}h(\boldsymbol{X}_i, \boldsymbol{X}_j) = \boldsymbol{\Sigma}$, $i \neq j$ (see equation (5.2)). From the iid sample $\boldsymbol{X}_1, ...\boldsymbol{X}_n$ we define the random matrices $\mathbf{Z}_{i,j}$, $1 \leq i < j \leq n$, as

$$\mathbf{Z}_{i,j} = \psi_\tau\left(h(\boldsymbol{X}_i, \boldsymbol{X}_j)\right), \quad \tau > 0.$$

Then we rewrite $\widehat{\boldsymbol{\Sigma}}^{\tau}$ in the following way:

$$\widehat{\boldsymbol{\Sigma}}^{\tau} = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \mathbf{Z}_{i,j}. \tag{5.8}$$

Now, let $\mathcal{P}$ the set of permutations of $\{1, ..., n\}$. For $\pi \in \mathcal{P}$ define the matrices $\mathbf{Z}_{\pi,j}$, $j = 1, ..., m$, $m = \lfloor n/2 \rfloor$, as

$$\mathbf{Z}_{\pi,j} = \mathbf{Z}_{\pi(2j-1),\pi(2j)}.$$

Finally, defining

$$\mathbf{V}_\pi = \frac{\mathbf{Z}_{\pi,1} + \cdots + \mathbf{Z}_{\pi,m}}{m}, \quad \pi \in \mathcal{P},$$

we can write

$$\widehat{\boldsymbol{\Sigma}}^{\tau} = \sum_{\pi \in \mathcal{P}} \frac{1}{n!} \mathbf{V}_\pi. \tag{5.9}$$

$$\pi = \{i_1, ..., i_n\} \in \mathcal{P}$$

$$\Downarrow$$

$$\underbrace{\underbrace{\boldsymbol{X}_{i_1}, \boldsymbol{X}_{i_2}}_{\mathbf{Z}_{\pi,1}}, \underbrace{\boldsymbol{X}_{i_3}, \boldsymbol{X}_{i_4}}_{\mathbf{Z}_{\pi,2}}, ..., \underbrace{\boldsymbol{X}_{i_{n-1}}, \boldsymbol{X}_{i_n}}_{\mathbf{Z}_{\pi,m}}}_{\mathbf{V}_\pi = m^{-1}\sum_{j=1}^{m} \mathbf{Z}_{\pi,j}}$$

$$\Downarrow$$

$$\widehat{\mathbf{\Sigma}}^{\tau} = \frac{1}{n!}\sum_{\pi\in\mathcal{P}} \mathbf{V}_\pi$$

Figure 5.2: How to compute the Hoeffding decomposition of $\widehat{\mathbf{\Sigma}}^{\tau}$.

Figure 5.2 presents a description on how to construct this decomposition. The sums (5.8) and (5.9) are indeed equal since $h(\boldsymbol{x},\boldsymbol{y}) = h(\boldsymbol{y},\boldsymbol{x})$ imply that $\mathbf{Z}_{i,j} = \mathbf{Z}_{j,i}$, so in the sum (5.9) each $\mathbf{Z}_{i,j}$ is repeated $2(n-2)!m$ times[3]. Then,

$$\sum_{\pi\in\mathcal{P}} \frac{1}{n!}\mathbf{V}_\pi = \frac{1}{n!m}\sum_{1\le i<j\le n} 2(n-2)!m\mathbf{Z}_{i,j} = \frac{1}{\binom{n}{2}}\sum_{1\le i<j\le n} \mathbf{Z}_{i,j}.$$

We refer to (5.9) as the *Hoeffding decomposition* of $\widehat{\mathbf{\Sigma}}^{\tau}$ and we'll use it just for its theoretical advantages. This is different from the classic definition of Hoeffding decomposition that is made under projections of the function h (see for example [12].) As we mentioned in the beginning of this section, we have been able to write $\widehat{\mathbf{\Sigma}}^{\tau}$ in the form (5.4), so in the next result we'll prove a concentration bound based on Theorem 5.1.1.

Theorem 5.2.1. *Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n \in \mathbb{R}^p$ be iid random vectors with*

$$v^2 = \frac{1}{4}\left|\left|\left|\mathbb{E}\left((\boldsymbol{X}_1 - \boldsymbol{X}_2)(\boldsymbol{X}_1 - \boldsymbol{X}_2)^{\intercal}\right)^2\right|\right|\right| < \infty.$$

For $\tau > 0$ define $\widehat{\mathbf{\Sigma}}^{\tau}$ as (5.3). Then, for any $s \ge 0$,

$$|||\widehat{\mathbf{\Sigma}}^{\tau} - \mathbf{\Sigma}||| \ge \frac{s}{\sqrt{m}}$$

[3] There are $2(n-2)!$ permutations in which element i is next to element j in the first two places and m places to put this elements together.

with probability at most

$$2p \exp\left(-\frac{s\sqrt{m}}{\tau} + \frac{mv^2}{\tau^2}\right), \tag{5.10}$$

where $m = \lfloor n/2 \rfloor$.

Proof. Analogously to the proof of Theorem 4.3.2, we'll prove that the two probabilities

$$\mathbb{P}\left(\lambda_1\left(\widehat{\mathbf{\Sigma}}^{\tau} - \mathbf{\Sigma}\right) \geq \frac{s}{\sqrt{m}}\right) \tag{5.11}$$

and

$$\mathbb{P}\left(-\lambda_p\left(\widehat{\mathbf{\Sigma}}^{\tau} - \mathbf{\Sigma}\right) \geq \frac{s}{\sqrt{m}}\right), \tag{5.12}$$

have the same bound equal to

$$p \exp\left(-\frac{s\sqrt{m}}{\tau} + \frac{mv^2}{\tau^2}\right).$$

Once this is done, the result follows from Proposition 2.3.2.

First, for (5.11) we have that for any $s \geq 0$

$$\mathbb{P}\left(\lambda_1\left(\widehat{\mathbf{\Sigma}}^{\tau} - \mathbf{\Sigma}\right) \geq \frac{s}{\sqrt{m}}\right) = \mathbb{P}\left(\lambda_1\left(m\widehat{\mathbf{\Sigma}}^{\tau} - m\mathbf{\Sigma}\right) \geq s\sqrt{m}\right),$$

Using the Hoeffding decomposition of $\widehat{\mathbf{\Sigma}}^{\tau}$ given in (5.9), note that

$$\begin{aligned} m\widehat{\mathbf{\Sigma}}^{\tau} &= \frac{1}{n!}\sum_{\pi\in\mathcal{P}} m\mathbf{V}_{\pi} \\ &= \sum_{\pi\in\mathcal{P}} \frac{1}{n!}\sum_{j=1}^{m} \mathbf{Z}_{\pi,j}. \end{aligned}$$

So $m\widehat{\mathbf{\Sigma}}^{\tau}$ is of the form (5.4) with $q_{\pi} = 1/n!$

Now, following the notation of Theorem 5.1.1 and the Hoeffding decomposition $\widehat{\mathbf{\Sigma}}^{\tau}$, define $\theta = \tau^{-1}$ and

$$\mathbf{X}^{*}_{\pi,j} = \mathbf{Z}_{\pi,j}, \quad \mathbf{H}_j = -\mathbf{\Sigma}, \quad \mathbf{H} = -m\mathbf{\Sigma},$$

$$\mathbf{W}_{\pi,j}(\theta) = \log\left(\mathbf{I} + \tau^{-1}\mathbf{U}_{\pi,j} + \tau^{-2}\mathbf{U}_{\pi,j}^2\right), \quad \mathbf{M}_{\pi,j}(\theta) = \tau^{-2}\mathbb{E}\mathbf{U}_{\pi,j}^2,$$

where $\mathbf{U}_{\pi,j} = h(\boldsymbol{X}_{\pi(2j-1)}, \boldsymbol{X}_{\pi(2j)}) = (\boldsymbol{X}_{\pi(2j-1)} - \boldsymbol{X}_{\pi(2j)})(\boldsymbol{X}_{\pi(2j-1)} - \boldsymbol{X}_{\pi(2j)})^\intercal/2$.

Note that unlike Theorem 5.1.1 we are using the sub index $\pi \in \mathcal{P}$ instead of $i \in \{1, ..., M\}$. This doesn't change anything from the conclusion of Theorem 5.1.1 since it is only a notation issue.

From Remark 4.3.1 we derive that

$$\theta\mathbf{X}_{\pi,j}^* = \psi_1\left(\tau^{-1}\mathbf{U}_{\pi,j}\right) \preceq \log\left(\mathbf{I} + \tau^{-1}\mathbf{U}_{\pi,j} + \tau^{-2}\mathbf{U}_{\pi,j}^2\right) = \mathbf{W}_{\pi,j}(\theta)$$

for $\pi \in \mathcal{P}$ and $j = 1, ..., m$. Also, by Lemma 4.3.1 (taking $\mathbf{A} = \tau^{-1}\mathbf{\Sigma} + \tau^{-2}\mathbb{E}\mathbf{U}_{\pi,j}^2$) and rearranging terms we get that

$$\begin{aligned}\log\mathbb{E}e^{\mathbf{W}_{\pi,j}(\theta)} + \theta\mathbf{H}_j &= \log\left(\mathbf{I} + \tau^{-1}\mathbf{\Sigma} + \tau^{-2}\mathbb{E}\mathbf{U}_{\pi,j}^2\right) - \tau^{-1}\mathbf{\Sigma} \\ &\preceq \tau^{-2}\mathbb{E}\mathbf{U}_{\pi,j}^2 \\ &= \mathbf{M}_{\pi,j}(\theta).\end{aligned}$$

Since for all $\pi \in \mathcal{P}$ it is true that

$$\begin{aligned}\left|\left|\left|\sum_{j=1}^m \mathbf{M}_{\pi,j}(\theta)\right|\right|\right| &= \tau^{-2}\left|\left|\left|\sum_{j=1}^m \mathbb{E}\mathbf{U}_{\pi,j}^2\right|\right|\right| \\ &= \frac{\tau^{-2}}{4}\left|\left|\left|\sum_{j=1}^m \mathbb{E}\left((\boldsymbol{X}_1 - \boldsymbol{X}_2)(\boldsymbol{X}_1 - \boldsymbol{X}_2)^\intercal\right)^2\right|\right|\right| \\ &= m\tau^{-2}v^2,\end{aligned}$$

then, by Theorem 5.1.1, we conclude that

$$\begin{aligned}\mathbb{P}\left(\lambda_1\left(\widehat{\mathbf{\Sigma}}^\tau - \mathbf{\Sigma}^\tau\right) \geq \frac{s}{m}\right) &= \mathbb{P}\left(\lambda_1\left(m\widehat{\mathbf{\Sigma}}^\tau - m\mathbf{\Sigma}^\tau\right) \geq s\sqrt{m}\right) \\ &\leq p\exp\left(-\frac{s\sqrt{m}}{\tau} + \frac{mv^2}{\tau^2}\right).\end{aligned}$$

Finally, for probability (5.12) we proceed in the same way by redefining the random matrices of Theorem 5.1.1:

$$\begin{aligned}&\mathbf{X}_{\pi,j}^* = -\mathbf{Z}_{\pi,j}, \quad \mathbf{H}_j = \mathbf{\Sigma}, \quad \mathbf{H} = m\mathbf{\Sigma}, \\ &\mathbf{W}_{\pi,j}(\theta) = \log\left(\mathbf{I} - \tau^{-1}\mathbf{U}_{\pi,j} + \tau^{-2}\mathbf{U}_{\pi,j}^2\right), \quad \mathbf{M}_{\pi,j}(\theta) = \tau^{-2}\mathbb{E}\mathbf{U}_{\pi,j}^2.\end{aligned}$$

Since

$$-l - og\left(\mathbf{I} - \tau^{-1}\mathbf{U}_{\pi,j} + \tau^{-2}\mathbf{U}_{\pi,j}^2\right) \preceq \mathbf{Z}_{\pi,j},$$

the procedure is analogous to the one done for (5.11). □

The result of Theorem 5.2.1 is remarkable since we've been able to obtain a concentration inequality for the covariance matrix estimation problem by just assuming that

$$\left\|\mathbb{E}\left((\boldsymbol{X}_1 - \boldsymbol{X}_2)(\boldsymbol{X}_1 - \boldsymbol{X}_2)^\intercal\right)^2\right\| < \infty.$$

This should be compared with the results presented in Chapter 3, in which we placed distributional assumptions on $\boldsymbol{X}$. Additionally, this concentration bound is almost identical to the one obtained in Theorem 4.3.2 with the main differences being n substituted by $m = \lfloor n/2 \rfloor$ and the 1/2 factor inside the exponential.

As we did in Chapter 4, optimizing the probability bound (5.10) over $\tau > 0$ we get that it is minimized in $\tilde{\tau} = 2\sqrt{m}v^2/s$. Taking $\tau = \tilde{\tau}$ and $s = 2v\sqrt{\log(2p) + t}$, $t \geq 0$, we get that

$$\tau = v\sqrt{\frac{m}{\log(2p) + t}}. \tag{5.13}$$

and

$$\left\|\widehat{\boldsymbol{\Sigma}}^\tau - \boldsymbol{\Sigma}\right\| \geq 2v\sqrt{\frac{\log(2p) + t}{m}} \tag{5.14}$$

with probability at most e^{-t}. The bound (5.14) increases slowly with the dimension p and decreases with the sample size n. It could be only seriously affected by the unknown parameter v. Note that we are presenting an ideal choice of the hyperparameter τ in equation (5.13) and this choice depends directly on v. In the next section we show an heuristic method to determine τ based on this observations.

5.3 How to choose the hyperparameter

In this section we present three different methods to choose the hyperparameter τ. The first has a theoretical guarantee whereas the other two doesn't. The objective is to show that there exist different criteria to overcome this difficulty and the selection of a method can depend in the context of the application.

5.3.1 Lepskii's method

One way to choose τ for $\widehat{\mathbf{\Sigma}}^{\tau}$ is with the Lepskii method presented in Chapter 4. Indeed, by defining $\theta = \sqrt{2}/\tau$ and $t = s/\sqrt{2}$ we get from Theorem 5.2.1 that

$$\mathbb{P}\left(\|\widehat{\mathbf{\Sigma}}^{\sqrt{2}/\theta} - \mathbf{\Sigma}\| \geq \frac{\sqrt{2}t}{\sqrt{m}}\right) \leq 2p\exp\left(-t\sqrt{m}\theta + \frac{mv^2\theta}{2}\right),$$

which is the same concentration bound obtained in Theorem 4.3.2 with n substituted by $m = \lfloor n/2 \rfloor$ and σ_n^2 substituted by mv^2, and the additional constant $\sqrt{2}$ in the norm lower bound. Then, by Remark 4.3.2, taking $\theta = \sqrt{\frac{2t}{m}\frac{1}{z}}$, with $z > 0$ such that $z > v$, and the mapping $t \mapsto z\sqrt{2t}$ we get that

$$\mathbb{P}\left(\|\widehat{\mathbf{\Sigma}}^{\sqrt{2}/\theta} - \mathbf{\Sigma}\| \geq z2\sqrt{\frac{t}{m}}\right) \leq 2pe^{-t}$$

or

$$\mathbb{P}\left(\|\widehat{\mathbf{\Sigma}}^{\sqrt{2}/\theta} - \mathbf{\Sigma}\| \geq z\sqrt{\frac{2t}{m}}\right) \leq 2pe^{-t/2},$$

which is almost the same concentration bound of Remark 4.5.1, with the only difference being the term $1/2$ in the exponential. Therefore, the same procedure of Lepskii from the previous chapter can be applied to $\widehat{\mathbf{\Sigma}}^{\tau}$. We define the robust estimator $\widehat{\mathbf{\Sigma}}^{*}$ as

$$\widehat{\mathbf{\Sigma}}^{*} = \widehat{\mathbf{\Sigma}}^{\tau_{j*}}, \quad \tau_{j_*} = \frac{\sqrt{2}}{\theta_{j_*}},$$

where $\theta_j = \sqrt{\frac{2t}{m}\frac{1}{z_j}}$ and $z_j = \gamma^j v_{\min}$, $\gamma \in (1, 2)$, and j_* is defined in (4.14) with $\mathbf{T}_j$ replaced with $\widehat{\mathbf{\Sigma}}^{\tau_j}$. A resume of the algorithm is shown in Figure 5.3. We present the next result as a corollary of Theorem 4.5.1 without proof since the procedure is the same.

Corollary 5.3.1. *Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ be an iid sample with same assumptions of Theorem 5.2.1. Then, for any $\epsilon > 0$,*

$$\|\widehat{\mathbf{\Sigma}}^{*} - \mathbf{\Sigma}\| \geq (3 + \epsilon)v\sqrt{\frac{2t}{m}}$$

with probability at most

$$2p\frac{\log(\gamma v_{\max}/v_{\min})}{\log(\gamma)}e^{-t/2}.$$

```
**Input**: sample $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ and parameters $v_{\max}, v_{\min}, \gamma, t$.
**Output**: robust estimator $\widehat{\boldsymbol{\Sigma}}^*$.

1. Create the sample $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_N$.
2. Calculate the cardinality $c = \lfloor 1 + \log(v_{\max}/v_{\min})/\log(\gamma) \rfloor$ of the set $|\mathcal{J}|$ and
   the vector $\boldsymbol{z} = (z_j)_{j=0,...,c-1}$, $z_j = \gamma^j v_{\min}$.
3. For $j = 0, ..., c-1$, choose the first $j$ such that
       $\|\widehat{\boldsymbol{\Sigma}}^{\tau_k} - \widehat{\boldsymbol{\Sigma}}^{\tau_j}\| \leq 2 z_k \sqrt{\frac{2t}{m}},$
   for all $k = j+1, ..., c-1$, where $\tau_j = \sqrt{2}/\theta_j$ and $\theta_j = \sqrt{\frac{2t}{m}\frac{1}{z_j}}$. Define this
   minimum $j$ as $j_*$. If the minimum doesn't exist, define $j_* = j_{c-1}$.
4. Return the matrix $\widehat{\boldsymbol{\Sigma}}^* = \widehat{\boldsymbol{\Sigma}}^{\tau_{j_*}}$.
```

Figure 5.3: Lepskii's algorithm to obtain the robust estimator $\widehat{\boldsymbol{\Sigma}}^*$.

The previous result is almost identical to the one obtained in Chapter 4, but with the substitution of n by m and e^{-t} by $e^{-t/2}$. This is the price to pay for the non-independence if the addends of the estimator.

5.3.2 Forth moment estimation

In [21] the authors presented a more intuitive method to establish τ. The argument is as follows. According to equation (5.13), for a fixed $t \geq 0$, in order to obtain a good concentration with high probability, we need to chose

$$\tau = v\sqrt{\frac{m}{\log(2p) + t}}, \tag{5.15}$$

where

$$v^2 = \frac{1}{4}\|\mathbb{E}\left(\boldsymbol{Y}_1\boldsymbol{Y}_1^\intercal\right)^2\| = \|\mathbb{E}\left(\boldsymbol{Y}_1\boldsymbol{Y}_1^\intercal/2\right)^2\|.$$

Of course v^2 is unknown, but it is natural to think that for a well chosen τ, a good estimator of $\mathbb{E}\left(\boldsymbol{Y}_1\boldsymbol{Y}_1^\intercal/2\right)^2$ is

$$\frac{1}{N}\sum_{i=1}^{N}\left[\psi_\tau\left(\boldsymbol{Y}_i\boldsymbol{Y}_i^\intercal/2\right)\right]^2 = \frac{1}{N}\sum_{i=1}^{N}\psi_\tau^2\left(\frac{1}{2}\|\boldsymbol{Y}_i\|_2^2\right)\frac{\boldsymbol{Y}_i\boldsymbol{Y}_i^\intercal}{\|\boldsymbol{Y}_i\|_2^2}. \tag{5.16}$$

Therefore, substituting (5.16) into (5.15), a good choice of τ is such that it satisfies the equality

$$\left|\left|\left|\frac{1}{\tau^2 N}\sum_{i=1}^{N}\psi_\tau^2\left(\frac{1}{2}\|\boldsymbol{Y}_i\|_2^2\right)\frac{\boldsymbol{Y}_i\boldsymbol{Y}_i^\intercal}{\|\boldsymbol{Y}_i\|_2^2}\right|\right|\right| = \frac{\log(2p)+t}{m}. \tag{5.17}$$

Following [49] for the univariate case, we present in the next theorem a condition under which equation (5.17) has a unique solution.

Theorem 5.3.2. *Suppose that the iid sample $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ satisfies the following two conditions:*

1. *$\boldsymbol{X}_i$ is a random vector with continuous coordinates.*
2. *Let $\lambda = \|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_2^2$. Then, for any $\tau \in (0,\infty)$, $\mathbb{P}(\lambda > \tau) < 1$.*

Then, if

$$\left|\left|\left|\frac{m}{\log(2p)+t}\sum_{i=1}^{N}\frac{\boldsymbol{Y}_i\boldsymbol{Y}_i^\intercal}{\|\boldsymbol{Y_i}\|_2^2}\right|\right|\right| > 1,$$

there exists a unique τ such that equality (5.17) is satisfied.

The proof of Theorem 5.3.2 is presented in appendix C.

5.3.3 Cross-validation

Another method to determine τ is by *leave-one-one cross-validation*. The idea behind cross validation is to minimize a loss function that depends on a hyperparameter and choose this hyperparameter as the best one. But, since this loss function depends on an unknown distribution we use an empirical version to estimate it. See the survey [2] for a general review on cross-validation methods. This methodology applied to covariance matrix estimation is presented in [41] and it is bases on the following proposition.

Proposition 5.3.3. *Let $\boldsymbol{Z} \in \mathbb{R}^p$ be a random vector with $\mathbb{E}\boldsymbol{Z} = \mathbf{0}$ and $\mathrm{Cov}\boldsymbol{Z} = \boldsymbol{\Sigma}$. Then,*

$$\boldsymbol{\Sigma} = \underset{\mathbf{S} \in \mathcal{S}_p}{\arg\min}\, \mathbb{E}|||\mathbf{S} - \boldsymbol{Z}\boldsymbol{Z}^\intercal|||_2^2.$$

The proof of Proposition 5.3.3 can be found in appendix C. Since

$$\mathbb{E}\left[(\boldsymbol{X}_1 - \bar{\boldsymbol{X}})(\boldsymbol{X}_1 - \bar{\boldsymbol{X}})^\intercal\right] = \frac{n-1}{n}\boldsymbol{\Sigma},$$

(see appendix C) this result indicates that a good choice of τ is such that the expectation

$$\mathbb{E}|||\widehat{\boldsymbol{\Sigma}}^\tau - a_n(\boldsymbol{X}_1 - \bar{\boldsymbol{X}})(\boldsymbol{X}_1 - \bar{\boldsymbol{X}})^\intercal|||_2^2, \quad a_n = \frac{n}{n-1}.$$

is minimized, but because we don't have access to this expected value we use a sample-based approach. Let $\widehat{\boldsymbol{\Sigma}}^\tau_{(j)}$ be the robust estimator (5.3) calculated without the observation $\boldsymbol{X}_j$. The leave-one-out procedure defined in [41] consists in calculating the loss function

$$L(\tau) = \frac{1}{n}\sum_{j=1}^{n} |||\widehat{\boldsymbol{\Sigma}}^\tau_{(j)} - a_n(\boldsymbol{X}_j - \bar{\boldsymbol{X}})(\boldsymbol{X}_j - \bar{\boldsymbol{X}})^\intercal|||_2^2,$$

where $\tau \in \mathcal{C}$, where $\mathcal{C}$ is some subset of $\mathbb{R}$, and choosing the hyperparameter

$$\hat{\tau} = \underset{\tau \in \mathcal{C}}{\arg\min}\, L(\tau).$$

5.4 Simulation study

In this section we derive a simulation study to assess the performace of the robust estimator $\widehat{\boldsymbol{\Sigma}}^*$, obtained by the Lepskii method of Figure 5.3, against the classical empirical estimator $\widehat{\boldsymbol{\Sigma}}$ of Chapter 3.

The sample was obtained as follows: simulate independently $\boldsymbol{X}_1^0, ..., \boldsymbol{X}_n^0$ from a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}^0$. Let $S_1, ..., S_n$ be iid χ^2_ν random variables with $\nu = 3$ (also independent of the $\boldsymbol{X}_j^0$). Define $\boldsymbol{X}_j = \boldsymbol{X}_j^0/\sqrt{S_j/\nu}$. The distribution of $\boldsymbol{X}_j$ is said to be *multivariate-T* and has mean $\boldsymbol{\mu}$ and covariance matrix given by $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^0\nu/(\nu-2)$.This distribution is the multivariate version of the classical T distribution which is known for having heavy tails. Figure 5.4 presents the univariate T density function given by

$$f(x;\nu) = \frac{\gamma((\nu+1)/2)}{\sqrt{\nu\pi}\gamma(\nu/2)}\left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}.$$

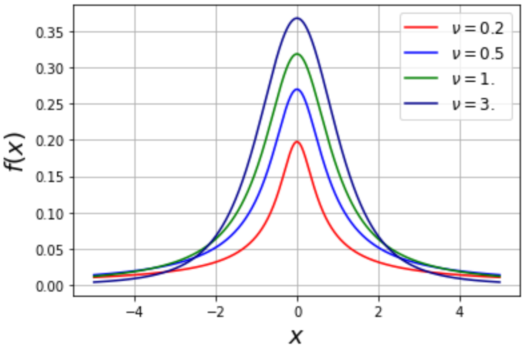


Figure 5.4: Univariate T density function for different values of ν.

One can observe that for smaller values of ν this density function has heavier tails. One major caution is that a random variable with T distribution has finite variance only when $\nu > 2$. See [23, Chapter 1] for these and more properties of the multivariate-T distribution.

In Figure 5.5 we present four different data sets from this distribution. One can observe that the appearance of atypical points is the rule rather than the exception. The mean vector $\boldsymbol{\mu}$ was fixed by simulating a $p \times 1$ vector with iid $\mathcal{N}(0, 100)$ entries, and the matrix $\boldsymbol{\Sigma}$ was fixed by simulating a matrix $\mathbf{B}$ with iid $\mathcal{N}(0, 5)$ entries and defining $\boldsymbol{\Sigma} = \mathbf{B}^\intercal\mathbf{B}$.

We constructed the sample $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_N$, $N = n(n-1)/2$, as (5.7) and calculated the estimator $\widehat{\boldsymbol{\Sigma}}^*$ of Figure 5.3. The parameters were chosen as $\gamma = 1.5$ and $v_{\max}$ and $v_{\min}$ where estimated as

$$v_{\max} = 2\sqrt{\frac{1}{4N}\sum_{j=1}^{N} |||(\boldsymbol{Y}_j\boldsymbol{Y}_j^\intercal)^2|||}, \quad v_{\min} = \frac{v_{\max}}{100}$$

i.e., $v_{\max}$ is two times the empirical estimator of $|||\mathbb{E}\,(\boldsymbol{Y}\boldsymbol{Y}^\intercal/2)^2\,|||$. For different dimensions p, we evaluated the difference $|||\widehat{\boldsymbol{\Sigma}}^* - \boldsymbol{\Sigma}|||$ and compare it with $|||\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}|||$. Different values of the parameter t were takin to observe its effect in the estimation. Figure 5.6 shows that the robust estimator $\widehat{\boldsymbol{\Sigma}}^*$ outperforms the classical estimator $\widehat{\boldsymbol{\Sigma}}$ for every dimension p and every choice of t.

The values of the robustification parameter τ chosen by the algorithm are shown in Figure 5.7. We can observe that the value of τ does not change drastically when changing t for $t \geq 0.5$. This is reflected in Figure 5.6 were it is clear that the variability of the estimator is reduced. This is due to the fact that for small τ (provoked by big t) a lot of observations are truncated so they are given the same value to calculate the estimator.

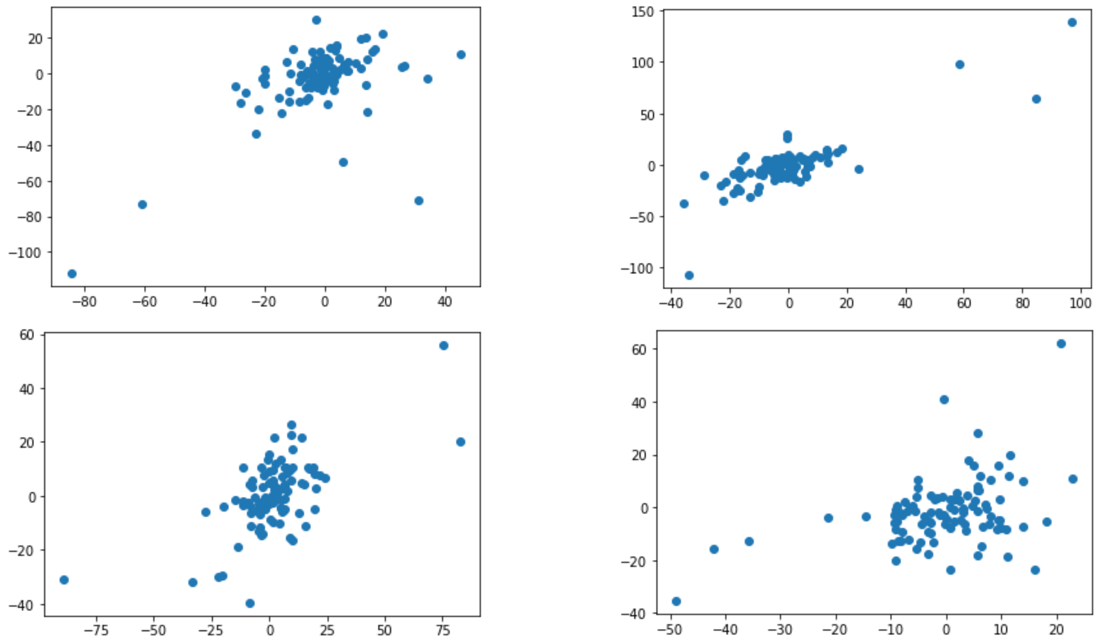
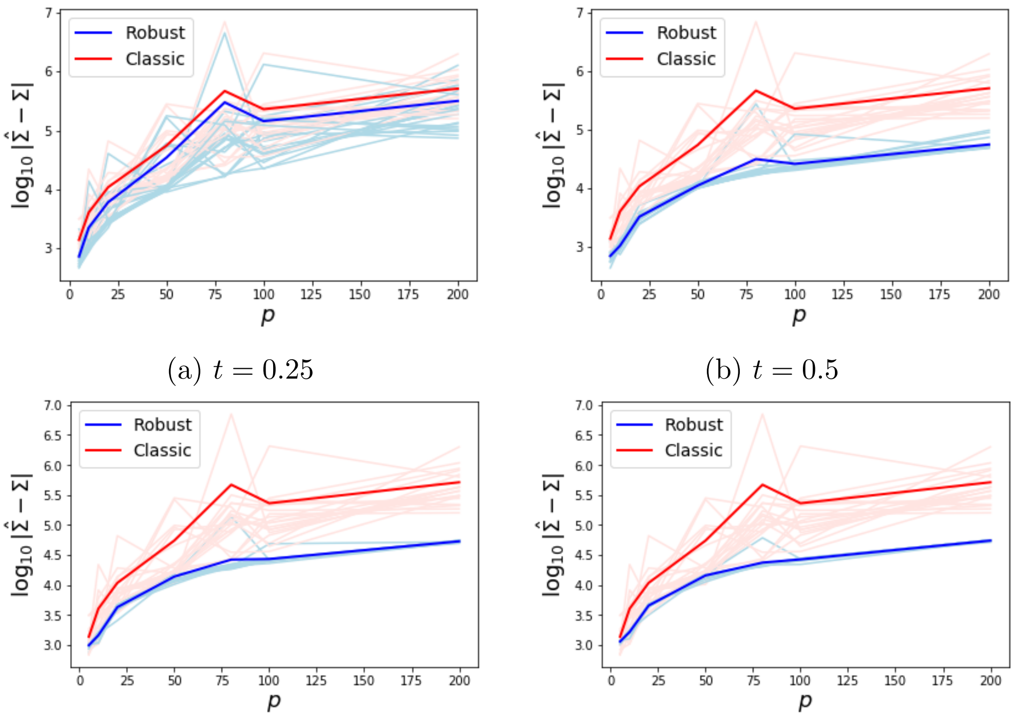


Figure 5.5: Four different data sets of sample size $n = 100$ simulated from the t-multivariate distribution in the case $p = 2$.



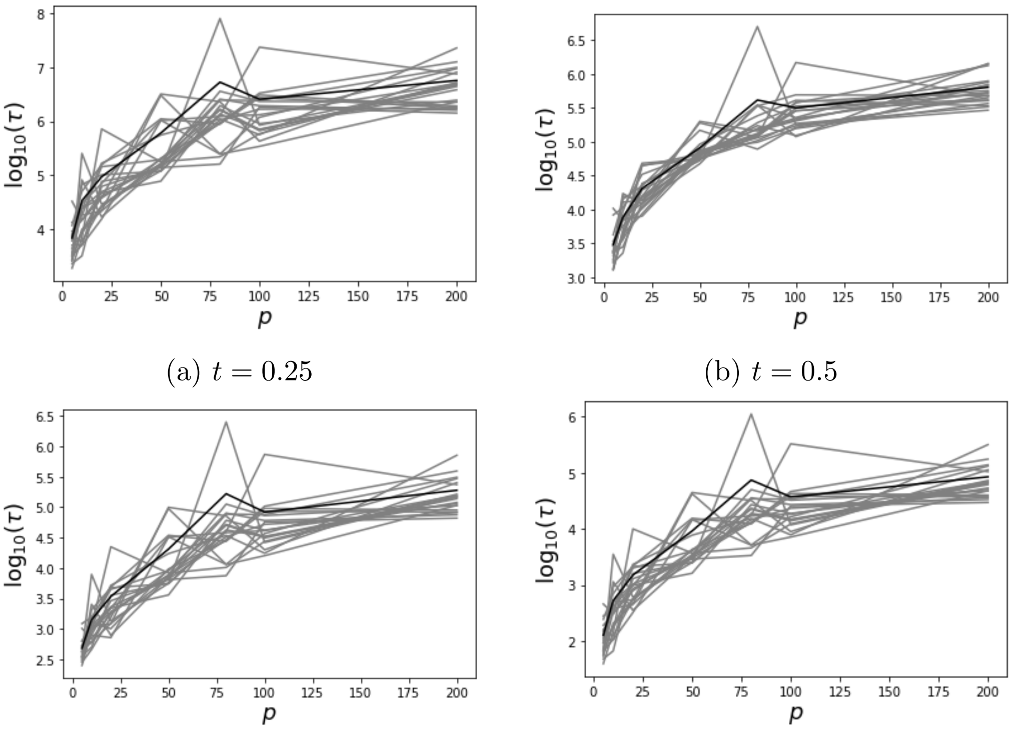
(a) $t = 0.25$

(b) $t = 0.5$

(c) $t = 2$

(d) $t = 10$

Figure 5.6: Performance evaluation of $\widehat{\boldsymbol{\Sigma}}^*$ vs $\widehat{\boldsymbol{\Sigma}}$. The values of p are in $\{5, 10, 20, 50, 80, 100, 200\}$. The sample size is $n = 50$. For every p there are 20 simulations of the estimator and the mean of them is shown with stronger color. The scale of the y-axis is presented in $\log_{10}$.



(a) $t = 0.25$

(b) $t = 0.5$

(c) $t = 2$

(d) $t = 10$

Figure 5.7: Values of τ obtained in the simulations of Figure 5.6.

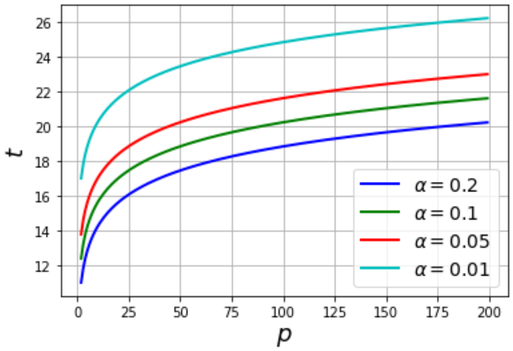


Figure 5.8: Lower bounds for t given p for different values of α. We take $\gamma = 1.5$ and $v_{\min} = v_{\max}/100$.

If one wants to choose t such that

$$\mathbb{P}\left(|||\widehat{\mathbf{\Sigma}}^* - \mathbf{\Sigma}||| \geq (3+\epsilon)v\sqrt{\frac{2t}{m}} \right) \leq \alpha$$

for some $\alpha \in (0,1)$, by Corollary 5.3.1, its is sufficient to choose

$$t \geq 2\log\left(\frac{2p\kappa}{\alpha}\right), \tag{5.18}$$

where $\kappa = \log(\gamma v_{\max}/v_{\min})/\log(\gamma)$. Figure 5.8 shows the lower bound (5.18) for different dimensions p and different values of α. Despite that the bound increases slowly after certain point, as we saw in Figure 5.6 a big value of t can cause two much regularization (small τ) provoking that a lot of sampled points are assigned the same small value. Therefore, the election of t have to be done with caution.

As we discussed earlier, when the parameter ν of the T distribution is big, the tails of the distribution are lighter. In order to observe the performance of $\widehat{\mathbf{\Sigma}}^*$ in a light-tail case, we performed the same simulation study choosing the parameter $\nu = 30$. Figure 5.9 shows that the performance of $\widehat{\mathbf{\Sigma}}^*$ is similar to $\widehat{\mathbf{\Sigma}}$ since there are fewer outliers that make the performance of $\widehat{\mathbf{\Sigma}}$ worse. Nevertheless, the estimator $\widehat{\mathbf{\Sigma}}^*$ always has smaller error even in higher dimensions.

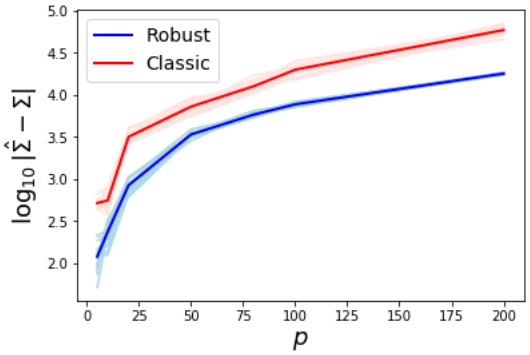


Figure 5.9: Same simulation study with $\nu = 30$ and $t = 0.5$.

Summary and observations

In this chapter we presented a straightforward way to estimate the covariance matrix of a random vector making few assumptions abound the moments. By performing a combinatorial decomposition of the estimator we've been able to prove a concentration bound that is similar to the one obtained in Chapter 4 and without assuming that the random vector is centered, in contrast with Chapter 3. Additionally, we mentioned three methods to choose the robustification parameter τ, one of which has a concentration guarantee. The simulation study shows that the robust estimator presented outperforms the classical empirical estimator. Nevertheless, the Lespkii's procedure used depends on the hyperparameter t that can affect the performance of the estimations.

Chapter 6

Robust matrix completion

In this chapter we present a different application of the concentration results obtained in Chapter 4. The problem of matrix completion has been a prolific research subject since the last decade. See for example the articles [7], [8] and [5] and references therein. This context differs from what we have presented so far in the sense that the estimator has no explicit form because it is obtained from an optimization problem. This yields a clear difficulty: how can we guarantee concentration results in matrix completion since we can not manipulate the estimator directly? Even more, how can we adapt the known estimation methods to make the estimators robust? This will be addressed in what follows.

Section 1 gives the context of matrix completion that we'll be working on. Then, in Section 2 we present some methods for estimation in this context, and in Section 3 we show how to adapt these methods to obtain a robust estimator with concentration guarantees that are fully proved. Finally, Section 4 describes how to compute the robust estimator presented.

6.1 What is matrix completion?

Let $\mathbf{B}_0 \in \mathcal{M}_{m_1,m_2}$ be an unknown matrix. We observe only a subset of its entries with noise, and we want to estimate $\mathbf{B}_0$ through this partial and noisy observations. Stated more formally, define the set of $m_1 \times m_2$ matrices $\mathcal{X}$ as

$$\mathcal{X} = \left\{ \boldsymbol{e}_j(m_1)\boldsymbol{e}_k^{\intercal}(m_2),\ j = 1, ..., m_1,\ k = 1, ..., m_2 \right\},$$

where $\{\boldsymbol{e}_1(m_1), ..., \boldsymbol{e}_{m_1}(m_1)\}$ and $\{\boldsymbol{e}_1(m_2), ..., \boldsymbol{e}_{m_2}(m_2)\}$ are the canonical bases of $\mathbb{R}^{m_1}$ and $\mathbb{R}^{m_2}$, respectively. Let[1] $\mathbf{X} \sim \mathcal{U}(\mathcal{X})$ and Y be the random variable

$$Y = \text{tr }(\mathbf{X}^\intercal \mathbf{B}_0) + \xi, \tag{6.1}$$

where ξ is a random noise such that $\mathbb{E}[\xi|\mathbf{X}] = 0$. The model 6.1 is called Trace Regression Model. Suppose that $(Y_1, \mathbf{X}_1), ..., (Y_n, \mathbf{X}_n)$ are iid copies of $(Y, \mathbf{X})$. The problem of matrix completion consists on developing a methodology to estimate $\mathbf{B}_0$ from the sample $\{(Y_j, \mathbf{X}_j)\}$.

As pointed out in [38] and [22], this model is called Uniform Sampling at Random (USR), which differs from the popular work of [7], called Collaborative Sampling (CS). The main difference is that in USR we can have repetitions in the matrices $\mathbf{X}_1, ..., \mathbf{X}_n$, while in CS this is not possible. The CS matrix completion model is used to describe recommendation systems in which a customer rates an item only once. Meanwhile, the USR matrix completion model can be used to describe the transmission of large matrices through a noisy communication channel. In this thesis we work with the USR model and present the concentration results from Koltchisnskii et al. in [22] and the robust procedure from Minsker in [32].

It is necessary to mention that important work has been done in the CS model within a robust context. The seminal work of [6] presents a generic method of matrix estimation that translates into robust matrix completion and robust PCA (something that deviates broadly from what we did in Chapter 4.)

6.2 Nuclear norm penalization

As pointed out in [7], in many instances the matrix $\mathbf{B}_0$ can be thought to have low rank or be approximately low rank. With this assumption, one wants to perform an estimation procedure that captures this low-rank structure. This is performed typically by penalizing with the Schatten 1-norm also called the nuclear norm. This is done since the nuclear norm is the Schatten p-norm that is convex and closest to the Shatten 0-norm[2], which gives the rank of a matrix. Before presenting the penalized estimator let us derive an unbiased estimator of $\mathbf{B}_0$.

[1] This means that $\mathbf{X}$ is uniformly distributed in the set $\mathcal{X}$. More specifically, for any j, k, with probability $\frac{1}{m_1 m_2}$ we obtain the matrix $\mathbf{X} = \boldsymbol{e}_j(m_1)\boldsymbol{e}_k^\intercal(m_2)$.

[2] The Schatten p-norms satisfy being a norm only when $p \geq 1$, but are called *norm* for any $p \geq 0$ just for simplicity.

With the assumptions made earlier, it's easy to verify that

$$\mathbb{E}Y\mathbf{X} = \frac{1}{m_1 m_2}\mathbf{B}_0.$$

Indeed, since $\mathbf{X} \sim \mathcal{U}(\mathcal{X})$ and $\mathbb{E}[\xi|\mathbf{X}] = 0$,

$$\begin{aligned}
\mathbb{E}Y\mathbf{X} &= \mathbb{E}\left[(\operatorname{tr}(\mathbf{X}^\intercal \mathbf{B}_0) + \xi)\mathbf{X}\right] \\
&= \mathbb{E}\left[\operatorname{tr}(\mathbf{X}^\intercal \mathbf{B}_0)\mathbf{X}\right] + \mathbb{E}\left[\mathbf{X}\mathbb{E}[\xi|\mathbf{X}]\right] \\
&= \frac{1}{m_1 m_2}\sum_{j=1}^{m_1}\sum_{k=1}^{m_2}(\mathbf{B}_0)_{jk}\boldsymbol{e}_j(m_1)\boldsymbol{e}_k^\intercal(m_2) \\
&= \frac{1}{m_1 m_2}\mathbf{B}_0,
\end{aligned}$$

where the last inequality follows because $\boldsymbol{e}_j(m_1)\boldsymbol{e}_k^\intercal(m_2)$ is a matrix of zeros except that in the j, k-entry it has a 1. This indicates that an unbiased estimator of $\mathbf{B}_0$ is

$$\widehat{\mathbf{B}} = \frac{m_1 m_2}{n}\sum_{j=1}^{n} Y_j \mathbf{X}_j.$$

Therefore, in [22] the authors proposed the penalized estimator $\widehat{\mathbf{B}}^\tau$, $\tau > 0$, defined as

$$\widehat{\mathbf{B}}^\tau = \operatorname*{arg\,min}_{\mathbf{B}\in\mathcal{M}_{m_1,m_2}} \left\{ \frac{1}{m_1 m_2} |||\mathbf{B} - \widehat{\mathbf{B}}|||_2^2 + \tau |||\mathbf{B}|||_1 \right\},$$

i.e., $\widehat{\mathbf{B}}^\tau$ is the closest matrix to $\mathbf{B}$ in Frobenius norm with penalized rank. The next theorem, proved in [22] (Theorem 1), gives us a first upper bound for the performance of this estimator.

Theorem 6.2.1. *Define* $\mathbf{M} \in \mathcal{M}_{m_1,m_2}$ *as*

$$\mathbf{M} = \widehat{\mathbf{B}} - \mathbb{E}Y\mathbf{X}.$$

If $\tau \geq 2|||\mathbf{M}|||$*, then*

$$\frac{1}{m_1 m_2}|||\widehat{\mathbf{B}}^\tau - \mathbf{B}_0|||_2^2 \leq \inf_{\mathbf{B}\in\mathcal{M}_{m_1,m_2}} \left\{ \frac{1}{m_1 m_2}|||\mathbf{B} - \mathbf{B}_0|||_2^2 + \left(\frac{1+\sqrt{2}}{2}\right)^2 m_1 m_2 \tau^2 \mathit{rank}(\mathbf{B}) \right\}.$$

As we see, the condition $\tau \geq 2|||\mathbf{M}|||$ is not certain, so to obtain a concentration inequality one has to calculate

$$\mathbb{P}(\tau \geq 2|||\mathbf{M}|||).$$

In order to do so it is common to make distributional assumptions on the noise ξ. For example, in [22] the authors assume that ξ follows a sub-exponential distribution (see Appendix B for a definition of sub-exponential distributions.) We'll prove a version of Theorem 6.2.1 in form of a lemma (Lemma 6.3.2) in the context of robust estimation. More specifically, we'll just assume that

$$\mathrm{Var}\xi < \infty \quad \text{and} \quad |||\mathbf{B}_0|||_{\max} < \infty,$$

where $|||\mathbf{B}_0|||_{\max} = \max_{i,j} |(\mathbf{B}_0)_{ij}|$. The proof is essentially the same as the one presented in [22], but some careful has to be taken because we'll incorporate the matrix dilation defined on previous chapters.

6.3 Robust penalized estimation

The matrices $\mathbf{B}_0$ and $\widehat{\mathbf{B}}$ are rectangular (and non-symmetric), so to use the concentration results of Chapter 4 we need to incorporate the matrix dilation $\mathcal{H}$ into the equation. In [32] Minsker proposed the estimator $\widehat{\mathbf{R}} \in \mathcal{S}_{m_1+m_2}$ of $\mathcal{H}(\mathbf{B}_0)$ defined as

$$\widehat{\mathbf{R}} = \frac{1}{n\theta} \sum_{j=1}^{n} \psi\left(\theta m_1 m_2 Y_j \mathcal{H}(\mathbf{X}_j)\right),$$

with the specific choice of θ given by

$$\theta = \theta(t, n, \mathbf{B}_0) = \frac{1}{\sqrt{\mathrm{Var}\xi} \vee |||\mathbf{B}_0|||_{\max}} \sqrt{\frac{t + \log(2(m_1 + m_2))}{n m_1 m_2 (m_1 \vee m_2)}}.$$

The estimator $\widehat{\mathbf{R}}$ is a robust surrogate of $\widehat{\mathbf{B}}$. To incorporate the nuclear norm penalization, Minsker defined the following penalized robust estimator:

$$\widehat{\mathbf{R}}^{\tau} = \underset{\mathbf{B} \in \mathcal{M}_{m_1,m_2}}{\arg\min} \left\{ \frac{1}{m_1 m_2} |||\mathcal{H}(\mathbf{B}) - \widehat{\mathbf{R}}|||_2^2 + 2\tau |||\mathbf{B}|||_1 \right\}$$

Note that by Theorem A.3.5 $2|||\mathbf{B}|||_p^p = |||\mathcal{H}(\mathbf{B})|||_p^p$ for any $p > 0$. Define $\mathbb{A} \subset \mathcal{S}_{m_1+m_2}$ as

$$\mathbb{A} = \{\mathbf{A} \in \mathcal{S}_{m_1+m_2} \ : \ \mathbf{A} = \mathcal{H}(\mathbf{B}) \text{ for some } \mathbf{B} \in \mathcal{M}_{m_1,m_2}\}.$$

Then, we can write

$$\mathcal{H}(\widehat{\mathbf{R}}^{\tau}) = \underset{\mathbf{A}\in\mathbb{A}}{\arg\min} \left\{ \frac{1}{m_1 m_2} |||\mathbf{A} - \widehat{\mathbf{R}}|||_2^2 + \tau |||\mathbf{A}|||_1 \right\}.$$

Note that $\widehat{\mathbf{R}}^{\tau} \in \mathcal{M}_{m_1,m_2}$ but $\mathcal{H}(\widehat{\mathbf{R}}^{\tau}) \in \mathcal{S}_{m_1+m_2}$. The next theorem, provided in [32], gives us a concentration bound for the performance of $\widehat{\mathbf{R}}^{\tau}$ in Frobenius norm.

Theorem 6.3.1. *Define the matrix* $\mathbf{M} \in \mathcal{S}_{m_1+m_2}$ *as*

$$\begin{aligned} \mathbf{M} &= \widehat{\mathbf{R}} - \mathbb{E}\left[m_1 m_2 Y \mathcal{H}(\mathbf{X})\right] \\ &= \widehat{\mathbf{R}} - \mathcal{H}(\mathbf{B}_0). \end{aligned}$$

If ξ_j *and* $\mathbf{X}_j$ *are independent,* $j = 1, ..., n$*, and* $\operatorname{Var}\xi < \infty$*, then for any*

$$\tau \geq 8 m_1 m_2 (m_1 \vee m_2) \left(|||\mathbf{B}_0|||_{\max} \vee \sqrt{\operatorname{Var}\xi} \right) \sqrt{\frac{t + \log(2(m_1 + m_2))}{n}},$$

we have that

$$\frac{1}{m_1 m_2} |||\widehat{\mathbf{R}}^{\tau} - \mathbf{B}_0|||_2^2 \leq \inf_{\mathbf{B}\in\mathcal{M}_{m_1,m_2}} \left\{ \frac{1}{m_1 m_2} |||\mathbf{B} - \mathbf{B}_0|||_2^2 + \left(\frac{1+\sqrt{2}}{2}\right)^2 m_1 m_2 \tau^2 \mathit{rank}(\mathbf{B}) \right\},$$

with probability at least $1 - e^{-t}$.

The result of Theorem 6.3.1 differs from what we obtained in previous chapters:

1. The norm bound has no explicit form. This is due to the fact that the estimator also has no explicit form and we have to use different machinery to control the norm of the difference.

2. We are working with the 2-Schatten norm (or Frobenius norm) instead of the usual ∞-Schatten norm (or spectral norm.) This follows from the definition of the estimator, since we are trying to minimize the Frobenius norm instead of the nuclear norm.

3. The main information that we can get from this theorem is the range of values of the penalization τ for which we can obtain a “good concentration bound.”

To prove Theorem 6.3.1 we need the following lemma. As mentioned earlier, this is a version of Theorem 6.2.1 presented in [22]. The main idea behind the proof is that the minimization problem involves a convex function in a convex set so we can use the methods of subdifferential calculus presented in Appendix C.

Lemma 6.3.2. *Define* $\mathbf{M}$ *as in Theorem 6.3.1. If* $\tau \geq 4|||\mathbf{M}|||$*, then,*

$$\frac{1}{m_1 m_2}|||\mathcal{H}(\widehat{\mathbf{R}}^{\tau}) - \mathcal{H}(\mathbf{B}_0)|||_2^2 \leq \inf_{\mathbf{A}\in\mathbb{A}} \left\{ \frac{1}{m_1 m_2}|||\mathbf{A} - \mathcal{H}(\mathbf{B}_0)|||_2^2 + \left(\frac{1+\sqrt{2}}{2}\right)^2 m_1 m_2 \tau^2 rank(\mathbf{A}) \right\}.$$

To prove Lemma 6.3.2 it will be necessary to recall the trace duality property of Theorem A.5.9, which states that for matrices $\mathbf{B}, \mathbf{B}' \in \mathcal{M}_{n,m}$ we have that $|\langle \mathbf{B}, \mathbf{B}' \rangle| \leq |||\mathbf{B}|||_1 |||\mathbf{B}'|||$. Throughout this chapter, $\langle \mathbf{B}, \mathbf{B}' \rangle = \operatorname{tr}(\mathbf{B}^{\intercal}\mathbf{B}')$.

Proof of Lemma 6.3.2. Write $\widehat{\mathbf{A}}^{\tau} = \mathcal{H}(\widehat{\mathbf{R}}^{\tau})$, $\mathbf{A}_0 = \mathcal{H}(\mathbf{B}_0)$ and

$$F_\tau(\mathbf{A}) = \mu|||\mathbf{A} - \widehat{\mathbf{R}}|||_2^2 + \tau|||\mathbf{A}|||_1, \quad \mu = (m_1 m_2)^{-1},\ \mathbf{A} \in \mathcal{S}_{m_1+m_2}.$$

Then, we want to prove that whenever $\tau \geq 4|||\mathbf{M}|||$,

$$\mu|||\widehat{\mathbf{A}}^{\tau} - \mathbf{A}_0|||_2^2 \leq \inf_{\mathbf{A}\in\mathbb{A}} \left\{ \mu|||\mathbf{A} - \mathbf{A}_0|||_2^2 + \left(\frac{1+\sqrt{2}}{2}\right)^2 \mu^{-1}\tau^2 \operatorname{rank}(\mathbf{A}) \right\},$$

where

$$\widehat{\mathbf{A}}^{\tau} = \arg\min_{\mathbf{A}\in\mathbb{A}} F_\tau(\mathbf{A}).$$

It is important to mention that, due to the form of $\widehat{\mathbf{R}}$, this lemma is not a consequence of Theorem 6.2.1, but it can be derived following the same steps.

The proof is divided in four parts. Part I is dedicated to obtain a bound for the inner product $\langle \widehat{\mathbf{A}}^{\tau} - \mathbf{A}_0, \widehat{\mathbf{A}}^{\tau} - \mathbf{A} \rangle$ for some $\mathbf{A} \in \mathbb{A}$. Part II develops a generic bound on the norm difference $|||\widehat{\mathbf{A}}^{\tau} - \mathbf{A}_0|||_2^2$. Part III focuses on bounding the inner product $\langle \mathbf{M}, \widehat{\mathbf{A}}^{\tau} - \mathbf{A} \rangle$. Part IV joins everything and finally obtains the desired bound for $|||\widehat{\mathbf{A}}^{\tau} - \mathbf{A}_0|||_2^2$.

Part I. By Proposition C.10.4 (b) and Theorem C.10.5 which gives the form of $\partial F_\tau(\widehat{\mathbf{A}}^{\tau})$, there exists a $\widehat{\mathbf{V}} \in \partial|||\widehat{\mathbf{A}}^{\tau}|||_1$ such that, for all $\mathbf{A} \in \mathbb{A}$,

$$\langle 2\mu(\widehat{\mathbf{A}}^{\tau} - \widehat{\mathbf{R}}) + \tau\widehat{\mathbf{V}}, \mathbf{A} - \widehat{\mathbf{A}}^{\tau}\rangle \geq 0,$$

or , by linearity and changing signs,

$$2\mu\langle\widehat{\mathbf{A}}^{\tau}, \widehat{\mathbf{A}}^{\tau} - \mathbf{A}\rangle - 2\mu\langle\widehat{\mathbf{R}}, \widehat{\mathbf{A}}^{\tau} - \mathbf{A}\rangle + \tau\langle\widehat{\mathbf{V}}, \widehat{\mathbf{A}}^{\tau} - \mathbf{A}\rangle \leq 0. \tag{6.2}$$

Fix an arbitrary $\mathbf{A} \in \mathbb{A}$ of rank r with SVD $\mathbf{A} = \sum_{j=1}^{r} s_j \boldsymbol{u}_j \boldsymbol{v}_j^{\intercal}$ and support[3] (S_1, S_2), and consider an arbitrary $\mathbf{V} \in \partial|||\mathbf{A}|||_1$. Recall that $\mathbf{M} = \widehat{\mathbf{R}} - \mathbf{A}_0$, so by adding and subtracting

$$-2\mu\langle\mathbf{A}_0, \widehat{\mathbf{A}}^{\tau} - \mathbf{A}\rangle + \tau\langle\mathbf{V}, \widehat{\mathbf{A}}^{\tau} - \mathbf{A}\rangle$$

to inequality (6.2) we get that

$$2\mu\langle\widehat{\mathbf{A}}^{\tau} - \mathbf{A}_0, \widehat{\mathbf{A}}^{\tau} - \mathbf{A}\rangle + \tau\langle\widehat{\mathbf{V}} - \mathbf{V}, \widehat{\mathbf{A}}^{\tau} - \mathbf{A}\rangle \leq -\tau\langle\mathbf{V}, \widehat{\mathbf{A}}^{\tau} - \mathbf{A}\rangle + 2\mu\langle\mathbf{M}, \widehat{\mathbf{A}}^{\tau} - \mathbf{A}\rangle. \tag{6.3}$$

By the monotonicity of the subdifferential of convex functions presented in Proposition C.10.3, we have that $\langle\widehat{\mathbf{V}} - \mathbf{V}, \widehat{\mathbf{A}}^{\tau} - \mathbf{A}\rangle \geq 0$, which implies from (6.3) that

$$2\mu\langle\widehat{\mathbf{A}}^{\tau} - \mathbf{A}_0, \widehat{\mathbf{A}}^{\tau} - \mathbf{A}\rangle \leq -\tau\langle\mathbf{V}, \widehat{\mathbf{A}}^{\tau} - \mathbf{A}\rangle + 2\mu\langle\mathbf{M}, \widehat{\mathbf{A}}^{\tau} - \mathbf{A}\rangle. \tag{6.4}$$

On the other hand, by Theorem C.10.5, for an arbitrary matrix $\mathbf{W}$ with $|||\mathbf{W}||| \leq 1$, we can write

$$\mathbf{V} = \sum_{j=1}^{r} \boldsymbol{u}_j \boldsymbol{v}_j^{\intercal} + \mathbf{P}_{S_1^{\perp}} \mathbf{W} \mathbf{P}_{S_2^{\perp}}, \tag{6.5}$$

where $\mathbf{P}_{S_1^{\perp}}, \mathbf{P}_{S_2^{\perp}} \in \mathcal{S}_{m_1+m_2}$ are the orthogonal projectors in $S_1^{\perp}$ and $S_2^{\perp}$, respectively[4]. But, observe that

$$\begin{aligned}
\langle\mathbf{P}_{S_1^{\perp}} \mathbf{W} \mathbf{P}_{S_2^{\perp}}, \widehat{\mathbf{A}}^{\tau} - \mathbf{A}\rangle &= \operatorname{tr}\left[\mathbf{P}_{S_2^{\perp}} \mathbf{W}^{\intercal} \mathbf{P}_{S_1^{\perp}} (\widehat{\mathbf{A}}^{\tau} - \mathbf{A})\right] \\
&= \operatorname{tr}\left[\mathbf{P}_{S_2^{\perp}} \mathbf{W}^{\intercal} \mathbf{P}_{S_1^{\perp}} \widehat{\mathbf{A}}^{\tau}\right] \\
&= \operatorname{tr}\left[\mathbf{W}^{\intercal} \mathbf{P}_{S_1^{\perp}} \widehat{\mathbf{A}}^{\tau} \mathbf{P}_{S_2^{\perp}}\right]
\end{aligned}$$

[3] For $\mathbf{A}$ of rank r with SVD $\mathbf{A} = \sum_{j=1}^{r} s_j \boldsymbol{u}_j \boldsymbol{v}_j^{\intercal}$ we define $S_1 = \operatorname{span}\{\boldsymbol{u}_1, ..., \boldsymbol{u}_r\}$ and $S_2 = \operatorname{span}\{\boldsymbol{v}_1, ..., \boldsymbol{v}_r\}$, and the pair (S_1, S_2) is called the support of $\mathbf{A}$.

[4] We also define $\mathbf{P}_{S_1} \in \mathcal{S}_{m_1}$ and $\mathbf{P}_{S_2} \in \mathcal{S}_{m_2}$ as $\mathbf{P}_{S_1} = \mathbf{I} - \mathbf{P}_{S_1^{\perp}}$ and $\mathbf{P}_{S_2} = \mathbf{I} - \mathbf{P}_{S_2^{\perp}}$

$$= \langle \mathbf{W}, \mathbf{P}_{S_1^\perp} \widehat{\mathbf{A}}^\tau \mathbf{P}_{S_2^\perp} \rangle,$$

where the second equality follows from Theorem A.3.4, i.e., since S_1 is the columns space of $\mathbf{A}$ then $\mathbf{P}_{S_2^\perp}\mathbf{A} = \mathbf{0}$. Now, by Proposition A.5.10 of equality in trace duality, there exists $\mathbf{W}$ with $|||\mathbf{W}||| \leq 1$ such that

$$\langle \mathbf{P}_{S_1^\perp} \mathbf{W} \mathbf{P}_{S_2^\perp}, \widehat{\mathbf{A}}^\tau - \mathbf{A} \rangle = \langle \mathbf{W}, \mathbf{P}_{S_1^\perp} \widehat{\mathbf{A}}^\tau \mathbf{P}_{S_2^\perp} \rangle = |||\mathbf{P}_{S_1^\perp} \widehat{\mathbf{A}}^\tau \mathbf{P}_{S_2^\perp}|||_1.$$

For this choice of $\mathbf{W}$, inequality (6.4) and equation (6.5) implies

$$2\mu \langle \widehat{\mathbf{A}}^\tau - \mathbf{A}_0, \widehat{\mathbf{A}}^\tau - \mathbf{A} \rangle + \tau |||\mathbf{P}_{S_1^\perp} \widehat{\mathbf{A}}^\tau \mathbf{P}_{S_2^\perp}|||_1 \leq -\tau \left\langle \sum_{j=1}^r \boldsymbol{u}_j \boldsymbol{v}_j^\intercal, \widehat{\mathbf{A}}^\tau - \mathbf{A} \right\rangle + 2\mu \langle \mathbf{M}, \widehat{\mathbf{A}}^\tau - \mathbf{A} \rangle. \tag{6.6}$$

Part II. On the other hand, we have the following three equalities:

$$|||\mathbf{A} - \mathbf{A}_0|||_2^2 = |||\widehat{\mathbf{A}}^\tau - \mathbf{A}|||_2^2 + |||\widehat{\mathbf{A}}^\tau - \mathbf{A}_0|||_2^2 - 2\langle \widehat{\mathbf{A}}^\tau - \mathbf{A}, \widehat{\mathbf{A}}^\tau - \mathbf{A}_0 \rangle;$$

$$\left|\left|\left| \sum_{j=1}^r \boldsymbol{u}_j \boldsymbol{v}_j^\intercal \right|\right|\right| = 1,$$

which is due to the fact that $\sum_{j=1}^r \boldsymbol{u}_j \boldsymbol{v}_j^\intercal$ defines the SVD of a matrix with an identity in the middle; and

$$\left\langle \sum_{j=1}^r \boldsymbol{u}_j \boldsymbol{v}_j^\intercal, \widehat{\mathbf{A}}^\tau - \mathbf{A} \right\rangle = \left\langle \sum_{j=1}^r \boldsymbol{u}_j \boldsymbol{v}_j^\intercal, \mathbf{P}_{S_1}(\widehat{\mathbf{A}}^\tau - \mathbf{A})\mathbf{P}_{S_2} \right\rangle,$$

which emerges from $\mathbf{P}_{S_1}\mathbf{A}\mathbf{P}_{S_2} = \mathbf{A}$, $\boldsymbol{u}_j^\intercal \mathbf{P}_{S_1} = \boldsymbol{u}_j^\intercal$ and $\mathbf{P}_{S_1}\boldsymbol{v}_j = \boldsymbol{v}_j$. Then, from (6.6) and trace duality we get

$$\begin{aligned} &\mu |||\widehat{\mathbf{A}}^\tau - \mathbf{A}_0|||_2^2 + \mu |||\widehat{\mathbf{A}}^\tau - \mathbf{A}|||_2^2 + \tau |||\mathbf{P}_{S_1^\perp} \widehat{\mathbf{A}}^\tau \mathbf{P}_{S_2^\perp}|||_1 \\ &\qquad \leq \mu |||\mathbf{A} - \mathbf{A}_0|||_2^2 + \tau |||\mathbf{P}_{S_1}(\widehat{\mathbf{A}}^\tau - \mathbf{A})\mathbf{P}_{S_2}|||_1 + 2\mu \langle \mathbf{M}, \widehat{\mathbf{A}}^\tau - \mathbf{A} \rangle \end{aligned} \tag{6.7}$$

Part III. Now, we proceed to bound $\langle \mathbf{M}, \widehat{\mathbf{A}}^\tau - \mathbf{A} \rangle$. Define the linear map $\mathcal{P}_{\mathbf{A}} : \mathcal{M}_{m_1+m_2} \to \mathcal{M}_{m_1+m_2}$ as

$$\mathcal{P}_{\mathbf{A}}(\mathbf{A}') = \mathbf{A}' - \mathbf{P}_{S_1^\perp} \mathbf{A}' \mathbf{P}_{S_2^\perp}.$$

Then, for any $\mathbf{A}' \in \mathcal{M}_{m_1+m_2}$

$$\begin{aligned}\langle \mathcal{P}_{\mathbf{A}}(\mathbf{M}), \mathcal{P}_{\mathbf{A}}(\mathbf{A}')\rangle &= \operatorname{tr}\,[\mathcal{P}_{\mathbf{A}}(\mathbf{M})^\intercal \mathcal{P}_{\mathbf{A}}(\mathbf{A}')] \\ &= \operatorname{tr}\left[\mathbf{M}^\intercal\mathbf{A}' - \mathbf{M}^\intercal\mathbf{P}_{S_1^\perp}\mathbf{A}'\mathbf{P}_{S_2^\perp} - \mathbf{P}_{S_2^\perp}\mathbf{M}^\intercal\mathbf{P}_{S_1^\perp}\mathbf{A}' + \mathbf{P}_{S_2^\perp}\mathbf{M}^\intercal\mathbf{P}_{S_1^\perp}\mathbf{A}'\mathbf{P}_{S_1^\perp}\right] \\ &= \operatorname{tr}\left[\mathbf{M}^\intercal\mathbf{A}' - \mathbf{M}^\intercal\mathbf{P}_{S_1^\perp}\mathbf{A}'\mathbf{P}_{S_2^\perp}\right] \\ &= \langle \mathcal{P}_{\mathbf{A}}(\mathbf{M}), \mathbf{A}'\rangle,\end{aligned}$$

from which we obtain that

$$\begin{aligned}\langle \mathbf{M}, \widehat{\mathbf{A}}^\tau - \mathbf{A}\rangle &= \langle \mathcal{P}_{\mathbf{A}}(\mathbf{M}), \widehat{\mathbf{A}}^\tau - \mathbf{A}\rangle + \langle \mathbf{P}_{S_1^\perp}\mathbf{M}\mathbf{P}_{S_2^\perp}, \widehat{\mathbf{A}}^\tau - \mathbf{A}\rangle \\ &= \langle \mathcal{P}_{\mathbf{A}}(\mathbf{M}), \mathcal{P}_{\mathbf{A}}(\widehat{\mathbf{A}}^\tau - \mathbf{A})\rangle + \langle \mathbf{P}_{S_1^\perp}\mathbf{M}\mathbf{P}_{S_2^\perp}, \widehat{\mathbf{A}}^\tau\rangle \\ &= \langle \mathcal{P}_{\mathbf{A}}(\mathbf{M}), \mathcal{P}_{\mathbf{A}}(\widehat{\mathbf{A}}^\tau - \mathbf{A})\rangle + \langle \mathbf{P}_{S_1^\perp}\mathbf{M}\mathbf{P}_{S_2^\perp}, \mathbf{P}_{S_1^\perp}\widehat{\mathbf{A}}^\tau\mathbf{P}_{S_2^\perp}\rangle.\end{aligned}$$

Define $\Lambda = 2\mu|||\mathcal{P}_{\mathbf{A}}(\mathbf{M})|||_2$ and $\Gamma = 2\mu|||\mathbf{P}_{S_1^\perp}\mathbf{M}\mathbf{P}_{S_2^\perp}|||$. By Cauchy-Schwarz and trace duality,

$$\begin{aligned}2\mu\langle \mathbf{M}, \widehat{\mathbf{A}}^\tau - \mathbf{A}\rangle &\leq \Lambda|||\mathcal{P}_{\mathbf{A}}(\widehat{\mathbf{A}}^\tau - \mathbf{A})|||_2 + \Gamma|||\mathbf{P}_{S_1^\perp}\widehat{\mathbf{A}}^\tau\mathbf{P}_{S_2^\perp}|||_1 \\ &\leq \Lambda|||\widehat{\mathbf{A}}^\tau - \mathbf{A}|||_2 + \Gamma|||\mathbf{P}_{S_1^\perp}\widehat{\mathbf{A}}^\tau\mathbf{P}_{S_2^\perp}|||_1.\end{aligned}$$

The second inequality follows from Proposition A.5.12 since $\mathcal{P}_{\mathbf{A}}$ is a orthogonal projection operator. Now, by submultiplicity of the norm and the fact that $|||\mathbf{P}||| = 1$ for any orthogonal projection matrix $\mathbf{P}$, we have that

$$\Gamma \leq 2\mu|||\mathbf{M}|||\left(|||\mathbf{P}_{S_1^\perp}||| \cdot |||\mathbf{P}_{S_1^\perp}|||\right) \leq \tau\mu.$$

Additionally, note that

$$\mathcal{P}_{\mathbf{A}}(\mathbf{M}) = \mathbf{M} - (\mathbf{I} - \mathbf{P}_{S_1})\mathbf{M}(\mathbf{I} - \mathbf{P}_{S_2}) = \mathbf{P}_{S_1^\perp}\mathbf{M}\mathbf{P}_{S_2} + \mathbf{P}_{S_1}\mathbf{M},$$

and because $\operatorname{rank}(\mathbf{P}_{S_j}) \leq \operatorname{rank}(\mathbf{A})$, $j = 1, 2$ and $|||\mathbf{A}'|||_2^2 \leq \operatorname{rank}(\mathbf{A}')|||\mathbf{A}'|||^2$, for any matrix $\mathbf{A}'$, we get that

$$\begin{aligned}\Lambda &\leq 2\mu\sqrt{\operatorname{rank}(\mathcal{P}_{\mathbf{A}}(\mathbf{M}))}|||\mathcal{P}_{\mathbf{A}}(\mathbf{M})||| \\ &\leq 2\mu\sqrt{\operatorname{rank}(\mathbf{P}_{S_1^\perp}\mathbf{M}\mathbf{P}_{S_2} + \mathbf{P}_{S_1}\mathbf{M})}|||\mathbf{P}_{S_1^\perp}\mathbf{M}\mathbf{P}_{S_2} + \mathbf{P}_{S_1}\mathbf{M}||| \\ &\leq 4\mu\sqrt{2\operatorname{rank}(\mathbf{A})}|||\mathbf{M}||| \\ &\leq \tau\mu\sqrt{2\operatorname{rank}(\mathbf{A})},\end{aligned}$$

where we used that $\operatorname{rank}(\mathbf{A}'+\mathbf{A}'') \leq \operatorname{rank}(\mathbf{A}')+\operatorname{rank}(\mathbf{A}')$ and $\operatorname{rank}(\mathbf{A}'\mathbf{A}'') \leq \min\{\operatorname{rank}(\mathbf{A}'), \operatorname{rank}(\mathbf{A}'')\}$ for any matrices $\mathbf{A}'$ and $\mathbf{A}''$, and the triangle inequality.

Also, by Cauchy-Schwarz inequality, $|||\mathbf{A}'|||_1 \leq \sqrt{\operatorname{rank}(\mathbf{A}')}|||\mathbf{A}'|||_2$ for any matrix $\mathbf{A}'$, so we have that

$$\begin{aligned}|||\mathbf{P}_{S_1}(\widehat{\mathbf{A}}^{\tau} - \mathbf{A})\mathbf{P}_{S_2}|||_1 &\leq \sqrt{\operatorname{rank}(\mathbf{A})}|||\mathbf{P}_{S_1}(\widehat{\mathbf{A}}^{\tau} - \mathbf{A})\mathbf{P}_{S_2}|||_2 \\ &\leq \sqrt{\operatorname{rank}(\mathbf{A})}|||\widehat{\mathbf{A}}^{\tau} - \mathbf{A}|||_2.\end{aligned}$$

In the first inequality we use that $\operatorname{rank}(\mathbf{A}'\mathbf{A}'') \leq \min\{\operatorname{rank}(\mathbf{A}'), \operatorname{rank}(\mathbf{A}'')\}$, and in the last inequality we used Proposition A.5.12.

Part IV. Therefore, by (6.7) and the fact that $\tau \geq 4|||\mathbf{M}|||$, we arrive at

$$\begin{aligned}\mu|||\widehat{\mathbf{A}}^{\tau} - \mathbf{A}_0|||_2^2 + \mu|||\widehat{\mathbf{A}}^{\tau} - \mathbf{A}|||_2^2 + \tau|||\mathbf{P}_{S_1^\perp}\widehat{\mathbf{A}}^{\tau}\mathbf{P}_{S_2^\perp}|||_1 \\ \leq \mu|||\mathbf{A} - \mathbf{A}_0|||_2^2 + \tau\sqrt{\operatorname{rank}(\mathbf{A})}|||\widehat{\mathbf{A}}^{\tau} - \mathbf{A}|||_2 \\ + \tau\mu\sqrt{2\operatorname{rank}(\mathbf{A})}|||\widehat{\mathbf{A}}^{\tau} - \mathbf{A}|||_2 + \tau\mu|||\mathbf{P}_{S_1^\perp}\widehat{\mathbf{A}}^{\tau}\mathbf{P}_{S_2^\perp}|||_1.\end{aligned} \tag{6.8}$$

Since $\mu \leq 1$, we have that

$$\tau\mu\sqrt{2\operatorname{rank}(\mathbf{A})}|||\widehat{\mathbf{A}}^{\tau} - \mathbf{A}|||_2 \leq \tau\sqrt{2\operatorname{rank}(\mathbf{A})}|||\widehat{\mathbf{A}}^{\tau} - \mathbf{A}|||_2. \tag{6.9}$$

Rearranging terms and multiplying by μ^{-1} in (6.8) we obtain that

$$\begin{aligned}|||\widehat{\mathbf{A}}^{\tau} - \mathbf{A}_0|||_2^2 + \tau(\mu^{-1} - 1)|||\mathbf{P}_{S_1^\perp}\widehat{\mathbf{A}}^{\tau}\mathbf{P}_{S_2^\perp}|||_1 \leq |||\mathbf{A} - \mathbf{A}_0|||_2^2 \\ + \mu^{-1}\tau\sqrt{\operatorname{rank}(\mathbf{A})}(\sqrt{2}+1)|||\widehat{\mathbf{A}}^{\tau} - \mathbf{A}|||_2 - |||\widehat{\mathbf{A}}^{\tau} - \mathbf{A}|||_2^2.\end{aligned} \tag{6.10}$$

Note that we used inequality (6.9). Since the function $x \mapsto \alpha x - x^2$, $\alpha > 0$, is dominated by $\alpha^2/4$, we get that

$$\mu^{-1}\tau\sqrt{\operatorname{rank}(\mathbf{A})}(\sqrt{2}+1)|||\widehat{\mathbf{A}}^{\tau} - \mathbf{A}|||_2 - |||\widehat{\mathbf{A}}^{\tau} - \mathbf{A}|||_2^2 \leq \mu^{-2}\tau^2\operatorname{rank}(\mathbf{A})(\sqrt{2}+1)^2\frac{1}{4},$$

and finally, since

$$|||\widehat{\mathbf{A}}^{\tau} - \mathbf{A}_0|||_2^2 \leq |||\widehat{\mathbf{A}}^{\tau} - \mathbf{A}_0|||_2^2 + \tau(\mu^{-1} - 1)|||\mathbf{P}_{S_1^\perp}\widehat{\mathbf{A}}^{\tau}\mathbf{P}_{S_2^\perp}|||_1,$$

we conclude from (6.10) that

$$|||\widehat{\mathbf{A}}^{\tau} - \mathbf{A}_0|||_2^2 \leq |||\mathbf{A} - \mathbf{A}_0|||_2^2 + \mu^{-2}\tau^2\operatorname{rank}(\mathbf{A})\left(\frac{1+\sqrt{2}}{2}\right)^2.$$

Multiplying by μ and taking infimum on $\mathbb{A}$ ends the proof. □

With the aid of Lemma 6.3.2, we just need to find a bound for $\mathbb{P}(\tau \geq 4|||\mathbf{M}|||)$ in order to prove Theorem 6.3.1. Before doing it, we need a technical straightforward lemma.

Lemma 6.3.3. *Define σ^2 as*

$$\sigma^2 = (m_1 m_2)^2 \max\left\{|||\mathbb{E}\left[Y^2 \mathbf{X}\mathbf{X}^\intercal\right]|||,\, |||\mathbb{E}\left[Y^2 \mathbf{X}^\intercal \mathbf{X}\right]|||\right\}.$$

If ξ is independent of $\mathbf{X}$, then,

$$\sigma^2 \leq (m_1 m_2)^2 (\mathrm{Var}\xi \vee |||\mathbf{B}_0|||_{\max}^2) \frac{2}{m_1 \wedge m_2} = 2 m_1 m_2 (m_1 \vee m_2)(\mathrm{Var}\xi \vee |||\mathbf{B}_0|||_{\max}^2)$$

Proof. By definition of Y,

$$\begin{aligned}\mathbb{E}\left[Y^2 \mathbf{X}\mathbf{X}^\intercal\right] &= \mathbb{E}\left[(\mathrm{tr}\,(\mathbf{X}^\intercal \mathbf{B}_0) + \xi)^2 \mathbf{X}\mathbf{X}^\intercal\right] \\ &= \mathbb{E}\left[\xi^2 \mathbf{X}\mathbf{X}^\intercal\right] + \mathbb{E}\left[\mathrm{tr}\,(\mathbf{X}^\intercal \mathbf{B}_0)^2 \mathbf{X}\mathbf{X}^\intercal\right].\end{aligned}$$

Observe that $\mathrm{tr}\,(\mathbf{X}^\intercal \mathbf{B}_0) \leq |||\mathbf{B}_0|||_{\max}$ and

$$\begin{aligned}\mathbb{E}\mathbf{X}\mathbf{X}^\intercal &= \frac{1}{m_1 m_2} \sum_{j=1}^{m_1} \sum_{k=1}^{m_2} \boldsymbol{e}_j(m_1) \boldsymbol{e}_k^\intercal(m_2) \boldsymbol{e}_k(m_2) \boldsymbol{e}_j^\intercal(m_1) \\ &= \frac{1}{m_1} \sum_{j=1}^{m_1} \boldsymbol{e}_j(m_1) \boldsymbol{e}_j^\intercal(m_1) \\ &= \frac{1}{m_1} \mathbf{I},\end{aligned}$$

so $|||\mathbb{E}\mathbf{X}\mathbf{X}^\intercal||| = 1/m_1$. By independence of ξ and $\mathbf{X}$, $\mathbb{E}\left[\xi^2 \mathbf{X}\mathbf{X}^\intercal\right] = (\mathrm{Var}\xi)(\mathbb{E}\mathbf{X}\mathbf{X}^\intercal)$. Therefore,

$$|||\mathbb{E}\left[Y^2 \mathbf{X}\mathbf{X}^\intercal\right]||| \leq \frac{1}{m_1} \mathrm{Var}\xi + \frac{1}{m_1} |||\mathbf{B}_0|||_{\max}.$$

In the same way,

$$|||\mathbb{E}\left[Y^2 \mathbf{X}^\intercal \mathbf{X}\right]||| \leq \frac{1}{m_2} \mathrm{Var}\xi + \frac{1}{m_2} |||\mathbf{B}_0|||_{\max}.$$

This ends the proof of the lemma. □

Now we are ready to prove the main result.

Proof of Theorem 6.3.1. By Theorem A.3.5

$$\|\|\mathcal{H}(\widehat{\mathbf{R}}^{\tau}) - \mathcal{H}(\mathbf{B}_0)\|\|_2^2 = \|\|\mathcal{H}(\widehat{\mathbf{R}}^{\tau} - \mathbf{B}_0)\|\|_2^2 = 2\|\|\widehat{\mathbf{R}}^{\tau} - \mathbf{B}_0\|\|_2^2.$$

Also $\operatorname{rank}(\mathcal{H}(\mathbf{B})) = 2\operatorname{rank}(\mathbf{B})$, so

$$\inf_{\mathbf{A}\in\mathbb{A}} \left\{ \frac{1}{m_1 m_2} \|\|\mathbf{A} - \mathcal{H}(\mathbf{B}_0)\|\|_2^2 + \left(\frac{1+\sqrt{2}}{2}\right)^2 m_1 m_2 \tau^2 \operatorname{rank}(\mathbf{A}) \right\}$$
$$= 2 \inf_{\mathbf{B}\in\mathcal{M}_{m_1,m_2}} \left\{ \frac{1}{m_1 m_2} \|\|\mathbf{B} - \mathbf{B}_0\|\|_2^2 + 2\left(\frac{1+\sqrt{2}}{2}\right)^2 m_1 m_2 \tau^2 \operatorname{rank}(\mathbf{B}) \right\}.$$

Then, Lemma 6.3.2 imply that whenever $\tau \geq 4\|\|\mathbf{M}\|\|$,

$$\|\|\widehat{\mathbf{R}}^{\tau} - \mathbf{B}_0\|\|_2^2 \leq \inf_{\mathbf{B}\in\mathcal{M}_{m_1,m_2}} \left\{ \frac{1}{m_1 m_2} \|\|\mathbf{B} - \mathbf{B}_0\|\|_2^2 + 2\left(\frac{1+\sqrt{2}}{2}\right)^2 m_1 m_2 \tau^2 \operatorname{rank}(\mathbf{B}) \right\}.$$

Now we just need to verify that for the specified range of τ we have that $\mathbb{P}(\tau \geq 4\|\|\mathbf{M}\|\|) \geq 1 - e^{-t}$, $t \geq 0$.

Recall that $\mathbf{M}$ is defined as

$$\mathbf{M} = \frac{1}{n\theta} \sum_{j=1}^{n} \psi(\theta m_1 m_2 Y_j \mathbf{X}_j) - \mathbb{E}\left[m_1 m_2 Y \mathbf{X}\right].$$

This is of the same form $\mathbf{T} - \mathbb{E}\mathbf{Y}$ of Theorem 4.3.2. Observe that by Lemma A.5.2,

$$\mathbb{E}\|\|(m_1 m_2)^2 Y^2 \mathcal{H}(\mathbf{X})^2\|\| = \sigma^2,$$

where σ^2 is defined in Lemma 6.3.3. Now, by Remark 4.3.2, taking

$$t \mapsto t + \log(2(m_1 + m_2)) \quad \text{and} \quad s = \sqrt{2 m_1 m_2 (m_1 \vee m_2)(\operatorname{Var}\xi \vee \|\|\mathbf{B}_0\|\|_{\max}^2)},$$

we get that

$$\mathbb{P}\left(\|\|\mathbf{M}\|\| \leq 2\sqrt{m_1 m_2 (m_1 \vee m_2)} (\|\|\mathbf{B}_0\|\|_{\max} \vee \sqrt{\operatorname{Var}\xi}) \sqrt{\frac{t + \log(2(m_1 + m_2))}{n}} \right) \geq 1 - e^{-t},$$

where the parameter θ is

$$\theta = \sqrt{\frac{2t}{n}\frac{1}{s}} = \sqrt{\frac{t+\log(2(m_1+m_2))}{nm_1m_2(m_1\vee m_2)(\mathrm{Var}\xi \vee \|\mathbf{B}_0\|_{\max}^2)}}.$$

This end the proof since,

$$\mathbb{P}\left(\tau \geq 4\|\mathbf{M}\|\right) \geq \mathbb{P}\left(4\|\mathbf{M}\| \leq 8m_1m_2(m_1\vee m_2)(\|\mathbf{B}_0\|_{\max} \vee \sqrt{\mathrm{Var}\xi})\sqrt{\frac{t+\log(2(m_1+m_2))}{n(m_1\wedge m_2)}}\right).$$

□

Now that we have a concentration inequality for $\widehat{\mathbf{R}}^{\tau}$ we can proceed to find a method of calculation. This is the topic of the next section.

6.4 Approximate computation of the estimator

Unlike [22], the estimator $\widehat{\mathbf{R}}^{\tau}$ has no explicit form due to the incorporation of the matrix dilation that helped us obtain a robust concentration bound. Nevertheless, the work done in [33] motivates an algorithmic procedure to calculate $\widehat{\mathbf{R}}^{\tau}$. Even more, the method of [33] allow us to calculate a more general estimator that will define as

$$\widehat{\mathbf{R}}_p^{\tau} = \underset{\mathbf{B}\in\mathcal{M}_{m_1,m_2}}{\arg\min}\left\{\frac{1}{m_1m_2}\|\mathcal{H}(\mathbf{B}) - \widehat{\mathbf{R}}\|_2^2 + 2\tau\|\mathbf{B}\|_p^p\right\}, \quad p\in(0,1]. \tag{6.11}$$

Note that $\widehat{\mathbf{R}}^{\tau} = \widehat{\mathbf{R}}_1^{\tau}$. This estimator is promising because taking $p<1$ let us to get an estimator that has a more accurate penalization that resembles $\mathrm{rank}(\mathbf{B})$. Unfortunately, when $p<1$ we can not obtain a concentration guarantee with the same method of Lemma 6.3.2 since the function $\mathbf{B}\to\|\mathbf{B}\|_p$ turns out to be not convex and we can not define a subdifferential. However, being able to calculate $\widehat{\mathbf{R}}_p^{\tau}$ for any $p\in(0,1]$ can be a benefit for future research.

The optimization problem (6.11) can be rewritten as

$$\min_{\substack{\mathbf{B},\mathbf{A}\\ \mathbf{A}=\mathcal{H}(\mathbf{B})}}\left\{\frac{1}{m_1m_2}\|\mathbf{A}-\widehat{\mathbf{R}}\|_2^2 + \tau\|\mathbf{A}\|_p^p\right\}. \tag{6.12}$$

```
Input: $\rho \in (1,2)$ and some initializations $\mu > 0$ and $\boldsymbol{\Omega} \in \mathbb{R}^n$.
Output: An $\boldsymbol{x}^* \in \mathbb{R}^n$ that minimizes (6.13).

While not converge do
    1. Update $\boldsymbol{x} \leftarrow \min_{\boldsymbol{x}\in\mathbb{R}^n} \left\{ f(\boldsymbol{x}) + \frac{\mu}{2}\|h(\boldsymbol{x}) + \frac{1}{\mu}\boldsymbol{\Omega}\|_2^2 \right\}$.
    2. Update $\boldsymbol{\Omega} \leftarrow \boldsymbol{\Omega} + \mu h(\boldsymbol{x})$.
    3. Update $\mu \leftarrow \rho\mu$.
End while.
```

Figure 6.1: ALM algorithm to solve (6.13).

This last problem is similar to the problem

$$\min_{\substack{\boldsymbol{x}\in\mathbb{R}^n \\ h(\boldsymbol{x})=0}} f(\boldsymbol{x}), \tag{6.13}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is the objective function and $h : \mathbb{R}^n \to \mathcal{V} \subset \mathbb{R}^n$ is some constraint. To solve problem (6.13), in [33] the authors proposed to use the Augmented Lagrangian Method (ALM) described in Figure 6.1. Of course, as pointed out in [33], this procedure can be applied to the matrix case substituting the ℓ_2 vector norm by the 2-Schatten matrix norm, i.e., the Frobenius norm.

In the case at hand we have $f : \mathcal{M}_{m_1,m_2} \times \mathcal{S}_{m_1+m_2} \to \mathbb{R}$ defined as

$$f(\mathbf{B}, \mathbf{A}) = \frac{1}{m_1 m_2} \|\!|\mathbf{A} - \widehat{\mathbf{R}}\|\!|_2^2 + \tau \|\!|\mathbf{A}\|\!|_p^p,$$

and $h : \mathcal{M}_{m_1,m_2} \times \mathcal{S}_{m_1+m_2} \to \mathcal{S}_{m_1+m_2}$ defined as

$$h(\mathbf{B}, \mathbf{A}) = \mathbf{A} - \mathcal{H}(\mathbf{B}).$$

For problem (6.12) the step one of ALM algorithm is

$$\min_{\mathbf{B},\mathbf{A}} g(\mathbf{B}, \mathbf{A}), \tag{6.14}$$

where,

$$g(\mathbf{B},\mathbf{A}) = \frac{1}{m_1m_2}|||\mathbf{A}-\widehat{\mathbf{R}}|||_2^2 + \tau|||\mathbf{A}|||_p^p + \frac{\mu}{2}\left|\left|\left|h(\mathbf{B},\mathbf{A})+\frac{1}{\mu}\mathbf{\Omega}\right|\right|\right|_2^2,$$

where $\mathbf{\Omega} \in \mathcal{M}_{m_1+m_2}$. Up to a constant factor we get with a little a algebra that

$$\begin{aligned} g(\mathbf{B},\mathbf{A}) = &\left(\frac{1}{m_1m_2}+\frac{\mu}{2}\right)|||\mathbf{A}|||_2^2 - \frac{2}{m_1m_2}\langle\mathbf{A},\widehat{\mathbf{R}}\rangle + \tau|||\mathbf{A}|||_p^p \\ &+ \frac{\mu}{2}|||\mathcal{H}(\mathbf{B})|||_2^2 + \frac{1}{2}\langle\mathcal{H}(\mathbf{B}),\mathbf{\Omega}\rangle + \mu\langle\mathbf{A},\mathcal{H}(\mathbf{B})-\mu^{-1}\mathbf{\Omega}\rangle. \end{aligned}$$

As in [33] we use Alternating Direction Method (ADM) to solve (6.14). First, fixing $\mathbf{A}$ we need to solve

$$\min_{\mathbf{B}\in\mathcal{M}_{m_1,m_2}}\left\{|||\mathcal{H}(\mathbf{B})|||_2^2 + \langle\mathcal{H}(\mathbf{B}),\mu^{-1}\mathbf{\Omega}+2\mathbf{A}\rangle\right\} = \min_{\mathbf{B}\in\mathcal{M}_{m_1,m_2}}|||\mathcal{H}(\mathbf{B})+(2\mu)^{-1}\mathbf{\Omega}+\mathbf{A}|||_2^2 \tag{6.15}$$

By an analogous procedure to Proposition C.9.1, we can verify that for any $\mathbf{C}\in\mathcal{M}_{m_1+m_2}$, the function $\mathbf{B}\mapsto|||\mathcal{H}(\mathbf{B})-\mathbf{C}|||_2^2$ is strictly convex. Then, there exist a unique minimum and we can find it by differentiation. Let $\mathbf{C}_{12}\in\mathcal{M}_{m_1,m_2}$ and $\mathbf{C}_{21}\in\mathcal{M}_{m_2,m_1}$ be such that

$$\mathbf{C} = \begin{pmatrix} \star & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \star \end{pmatrix}.$$

Then,

$$|||\mathcal{H}(\mathbf{B})|||_2^2 = 2|||\mathbf{B}|||_2^2 = 2\mathrm{tr}\left(\mathbf{B}^\intercal\mathbf{B}\right)$$

and

$$\langle\mathcal{H}(\mathbf{B}),\mathbf{C}\rangle = \mathrm{tr}\left(\mathbf{B}\mathbf{C}_{21}\right) + \mathrm{tr}\left(\mathbf{B}^\intercal\mathbf{C}_{12}\right).$$

Therefore, by the matrix derivation techniques presented in [36],

$$\frac{\partial}{\partial\mathbf{B}}|||\mathcal{H}(\mathbf{B})-\mathbf{C}|||_2^2 = 4\mathbf{B}-2\mathbf{C}_{21}^\intercal-2\mathbf{C}_{12}.$$

So the value of $\mathbf{B}$ that minimizes (6.15) is

$$\mathbf{B}_{\text{opt}} = \frac{1}{2}\left(\mathbf{C}_{21}^\intercal+\mathbf{C}_{12}\right) \tag{6.16}$$

where in this case

$$\mathbf{C} = -(2\mu)^{-1}\mathbf{\Omega} - \mathbf{A}.$$

Now, fixing $\mathbf{B}$ we need to solve

$$\min_{\mathbf{A}\in\mathcal{S}_{m_1+m_2}} \left\{ \frac{1}{2}|||\mathbf{A}-\mathbf{Q}|||_2^2 + \gamma|||\mathbf{A}|||_p^p \right\}, \tag{6.17}$$

where

$$\mathbf{Q} = \left(\frac{1}{m_1m_2} + \frac{\mu}{2}\right)^{-1/2} \left(\frac{1}{m_1m_2}\widehat{\mathbf{R}} + \frac{1}{2}\mathbf{\Omega} - \frac{\mu}{2}\mathcal{H}(\mathbf{B})\right), \tag{6.18}$$

$$\gamma = \frac{\tau}{2}\left(\frac{1}{m_1m_2} + \frac{\mu}{2}\right)^{-\frac{p+1}{2}}.$$

The restriction $\mathbf{A} \in \mathcal{S}_{m_1+m_2}$ in (6.17) is a major drawback since there is no standard method to solve this problem. As we show in a few lines, this is provoked by the fact that $\mathbf{Q}$ is not necessarily symmetric. To keep things simple, we relax (6.17) and instead solve

$$\mathbf{A}_* = \operatorname*{arg\,min}_{\mathbf{A}\in\mathcal{M}_{m_1+m_2}} \left\{ \frac{1}{2}|||\mathbf{A}-\mathbf{Q}|||_2^2 + \gamma|||\mathbf{A}|||_p^p \right\}, \tag{6.19}$$

Fortunately, problem (6.19) has a unique solution. Before presenting this solution, to get a symmetric result we choose the symmetric matrix that is closest to $\mathbf{A}_*$ in Frobenius norm, i.e., we choose as the optimum

$$\mathbf{A}_{\text{opt}} = \operatorname*{arg\,min}_{\mathbf{A}\in\mathcal{S}_{m_1+m_2}} |||\mathbf{A}-\mathbf{A}_*|||_2^2.$$

By Proposition C.11.1, this problem has a unique solution given by

$$\mathbf{A}_{\text{opt}} = \left(\frac{1}{2}\mathbf{J} + \frac{1}{2}\mathbf{I}\right) \circ \left(\mathbf{A}_* + \mathbf{A}_*^\intercal + \mathbf{A}_* \circ \mathbf{I}\right), \tag{6.20}$$

where $\circ$ is the Hadamard or entry-wise product and $\mathbf{J}$ is the all ones matrix.

Now, we need to obtain $\mathbf{A}_*$. Let $\mathbf{Q} = \mathbf{U}\mathbf{D}\mathbf{V}^\intercal$ be a full SVD of $\mathbf{Q}$ where $\mathbf{D} = \operatorname{diag}(s_1, ..., s_{m_1+m_2})$ and $s_1 \geq \cdots \geq s_{m_1+m_2}$ are the singular values of $\mathbf{Q}$. For $p = 1$ the problem (6.19) has a unique minimum attained at

$$\mathbf{A}_* = \mathbf{U}T_\gamma(\mathbf{D})\mathbf{V}^\intercal,$$

where

$$T_\gamma(\mathbf{D}) = \operatorname{diag}\left(\max\left\{0, s_1 - \gamma\right\}, ..., \max\left\{0, s_{m_1+m_2} - \gamma\right\}\right).$$

A proof of this fact can be consulted in [5]. As we mentioned, $\mathbf{A}_*$ is not necessarily symmetric since in general $\mathbf{U} \neq \mathbf{V}$ for squared or even symmetric matrices. In [33] the authors prove more generally that (6.19) has a unique solution for any $p \in (0, 1]$. For $p < 1$ this solution is not explicit, but straightforward to find with a simple root finder like the Newton method. To state the result, lets define some functions and quantities:

$$\begin{aligned} H(x; a) &= \frac{1}{2}(x - a)^2 + \gamma |x|^p, \\ G(x; a) &= x - a + \gamma p |x|^{p-1} \operatorname{sgn}(x), \\ v &= \left[\gamma p(1-p)\right]^{\frac{1}{2-p}} + \gamma p \left[\gamma p(1-p)\right]^{\frac{p-1}{2-p}}. \end{aligned}$$

For $a > v$ denote $x(a)$ as the root of $G(x; a)$ on the interval (v, a), i.e., $G(x(a); a) = 0$ and $x(a) \in (v, a)$. This root can be found with the Newton method.

Theorem 6.4.1. *Recall the SVD* $\mathbf{Q} = \mathbf{U}\mathbf{D}\mathbf{V}^\intercal$*. For* $p \in (0, 1]$ *the unique solution of problem* (6.19) *is*

$$\mathbf{A}_* = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^\intercal,$$

where $\boldsymbol{\Delta} = \operatorname{diag}(\delta_1, ..., \delta_{m_1+m_2})$ *and*

$$\delta_j = \begin{cases} 0, & s_j \leq v \\ \arg\min_{x \in \{0, x(s_j)\}} H(x; s_j), & s_j > v. \end{cases}$$

As in [33] one can use for the convergence criteria the function

$$C(k, k-1) = \frac{|||\mathbf{R}^{(k)} - \mathbf{R}^{(k-1)}|||_2}{\max\left\{1, |||\mathbf{R}^{(k)}|||_2\right\}}, \quad k \geq 1,$$

and stop the process when $C(k, k-1)$ is less than some tolerance. The algorithm to find $\widehat{\mathbf{R}}^\tau$ is presented in Figure 6.2.

```
**Input**: $\rho \in (1,2)$, $\tau > 0$, a tolerance $\epsilon$ and some initializations $\mu > 0$, $\mathbf{A} \in \mathcal{S}_{m_1+m_2}$ and $\boldsymbol{\Omega} \in \mathcal{M}_{m_1+m_2}$.
**Output**: Robust estimator $\widehat{\mathbf{R}}^{\tau} = \widehat{\mathbf{R}}_1^{\tau}$.

$k = -1$; $C(0,-1) = C(-1,-2) = \epsilon + 1$
**While** $C(k, k-1) > \epsilon$ **do**
    1. Update $\mathbf{B} \leftarrow \frac{1}{2}\left(\mathbf{C}_{21}^{\intercal} + \mathbf{C}_{12}\right)$, where $\mathbf{C} = -(2\mu)^{-1}\boldsymbol{\Omega} - \mathbf{A}$.
    2. Calculate the SVD $\mathbf{Q} = \mathbf{U}\mathbf{D}\mathbf{V}^{\intercal}$ for $\mathbf{Q}$ defined in (6.18) and take $\mathbf{A}_* = \mathbf{U}T_{\gamma}(\mathbf{D})\mathbf{V}^{\intercal}$.
    3. Update $\mathbf{A} \leftarrow \left(\frac{1}{2}\mathbf{J} + \frac{1}{2}\mathbf{I}\right) \circ \left(\mathbf{A}_* + \mathbf{A}_*^{\intercal} + \mathbf{A}_* \circ \mathbf{I}\right)$.
    4. Update $\boldsymbol{\Omega} \leftarrow \boldsymbol{\Omega} + \mu\left(\mathbf{A} - \mathcal{H}(\mathbf{B})\right)$ and $\mu \leftarrow \rho\mu$.
    5. $k \leftarrow k+1$; $\mathbf{R}^{(k)} = \mathbf{B}$.
**End while**.
**Return** $\mathbf{R}^{(k)}$.
```

Figure 6.2: Calculation of $\widehat{\mathbf{R}}_p^{\tau}$ defined in (6.11) for the case $p = 1$.

Summary and observations

In this chapter we gave an application of the robust methodology of this thesis that differs broadly from what we have presented in previous chapter since the form of the estimator is not explicit and we use non-trivial methods to obtain concentration results. The problem of matrix completion consisted in the estimation of a matrix that is partially observed with noise. To do so, the estimator used in the literature finds the closest matrix in Frobenius norm with reduced rank, where this is reflected through nuclear norm penalization. We gave an overview of the work of Koltchinskii et al. and presented the robust methodology from Minsker. The main theorem (Theorem 6.3.1) provided us with useful information to choose the penalization parameter τ. Finally, we gave a novel approach to approximately compute the estimator $\widehat{\mathbf{R}}_p^{\tau}$, where we are able to penalize with any p-Schatten norm with $p \in (0, 1]$ obtaining a better approximation to the rank.

References

[1] Theodore Wilbur Anderson. *An Introduction to Multivariate Statistical Analysis.* 3rd edition, 2003.

[2] Sylvain Arlot, Alain Celisse, et al. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

[3] Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 1997.

[4] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence.* Oxford university press, 2013.

[5] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982, 2010.

[6] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

[7] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

[8] Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[9] Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.

[10] R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001.

[11] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[12] Victor De la Peña and Evarist Giné. *Decoupling: from dependence to independence.* Springer-Verlag New York, 1999.

[13] Anders Eklund, Thomas E Nichols, and Hans Knutsson. Cluster failure: Why fmri inferences for spatial extent have inflated false-positive rates. *Proceedings of the national academy of sciences*, 113(28):7900–7905, 2016.

[14] Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics reports*, 659:1–44, 2016.

[15] Christophe Giraud. *Introduction to high-dimensional statistics*, volume 138. CRC Press, 2014.

[16] Allan Gut. *An Intermediate Course in Probability.* Springer-Verlag New York, 2nd edition, 2009.

[17] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

[18] Roger A Horn and Charles R Johnson. *Matrix analysis.* Cambridge university press, 2012.

[19] Peter J Huber and Elvezio M Ronchetti. *Robust statistics.* John Wiley & Sons, 2nd edition, 2009.

[20] Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327, 2001.

[21] Yuan Ke, Stanislav Minsker, Zhao Ren, Qiang Sun, and Wen-Xin Zhou. User-friendly covariance estimation for heavy-tailed distributions. *Statistical Science*, 34(3):454–471, 2019.

[22] Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

[23] Samuel Kotz and Saralees Nadarajah. *Multivariate t-distributions and their applications*. Cambridge University Press, 2004.

[24] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2002.

[25] OV Lepskii. Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications*, 36(4):682–697, 1992.

[26] Li Liu, Douglas M Hawkins, Sujoy Ghosh, and S Stanley Young. Robust singular value decomposition analysis of microarray data. *Proceedings of the National Academy of Sciences*, 100(23):13167–13172, 2003.

[27] Karim Lounici et al. High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058, 2014.

[28] Gábor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. *The annals of statistics*, 47(2):783–794, 2019.

[29] Gabor Lugosi and Shahar Mendelson. Risk minimization by median-of-means tournaments. *Journal of the European Mathematical Society*, 22:925–965, 2020.

[30] Gabor Lugosi, Shahar Mendelson, et al. Regularization, sparse recovery, and median-of-means tournaments. *Bernoulli*, 25(3):2075–2106, 2019.

[31] Per-Gunnar Martinsson and Joel Tropp. Randomized numerical linear algebra: Foundations & algorithms. *arXiv preprint arXiv:2002.01387*, 2020.

[32] Stanislav Minsker. Sub-gaussian estimators of the mean of a random matrix with heavy-tailed entries. *The Annals of Statistics*, 46(6A):2871–2903, 2018.

[33] Feiping Nie, Hua Wang, Heng Huang, and Chris Ding. Joint schatten p-norm and ℓ_p-norm robust matrix comletion for missing value recovery. *Knowledge and Information Systems*, 42(3):525–544, 2013.

[34] Roberto Oliveira. Sums of random hermitian matrices and an inequality by rudelson. *Electronic Communications in Probability*, 15:203–212, 2010.

[35] Roberto Imbuzeiro Oliveira. Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges. *arXiv preprint arXiv:0911.0600*, 2009.

[36] KB Petersen and MS Pedersen. The matrix cookbook, vol. 7. *Technical University of Denmark*, 15, 2008.

[37] Mohsen Pourahmadi. *High-dimensional covariance estimation: with high-dimensional data*, volume 882. John Wiley & Sons, 2013.

[38] Angelika Rohde and Alexandre B Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

[39] Robert J Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

[40] Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

[41] Jun Tong, Rui Hu, Jiangtao Xi, Zhitao Xiao, Qinghua Guo, and Yanguang Yu. Linear shrinkage estimation of covariance matrices using low-complexity cross-validation. *Signal Processing*, 148:223–233, 2018.

[42] Lloyd N Trefethen and David Bau III. *Numerical linear algebra*, volume 50. Siam, 1997.

[43] Joel Tropp. From joint convexity of quantum relative entropy to a concavity theorem of lieb. *Proceedings of the American Mathematical Society*, 140(5):1757–1760, 2012.

[44] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

[45] Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

[46] Hoang Tuy. *Convex analysis and global optimization.* Springer International Publishing, 2nd edition, 2016.

[47] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[48] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[49] Lili Wang, Chao Zheng, Wen Zhou, and Wen-Xin Zhou. A new principle for tuning-free huber regression. *Preprint*, 2018.

[50] John Watrous. *The theory of quantum information.* Cambridge university press, 2018.

[51] G Alistair Watson. Characterization of the subdifferential of some matrix norms. *Linear algebra and its applications*, 170:33–45, 1992.

[52] Jianfeng Yao, Shurong Zheng, and ZD Bai. *Sample covariance matrices and high-dimensional data analysis.* Cambridge University Press Cambridge, 2015.

[53] Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

APPENDICES

Appendix A

Some matrix analysis results

A.1 Definition of symmetric matrix operator

Let $\mathbf{A}$ be a $p \times p$ symmetric matrix with spectral decomposition

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^\intercal.$$

Suppose that $\lambda_j(\mathbf{A}) \in \mathcal{C} \subset \mathbb{R}$ for all j, and that $f : \mathcal{C} \to \mathbb{R}$ is a real-valued function. In Chapter 2 we defined the matrix $f(\mathbf{A})$ as

$$f(\mathbf{A}) = \mathbf{U}f(\mathbf{D})\mathbf{U}^\intercal,$$

where

$$f(\mathbf{D}) = \begin{pmatrix} f(\lambda_1(\mathbf{A})) & 0 & \cdots & 0 \\ 0 & f(\lambda_2(\mathbf{A})) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f(\lambda_p(\mathbf{A})) \end{pmatrix}.$$

To see that $f(\mathbf{A})$ is well defined[1], we first adopt the convention that for any polynomial $P(x) = \sum_j \beta_j x^j$, the matrix $P(\mathbf{A})$ is equal to $\sum_j \beta_j \mathbf{A}^j$.

By simplicity, denote $\lambda_j = \lambda_j(\mathbf{A})$. Also, without loss of generality suppose that $\lambda_i \neq \lambda_j$ for all $i \neq j$. Now, define the Lagrange polynomial interpolation L as

$$L(x) = \sum_{j=1}^{p} f(\lambda_j)\ell_j(x), \quad \ell_j(x) = \prod_{k \neq j} \frac{x - \lambda_k}{\lambda_j - \lambda_k}.$$

[1] The matrix $\mathbf{U}$ is not unique.

It is easy to see that

$$L(\lambda_i) = f(\lambda_i), \quad \forall i,$$

so

$$L(\mathbf{D}) = f(\mathbf{D}),$$

and

$$\mathbf{U}L(\mathbf{D})\mathbf{U}^\intercal = \mathbf{U}f(\mathbf{D})\mathbf{U}^\intercal.$$

On the other hand, L is a polynomial of degree p, so it has the form

$$L(x) = \sum_{j=0}^{p} \alpha_j x^j.$$

Then,

$$\begin{aligned}
\mathbf{U}L(\mathbf{D})\mathbf{U}^\intercal &= \sum_{j=0} \alpha_j \mathbf{U}\mathbf{D}^j\mathbf{U}^\intercal \\
&= \sum_{j=0}^{p} \alpha_j (\mathbf{U}\mathbf{D}\mathbf{U}^\intercal)\cdots(\mathbf{U}\mathbf{D}\mathbf{U}^\intercal) \\
&= \sum_{j=0}^{p} \alpha_j (\mathbf{U}\mathbf{D}\mathbf{U}^\intercal)^j \\
&= L(\mathbf{A}).
\end{aligned}$$

Therefore it doesn't matter the choice of $\mathbf{U}$, the matrix product $\mathbf{U}f(\mathbf{D})\mathbf{U}^\intercal$ will be equal to $L(\mathbf{A})$.

In the case that $\lambda_i = \lambda_j$ for some $i \neq j$, one can apply the same reasoning taking only different eigenvalues. And in the case $\lambda_1 = \cdots = \lambda_p$, we follow the same steps with the constant polynomial $L(x) = \alpha_0$.

A.2 Variational principles for eigenvalues

In this section we study some properties of the eigenvalues of symmetric matrices. Most of the results were taken from [3] and [18].

Fact A.2.1. *Let V_1 and V_2 be two subspaces of the vector space V, where $\dim(V) = p$. Then,*

$$\begin{aligned}\dim(V_1 \cap V_2) &= \dim(V_1) + \dim(V_2) - \dim(V_1 \oplus V_2)\\ &\geq \dim(V_1) + \dim(V_2) - p.\end{aligned}$$

From this, we can deduce that if V_3 is another subspace of V, then

$$\dim(V_1 \cap V_2 \cap V_3) \geq \dim(V_1) + \dim(V_2) + \dim(V_3) - 2p.$$

The next theorem and it's corollary were taken from [3, p. 58].

Theorem A.2.2 (Poncairé's inequality)**.** *For $\mathbf{A} \in \mathcal{S}_p$ and any subspace $W \subset \mathbb{R}^p$ such that $\dim(W) = k$, $1 \leq k \leq p$, there exit $\boldsymbol{x}, \boldsymbol{y} \in W$ unitary such that*

$$\boldsymbol{x}^\intercal \mathbf{A} \boldsymbol{x} \leq \lambda_k(\mathbf{A}) \quad \textit{and} \quad \boldsymbol{y}^\intercal \mathbf{A} \boldsymbol{y} \geq \lambda_{n-k+1}(\mathbf{A}).$$

Proof. Let $\boldsymbol{u}_j$ be the eigenvector associated to $\lambda_j(\mathbf{A})$ and suppose that $\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_p$ are orthonormal. Also let W^* be the subspace spanned by $\boldsymbol{u}_k, ..., \boldsymbol{u}_p$. Then $\dim(W) + \dim(W^*) = k + p - k + 1 = p + 1$ and by Fact A.2.1 we have that $\dim(W \cap W^*) \geq 1$ and $W \cap W^* \neq \emptyset$.

Now let $\boldsymbol{x} \in W \cap W^*$ then $\boldsymbol{x} \in W^*$ and for some reals $\alpha_k, ..., \alpha_p$ we can write

$$\boldsymbol{x} = \sum_{j=k}^{p} \alpha_j \boldsymbol{u}_j.$$

We can suppose that $\sum_{j=k}^{p} \alpha_j^2 = 1$. This is because the orthonormality of $\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_p$ implies that $\|\boldsymbol{x}\|_2^2 = \sum_{j=k}^{p} \alpha_j^2$ and by normalizing $\boldsymbol{x}$ we get that $\sum_{j=k}^{p} \alpha_j^2 = 1$. Hence,

$$\begin{aligned}\boldsymbol{x}^\intercal \mathbf{A} \boldsymbol{x} &= \langle \boldsymbol{x}, \mathbf{A}\boldsymbol{x} \rangle\\ &= \langle \sum_{j=k}^{p} \alpha_j \boldsymbol{u}_j,\ \mathbf{A} \sum_{j=k}^{p} \alpha_j \boldsymbol{u}_j \rangle\\ &= \langle \sum_{j=k}^{p} \alpha_j \boldsymbol{u}_j,\ \sum_{j=k}^{p} \alpha_j \lambda_j(\mathbf{A}) \boldsymbol{u}_j \rangle\\ &= \sum_{j=k}^{n} \alpha_j^2 \lambda_j(\mathbf{A})\end{aligned}$$

$$
\begin{aligned}
&\le \lambda_k(\mathbf{A}) \sum_{j=k}^{p} \alpha_j^2 \\
&= \lambda_k(\mathbf{A}).
\end{aligned}
$$

For the second inequality we proceed in the same way. Define $W^\dagger$ as the subspace spanned by $\boldsymbol{u}_1, ..., \boldsymbol{u}_{p-k+1}$. By the same arguments take $\boldsymbol{y} \in W \cap W^\dagger$ such that $\boldsymbol{y} = \sum_{j=1}^{p-k+1} \beta_j \boldsymbol{u}_j$ with $\sum_{j=1}^{p-k+1} \beta_j^2 = 1$. Then,

$$
\boldsymbol{y}^\intercal \mathbf{A} \boldsymbol{y} = \sum_{j=1}^{p-k+1} \beta_j^2 \lambda_j(\mathbf{A}) \ge \lambda_{p-k+1}(\mathbf{A}) \sum_{j=1}^{p-k+1} \alpha_j^2 = \lambda_{p-k+1}(\mathbf{A}).
$$

□

The poof of Poncaire's inequality illustrates how to verify the next fact: for every $\mathbf{A} \in \mathcal{S}_p$ and $\boldsymbol{x} \in \mathbb{R}^p$ such that $\|\boldsymbol{x}\|_2 = 1$ we have that

$$
\boldsymbol{x}^\intercal \mathbf{A} \boldsymbol{x} \le \lambda_1(\mathbf{A}). \tag{A.1}
$$

This is because there exist a orthonormal basis $\boldsymbol{v}_1, ..., \boldsymbol{v}_p$ of $\mathbb{R}^p$ formed by eigenvector of $\mathbf{A}$ and we can represent every normalized $\boldsymbol{x} \in \mathbb{R}^p$ as $\boldsymbol{x} = \sum_{j=1}^{p} \gamma_j \boldsymbol{v}_j$ with $\sum_{j=1}^{p} \gamma_j^2 = 1$. So, with the same technique as in the later proof, we get that

$$
\boldsymbol{x}^\intercal \mathbf{A} \boldsymbol{x} = \sum_{j=1}^{p} \gamma_j^2 \lambda_j(\mathbf{A}) \le \lambda_1(\mathbf{A}).
$$

Similarly, define $\boldsymbol{x} \in \{\boldsymbol{v}_{i_1}, ..., \boldsymbol{v}_{i_2}\}$, for $1 \le i_1 \le i_2 \le p$. We can write $\boldsymbol{x} = \sum_{j=i_1}^{i_2} \alpha_j \boldsymbol{v}_j$ with $\sum_{k=i_1}^{i_2} \alpha_k^2 = 1$. Then,

$$
\lambda_{i_2}(\mathbf{A}) \le \sum_{k=i_1}^{i_2} \alpha_k^2 \lambda_k(\mathbf{A}) = \boldsymbol{x}^\intercal \mathbf{A} \boldsymbol{x}, \tag{A.2}
$$

and in the same way,

$$
\boldsymbol{x}^\intercal \mathbf{A} \boldsymbol{x} = \sum_{k=i_1}^{i_2} \alpha_k^2 \lambda_k(\mathbf{A}) \le \lambda_{i_1}(\mathbf{A}). \tag{A.3}
$$

Corollary A.2.3 (Fischer-Courant min-max principle)**.** *For* $\mathbf{A} \in \mathcal{S}_p$ *and* $1 \leq k \leq p$

$$\lambda_k(\mathbf{A}) = \max_{\substack{W \subset \mathbb{R}^p \\ \dim(W)=k}} \min_{\substack{\boldsymbol{x} \in W \\ \|x\|_2=1}} \boldsymbol{x}^\intercal \mathbf{A} \boldsymbol{x} = \min_{\substack{W \subset \mathbb{R}^p \\ \dim(W)=p-k+1}} \max_{\substack{\boldsymbol{x} \in W \\ \|x\|_2=1}} \boldsymbol{x}^\intercal \mathbf{A} \boldsymbol{x},$$

where W *is taken as a subspace of* $\mathbb{R}^p$*.*

Proof. For any k-dimensional subspace W, the Poncaire's inequality implies that

$$\min_{\substack{\boldsymbol{x} \in W \\ \|x\|_2=1}} \boldsymbol{x}^\intercal \mathbf{A} \boldsymbol{x} \leq \lambda_k(\mathbf{A}),$$

then

$$\max_{\substack{W \subset \mathbb{R}^p \\ \dim(W)=k}} \min_{\substack{\boldsymbol{x} \in W \\ \|x\|_2=1}} \boldsymbol{x}^\intercal \mathbf{A} \boldsymbol{x} \leq \lambda_k(\mathbf{A}). \tag{A.4}$$

In particular taking W as the subspace spanned by $\boldsymbol{u}_k, ..., \boldsymbol{u}_p$, the orthonormal set of eigenvalues associated with $\lambda_k(\mathbf{A}), ..., \lambda_p(\mathbf{A})$ respectively, then since $\boldsymbol{u}_k^\intercal \mathbf{A} \boldsymbol{u}_k = \lambda_k(\mathbf{A})$ the inequality (A.4) turns into equality. For the second equality, define $k' = p - k + 1$ so that $p - k' + 1 = k$ and use the second inequality of Poncaire's theorem with k'-dimensional subspaces employing the same arguments as before. □

Another more basic result is the so called Rayleigh quotient theorem. We'll present it without a proof since the techniques used are essentially the same as for the two previous results. The statement and the proof can be found in [18, p. 234].

Theorem A.2.4 (Rayleigh quotient)**.** *Let* $\mathbf{A}$ *be a* $p \times p$ *symmetric matrix. Define some integers* $1 \leq i_1 < \cdots < i_k \leq p$*,* $k \leq p$*, and let* $\boldsymbol{v}_{i_1}, ..., \boldsymbol{v}_{i_k}$ *be orthonormal such that* $\mathbf{A}\boldsymbol{v}_{i_\ell} = \lambda_{i_\ell}(\mathbf{A})\boldsymbol{v}_{i_\ell}$*,* $\ell = 1, ..., k$*. Take* $S = span\{\boldsymbol{v}_{i_1}, ..., \boldsymbol{v}_{i_k}\}$*. Then,*

$$\lambda_{i_1}(\mathbf{A}) = \max_{\substack{\boldsymbol{x} \in S \\ \|\boldsymbol{x}\|_2=1}} \boldsymbol{x}^\intercal \mathbf{A} \boldsymbol{x}.$$

An immediate corollary of the previous theorem is that

$$\lambda_1(\mathbf{A}) = \max_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^\intercal \mathbf{A} \boldsymbol{x},$$

taking $k = p$, $i_1 = 1, ..., i_p = p$ and noticing that $S = \text{span}\{\boldsymbol{v}_1, ..., \boldsymbol{v}_p\} = \mathbb{R}^p$.

An important inequality that will be used later is the so called Weyl's inequality.

Theorem A.2.5 (Weyl's inequalities)**.** *Let* $\mathbf{A}$ *and* $\mathbf{H}$ *be* $p \times p$ *symmetric matrices. Then,*

$$\begin{aligned} \lambda_j(\mathbf{A}+\mathbf{H}) &\le \lambda_i(\mathbf{A}) + \lambda_{j-i+1}(\mathbf{H}), \quad i \le j, \\ \lambda_j(\mathbf{A}+\mathbf{H}) &\ge \lambda_i(\mathbf{A}) + \lambda_{j-i+p}(\mathbf{H}), \quad i \ge j. \end{aligned}$$

Proof. For the first inequality, let $\boldsymbol{u}_k$, $\boldsymbol{v}_k$ and $\boldsymbol{w}_k$, $k = 1, ..., p$, be the eigenvectors of $\mathbf{A}$, $\mathbf{H}$ and $\mathbf{A}+\mathbf{H}$, respectively. Let $W_1 = \operatorname{span}\{\boldsymbol{u}_i, ..., \boldsymbol{u}_p\}$, $W_2 = \operatorname{span}\{\boldsymbol{v}_{j-i+1}, ..., \boldsymbol{v}_p\}$ and $W_3 = \operatorname{span}\{\boldsymbol{w}_1, ..., \boldsymbol{w}_j\}$. Because $\dim(W_1) + \dim(W_2) + \dim(W_3) = 2p + 1$, by Fact A.2.1 we can define $\boldsymbol{x} \in W_1 \cap W_2 \cap W_3$, and by equations (A.2) and (A.3) we get

$$\begin{aligned} \lambda_j(\mathbf{A}+\mathbf{H}) &\le \boldsymbol{x}^\intercal(\mathbf{A}+\mathbf{H})\boldsymbol{x} \\ &= \boldsymbol{x}^\intercal\mathbf{A}\boldsymbol{x} + \boldsymbol{x}^\intercal\mathbf{H}\boldsymbol{x} \\ &\le \lambda_i(\mathbf{A}) + \lambda_{j-i+1}(\mathbf{H}). \end{aligned}$$

The second inequality is proved in the same way by redefining $W_1 = \operatorname{span}\{\boldsymbol{u}_1, ..., \boldsymbol{u}_i\}$, $W_2 = \operatorname{span}\{\boldsymbol{v}_1, ..., \boldsymbol{v}_{j-i+p}\}$ and $W_3 = \operatorname{span}\{\boldsymbol{w}_j, ..., \boldsymbol{w}_p\}$. □

An immediate consequence of Weyl's inequalities is the next double inequality.

Corollary A.2.6. *For* $\mathbf{A}, \mathbf{H} \in \mathcal{S}_p$*,*

$$\lambda_p(\mathbf{H}) \le \lambda_k(\mathbf{A}+\mathbf{H}) - \lambda_k(\mathbf{A}) \le \lambda_1(\mathbf{H}), \quad 1 \le k \le p.$$

Proof. Choose $k \in \{1, ..., p\}$ and let $i = j = k$ in Weyl's inequalities. □

Theorem A.2.7 (Cauchy's Interlacing Theorem)**.** *Let* $\mathbf{A} \in \mathcal{S}_p$ *and define the matrix* $\mathbf{A}_r$*,* $r \le p$*, as the* $r \times r$ *northwest corner of* $\mathbf{A}$*, i.e.,*

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_r & \star \\ \star & \star \end{pmatrix}. \tag{A.5}$$

Then, for $j = 1, ..., r$*,*

$$\lambda_j(\mathbf{A}) \ge \lambda_j(\mathbf{A}_r) \ge \lambda_{p-r+j}(\mathbf{A}). \tag{A.6}$$

Proof. Let $\boldsymbol{v}_1, ..., \boldsymbol{v}_p$ and $\boldsymbol{u}_1, ..., \boldsymbol{u}_r$ be the eigenvectors of associated to $\lambda_1(\mathbf{A}) \ge \cdots \ge \lambda_p(\mathbf{A})$ and $\lambda_1(\mathbf{A}_r) \ge \cdots \ge \lambda_r(\mathbf{A}_r)$, respectively. Define the vectors $\hat{\boldsymbol{u}}_i \in \mathbb{R}^p$ as

$$\hat{\boldsymbol{u}}_i = \begin{pmatrix} \boldsymbol{u}_i \\ \mathbf{0} \end{pmatrix}, \quad i = 1, ..., r.$$

Choose $j \in \{1, ..., r\}$ and define the sets $S_1 = \mathrm{span}\{\boldsymbol{v}_j, ..., \boldsymbol{v}_p\}$ and $S_2 = \mathrm{span}\{\hat{\boldsymbol{u}}_1, ..., \hat{\boldsymbol{u}}_j\}$. Since $\dim(S_1) + \dim(S_2) = p + 1$, the intersection $S_1 \cap S_2$ is non-trivial. Take $\boldsymbol{x} \in S_1 \cap S_2$. Then $\boldsymbol{x} \in S_2$ and is of the form $\boldsymbol{x}^\intercal = (\boldsymbol{y}^\intercal, \mathbf{0}^\intercal)$, where $\boldsymbol{y} \in \mathrm{span}\{\boldsymbol{u}_1, ..., \boldsymbol{u}_j\}$. Then,

$$\boldsymbol{x}^\intercal \mathbf{A} \boldsymbol{x} = (\boldsymbol{y}^\intercal,\ \mathbf{0}^\intercal) \begin{pmatrix} \mathbf{A}_r & \star \\ \star & \star \end{pmatrix} \begin{pmatrix} \boldsymbol{y} \\ \mathbf{0} \end{pmatrix} = \boldsymbol{y}^\intercal \mathbf{A}_r \boldsymbol{y}.$$

Hence, from inequalities (A.2) and (A.3) we have that

$$\lambda_j(\mathbf{A}_r) \leq \boldsymbol{y}^\intercal \mathbf{A}_r \boldsymbol{y} = \boldsymbol{x}^\intercal \mathbf{A} \boldsymbol{x} \leq \lambda_j(\mathbf{A}).$$

This proves the first inequality. The second inequality is proved analogously redefining S_1 and S_2 as $S_1 = \mathrm{span}\{\boldsymbol{v}_1, ..., \boldsymbol{v}_{p-r+j}\}$ and $S_2 = \mathrm{span}\{\hat{\boldsymbol{u}}_j, ..., \hat{\boldsymbol{u}}_r\}$. □

Theorem A.2.7 implies that if $\mathbf{A} \in \mathcal{S}_p$ is of the form (A.5), then

$$\lambda_1(\mathbf{A}) \geq \lambda_1(\mathbf{A}_r) \quad \text{and} \quad \lambda_p(\mathbf{A}) \leq \lambda_r(\mathbf{A}_r).$$

Therefore, since $|||\mathbf{A}||| = \max\{|\lambda_1(\mathbf{A})|, |\lambda_p(\mathbf{A})|\}$[2], we get that

$$|||\mathbf{A}||| \geq |||\mathbf{A}_r|||.$$

Other important observation is that the inequalities (A.6) are obtained if we have a matrix $\mathbf{A} \in \mathcal{S}_p$ of the form

$$\mathbf{A} = \begin{pmatrix} \star & \star \\ \star & \mathbf{A}_r. \end{pmatrix}.$$

The proof follows analogous steps.

One interesting property of the maximum singular value as a function $\lambda_1 : \mathcal{S}_p \to \mathbb{R}$ is that it is convex. This is stated in the next Lemma.

Lemma A.2.8. *The map $\mathbf{A} \mapsto \lambda_1(\mathbf{A})$ is convex in $\mathcal{S}_p$.*

Proof. Let $\mathbf{A}, \mathbf{H} \in \mathcal{S}_p$ and $t \in [0, 1]$. Using equation (A.1) we have that for every $\boldsymbol{x} \in \mathbb{R}^p$, $\lambda_1(\mathbf{A}) \geq \boldsymbol{x}^\intercal \mathbf{A} \boldsymbol{x} / \boldsymbol{x}^\intercal \boldsymbol{x}$ which implies that $\boldsymbol{x}^\intercal \lambda_1(\mathbf{A}) \mathbf{I} \boldsymbol{x} - \boldsymbol{x}^\intercal \mathbf{A} \boldsymbol{x} \geq 0$. Then

$$\lambda_1(\mathbf{A})\mathbf{I} - \mathbf{A} \succeq 0 \quad \text{and} \quad \lambda_1(\mathbf{H})\mathbf{I} - \mathbf{H} \succeq 0.$$

Multiplying by t and $1 - t$ we obtain by Lemma 2.1.2 that

$$t\mathbf{A} + (1 - t)\mathbf{H} \preceq (t\lambda_1(\mathbf{A}) + (1 - t)\lambda_1(\mathbf{H}))\mathbf{I}.$$

Finally by Lemma 2.1.4 $\lambda_1(t\mathbf{A} + (1 - t)\mathbf{H}) \leq t\lambda_1(\mathbf{A}) + (1 - t)\lambda_1(\mathbf{H})$. □

[2] See the definition of the spectral norm $|||\cdot||| = |||\cdot|||_\infty$ in the Section A.3.

A.3 The singular value decomposition

Let $\mathbf{B} \in \mathcal{M}_{n,p}$. A singular value decomposition (SVD) of $\mathbf{B}$ is a factorization

$$\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^\intercal, \tag{A.7}$$

where $\mathbf{U} \in \mathcal{M}_n$ is orthogonal, $\mathbf{D} \in \mathcal{M}_{n,p}$ is diagonal and $\mathbf{V} \in \mathcal{M}_p$ is orthogonal. This is called a *full* or *complete* SVD of $\mathbf{B}$. The diagonal elements of $\mathbf{D}$ are called singular values and are denoted $s_1(\mathbf{B}) \geq \cdots s_m(\mathbf{B}) \geq 0$, $m = n \wedge p$. Suppose that $n \geq p$, then the SVD of $\mathbf{B}$ can be expressed in *reduced* form as

$$\mathbf{B} = \mathbf{U}_1\mathbf{D}_p\mathbf{V}^\intercal, \tag{A.8}$$

where $\mathbf{U}_1 \in \mathcal{M}_{n,p}$ has orthogonal columns and $\mathbf{D}_p \in \mathcal{M}_p$ is diagonal. Stated loosely, the reduced SVD is obtained by dropping off the columns of $\mathbf{U}$ that are multiplied by zero and the rows of $\mathbf{D}$ that are zero. Conversely, if one has the reduced SVD (A.8), then define $\mathbf{U}_2 \in \mathcal{M}_{n,(n-p)}$ such that its columns are orthogonal to the ones of $\mathbf{U}_1$ and set $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2)$. An by completing $\mathbf{D}$ with zeros, we get the full SVD (A.7). The case $n < p$ can be done similarly by taking transpose.

We'll mention the existence and uniqueness theorem for SVD without proof. The proof can be found, e. g., in [42, p. 29].

Theorem A.3.1 (SDV existence and uniqueness)**.** *Every matrix $\mathbf{B} \in \mathcal{M}_{n,p}$ has an SVD (A.7). Furthermore, the singular values $s_i(\mathbf{B})$ are uniquely determined, and if $n = p$ and $s_i(\mathbf{B})$ are distinct, the columns of the matrices $\mathbf{U}$ and $\mathbf{V}$ are uniquely determined up to a sign.*

From the SVD we can find the rank of a matrix. This is stated in the next theorem.

Theorem A.3.2. *The rank of $\mathbf{B} \in \mathcal{M}_{n,p}$ is the number of non-singular values.*

Proof. Without loss of generality suppose $n \geq p$. Let $\mathbf{B} = \mathbf{U}_1\mathbf{D}_p\mathbf{V}^\intercal$ be the reduced SVD of $\mathbf{B}$. Then, $\mathbf{B}^\intercal\mathbf{B} = \mathbf{V}\mathbf{D}_p^2\mathbf{V}^\intercal$, so the eigenvalues of $\mathbf{B}^\intercal\mathbf{B}$ are $s_j(\mathbf{B})^2$, $j = 1, ..., p$. Since[3] $\text{rank}(\mathbf{B}) = \text{rank}(\mathbf{B}^\intercal\mathbf{B})$, and the rank of any symmetric matrix are the number of non-zero eigenvalues, the result follows. □

[3] This comes from the facts $p = \text{rank}(\mathbf{B}) + \dim \text{Ker}(\mathbf{B})$, $p = \text{rank}(\mathbf{B}^\intercal\mathbf{B}) + \dim \text{Ker}(\mathbf{B}^\intercal\mathbf{B})$ and $\text{Ker}(\mathbf{B}) = \text{Ker}(\mathbf{B}^\intercal\mathbf{B})$.

The next theorem indicates an special relationship between eigenvalues and singular values of a symmetric matrix.

Theorem A.3.3. *Let $\mathbf{A} \in \mathcal{S}_p$. Then, the singular values of $\mathbf{A}$ are the absolute values of its eigenvalues.*

Proof. Let $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^\intercal$ be the spectral decomposition of $\mathbf{A}$. We can write

$$\mathbf{A} = \mathbf{P}|\mathbf{D}|\text{sign}(\mathbf{D})\mathbf{P}^\intercal,$$

where $|\mathbf{D}| = \text{diag}(|\lambda_i(\mathbf{A})|)$ and $\text{sign}(\mathbf{D}) = \text{diag}\left[\text{sign}(\lambda_i(\mathbf{A}))\right]$. This is an SVD of $\mathbf{A}$ since $\text{sign}(\mathbf{D})\mathbf{P}^\intercal$ is orthogonal. This finishes the proof. □

Theorem A.3.4. *For $n \geq p$ define $\mathbf{B} \in \mathcal{M}_{n,p}$ with rank r and full SVD given by $\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^\intercal$. The column space (or range) of $\mathbf{B}$ is $C(\mathbf{B}) := \text{span}\{\boldsymbol{u}_1, ..., \boldsymbol{u}_r\}$, and the row space (or null space) of $\mathbf{B}$ is $C(\mathbf{B}^\intercal) := \text{span}\{\boldsymbol{v}_1, ..., \boldsymbol{v}_r\}$.*

Proof. This is a direct consequence of the fact that $C(\mathbf{D}) = \text{span}\{\boldsymbol{e}_1, ..., \boldsymbol{e}_r\}$ and $C(\mathbf{D}^\intercal) = \text{span}\{\boldsymbol{e}_{r+1}, ..., \boldsymbol{e}_p\}$. □

Theorem A.3.4 also holds for $p \geq n$ by taking transpose.

There is no definition of eigenvalues for rectangular matrices, but we can always obtain singular values for any matrix. Then, it is natural to think that by taking the symmetric dilation of a rectangular matrix we can find its eigenvalues from the singular values of the original matrix. The next Theorem, taken from [18, p. 450], indicates how to do this.

Theorem A.3.5. *Let $\mathbf{B} \in \mathcal{M}_{np}$ with $m = n \wedge p$. The eigenvalues of the symmetric dilation*

$$\mathcal{H}(\mathbf{B}) = \begin{pmatrix} \mathbf{0} & \mathbf{B} \\ \mathbf{B}^\intercal & \mathbf{0} \end{pmatrix}$$

are

$$s_1(\mathbf{B}) \geq \cdots s_m(\mathbf{B}) \geq \underbrace{0 = \cdots = 0}_{|n-p|} \geq -s_m(\mathbf{B}) \geq \cdots \geq -s_1(\mathbf{B}).$$

Proof. Without loss of generality[4], suppose that $n \geq p$. Let

$$\mathbf{B} = \mathbf{U}_1\mathbf{D}_p\mathbf{V}^\intercal$$

[4] Since the singular values of $\mathbf{B}$ and $\mathbf{B}^\intercal$ are the same, if $p \geq n$ do the same procedure but with $\mathcal{H}(\mathbf{B}^\intercal)$.

be the reduced singular value decomposition of $\mathbf{B}$, where $\mathbf{U}_1$ is a $n \times p$ matrix with orthonormal columns, $\mathbf{D}_p = \text{diag}(s_1(\mathbf{A}), ..., s_p(\mathbf{A}))$ and $\mathbf{V}$ is a $p \times p$ orthogonal matrix. And let

$$\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^\intercal$$

be the complete singular value decomposition of $\mathbf{B}$, where $\mathbf{U} = (\mathbf{U}_1, \mathbf{U}_2)$ is a $n \times n$ unitary matrix and $\mathbf{D}$ is $n \times p$ and is such that $\mathbf{D}^\intercal = (\mathbf{D}_p, \mathbf{0}_{p\times(n-p)})^\intercal$.

Define the matrices, $\widehat{\mathbf{U}} = \mathbf{U}_1/\sqrt{2}$ and $\widehat{\mathbf{V}} = \mathbf{V}/\sqrt{2}$, and the $(n+p) \times (n+p)$ matrix

$$\mathbf{W} = \begin{pmatrix} \widehat{\mathbf{U}} & -\widehat{\mathbf{U}} & \mathbf{U}_2 \\ \widehat{\mathbf{V}} & \widehat{\mathbf{V}} & \mathbf{0}_{p\times(n-p)} \end{pmatrix}.$$

This construction implies that $\mathbf{W}$ is an orthogonal matrix and that

$$\mathcal{H}(\mathbf{B}) = \mathbf{W} \begin{pmatrix} \mathbf{D}_p & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{D}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0}_{(n-p)\times(n-p)} \end{pmatrix} \mathbf{W}^\intercal$$

which is an spectral decomposition of $\mathcal{H}(\mathbf{B})$. □

A.4 Schatten norms

The classical way to characterize the Schatten k-norms are as follows (see [3, Chapter 4], [18, Chapter 7] and [51].) First we need to define *symmetric gauge functions.*

Definition A.4.1 (Symmetric gauge function)**.** *The function $\phi : \mathbb{R}^m \to \mathbb{R}$ is said to be symmetric gauge if it satisfies the following conditions:*

1. $\phi(\boldsymbol{x}) > 0$ *for* $\boldsymbol{x} \neq \mathbf{0}$*,*
2. $\phi(\alpha\boldsymbol{x}) = |\alpha|\phi(\boldsymbol{x})$*,*
3. $\phi(\boldsymbol{x} + \boldsymbol{y}) \leq \phi(\boldsymbol{x}) + \phi(\boldsymbol{y})$*,*
4. $\phi(\epsilon_1 x_{i_1}, ..., \epsilon_m x_{i_m}) = \phi(\boldsymbol{x})$*,*

where α is a scalar, $\epsilon_i = \pm 1$ for all i and $\{i_1, ..., imn\}$ is a permutation of $\{1, ..., m\}$.

One immediate example of symmetric gauge function is $\phi(\boldsymbol{x}) = \|\boldsymbol{x}\|_k$ where $\|\cdot\|_k$ is the classic ℓ_k vector norm with $k \geq 1$. The following theorem indicates how to obtain a matrix norm form a symmetric gauge function. The proof can be found in [3, Chapter 4].

Theorem A.4.2. *Let $m = n \wedge p$. Given a symmetric gauge function ϕ in $\mathbb{R}^m$, define the function on $\mathcal{M}_{n,p}$ as*

$$|||\mathbf{A}|||_\phi = \phi\left(\boldsymbol{s}(\mathbf{A})\right),$$

where $\boldsymbol{s}(\mathbf{A})$ is the vector of singular values of $\mathbf{A}$. Then, this defines a unitary invariant norm[5] *on $\mathcal{M}_{n,p}$.*

From Theorem A.4.2, we define the Schatten k-norm, $k \geq 1$, as

$$|||\mathbf{A}|||_k = \|\boldsymbol{s}(\mathbf{A})\|_k.$$

We call $|||\cdot||| = |||\cdot|||_\infty$ the operator norm, $|||\cdot|||_2$ the Frobenius norm and $|||\cdot|||_1$ the nuclear norm. In particular $|||\mathbf{A}|||$ gives us the largest eigenvalue of $\mathbf{A}$.

It is common to take $k \geq 0$, and for $k < 1$ we call $|||\cdot|||_k$ a quasi-norm. The following Proposition taken from [50, Chapter 1] enlists some properties of Schatten norms.

Proposition A.4.3 (Properties of Schatten p- norms)**.** *Let $\mathbf{A} \in \mathcal{M}_{n,p}$ and $1 \leq k \leq q \leq \infty$. Then,*

(a) $|||\mathbf{A}|||_k = \left(\operatorname{tr}\left[(\mathbf{A}^\intercal\mathbf{A})^{k/2}\right]\right)^{1/k}$*. In particular,*

$$|||\mathbf{A}|||_2^2 = \operatorname{tr}\left(\mathbf{A}^\intercal\mathbf{A}\right) = \sum_{i=1}^{n}\sum_{j=1}^{m}|a_{ij}|^2.$$

(b) $|||\mathbf{A}|||_k \geq |||\mathbf{A}|||_q$ *(monotonicity).*

(c) $|||\mathbf{A}|||_k \leq \mathit{rank}(\mathbf{A})^{1/k-1/q}|||\mathbf{A}|||_q$.

(d) For $\mathbf{B} \in \mathcal{M}_{p,r}$, $|||\mathbf{AB}|||_k \leq |||\mathbf{A}|||_k|||\mathbf{B}|||_k$ (submultiplicity).

[5] A matrix norm $|||\cdot|||'$ is called *unitary invariant* if $|||\mathbf{UAV}|||' = |||\mathbf{A}|||'$ for any orthogonal matrices $\mathbf{U}, \mathbf{V}$.

A.5 Properties of singular values and norms

Since $\mathcal{H}(\mathbf{B})$ is symmetric, its singular values are the same as $\mathbf{B}$ but with each value repeated at least once. Therefore, from Weyl's inequality and Theorem A.3.5 we obtain the next proposition.

Proposition A.5.1 (Singular values are Lipschitz)**.** *Let* $\mathbf{F}, \mathbf{G} \in \mathcal{M}_{n,p}$ *and* $m = n \wedge p$. *Then,*

$$|s_i(\mathbf{F}) - s_i(\mathbf{G})| \leq |||\mathbf{F} - \mathbf{G}|||, \quad i = 1, ..., m.$$

Proof. Its easy to note that the matrix dilation $\mathcal{H} : \mathcal{M}_{n,p} \to \mathcal{S}_{(n+p),(n+p)}$ is a linear operator. Then, $\mathcal{H}(\mathbf{F} - \mathbf{G}) + \mathcal{H}(\mathbf{G}) = \mathcal{H}(\mathbf{F})$. Now, by Corollary A.2.6 we have that

$$\lambda_m \left(\mathcal{H}(\mathbf{F} - \mathbf{G})\right) \leq \lambda_i \left(\mathcal{H}(\mathbf{F} - \mathbf{G}) + \mathcal{H}(\mathbf{G})\right) - \lambda_i \left(\mathcal{H}(\mathbf{G})\right) \leq \lambda_1 \left(\mathcal{H}(\mathbf{F} - \mathbf{G})\right),$$

for $i = 1, ..., m$. Then, by Theorem A.3.5 we have that

$$s_m \left(\mathbf{F} - \mathbf{G}\right) \leq s_i \left(\mathbf{F}\right) - s_i \left(\mathbf{G}\right) \leq s_1 \left(\mathbf{F} - \mathbf{G}\right).$$

Since singular values are non-negative, $s_m \left(\mathbf{F} - \mathbf{G}\right) \geq -s_1 \left(\mathbf{F} - \mathbf{G}\right)$ and

$$\left|s_i \left(\mathbf{F}\right) - s_i \left(\mathbf{G}\right)\right| \leq s_1 \left(\mathbf{F} - \mathbf{G}\right) = |||\mathbf{F} - \mathbf{G}|||.$$

□

The next theorem gives us a way to calculate the the spectral norm of a block diagonal symmetric matrix.

Lemma A.5.2. *Define* $\mathbf{A} \in \mathcal{S}_{p_1+p_2}$*,* $\mathbf{A}_1 \in \mathcal{S}_{p_1}$ *and* $\mathbf{A}_2 \in \mathcal{S}_{p_2}$ *such that*

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 \end{pmatrix}.$$

Then,

$$|||\mathbf{A}||| = \max\left\{|||\mathbf{A}_1|||, |||\mathbf{A}_2|||\right\}.$$

Proof. We just need to prove that the set of eigenvalues of $\mathbf{A}$ is equal to the set of eigenvalues of $\mathbf{A}_1$ and $\mathbf{A}_2$. If we prove that, we obtain that

$$\begin{aligned}
|||\mathbf{A}||| &= \max\left\{|\lambda_1(\mathbf{A}_1)|, ..., |\lambda_{p_1}(\mathbf{A}_1)|, |\lambda_1(\mathbf{A}_2)|, ..., |\lambda_{p_2}(\mathbf{A}_2)|\right\} \\
&= \max\left\{\max\left(|\lambda_1(\mathbf{A}_1)|, |\lambda_{p_1}(\mathbf{A}_1)|\right), \max\left(|\lambda_1(\mathbf{A}_2)|, ..., |\lambda_{p_2}(\mathbf{A}_2)|\right)\right\} \\
&= \max\left\{|||\mathbf{A}_1|||, |||\mathbf{A}_2|||\right\}.
\end{aligned}$$

To see that the spectrum of $\mathbf{A}$ corresponds to that of $\mathbf{A}_1$ and $\mathbf{A}_2$, let $\boldsymbol{v}_1, ..., \boldsymbol{v}_{p_1}$ and $\boldsymbol{u}_1, ..., \boldsymbol{u}_{p_2}$ be a sets of ordered[6] eigenvectors of $\mathbf{A}_{p_1}$ and $\mathbf{A}_{p_2}$, respectively. Then, for some i,

$$\mathbf{A}\begin{pmatrix}\boldsymbol{v}_i \\ \mathbf{0}\end{pmatrix} = \begin{pmatrix}\mathbf{A}_1\boldsymbol{v}_i \\ \mathbf{0}\end{pmatrix} = \lambda_i(\mathbf{A}_1)\begin{pmatrix}\boldsymbol{v}_i \\ \mathbf{0}\end{pmatrix},$$

and similarly

$$\mathbf{A}\begin{pmatrix}\mathbf{0} \\ \boldsymbol{u}_i\end{pmatrix} = \begin{pmatrix}\mathbf{0} \\ \mathbf{A}_2\boldsymbol{u}_i\end{pmatrix} = \lambda_i(\mathbf{A}_2)\begin{pmatrix}\mathbf{0} \\ \boldsymbol{u}_i\end{pmatrix}.$$

This characterize all the eigenvalues of $\mathbf{A}$. □

Corollary A.5.3. *Let $\mathbf{B} \in \mathcal{M}_{n,p}$. Then, $|||\mathcal{H}(\mathbf{B})||| = |||\mathbf{B}|||$.*

Proof. Note that

$$\mathcal{H}(\mathbf{B})^2 = \begin{pmatrix}\mathbf{B}\mathbf{B}^\intercal & \mathbf{0} \\ \mathbf{0} & \mathbf{B}^\intercal\mathbf{B}\end{pmatrix}.$$

By Lemma A.5.2 we get

$$|||\mathcal{H}(\mathbf{B})^2||| = \max\left\{|||\mathbf{B}\mathbf{B}^\intercal|||, |||\mathbf{B}^\intercal\mathbf{B}|||\right\}.$$

But, by taking any SVD representation of $\mathbf{B}$ with $m = n \wedge p$ it's easy to see that

$$|||\mathbf{B}\mathbf{B}^\intercal||| = \max_{1\leq i\leq m} s_i^2(\mathbf{B}) = |||\mathbf{B}^\intercal\mathbf{B}|||$$

Then $|||\mathcal{H}(\mathbf{B})^2||| = \left(\max_{1\leq i\leq m} s_i(\mathbf{B})\right)^2 = |||\mathbf{B}|||^2$. Observing that $|||\mathcal{H}(\mathbf{B})^2||| = |||\mathcal{H}(\mathbf{B})|||^2$ ends the proof. The last equality is true because for any symmetric matrix $\mathbf{A} \in \mathcal{S}_p$ we have that

$$|||\mathbf{A}^2||| = \max_{1\leq i\leq p} \lambda_i^2(\mathbf{A}) = \left(\max_{1\leq i\leq p} |\lambda_i(\mathbf{A})|\right)^2 = |||\mathbf{A}|||^2.$$

□

[6] Order corresponding to the eigenvalues $\lambda_i(\cdot)$, e.g., $\mathbf{A}_1\lambda_i(\mathbf{A}_1) = \boldsymbol{v}_i\lambda_i(\mathbf{A}_1)$.

Corollary A.5.3 could also be derived from Theorem A.3.5 that states that $\mathbf{B} \in \mathcal{M}_{n,p}$ and $\mathcal{H}(\mathbf{B})$ have the same eigenvalues, with repetitions. This indicates that the Schatten k-norm of $\mathcal{H}(\mathbf{B})$ is given by

$$|||\mathcal{H}(\mathbf{B})|||_k = \left(\sum_{j=1}^{2m} |s_j(\mathbf{B})|^k \right)^{1/k} = 2^{1/k} |||\mathbf{B}|||_k, \quad m = n \wedge p.$$

Recall that our definition of operator norm of $\mathbf{B} \in \mathcal{M}_{n,p}$ is

$$|||\mathbf{B}||| = |||\mathbf{B}|||_\infty = s_1(\mathbf{B}).$$

We'll prove a different representation of this norm.

Theorem A.5.4. *For any* $\mathbf{B} \in \mathcal{M}_{n,p}$,

$$|||\mathbf{B}||| = \max_{\|\boldsymbol{x}\|_2=1} \|\mathbf{B}\boldsymbol{x}\|_2 = \max_{\substack{\|\boldsymbol{x}\|_2=1 \\ \|\boldsymbol{y}\|_2=1}} \boldsymbol{x}^\intercal \mathbf{B} \boldsymbol{y},$$

where $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *are of the correct dimensions.*

Proof. Let $\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^\intercal$ be the complete singular value decomposition of $\mathbf{B}$. Note the following set equality:

$$\{\boldsymbol{y} \in \mathbb{R}^p \,|\, \boldsymbol{y} = \mathbf{V}^\intercal \boldsymbol{x}, \text{ for some } \boldsymbol{x} \text{ such that } \|\boldsymbol{x}\|_2 = 1\} = \{\boldsymbol{y} \in \mathbb{R}^p \,|\, \|\boldsymbol{y}\|_2 = 1\}, \tag{A.9}$$

i.e., we are just rewriting the know fact that the unitary sphere in $\mathbb{R}^p$ is invariant under orthogonal transformations. Then, we have that

$$\begin{aligned}
\max_{\|\boldsymbol{x}\|_2=1} \|\mathbf{B}\boldsymbol{x}\|_2 &= \max_{\|\boldsymbol{x}\|_2=1} \|\mathbf{U}\mathbf{D}\mathbf{V}^\intercal \boldsymbol{x}\|_2 \\
&= \max_{\|\boldsymbol{x}\|_2=1} \|\mathbf{D}\mathbf{V}^\intercal \boldsymbol{x}\|_2, \quad \text{(since } \mathbf{U} \text{ is orthogonal)} \\
&= \max_{\substack{\boldsymbol{y}=\mathbf{V}^\intercal\boldsymbol{x} \\ \|\boldsymbol{x}\|_2=1}} \|\mathbf{D}\boldsymbol{y}\|_2 \\
&= \max_{\|\boldsymbol{y}\|_2=1} \|\mathbf{D}\boldsymbol{y}\|_2, \quad \text{(from equality A.9)} \\
&\leq \max_{\|\boldsymbol{y}\|_2=1} \|s_1(\mathbf{B})\boldsymbol{y}\|_2 \\
&= s_1(\mathbf{B}).
\end{aligned}$$

But $\|\mathbf{D}\boldsymbol{y}\|_2 = s_1(\mathbf{B})$ for $\boldsymbol{y} = \boldsymbol{e}_1$. This establishes the first equality. The second equality follows in a similar way:

$$\begin{aligned}
\max_{\substack{\|\boldsymbol{x}\|_2=1\\ \|\boldsymbol{y}\|_2=1}} \boldsymbol{x}^\intercal \mathbf{B}\boldsymbol{y} &= \max_{\substack{\|\boldsymbol{x}\|_2=1\\ \|\boldsymbol{y}\|_2=1}} \boldsymbol{x}^\intercal \mathbf{U}\mathbf{D}\mathbf{V}^\intercal \boldsymbol{y} \\
&= \max_{\substack{\boldsymbol{z}=\mathbf{U}^\intercal\boldsymbol{x},\|\boldsymbol{x}\|_2=1\\ \boldsymbol{w}=\mathbf{V}^\intercal\boldsymbol{y},\|\boldsymbol{y}\|_2=1}} \boldsymbol{z}^\intercal \mathbf{D}\boldsymbol{w} \\
&= \max_{\substack{\|\boldsymbol{z}\|_2=1\\ \|\boldsymbol{w}\|_2=1}} \boldsymbol{z}^\intercal \mathbf{D}\boldsymbol{w} \\
&\leq \max_{\substack{\|\boldsymbol{z}\|_2=1\\ \|\boldsymbol{w}\|_2=1}} |\boldsymbol{z}^\intercal \mathbf{D}\boldsymbol{w}| \\
&\leq s_1(\mathbf{B}) \max_{\substack{\|\boldsymbol{z}\|_2=1\\ \|\boldsymbol{w}\|_2=1}} |\boldsymbol{z}^\intercal \boldsymbol{w}| \\
&\leq s_1(\mathbf{B}),
\end{aligned}$$

where the last inequality follows from $|\boldsymbol{z}^\intercal\boldsymbol{w}| \leq \|\boldsymbol{z}\|_2\|\boldsymbol{w}\|_2$. Since $\boldsymbol{z}^\intercal\mathbf{D}\boldsymbol{w} = s_1(\mathbf{B})$ taking $\boldsymbol{z} = \boldsymbol{w} = \boldsymbol{e}_1$, the second equality is established. □

Similarly, one can show the next representation of the smallest singular value.

Theorem A.5.5. *For any* $\mathbf{B} \in \mathcal{M}_{n,p}$, $m = n \wedge p$,

$$s_m(\mathbf{B}) = \min_{\|\boldsymbol{x}\|_2=1} \|\mathbf{B}\boldsymbol{x}\|_2 = \min_{\substack{\|\boldsymbol{x}\|_2=1\\ \|\boldsymbol{y}\|_2=1}} \boldsymbol{x}^\intercal\mathbf{B}\boldsymbol{y},$$

where $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *are of the correct dimensions.*

Lemma A.5.6. *Let* $\mathbf{B} \in \mathcal{M}_{n,p}$. *Then, for any matrices* $\mathbf{A} \in \mathcal{S}_n$ *and* $\mathbf{H} \in \mathcal{S}_p$,

$$\left|\!\left|\!\left| \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\intercal & \mathbf{H} \end{pmatrix} \right|\!\right|\!\right| \geq \left|\!\left|\!\left| \begin{pmatrix} \mathbf{0} & \mathbf{B} \\ \mathbf{B}^\intercal & \mathbf{0} \end{pmatrix} \right|\!\right|\!\right|$$

Proof. In terms of eigenvalues, the spectral norm of a symmetric matrix $\mathbf{M} \in \mathcal{S}_p$ is

$$|\!|\!|\mathbf{M}|\!|\!| = \max\{|\lambda_1(\mathbf{M})|, |\lambda_p(\mathbf{M})|\}.$$

So, if $\mathbf{Q}$ is also symmetric, $|\!|\!|\mathbf{M}|\!|\!| \geq |\!|\!|\mathbf{Q}|\!|\!|$ if and only if $|\!|\!|\mathbf{M}^2|\!|\!| \geq |\!|\!|\mathbf{Q}^2|\!|\!|$.

Now, with and easy calculation we get

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\intercal & \mathbf{H} \end{pmatrix}^2 = \begin{pmatrix} \mathbf{A}^2 + \mathbf{B}\mathbf{B}^\intercal & \mathbf{A}\mathbf{B} + \mathbf{B}\mathbf{H} \\ \mathbf{B}^\intercal\mathbf{A} + \mathbf{H}\mathbf{B}^\intercal & \mathbf{H}^2 + \mathbf{B}^\intercal\mathbf{B} \end{pmatrix}.$$

Since $\mathbf{A}^2 + \mathbf{B}\mathbf{B}^\intercal$ is a sub-matrix of the square, by Theorem A.2.7 and the comment bellow we have that

$$\left|\!\left|\!\left| \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\intercal & \mathbf{H} \end{pmatrix}^2 \right|\!\right|\!\right| \geq |||\mathbf{A}^2 + \mathbf{B}\mathbf{B}^\intercal||| \geq |||\mathbf{B}\mathbf{B}^\intercal|||,$$

where the second inequality follows because $\mathbf{A}^2$ and $\mathbf{B}\mathbf{B}^\intercal$ are non-negative definite and from the fact $\mathbf{A}^2 + \mathbf{B}\mathbf{B}^\intercal \succeq \mathbf{B}\mathbf{B}^\intercal$. Similarly,

$$\left|\!\left|\!\left| \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^\intercal & \mathbf{H} \end{pmatrix}^2 \right|\!\right|\!\right| \geq |||\mathbf{H}^2 + \mathbf{B}^\intercal\mathbf{B}||| \geq |||\mathbf{B}^\intercal\mathbf{B}|||.$$

Since

$$\begin{pmatrix} \mathbf{0} & \mathbf{B} \\ \mathbf{B}^\intercal & \mathbf{0} \end{pmatrix}^2 = \begin{pmatrix} \mathbf{B}\mathbf{B}^\intercal & \mathbf{0} \\ \mathbf{0} & \mathbf{B}^\intercal\mathbf{B} \end{pmatrix}$$

we get from Lemma A.5.2 that

$$\left|\!\left|\!\left| \begin{pmatrix} \mathbf{B}\mathbf{B}^\intercal & \mathbf{0} \\ \mathbf{0} & \mathbf{B}^\intercal\mathbf{B} \end{pmatrix} \right|\!\right|\!\right| = \max\left\{|||\mathbf{B}\mathbf{B}^\intercal|||, |||\mathbf{B}^\intercal\mathbf{B}|||\right\}.$$

Observing that $|||\mathbf{B}\mathbf{B}^\intercal||| = |||\mathbf{B}^\intercal\mathbf{B}|||$ ends the proof. □

From the Fischer-Courant min-max principle we can deduce the next characterization of singular values.

Proposition A.5.7. *For $\mathbf{B} \in \mathcal{M}_{n,p}$ and $j = 1, ..., ml$, $m = n \wedge p$,*

$$s_j(\mathbf{B}) = \max_{\substack{W \subset \mathbb{R}^p \\ dim(W) = j}} \min_{\substack{\boldsymbol{x} \in W \\ \|\boldsymbol{x}\| = 1}} \|\mathbf{B}\boldsymbol{x}\|_2.$$

Proof. Without loss of generality, suppose that $n \geq p$. Let $\mathbf{B} = \mathbf{U}\mathbf{D}\mathbf{V}^\intercal$ be a full SVD of $\mathbf{B}$. Hence, $\mathbf{B}^\intercal\mathbf{B} = \mathbf{V}\mathbf{D}^\intercal\mathbf{D}\mathbf{V}^\intercal$, where $\mathbf{D}^\intercal\mathbf{D} = \operatorname{diag}(s_1^2(\mathbf{B}), ..., s_m^2(\mathbf{B}), 0, ..., 0)$. Then, by the Fischer-Courant min-max principle,

$$s_1^2(\mathbf{B}) = \max_{\substack{W \subset \mathbb{R}^p \\ \dim(W) = j}} \min_{\substack{\boldsymbol{x} \in W \\ \|\boldsymbol{x}\| = 1}} \boldsymbol{x}^\intercal\mathbf{B}^\intercal\mathbf{B}\boldsymbol{x}$$

$$= \max_{\substack{W \subset \mathbb{R}^p \\ \dim(W)=j}} \min_{\substack{\boldsymbol{x} \in W \\ \|\boldsymbol{x}\|=1}} \|\mathbf{B}\boldsymbol{x}\|_2^2.$$

Taking square root completes the proof. □

The following lemma tells us how to bound the singular values of product of matrices.

Lemma A.5.8. *For any two matrices* $\mathbf{B} \in \mathcal{M}_{n,p}$ *and* $\mathbf{A} \in \mathcal{M}_{r,n}$ *we have that for* $j = 1, \dots, m$*,* $m = \min\{n, p, r\}$*,*

$$s_j(\mathbf{AB}) \leq |||\mathbf{A}||| s_j(\mathbf{B})$$
$$s_j(\mathbf{AB}) \leq |||\mathbf{B}||| s_j(\mathbf{A}).$$

The previous lemma is also true for $m = r \wedge p$, as long as $r \wedge p < n$.

Proof. We'll prove the first inequality. By Proposition A.5.7 we have that

$$s_j(\mathbf{AB}) = \max_{\substack{W \subset \mathbb{R}^p \\ \dim(W)=j}} \min_{\substack{\boldsymbol{x} \in W \\ \|\boldsymbol{x}\|=1}} \|\mathbf{AB}\boldsymbol{x}\|_2.$$

Observe that by Theorem A.2.4, for any $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{y} \neq \mathbf{0}$, we have that

$$\frac{\|\mathbf{A}\boldsymbol{y}\|_2^2}{\|\boldsymbol{y}\|_2^2} = \frac{\boldsymbol{y}^\intercal \mathbf{A}^\intercal \mathbf{A}\boldsymbol{y}}{\|\boldsymbol{y}\|_2^2} \leq \lambda_1(\mathbf{A}^\intercal \mathbf{A}) = s_1^2(\mathbf{A}).$$

Then, for any $\boldsymbol{y} \in \mathbb{R}^n$, $\|\mathbf{A}\boldsymbol{y}\|_2 \leq |||\mathbf{A}||| \|\boldsymbol{y}\|_2$. Therefore, we get that

$$\begin{aligned} s_j(\mathbf{AB}) &= \max_{\substack{W \subset \mathbb{R}^p \\ \dim(W)=j}} \min_{\substack{\boldsymbol{x} \in W \\ \|\boldsymbol{x}\|=1}} \|\mathbf{AB}\boldsymbol{x}\|_2 \\ &\leq \max_{\substack{W \subset \mathbb{R}^p \\ \dim(W)=j}} \min_{\substack{\boldsymbol{x} \in W \\ \|\boldsymbol{x}\|=1}} |||\mathbf{A}||| \|\mathbf{B}\boldsymbol{x}\|_2 \\ &= |||\mathbf{A}||| s_j(\mathbf{B}), \end{aligned}$$

where the last equality follows from Proposition A.5.7. The second inequality of the lemma can be proved analogously. □

With Lemma A.5.8 we can prove the next useful theorem.

Theorem A.5.9 (Trace duality)**.** *For any* $\mathbf{A}, \mathbf{B} \in \mathcal{M}_{n,p}$*,*

$$|\mathrm{tr}\,(\mathbf{A}^\intercal \mathbf{B})| \leq |||\mathbf{A}|||_1 |||\mathbf{B}|||.$$

Proof. First note that for any square matrix $\mathbf{C} \in \mathcal{M}_p$ with full SVD given by $\mathbf{C} = \mathbf{U}\mathbf{D}\mathbf{V}^\intercal$, we have that

$$\operatorname{tr} \mathbf{C} = \operatorname{tr} \left(\mathbf{U}\mathbf{D}\mathbf{V}^\intercal\right) = \operatorname{tr} \left(\mathbf{V}^\intercal\mathbf{U}\mathbf{D}\right).$$

Define $\mathbf{Q} = \mathbf{V}^\intercal\mathbf{U}$ and note that $\mathbf{Q}^\intercal\mathbf{Q} = \mathbf{Q}\mathbf{Q}^\intercal = \mathbf{I}$, i.e., $\mathbf{Q}$ is orthogonal. Hence,

$$|\operatorname{tr} \mathbf{C}| = |\operatorname{tr} \left(\mathbf{Q}\mathbf{D}\right)| = \Big|\sum_{j=1}^{p} q_{jj} s_j(\mathbf{C})\Big| \leq \sum_{j=1}^{p} |q_{jj}| s_j(\mathbf{C}).$$

Observe that $|q_{jj}| = |\boldsymbol{e}_j^\intercal \mathbf{Q} \boldsymbol{e}_j| \leq |\lambda_j(\mathbf{Q})|$ (Theorem A.2.4), and since the eigenvalues of orthogonal matrices are ± 1, we get that

$$|\operatorname{tr} \mathbf{C}| \leq \sum_{j=1}^{m} s_j(\mathbf{C}).$$

Then, applying this to the square matrix $\mathbf{A}^\intercal\mathbf{B}$, we get

$$|\operatorname{tr} (\mathbf{A}^\intercal\mathbf{B})| \leq \sum_{j=1}^{m} s_j(\mathbf{A}^\intercal\mathbf{B}).$$

And by Lemma A.5.8,

$$|\operatorname{tr} (\mathbf{A}^\intercal\mathbf{B})| \leq |||\mathbf{B}||| \sum_{j=1}^{m} s_j(\mathbf{A}^\intercal) = |||\mathbf{B}||| \, |||\mathbf{A}^\intercal|||_1 = |||\mathbf{B}||| \, |||\mathbf{A}|||_1,$$

where the last equality is due to the fact that $\mathbf{A}$ and $\mathbf{A}^\intercal$ have the same singular values. □

The next proposition establishes equality for trace duality.

Proposition A.5.10 (Equality in trace duality)**.** *For every $\mathbf{B} \in \mathcal{M}_{n,p}$ there exists a matrix $\mathbf{W} \in \mathcal{M}_{n,p}$ with $|||\mathbf{W}||| \leq 1$ for which*

$$\langle \mathbf{B}, \mathbf{W} \rangle = |||\mathbf{B}|||_1.$$

Proof. Suppose without loss of generality $n \geq p$. Let $\mathbf{U}\mathbf{D}\mathbf{V}^\intercal$ and $\mathbf{U}_1\mathbf{D}_p\mathbf{V}^\intercal$ be a full and reduced SVD of $\mathbf{B}$, respectively. Define $\mathbf{W} = \mathbf{U}_1\mathbf{V}^\intercal$. Then,

$$|||\mathbf{W}||| = |||\mathbf{U}_1\mathbf{V}^\intercal||| = |||\mathbf{U}\mathbf{I}_{n\times p}\mathbf{V}^\intercal||| = 1,$$

where $\mathbf{I}_{n\times p}$ is the $n \times p$ identity matrix. On the other hand,

$$\langle \mathbf{B}, \mathbf{W} \rangle = \operatorname{tr} \left(\mathbf{B}^\intercal\mathbf{W}\right) = \operatorname{tr} \left(\mathbf{V}\mathbf{D}_p\mathbf{U}_1^\intercal\mathbf{W}\right) = \operatorname{tr} \left(\mathbf{D}_p\mathbf{U}_1^\intercal\mathbf{W}\mathbf{V}\right) = \operatorname{tr} \mathbf{D}_p = |||\mathbf{B}|||_1.$$

Therefore, $\langle \mathbf{B}, \mathbf{W} \rangle = |||\mathbf{B}|||_1 = |||\mathbf{B}|||_1 |||\mathbf{W}|||$. □

A useful representation of the operator norm $|\!|\!| \cdot |\!|\!| = |\!|\!| \cdot |\!|\!|_\infty$ for symmetric matrices is the following.

Proposition A.5.11. *For $\mathbf{A} \in \mathcal{S}_p$,*

$$|\!|\!|\mathbf{A}|\!|\!| = \sup_{\|\boldsymbol{x}\|_2 = 1} |\boldsymbol{x}^\intercal \mathbf{A} \boldsymbol{x}|.$$

Proof. Let $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^\intercal$ be the spectral decomposition of $\mathbf{A}$. For $\boldsymbol{x}$ such that $\|\boldsymbol{x}\|_2 = 1$ define $\boldsymbol{y} = \mathbf{P}^\intercal \boldsymbol{x}$. Then $\|\boldsymbol{y}\|_2 = 1$ and

$$|\boldsymbol{x}^\intercal \mathbf{A} \boldsymbol{x}| = |\boldsymbol{y}^\intercal \mathbf{D} \boldsymbol{y}| = \Big| \sum_{j=1}^{p} \lambda_j(\mathbf{A}) y_j^2 \Big| \leq \max_j \{|\lambda_j(\mathbf{A})|\} \sum_{j=1}^{p} y_j^2 = |\!|\!|\mathbf{A}|\!|\!|,$$

so

$$|\!|\!|\mathbf{A}|\!|\!| \geq \sup_{\|\boldsymbol{x}\|_2 = 1} |\boldsymbol{x}^\intercal \mathbf{A} \boldsymbol{x}|.$$

Recall that $|\!|\!|\mathbf{A}|\!|\!| = \max\{\lambda_1(\mathbf{A}), |\lambda_p(\mathbf{A})|\}$, so by taking $\boldsymbol{x}$ as the corresponding unitary eigenvector of $\mathbf{A}$ associated with $\lambda_1(\mathbf{A})$ or $\lambda_p(\mathbf{A})$, the supremum achieves $|\!|\!|\mathbf{A}|\!|\!|$. □

The next proposition indicates that the Frobenius norm is invariant under projection operators.

Proposition A.5.12. *Let $T : \mathcal{M}_{n,m} \to \mathcal{M}_{n,m}$ be an orthogonal projection operator, i.e., an operator such that for $\mathbf{A}, \mathbf{B} \in \mathcal{M}_{n,m}$*

1. *T is linear,*
2. *$T(T(\mathbf{A})) = T(\mathbf{A})$,*
3. *$\langle T(\mathbf{A}), \mathbf{B} \rangle = \langle \mathbf{A}, T(\mathbf{B}) \rangle$.*

Also, define $\mathbf{P}_1 \in \mathcal{M}_n$ and $\mathcal{P} \in \mathcal{M}_m$ two orthogonal projection matrices, i.e., matrices such that $\mathbf{P}_i = \mathbf{P}_i^2 = \mathbf{P}_i^\intercal$. Then, for any $\mathbf{A} \in \mathcal{M}_{n,m}$

(a) $|\!|\!|T(\mathbf{A})|\!|\!|_2 \leq |\!|\!|\mathbf{A}|\!|\!|_2$.

(b) $|\!|\!|\mathbf{P}_1 \mathbf{A} \mathbf{P}_2|\!|\!|_2 \leq |\!|\!|\mathbf{A}|\!|\!|_2$.

(c) $|\!|\!|\mathbf{A} - \mathbf{P}_1\mathbf{A}\mathbf{P}_2|\!|\!|_2 \leq |\!|\!|\mathbf{A}|\!|\!|_2$.

Proof. (a) Since $\langle T(\mathbf{A}), (I-T)(\mathbf{A})\rangle = \langle \mathbf{A}, T(\mathbf{A} - T(\mathbf{A}))\rangle = \langle \mathbf{A}, T(\mathbf{A}) - T(\mathbf{A})\rangle = \mathbf{0}$, we have that

$$|\!|\!|\mathbf{A}|\!|\!|_2^2 = |\!|\!|T(\mathbf{A})|\!|\!|_2^2 |\!|\!|(I-T)(\mathbf{A})|\!|\!|_2^2$$

and $|\!|\!|T(\mathbf{A})|\!|\!|_2^2 \leq |\!|\!|\mathbf{A}|\!|\!|_2^2$.

(b) Define $T(\mathbf{A}) = \mathbf{P}_1\mathbf{A}\mathbf{P}_2$. Clearly, T is lineal and $T^2(\mathbf{A}) = T(\mathbf{A})$. Additionally, $\langle T(\mathbf{A}), \mathbf{B}\rangle = \operatorname{tr}(\mathbf{P}_2\mathbf{A}^{\mathsf{T}}\mathbf{P}_1\mathbf{B}) = \operatorname{tr}(\mathbf{A}^{\mathsf{T}}\mathbf{P}_1\mathbf{B}\mathbf{P}_2) = \langle \mathbf{A}, T(\mathbf{B})\rangle$. So the result follows from (a).

(c) It is obvious that if $T : \mathcal{M}_{n,m} \to \mathcal{M}_{n,m}$ is a orthogonal projection operator then $I - T$ is also an orthogonal projection operator. Then, the result follows from (a) and (b).

□

Appendix B

Stochastic processes and ϵ-nets

This appendix shows most of the technical details of Chapter 3. In the first section we work with Gaussian processes and connect the ideas with Gaussian matrices. The following sections are dedicated to the sub-Gaussian case and presents the classical theory of ϵ-nets.

B.1 Gaussian comparison inequalities

A collection $(X_t)_{t\in T}$ of real-valued random variables indexed by a non-empty set T, which is usually a subset of $\mathbb{R}^p$, $p \geq 1$, is called *random process* . We say that the process is centered if $\mathbb{E}X_t = 0$ for any $t \in T$.

Definition B.1.1 (Gaussian process)**.** *A random process $(X_t)_{t\in T}$ is called Gaussian process if the vector $(X_{t_1}, ..., X_{t_n})^\intercal$ has a Gaussian distribution for any $n \in \mathbb{N}$ and $t_1, ..., t_n \in T$.*

The next proposition assures that the maximum singular value of a standard Gaussian matrix, i.e., a matrix with iid standard Gaussian entries, is the maximum of a Gaussian process.

Proposition B.1.2. *Let U and V be subsets of $\mathbb{R}^n$ and $\mathbb{R}^p$, respectively. Define the random matrix $\mathbf{W} \sim \mathcal{N}_{n\times p}(\mathbf{I})$ and let $(Z_{\boldsymbol{u},\boldsymbol{v}})_{\boldsymbol{u},\boldsymbol{v}\in U\times V}$ be an stochastic process defined as*

$$Z_{\boldsymbol{u},\boldsymbol{v}} = \boldsymbol{u}^\intercal \mathbf{W} \boldsymbol{v}.$$

Then, $(Z_{\boldsymbol{u},\boldsymbol{v}})_{\boldsymbol{u},\boldsymbol{v}\in U\times V}$ is a centered Gaussian process in $U \times V$.

Proof. Observe that for any $(\boldsymbol{u}, \boldsymbol{v}) \in U \times V$, we have that

$$Z_{\boldsymbol{u},\boldsymbol{v}} = \sum_{i=1}^{n} \sum_{j=1}^{p} u_i v_j w_{ij},$$

where (w_{ij}) are the entries of $\mathbf{W}$. Let $a_1, ..., a_m$, $m \in \mathbb{N}$, be some arbitrary real numbers, and take $(\boldsymbol{u}^1, \boldsymbol{v}^1), ..., (\boldsymbol{u}^m, \boldsymbol{v}^m)$ in $U \times V$. Then,

$$\sum_{k=1}^{m} a_k Z_{\boldsymbol{u}^k,\boldsymbol{v}^k} = \sum_{k=1}^{m} \sum_{i=1}^{n} \sum_{j=1}^{p} a_k u_i^k v_j^k w_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{p} w_{ij} b_{ij},$$

where $b_{ij} = \sum_{k=1}^{m} a_k u_i^k v_j^k$. This, implies that $\sum_{k=1}^{m} a_k Z_{\boldsymbol{u}^k,\boldsymbol{v}^k}$ is a Gaussian random variable because (w_{ij}) are independent standard Gaussian. Since this holds for arbitrary m and $a_1, ..., a_m$, the vector

$$(Z_{\boldsymbol{u}^1,\boldsymbol{v}^1}, ..., Z_{\boldsymbol{u}^m,\boldsymbol{v}^m})^\intercal$$

is a Gaussian random vector for any $(\boldsymbol{u}^1, \boldsymbol{v}^1), ..., (\boldsymbol{u}^m, \boldsymbol{v}^m)$. Therefore, $(Z_{\boldsymbol{u},\boldsymbol{v}})_{\boldsymbol{u},\boldsymbol{v} \in U \times V}$ is a Gaussian process in $U \times V$. To see that it is centered, we just calculate the expectation directly:

$$\mathbb{E} Z_{\boldsymbol{u},\boldsymbol{v}} = \mathbb{E} \boldsymbol{u}^\intercal \mathbf{W} \boldsymbol{v} = \boldsymbol{u}^\intercal (\mathbb{E}\mathbf{W}) \boldsymbol{v} = 0, \quad \forall (\boldsymbol{u}, \boldsymbol{v}) \in U \times V.$$

□

The next two classical theorems are referred as *Gaussian comparison inequalities.* The proofs can be consulted in [47, Chapter 7].

Theorem B.1.3 (Sudakov-Fernique inequality)**.** *Let $(X_t)_{t \in T}$ and $(Y_t)_{t \in T}$ be two centered Gaussian processes, such that for all $t, s \in T$ we have that*

$$\mathbb{E} \left(X_t - X_s\right)^2 \leq \mathbb{E} \left(Y_t - Y_s\right)^2.$$

Then,

$$\mathbb{E} \sup_{t \in T} X_t \leq \mathbb{E} \sup_{t \in T} Y_t.$$

Theorem B.1.4 (Gordon's inequality)**.** *Let $(X_{u,t})_{u \in U, v \in V}$ and $(X_{u,t})_{u \in U, v \in V}$ be two centered Gaussian processes indexed by $U \times V$. Assume that*

1. $\mathbb{E}(X_{u,v} - X_{u',v'})^2 \leq \mathbb{E}(Y_{u,v} - Y_{u',v'})^2$ *for all* v, v' *and* $u \neq u'$*;*

2. $\mathbb{E}(X_{u,v} - X_{u,v'})^2 \leq \mathbb{E}(Y_{u,v} - Y_{u,v'})^2$ *for all* u, v, v'*.*

Then,

$$\mathbb{E} \sup_{u \in U} \inf_{v \in V} X_{u,v} \leq \mathbb{E} \sup_{u \in U} \inf_{v \in V} Y_{u,v}.$$

With this two results we are ready to prove Lemma 3.2.3. Remember that we want to prove that for $\mathbf{X} \sim \mathcal{N}_{n \times p}(\mathbf{\Sigma})$ with $\mathbf{\Sigma} \succ \mathbf{0}$, then

$$\mathbb{E} s_1(\mathbf{X}) \leq \sqrt{n} \lambda_1(\sqrt{\mathbf{\Sigma}}) + \sqrt{\operatorname{tr} \mathbf{\Sigma}},$$

and if $n \geq p$,

$$\mathbb{E} \left[\min_{\boldsymbol{v} \in V(R)} \frac{\|\mathbf{X}\boldsymbol{v}\|_2}{\sqrt{n}} \right] \geq 1 - R \sqrt{\frac{\operatorname{tr} \mathbf{\Sigma}}{n}},$$

where $R = 1/\lambda_p(\sqrt{\mathbf{\Sigma}})$ and $V(R) = \{\boldsymbol{v} \in \mathbb{R}^p : \|\sqrt{\mathbf{\Sigma}}\boldsymbol{v}\|_2 = 1, \|\boldsymbol{v}\|_2 \leq R\}$.

Proof of Lemma 3.2.3.

(a) Define the random matrix $\mathbf{W} \sim \mathcal{N}_{n \times p}(\mathbf{I})$. Then, $\mathbf{X} \sim \mathbf{W}\sqrt{\mathbf{\Sigma}}$. Now, define the two subsets

$$U = \{\boldsymbol{u} \in \mathbb{R}^n : \|\boldsymbol{u}\|_2 = 1\} \quad \text{and} \quad V = \{\boldsymbol{v} \in \mathbb{R}^p : \|\mathbf{\Sigma}^{-1/2}\boldsymbol{v}\|_2 = 1\}.$$

Therefore, by Theorem A.5.4,

$$\begin{aligned} s_1(\mathbf{X}) &= \sup_{\|\boldsymbol{u}\|_2 =, \|\boldsymbol{y}\|_2 = 1} \boldsymbol{u}^\intercal \mathbf{X} \boldsymbol{y} \\ &\sim \sup_{\|\boldsymbol{u}\|_2 =, \|\boldsymbol{v}\|_2 = 1} \boldsymbol{u}^\intercal \mathbf{W} \sqrt{\mathbf{\Sigma}} \boldsymbol{y} \\ &= \sup_{\boldsymbol{u} \in U, \boldsymbol{v} \in V} \boldsymbol{u}^\intercal \mathbf{W} \boldsymbol{v}, \quad (\boldsymbol{v} = \sqrt{\mathbf{\Sigma}} \boldsymbol{y}). \end{aligned}$$

In this way, by Proposition B.1.2, finding an upper bound to $\mathbb{E} s_1(\mathbf{X})$ is equivalent to find and upper bound for the maximum of the centered Gaussian process $(Z_{\boldsymbol{u},\boldsymbol{v}})_{\boldsymbol{u},\boldsymbol{v} \in U \times V}$ defined as $Z_{\boldsymbol{u},\boldsymbol{v}} = \boldsymbol{u}^\intercal \mathbf{W} \boldsymbol{v}$. To do this, we'll use the Sudakov-Fernique inequality of

Theorem B.1.3, by defining another centered Gaussian process $(Y_{\boldsymbol{u},\boldsymbol{v}})_{\boldsymbol{u},\boldsymbol{v}\in U\times V}$ which satisfies that

$$\mathbb{E}\left(Z_{\boldsymbol{u},\boldsymbol{v}} - Z_{\hat{\boldsymbol{u}},\hat{\boldsymbol{v}}}\right)^2 \leq \mathbb{E}\left(Y_{\boldsymbol{u},\boldsymbol{v}} - Y_{\hat{\boldsymbol{u}},\hat{\boldsymbol{v}}}\right)^2, \tag{B.1}$$

for all $(\boldsymbol{u},\boldsymbol{v}),(\hat{\boldsymbol{u}},\hat{\boldsymbol{v}}) \in U \times V$. First, let us define $(Y_{\boldsymbol{u},\boldsymbol{v}})$ in order to prove the inequality (B.1). For $(\boldsymbol{u},\boldsymbol{v}) \in U \times V$, define

$$Y_{\boldsymbol{u},\boldsymbol{v}} = \lambda \boldsymbol{u}^\intercal \boldsymbol{g} + \boldsymbol{v}^\intercal \boldsymbol{h},$$

where $\lambda = \lambda_1(\sqrt{\boldsymbol{\Sigma}})$ and $\boldsymbol{g}, \boldsymbol{h}$ are independent with distribution $\mathcal{N}_n(\mathbf{0},\mathbf{I})$ and $\mathcal{N}_p(\mathbf{0},\mathbf{I})$, respectively. With the same procedure as in the proof of Proposition B.1.2, one can show easily that $(Y_{\boldsymbol{u},\boldsymbol{v}})$ is a centered Gaussian process. Now, for $(\boldsymbol{u},\boldsymbol{v}),(\hat{\boldsymbol{u}},\hat{\boldsymbol{v}}) \in U \times V$,

$$\begin{aligned}
\mathbb{E}\left(Y_{\boldsymbol{u},\boldsymbol{v}} - Y_{\hat{\boldsymbol{u}},\hat{\boldsymbol{v}}}\right)^2 &= \operatorname{Var}\left(Y_{\boldsymbol{u},\boldsymbol{v}} - Y_{\hat{\boldsymbol{u}},\hat{\boldsymbol{v}}}\right) \\
&= \operatorname{Cov}\left(\lambda(\boldsymbol{u}-\hat{\boldsymbol{u}})^\intercal \boldsymbol{g} + (\boldsymbol{v}-\hat{\boldsymbol{v}})^\intercal \boldsymbol{h}\right) \\
&= \lambda(\boldsymbol{u}-\hat{\boldsymbol{u}})^\intercal \left(\operatorname{Cov}\boldsymbol{g}\right)(\boldsymbol{u}-\hat{\boldsymbol{u}}) + (\boldsymbol{v}-\hat{\boldsymbol{v}})^\intercal \left(\operatorname{Cov}\boldsymbol{h}\right)(\boldsymbol{v}-\hat{\boldsymbol{v}}) \\
&= \lambda^2 \|\boldsymbol{u}-\hat{\boldsymbol{u}}\|_2^2 + \|\boldsymbol{v}-\hat{\boldsymbol{v}}\|_2^2.
\end{aligned}$$

On the other hand, let us define to elements $(\boldsymbol{u},\boldsymbol{v})$ and $(\hat{\boldsymbol{u}},\hat{\boldsymbol{v}})$ from $U \times V$ such that, without loss of generality[1], $\|\boldsymbol{v}\|_2 \leq \|\hat{\boldsymbol{v}}\|_2$. Then,

$$\begin{aligned}
\mathbb{E}\left(Z_{\boldsymbol{u},\boldsymbol{v}} - Z_{\hat{\boldsymbol{u}},\hat{\boldsymbol{v}}}\right)^2 &= \operatorname{Var}\left(Z_{\boldsymbol{u},\boldsymbol{v}} - Z_{\hat{\boldsymbol{u}},\hat{\boldsymbol{v}}}\right) \\
&= \operatorname{Var}\left(\boldsymbol{u}^\intercal \mathbf{W} \boldsymbol{v} - \hat{\boldsymbol{u}}^\intercal \mathbf{W} \hat{\boldsymbol{v}}\right) \\
&= \operatorname{Var}\left(\sum_{i=1}^n \sum_{j=1}^p (u_i v_j - \hat{u}_i \hat{v}_j) w_{ij}\right) \\
&= \sum_{i=1}^n \sum_{j=1}^p (u_i v_j - \hat{u}_i \hat{v}_j)^2 \\
&= \sum_{i=1}^n \sum_{j=1}^p (\boldsymbol{u}\boldsymbol{v}^\intercal - \hat{\boldsymbol{u}}\hat{\boldsymbol{v}}^\intercal)_{ij}^2 \\
&= |||\boldsymbol{u}\boldsymbol{v}^\intercal - \hat{\boldsymbol{u}}\hat{\boldsymbol{v}}^\intercal|||_2^2.
\end{aligned}$$

This last element can be decomposed in the following way[2],

$$|||\boldsymbol{u}\boldsymbol{v}^\intercal - \hat{\boldsymbol{u}}\hat{\boldsymbol{v}}^\intercal|||_2^2 = |||\boldsymbol{u}(\boldsymbol{v}-\hat{\boldsymbol{v}})^\intercal + (\boldsymbol{u}-\hat{\boldsymbol{u}})\hat{\boldsymbol{v}}^\intercal|||_2^2$$

[1] If this is not the case, we just use that $\left(Z_{\boldsymbol{u},\boldsymbol{v}} - Z_{\hat{\boldsymbol{u}},\hat{\boldsymbol{v}}}\right)^2 = \left(Z_{\hat{\boldsymbol{u}},\hat{\boldsymbol{v}}} - Z_{\boldsymbol{u},\boldsymbol{v}}\right)^2$.

[2] Recall that the inner product of two matrices is defined as $\langle \mathbf{A}, \mathbf{B}\rangle = \operatorname{tr}(\mathbf{A}^\intercal \mathbf{B})$, and that $|||\mathbf{A}|||_2^2 = \langle \mathbf{A}, \mathbf{A}\rangle$.

$$
\begin{aligned}
&= |\!|\!|\boldsymbol{u}(\boldsymbol{v}-\hat{\boldsymbol{v}})^{\mathsf{T}}|\!|\!|_2^2 + |\!|\!|(\boldsymbol{u}-\hat{\boldsymbol{u}})\hat{\boldsymbol{v}}^{\mathsf{T}}|\!|\!|_2^2 + 2\langle \boldsymbol{u}(\boldsymbol{v}-\hat{\boldsymbol{v}})^{\mathsf{T}}, (\boldsymbol{u}-\hat{\boldsymbol{u}})\hat{\boldsymbol{v}}^{\mathsf{T}}\rangle \\
&= \operatorname{tr}\left((\boldsymbol{v}-\hat{\boldsymbol{v}})\boldsymbol{u}^{\mathsf{T}}\boldsymbol{u}(\boldsymbol{v}-\hat{\boldsymbol{v}})^{\mathsf{T}}\right) + \operatorname{tr}\left(\hat{\boldsymbol{v}}(\boldsymbol{u}-\hat{\boldsymbol{u}})^{\mathsf{T}}(\boldsymbol{u}-\hat{\boldsymbol{u}})\hat{\boldsymbol{v}}^{\mathsf{T}}\right) \\
&\quad + 2\operatorname{tr}\left((\boldsymbol{v}-\hat{\boldsymbol{v}})\boldsymbol{u}^{\mathsf{T}}\hat{\boldsymbol{v}}(\boldsymbol{u}-\hat{\boldsymbol{u}})^{\mathsf{T}}\right) \\
&= \|\boldsymbol{u}\|_2^2\|\boldsymbol{v}-\hat{\boldsymbol{v}}\|_2^2 + \|\hat{\boldsymbol{v}}\|_2^2\|\boldsymbol{u}-\hat{\boldsymbol{u}}\|_2^2 + 2\left(\|\boldsymbol{u}\|_2^2 - \boldsymbol{u}^{\mathsf{T}}\hat{\boldsymbol{u}}^{\mathsf{T}}\right)\left(\boldsymbol{v}^{\mathsf{T}}\hat{\boldsymbol{v}} - \|\hat{\boldsymbol{v}}\|_2^2\right) \\
&= \|\boldsymbol{v}-\hat{\boldsymbol{v}}\|_2^2 + \|\hat{\boldsymbol{v}}\|_2^2\|\boldsymbol{u}-\hat{\boldsymbol{u}}\|_2^2 + 2\left(1 - \boldsymbol{u}^{\mathsf{T}}\hat{\boldsymbol{u}}\right)\left(\boldsymbol{v}^{\mathsf{T}}\hat{\boldsymbol{v}} - \|\hat{\boldsymbol{v}}\|_2^2\right).
\end{aligned}
$$

By Cauchy-Schwartz and the assumption $\|\boldsymbol{v}\|_2 \leq |\!|\!|\hat{\boldsymbol{v}}|\!|\!|_2$, we have that

$$
\boldsymbol{u}^{\mathsf{T}}\hat{\boldsymbol{u}} \leq \|\boldsymbol{u}\|_2\|\hat{\boldsymbol{u}}\|_2 = 1 \quad \text{and} \quad \boldsymbol{v}^{\mathsf{T}}\hat{\boldsymbol{v}} \leq \|\boldsymbol{v}\|_2\|\hat{\boldsymbol{v}}\|_2 \leq \|\hat{\boldsymbol{v}}\|_2^2,
$$

so $\left(\|\boldsymbol{u}\|_2^2 - \boldsymbol{u}^{\mathsf{T}}\hat{\boldsymbol{u}}^{\mathsf{T}}\right)\left(\boldsymbol{v}^{\mathsf{T}}\hat{\boldsymbol{v}} - \|\hat{\boldsymbol{v}}\|_2^2\right) \leq 0$, and in consequence

$$
|\!|\!|\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}} - \hat{\boldsymbol{u}}\hat{\boldsymbol{v}}^{\mathsf{T}}|\!|\!|_2^2 \leq \|\boldsymbol{v}-\hat{\boldsymbol{v}}\|_2^2 + \|\hat{\boldsymbol{v}}\|_2^2\|\boldsymbol{u}-\hat{\boldsymbol{u}}\|_2^2.
$$

Additionally, by definition of the set V and Rayleigh quotient (Theorem A.2.4),

$$
\|\hat{v}\|_2^2 \leq \max_{\boldsymbol{v}\in V}\|\boldsymbol{v}\|_2^2 = \max_{\|\boldsymbol{y}\|_2=1}\|\sqrt{\boldsymbol{\Sigma}}\|_2^2 = \max_{\|\boldsymbol{y}\|_2=1}\boldsymbol{y}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{y} = \lambda^2.
$$

Therefore,

$$
\mathbb{E}\left(Z_{\boldsymbol{u},\boldsymbol{v}} - Z_{\hat{\boldsymbol{u}},\hat{\boldsymbol{v}}}\right)^2 \leq \|\boldsymbol{v}-\hat{\boldsymbol{v}}\|_2^2 + \lambda^2\|\boldsymbol{u}-\hat{\boldsymbol{u}}\|_2^2 = \mathbb{E}\left(Y_{\boldsymbol{u},\boldsymbol{v}} - Y_{\hat{\boldsymbol{u}},\hat{\boldsymbol{v}}}\right)^2.
$$

Therefore, by Sudakov-Fenique inequality,

$$
\begin{aligned}
\mathbb{E}s_1(\mathbf{X}) = \mathbb{E}\sup_{(\boldsymbol{u},\boldsymbol{v})\in U\times V} Z_{\boldsymbol{u},\boldsymbol{v}} &\leq \mathbb{E}\sup_{(\boldsymbol{u},\boldsymbol{v})\in U\times V} Y_{\boldsymbol{u},\boldsymbol{v}} \\
&= \lambda^2\mathbb{E}\sup_{\boldsymbol{u}\in U}\boldsymbol{u}^{\mathsf{T}}\boldsymbol{g} + \mathbb{E}\sup_{\boldsymbol{v}\in V}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{h}.
\end{aligned}
$$

Note that $\boldsymbol{u}^{\mathsf{T}}\boldsymbol{g} \leq \|\boldsymbol{g}\|_2$ and $\boldsymbol{v}^{\mathsf{T}}\boldsymbol{h} \leq \boldsymbol{y}^{\mathsf{T}}\sqrt{\boldsymbol{\Sigma}}\boldsymbol{h} = \|\boldsymbol{y}\|_2\|\sqrt{\boldsymbol{\Sigma}}\boldsymbol{h}\|_2$, where $\|\boldsymbol{y}\|_2 = 1$. Then,

$$
\mathbb{E}s_1(\mathbf{X}) \leq \lambda^2\mathbb{E}\|\boldsymbol{g}\|_2 + \mathbb{E}\|\sqrt{\boldsymbol{\Sigma}}\boldsymbol{h}\|_2.
$$

Finally, the result follows by Jensen's inequality:

$$
\begin{aligned}
\mathbb{E}\|\boldsymbol{g}\|_2 &\leq \sqrt{\mathbb{E}\boldsymbol{g}^{\mathsf{T}}\boldsymbol{g}} = \sqrt{n} \\
\mathbb{E}\|\sqrt{\boldsymbol{\Sigma}}\boldsymbol{h}\|_2 &\leq \sqrt{\mathbb{E}(\boldsymbol{h}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{h})} = \sqrt{\operatorname{tr}\boldsymbol{\Sigma}}.
\end{aligned}
$$

(b) Recall that $R = 1/\lambda_p(\sqrt{\boldsymbol{\Sigma}})$ and $V(R) = \{\boldsymbol{v} \in \mathbb{R}^p : \|\sqrt{\boldsymbol{\Sigma}}\boldsymbol{v}\|_2 = 1, \|\boldsymbol{v}\|_2 \leq R\}$. We'll proceed like in (a) but instead of using Sudakov-Fernique inequality we'll use Gordon's inequality. The reason for this is that

$$-\min_{\boldsymbol{y}\in V(R)} \|\mathbf{X}\boldsymbol{y}\|_2 = \max_{\boldsymbol{y}\in V(R)} (-\|\mathbf{X}\boldsymbol{y}\|_2) = \max_{\boldsymbol{y}\in V(R)} \min_{\|\boldsymbol{u}\|_2=1} \boldsymbol{u}^\intercal \mathbf{X}\boldsymbol{y}.$$

The last equality follows because, by Cauchy-Schwarz, for any $\boldsymbol{x} \neq \mathbf{0}$ we have that $\boldsymbol{u}^\intercal\boldsymbol{x} \geq -\|\boldsymbol{x}\|_2$, and by taking $\boldsymbol{u} = -\boldsymbol{x}/\|\boldsymbol{x}\|_2$ the minimum of $\boldsymbol{u}^\intercal\boldsymbol{x}$ reaches $-\|\boldsymbol{x}\|_2$.

Define the set U as in (a) and the set $V'(R)$ as

$$V'(R) = \{\boldsymbol{v} \in \mathbb{R}^p : \|\boldsymbol{v}\|_2 = 1, \|\boldsymbol{\Sigma}^{-1/2}\|_2 \leq R\}.$$

As we argued in (a), for $\mathbf{W} \sim \mathcal{N}_{n\times p}$, $\mathbf{X} \sim \mathbf{W}\sqrt{\boldsymbol{\Sigma}}$ and

$$-\min_{\boldsymbol{v}\in V(R)} \|\mathbf{X}\boldsymbol{v}\|_2 \sim \max_{\boldsymbol{v}\in V'(R)} \min_{\boldsymbol{u}\in U} \boldsymbol{u}^\intercal \mathbf{W}\boldsymbol{v}, \quad (\boldsymbol{v} = \sqrt{\boldsymbol{\Sigma}}\boldsymbol{y}).$$

We define the Gaussian processes $(Z_{\boldsymbol{u},\boldsymbol{v}})_{(\boldsymbol{u},\boldsymbol{v})\in U\times V'(R)}$ as

$$Z_{\boldsymbol{u},\boldsymbol{v}} = \boldsymbol{u}^\intercal \mathbf{W}\boldsymbol{v}.$$

and $(Y_{\boldsymbol{u},\boldsymbol{v}})_{(\boldsymbol{u},\boldsymbol{v})\in U\times V'(R)}$ as

$$Y_{\boldsymbol{u},\boldsymbol{v}} = \boldsymbol{u}^\intercal\boldsymbol{g} + \boldsymbol{v}^\intercal\boldsymbol{h},$$

where $\boldsymbol{g} \in \mathbb{R}^n$ and $\boldsymbol{h} \in \mathbb{R}^p$ are independent standard Gaussian. Let $(\boldsymbol{u},\boldsymbol{v})$ and $(\hat{\boldsymbol{u}},\hat{\boldsymbol{v}})$ be two elements of $U \times V'(R)$. Then, $\|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = \|\hat{\boldsymbol{u}}\|_2 = \|\hat{\boldsymbol{v}}\|_2 = 1$, and by the same procedure of (a),

$$\mathbb{E}\left(Y_{\boldsymbol{u},\boldsymbol{v}} - Y_{\hat{\boldsymbol{u}},\hat{\boldsymbol{v}}}\right)^2 = \|\boldsymbol{u} - \hat{\boldsymbol{u}}\|_2 + \|\boldsymbol{v} - \hat{\boldsymbol{v}}\|_2,$$

and

$$\mathbb{E}\left(Z_{\boldsymbol{u},\boldsymbol{v}} - Z_{\hat{\boldsymbol{u}},\hat{\boldsymbol{v}}}\right)^2 = |||\boldsymbol{u}\boldsymbol{v}^\intercal - \hat{\boldsymbol{u}}\hat{\boldsymbol{v}}^\intercal|||_2^2,$$

so

$$\mathbb{E}\left(Z_{\boldsymbol{u},\boldsymbol{v}} - Z_{\hat{\boldsymbol{u}},\hat{\boldsymbol{v}}}\right)^2 \leq \mathbb{E}\left(Y_{\boldsymbol{u},\boldsymbol{v}} - Y_{\hat{\boldsymbol{u}},\hat{\boldsymbol{v}}}\right)^2,$$

and if $\boldsymbol{u} = \hat{\boldsymbol{u}}$, then

$$\mathbb{E}\left(Z_{\boldsymbol{u},\boldsymbol{v}} - Z_{\hat{\boldsymbol{u}},\hat{\boldsymbol{v}}}\right)^2 = \|\boldsymbol{u}\|_2^2\|\boldsymbol{v} - \hat{\boldsymbol{v}}\|_2^2 = \|\boldsymbol{v} - \hat{\boldsymbol{v}}\|_2^2 = \mathbb{E}\left(Y_{\boldsymbol{u},\boldsymbol{v}} - Y_{\hat{\boldsymbol{u}},\hat{\boldsymbol{v}}}\right)^2.$$

Therefore, by Gordon's inequality,

$$\begin{aligned}
\mathbb{E}\left[-\min_{\boldsymbol{y}\in V(R)}\|\mathbf{X}\boldsymbol{y}\|_2\right] &\leq \mathbb{E}\left[\max_{\boldsymbol{v}\in V'(R)}\min_{\boldsymbol{u}\in U} Y_{\boldsymbol{u},\boldsymbol{v}}\right]\\
&= \mathbb{E}\left[\max_{\boldsymbol{v}\in V'(R)}\min_{\boldsymbol{u}\in U}\left(\boldsymbol{u}^\intercal\boldsymbol{g}+\boldsymbol{v}^\intercal\boldsymbol{h}\right)\right]\\
&= \mathbb{E}\min_{\boldsymbol{u}\in U}\boldsymbol{u}^\intercal\boldsymbol{g} + \mathbb{E}\max_{\boldsymbol{v}\in V'(R)}\boldsymbol{v}^\intercal\boldsymbol{h}\\
&\leq -\mathbb{E}\|\boldsymbol{g}\|_2 + R\mathbb{E}\|\sqrt{\boldsymbol{\Sigma}}\boldsymbol{h}\|_2,
\end{aligned}$$

where the last equality arises from[3]

$$|\boldsymbol{u}^\intercal\boldsymbol{g}| \leq \|\boldsymbol{g}\|_2 \quad \text{and} \quad |\boldsymbol{v}^\intercal\boldsymbol{h}| = |(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{v})^\intercal\boldsymbol{\Sigma}^{1/2}\boldsymbol{h}| \leq \|\sqrt{\boldsymbol{\Sigma}}\boldsymbol{h}\|_2 R.$$

At the same time, we have that

$$\frac{\mathbb{E}\|\sqrt{\boldsymbol{\Sigma}}\boldsymbol{h}\|_2}{\sqrt{\operatorname{tr}\boldsymbol{\Sigma}}} \leq \frac{\mathbb{E}\|\boldsymbol{h}\|_2}{\sqrt{p}}.$$

Indeed, since the matrix $\boldsymbol{\Sigma}/\operatorname{tr}(\boldsymbol{\Sigma})$ is positive definite and the sum of its eigenvalues sum to one it can be diagonalized as $\boldsymbol{\Sigma}/\operatorname{tr}(\boldsymbol{\Sigma}) = \mathbf{P}\mathbf{D}\mathbf{P}^\intercal$, where $\lambda_j(\mathbf{D}) \geq 0$ for each j and $\sum_j \lambda_j(\mathbf{D}) = 1$. Also, since $\mathbf{P}^\intercal\boldsymbol{h} \sim \boldsymbol{h}$,

$$\mathbb{E}\|\sqrt{\boldsymbol{\Sigma}/\operatorname{tr}(\boldsymbol{\Sigma})}\boldsymbol{h}\|_2 = \mathbb{E}\|\mathbf{P}\sqrt{\mathbf{D}}\mathbf{P}^\intercal\boldsymbol{h}\|_2 = \mathbb{E}\|\sqrt{\mathbf{D}}\boldsymbol{h}\|_2,$$

and the result follows from Proposition C.6.1 of Appendix C.

Therefore,

$$\begin{aligned}
\mathbb{E}\left[-\min_{\boldsymbol{y}\in V(R)}\|\mathbf{X}\boldsymbol{y}\|_2\right] &\leq -\mathbb{E}\|\boldsymbol{g}\|_2 + R\sqrt{\operatorname{tr}\boldsymbol{\Sigma}}\frac{\mathbb{E}\|\boldsymbol{h}\|_2}{\sqrt{p}}\\
&= (-\mathbb{E}\|\boldsymbol{g}\|_2 + \mathbb{E}\|\boldsymbol{h}\|_2) + \left(\frac{\sqrt{\operatorname{tr}\boldsymbol{\Sigma}}}{\lambda_p(\sqrt{\boldsymbol{\Sigma}})\sqrt{p}} - 1\right)\mathbb{E}\|\boldsymbol{h}\|_2\\
&\leq -\sqrt{n} + \sqrt{p} + \left(\frac{\sqrt{\operatorname{tr}\boldsymbol{\Sigma}}}{\lambda_p(\sqrt{\boldsymbol{\Sigma}})\sqrt{p}} - 1\right)\sqrt{p}\\
&= -\sqrt{n} + \frac{\sqrt{\operatorname{tr}\boldsymbol{\Sigma}}}{\lambda_p(\sqrt{\boldsymbol{\Sigma}})},
\end{aligned}$$

where the inequality follows from Proposition C.6.2 and because $\operatorname{tr}\boldsymbol{\Sigma} \geq p\lambda_p(\boldsymbol{\Sigma})$. This ends the proof.

[3] For the first term we used again that $\min_{\|\boldsymbol{u}\|_2=1}\boldsymbol{u}^\intercal\boldsymbol{x} = -\|\boldsymbol{x}\|_2$.

□

B.2 Orlicz norms and sub-Gaussian random variables

In this section we define the concept of Orlicz norms, which are convenient quantities when working with sub-Gaussian and sub-exponential random variables.

Definition B.2.1 (Orlicz norms)**.** *A function $\psi : [0,\infty) \to [0,\infty)$ is called Orlicz if ψ is convex, increasing and satisfies*

$$\psi(0) = 0, \quad \psi(x) \to \infty,\, x \to \infty.$$

The ψ-Orlicz norm of a real-valued random variable X is defined as

$$\|X\|_\psi = \inf\left\{t > 0 : \mathbb{E}\psi(|X|/t) \le 1\right\}.$$

It is not difficult to see that $\|\cdot\|_\psi$ is in fact a norm for random variables. The only property that may seem not obvious is the triangle inequality. To see that it is true assume that $\|X\|_\psi, \|Y\|_\psi < \infty$ and note that because ψ is increasing and convex, for any $t, s > 0$,

$$\psi\left(\frac{|X+Y|}{t+s}\right) \le \frac{t}{t+s}\psi\left(\frac{|X|}{t}\right) + \frac{s}{t+s}\psi\left(\frac{|Y|}{s}\right). \tag{B.2}$$

Now, fix $\epsilon > 0$ and choose t, s such that $t < \|X\|_\psi + \epsilon/2$, $s < \|Y\|_\psi + \epsilon/2$ and

$$\max\left\{\psi(|X|/t), \psi(|Y|/s)\right\} \le 1.$$

Taking expectation in (B.2) we obtain that

$$\mathbb{E}\psi\left(\frac{|X+Y|}{t+s}\right) \le 1 \qquad \text{for all } t, s \text{ such that } t + s < \|X\|_\psi + \|Y\|_\psi + \epsilon.$$

Taking $\epsilon \downarrow 0$ we get by definition of the Orlicz norm that

$$\|X+Y\|_\psi \le \|X\|_\psi + \|Y\|_\psi.$$

The Orlicz function $\psi_2(x) = e^{x^2} - 1$ defines the sub-Gaussian norm of Chapter 3. Other Orlicz function is $\psi_1(x) = e^x - 1$ which defines the sub-exponential norm

$$\|X\|_{\psi_1} = \inf\left\{t > 0 : \mathbb{E}\exp(|X|/t) \le 2\right\}.$$

This norm characterizes sub-exponential random variables which are defined as follows.

Definition B.2.2 (Sub-exponential)**.** *A real-valued random variable X with $\mathbb{E}X = \mu$ is called sub-exponential if there exists parameters (α, β) such that*

$$\log \mathbb{E}e^{\theta(X-\mu)} \leq \frac{\alpha^2\theta^2}{2}, \quad \forall|\theta| < \frac{1}{\beta}.$$

The proof of the next Theorem can be found in [47, Chapter 2].

Theorem B.2.3 (Sub-Gaussian and sub-exponential random variables)**.** *Let X be a real-valued random variable. Then,*

(a) X is sub-Gaussian if and only if $\|X\|_{\psi_2} < \infty$.

(b) X is sub-exponential if and only if $\|X\|_{\psi_1} < \infty$.

(c) X is sub-Gaussian if and only if X^2 is sub-exponential.

(d) If $\psi \in \{\psi_1, \psi_2\}$ and $\|X\|_\psi < \infty$, then $\|X - \mathbb{E}X\|_\psi \leq C\|X\|_\psi$, where $C > 0$ is a constant that depends on ψ.

Theorem B.2.4 (Bernstein inequality)**.** *Let $X_1, ..., X_n$ be independent random variables such that $\mathbb{E}X_i = 0$ and $\|X_i\|_{\psi_1} < \infty$ for all i. Then for any $t \geq 0$,*

$$\mathbb{P}\left(\Big|\frac{1}{n}\sum_{i=1}^{n} X_i\Big| \geq t\right) \leq 2\exp\left(-cn\min\left\{\frac{t^2}{K^2}, \frac{t}{K}\right\}\right),$$

where $K = \max_i \|X_i\|_{\psi_1}$ and $c > 0$ is an absolute constant that depends on K.

B.3 ϵ-nets and matrices

Definition B.3.1 (ϵ-net)**.** *Let (T, d) be a metric space and consider $K \subset T$. For $\epsilon > 0$, the subset $N \subset K$ is called an ϵ-net of K if*

$$\forall x \in K, \, \exists x_0 \in N \, : \, d(x, x_0) \leq \epsilon.$$

Equivalently, N is an ϵ-net if an only if K can be covered by balls with centers in N and radius ϵ.

The smallest possible cardinality of an ϵ-net of K is called the covering number of K and is denoted $\mathcal{N}(\epsilon, K, d)$. Equivalently, $\mathcal{N}(\epsilon, K, d)$ is the smallest number of closed balls with centers in K with radius ϵ whose union covers K.

Proposition B.3.2. *The covering number of the unitary Euclidean sphere S^{p-1} in $\mathbb{R}^p$ defined as*

$$S^{p-1} = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_2 = 1\},$$

satisfies for any $\epsilon > 0$ that

$$\mathcal{N}(\epsilon, S^{p-1}, \|\cdot\|_2) \leq \left(\frac{2}{\epsilon} + 1\right)^p.$$

For the proof of Proposition B.3.2 we need the following definition: we say that a finite set $M \subset \mathbb{R}^p$ is ϵ-separate if $\|\boldsymbol{x} - \boldsymbol{x}\|_2 > \epsilon$ for any $\boldsymbol{x}, \boldsymbol{y} \in M$. We say that M is maximal if for any $\boldsymbol{x} \notin M$, $M \cup \{\boldsymbol{x}\}$ is not ϵ-separate.

Proof. Let $M \subset S^{p-1}$ be a maximal ϵ-separate set. Then, for any $\boldsymbol{x} \in S^{p-1}$, $M \cup \{\boldsymbol{x}\}$ is not ϵ-separate, which implies that $\|\boldsymbol{x} - \boldsymbol{x}_0\| \leq \epsilon$ for some $\boldsymbol{x}_0 \in M$, i.e., M is an ϵ-net of S^{p-1}. In the previous argument, if one chooses $\boldsymbol{x} \in M$, then it is obvious that $\boldsymbol{x}_0 = \boldsymbol{x}$. Then,

$$\mathcal{N}(\epsilon, S^{p-1}, \|\cdot\|_2) \leq |M|.$$

For any $\boldsymbol{x}, \boldsymbol{y} \in M$, $\boldsymbol{x} \neq \boldsymbol{y}$, the balls[4] $B(\boldsymbol{x}, \epsilon/2)$ and $B(\boldsymbol{y}, \epsilon/2)$ are disjoint. Indeed, suppose that $\boldsymbol{z} \in B(\boldsymbol{x}, \epsilon/2) \cap B(\boldsymbol{y}, \epsilon/2)$, then

$$\|\boldsymbol{x} - \boldsymbol{y}\|_2 \leq \|\boldsymbol{x} - \boldsymbol{z}\|_2 + \|\boldsymbol{y} - \boldsymbol{z}\|_2 = \epsilon/2 + \epsilon/2 = \epsilon,$$

which can't be possible since M is ϵ-separate. On the other hand, $\cup_{\boldsymbol{x} \in M} B(\boldsymbol{x}, \epsilon/2)$ is contained in $B(\mathbf{0}, 1 + \epsilon/2)$, since for any $\boldsymbol{y} \in \cup_{\boldsymbol{x} \in M} B(\boldsymbol{x}, \epsilon/2)$, there exists some $\boldsymbol{x}$ such that $\|\boldsymbol{y} - \boldsymbol{x}\|_2 \leq \epsilon/2$, so

$$\|\boldsymbol{y}\|_2 \leq \|\boldsymbol{y} - \boldsymbol{x}\|_2 + \|\boldsymbol{x}\|_2 = \epsilon/2 + 1.$$

Therefore,

$$|M| \left(\frac{\epsilon}{2}\right)^p = \mathrm{Vol}(\cup_{\boldsymbol{x} \in M} B(\boldsymbol{x}, \epsilon/2)) \leq \mathrm{Vol}(B(\mathbf{0}, 1 + \epsilon/2)) = \left(1 + \frac{\epsilon}{2}\right)^p.$$

The later implies that $|M| \leq (1 + \epsilon/2)^p$. This ends the proof. □

[4] $B(\boldsymbol{x}, \epsilon)$ is a ball with center $\boldsymbol{x}$ and radius ϵ.

Proposition B.3.3. *For $\epsilon \in [0, 1/2)$, let N be a ϵ-net of S^{p-1}. Then, for any $\mathbf{A} \in \mathcal{S}_p$,*

$$|||\mathbf{A}||| \leq \frac{1}{1-2\epsilon} \max_{x \in N} |\langle \mathbf{A}\boldsymbol{x}, \boldsymbol{x}\rangle|.$$

Proof. Fix $\boldsymbol{x} \in S^{p-1}$ such that

$$|\langle \mathbf{A}\boldsymbol{x}, \boldsymbol{x}\rangle| = |||\mathbf{A}|||.$$

Observe that such $\boldsymbol{x}$ exists by Proposition A.5.11. Let $\boldsymbol{x}_0 \in N$ be an element such that

$$\|\boldsymbol{x} - \boldsymbol{x}_0\|_2 \leq \epsilon.$$

Then, because $\|\mathbf{A}\boldsymbol{y}\|_2 \leq |||\mathbf{A}|||\|\boldsymbol{y}\|_2$ for any $\boldsymbol{y} \in \mathbb{R}^p$ (see the proof of Lemma A.5.8), we have by triangle inequality and Cauchy-Schwarz that

$$\begin{aligned}
|\langle \mathbf{A}\boldsymbol{x}, \boldsymbol{x} - \boldsymbol{x}_0\rangle + \langle \mathbf{A}(\boldsymbol{x} - \boldsymbol{x}_0), \boldsymbol{x}_0\rangle| &\leq |\langle \mathbf{A}\boldsymbol{x}, \boldsymbol{x} - \boldsymbol{x}_0\rangle| + |\langle \mathbf{A}(\boldsymbol{x} - \boldsymbol{x}_0), \boldsymbol{x}_0\rangle| \\
&\leq \|\mathbf{A}\boldsymbol{x}\|_2\|\boldsymbol{x} - \boldsymbol{x}_0\|_2 + \|\mathbf{A}(\boldsymbol{x} - \boldsymbol{x}_0)\|_2\|\boldsymbol{x}_0\|_2 \\
&\leq |||\mathbf{A}|||\|\boldsymbol{x}\|_2\epsilon + |||\mathbf{A}|||\|\boldsymbol{x} - \boldsymbol{x}_0\|_2 \\
&\leq |||\mathbf{A}|||\epsilon + |||\mathbf{A}|||\epsilon \\
&= 2\epsilon|||\mathbf{A}|||.
\end{aligned}$$

Finally, applying again the triangle inequality,

$$\begin{aligned}
|\langle \mathbf{A}\boldsymbol{x}_0, \boldsymbol{x}_0\rangle| &\geq |\langle \mathbf{A}\boldsymbol{x}, \boldsymbol{x}\rangle| - |\langle \mathbf{A}\boldsymbol{x}, \boldsymbol{x} - \boldsymbol{x}_0\rangle + \langle \mathbf{A}(\boldsymbol{x} - \boldsymbol{x}_0), \boldsymbol{x}_0\rangle| \\
&\geq |||\mathbf{A}||| - 2\epsilon|||\mathbf{A}||| \\
&= (1 - 2\epsilon)|||\mathbf{A}|||.
\end{aligned}$$

The can be obtained for each $\boldsymbol{x}_0 \in N$, so we conclude that

$$\sup_{\boldsymbol{x} \in N} |\langle \mathbf{A}\boldsymbol{x}, \boldsymbol{x}\rangle| \geq (1 - 2\epsilon)|||\mathbf{A}|||.$$

□

Now we are ready to prove Theorem 3.3.2 of Chapter 3.

Proof of Theorem 3.3.2. By Proposition B.3.2, we can choose a $(1/4)$-net N of the sphere S^{p-1} with cardinality

$$|N| \leq \left(\frac{2}{1/4} + 1\right)^p = 9^p.$$

Then, by Proposition B.3.3,

$$\left\|\frac{1}{n}\mathbf{Z}^\intercal\mathbf{Z}-\mathbf{I}\right\| \leq \frac{1}{1-2(1/4)}\max_{\boldsymbol{x}\in N}\left|\left\langle\left(\frac{1}{n}\mathbf{Z}^\intercal\mathbf{Z}-\mathbf{I}\right)\boldsymbol{x},\boldsymbol{x}\right\rangle\right| = 2\max_{\boldsymbol{x}\in N}\left|\frac{1}{n}\|\mathbf{Z}\boldsymbol{x}\|_2-1\right|.$$

By the previous display, we just need to obtain the desired probability for the event

$$2\max_{\boldsymbol{x}\in N}\left|\frac{1}{n}\|\mathbf{Z}\boldsymbol{x}\|_2-1\right| \leq \epsilon, \quad \epsilon>0.$$

To do so, observe that[5] for any $\boldsymbol{x}\in N$,

$$\|\mathbf{Z}\boldsymbol{x}\|_2^2 = \sum_{j=1}^n (\boldsymbol{Z}_j^\intercal\boldsymbol{x})^2.$$

Define the random variables $X_j = \boldsymbol{Z}_j^\intercal\boldsymbol{x}$. Recall that $\boldsymbol{Z}_j$ is sub-Gaussian[6] with norm bounded by $K=\max_j\|\boldsymbol{Z}_j\|_{\psi_2}$, so

$$\|X_j\|_{\psi_2} \leq \|\boldsymbol{Z}_j\|_{\psi_2} \leq K.$$

Then, X_j is sub-Gaussian. Even more, $\mathbb{E}X_j=0$ and, since $\boldsymbol{Z}_j$ are isotropic,

$$\mathbb{E}X_j^2 = \mathbb{E}[\boldsymbol{x}^\intercal\boldsymbol{Z}_j\boldsymbol{Z}_j^\intercal\boldsymbol{x}] = \boldsymbol{x}^\intercal\mathbb{E}[\boldsymbol{Z}_j\boldsymbol{Z}_j^\intercal]\boldsymbol{x} = 1.$$

Therefore, by Theorem B.2.3 the random variables X_j^2-1 are sub-exponential with mean zero and

$$\|X_j^2-1\|_{\psi_1} \leq C'K^2, \quad \text{for some absolute constant } C'>0.$$

Define $L=2C'$ and[7] $\epsilon = LK^2\max\{\delta,\delta^2\}$. With this definition,

$$\min\left\{\left(\frac{\epsilon}{LK^2}\right)^2, \frac{\epsilon}{LK^2}\right\} = \min\left\{\max(\delta^2,\delta^4),\max(\delta,\delta^2)\right\} = \delta^2.$$

Then, applying Bernstein inequality of Theorem B.2.4 we obtain that[8]

$$\mathbb{P}\left(2\left|\frac{1}{n}\|\mathbf{Z}\boldsymbol{x}\|_2-1\right| \geq \epsilon\right) = \mathbb{P}\left(\left|\frac{1}{n}\sum_{j=1}^n(X_j^2-1)\right| \geq \frac{\epsilon}{2}\right)$$

[5] $\mathbf{Z}$ is the $n\times p$ ensemble defined as $\mathbf{Z}^\intercal = (\boldsymbol{Z}_1,...,\boldsymbol{Z}_n)$.

[6] The sub-Gaussian norm of a random vector $\boldsymbol{X}$ is $\|\boldsymbol{X}\|_{\psi_2} = \sup_{\|\boldsymbol{x}\|_2=1}\|\langle\boldsymbol{X},\boldsymbol{x}\rangle\|_{\psi_2}$.

[7] Recall that $\delta = C\left(\sqrt{\frac{p}{n}}+\frac{t}{\sqrt{n}}\right)$.

[8] In this case observe that $C'K^2 \geq \max_j\|X_j^2-1\|_{\psi_1}$.

$$
\begin{aligned}
&\leq 2\exp\left(-cn\min\left\{\left(\frac{\epsilon/2}{C'K^2}\right)^2, \frac{\epsilon/2}{C'K^2}\right\}\right) \\
&= 2\exp\left(-cn\min\left\{\left(\frac{\epsilon}{LK^2}\right)^2, \frac{\epsilon}{LK^2}\right\}\right) \\
&= 2\exp\left(-cn\delta^2\right) \\
&\leq 2\exp\left(-cC^2(p+t^2)\right), \quad ((x+y)^2 \geq x^2+y^2 \text{ for all } x, y \geq 0).
\end{aligned}
$$

Here, $c > 0$ is an absolute constant that depends on LK^2. Finally, by the union bound,

$$
\begin{aligned}
\mathbb{P}\left(2\max_{\boldsymbol{x}\in N}\left|\frac{1}{n}\|\mathbf{Z}\boldsymbol{x}\|_2 - 1\right| \geq \epsilon\right) &= \mathbb{P}\left(\bigcup_{\boldsymbol{x}\in N}\left|\frac{1}{n}\|\mathbf{Z}\boldsymbol{x}\|_2 - 1\right| \geq \frac{\epsilon}{2}\right) \\
&\leq \sum_{\boldsymbol{x}\in N}\mathbb{P}\left(\left|\frac{1}{n}\|\mathbf{Z}\boldsymbol{x}\|_2 - 1\right| \geq \frac{\epsilon}{2}\right) \\
&\leq \sum_{\boldsymbol{x}\in N} 2\exp\left(-cC^2(p+t^2)\right) \\
&= 9^p 2\exp\left(-cC^2(p+t^2)\right).
\end{aligned}
$$

If we choose the constant C such that[9]

$$
C^2 = \sup_{t\geq 0}\frac{p\log(9)+t^2}{ct^2+cp} = \frac{\log(9)}{p},
$$

then we have that for any $t \geq 0$

$$
-cC^2(p+t^2) + p\log(9) \leq -t^2 \quad \text{and} \quad 9^p\exp\left(-cC^2(p+t^2)\right).
$$

Therefore, the final bound is

$$
\mathbb{P}\left(2\max_{\boldsymbol{x}\in N}\left|\frac{1}{n}\|\mathbf{Z}\boldsymbol{x}\|_2 - 1\right| \geq LK^2\max\{\delta,\delta^2\}\right) \leq 2e^{-t^2}.
$$

Since ,

$$
\left[\left\|\frac{1}{n}\mathbf{Z}^\intercal\mathbf{Z} - \mathbf{I}\right\| \geq LK^2\max\{\delta,\delta^2\}\right] \subset \left[2\max_{\boldsymbol{x}\in N}\left|\frac{1}{n}\|\mathbf{Z}\boldsymbol{x}\|_2 - 1\right| \geq LK^2\max\{\delta,\delta^2\}\right],
$$

the proof is done. □

[9] The function $t \mapsto (n\log(9)+t^2)/(ct^2+cn)$ has negative derivative in $[0,\infty)$ given by $t \mapsto (1-\log(9))t/(ct^2+cn)^2$, so it is decreasing and bounded by $(n\log(9)+0^2)/(c0^2+cn) = \log(9)/c$.

Appendix C

Miscellaneous results

The following results do not aim at a specific topic. Instead, they are presented to complement the development of the chapters of this thesis.

C.1 Integral Representation of the Logarithm

Proposition C.1.1. *For any $x > 0$,*

$$\log x = \int_0^\infty \left(\frac{1}{1+y} - \frac{1}{x+y} \right) \, dy.$$

Proof. First, for $a \geq 0$,

$$\int_0^a \left(\frac{1}{1+y} - \frac{1}{x+y} \right) \mathrm{d}y = \log(1) - \log(1+a) - \log(x) + \log(x+a) = \log(x) + \log\left(\frac{1+a}{x+a}\right).$$

Taking the limit when $a \to \infty$ yields the result. □

Corollary C.1.2. *For any $\mathbf{A} \succ \mathbf{0}$,*

$$\log \mathbf{A} = \int_0^\infty \left((1+y)^{-1}\mathbf{I} - (\mathbf{A} + y\mathbf{I})^{-1} \right) \, dy,$$

where the integral is taking entry-wise.

Proof. Is immediate from Proposition C.1.1, Definition 2.1.1 and Example 2.1.3. □

C.2 Measurability of eigenvalues

An equivalent definition of random matrix is the next one: if $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, a random matrix is a measurable map $\mathbf{Z} : \Omega \to \mathcal{M}_{p,q}$, i.e., every pre-image[1] $\mathbf{Z}^{-}B$ falls in $\mathcal{F}$, where $B \in \mathbb{B}(\mathcal{M}_{p,q})$ and $\mathbb{B}(\mathcal{M}_{p,q})$ is the σ-álgebra generated by open sets[2] of $\mathcal{M}_{p,q}$. One can find more details in [40, Chapter 1]. Therefore, if $f : \mathcal{M}_{p,q} \to \mathbb{R}$ is a continuous function, we can conclude that $f(\mathbf{Z})$ is measurable whenever $\mathbf{Z}$ is measurable.

Theorem C.2.1. *Let $\mathbf{X} \in \mathcal{S}_p$ be a random matrix. Then $\lambda_j(\mathbf{X})$, $j = 1, ..., p$ is measurable.*

Proof. Since the application of a continuous function to a measurable function preserves the measurability, we just need to prove that, for each j, $\lambda_j : \mathcal{S}_p \to \mathbb{R}$ is indeed continuous. Observe that from Corollary A.2.6 of Weyl's inequalities we get that, for any $\mathbf{A}, \mathbf{H} \in \mathcal{S}_p$,

$$\lambda_p(\mathbf{A} - \mathbf{H}) \leq \lambda_j(\mathbf{A} - \mathbf{H} + \mathbf{H}) - \lambda_j(\mathbf{H}) \leq \lambda_1(\mathbf{A} - \mathbf{H}).$$

Also, $\lambda_1(\mathbf{A} - \mathbf{H}) \leq |||\mathbf{A} - \mathbf{H}|||$ and $\lambda_p(\mathbf{A} - \mathbf{H}) \geq -|||\mathbf{A} - \mathbf{H}|||$, so

$$|\lambda_j(\mathbf{A}) - \lambda_j(\mathbf{H})| \leq |||\mathbf{A} - \mathbf{H}|||.$$

Then λ_j is Lipschitz (see definition 3.1.3 in Chapter 3) and in consequence continuous. This ends the proof. □

C.3 Jensen's inequality

Lemma C.3.1 (Jensen's inequality for matrices)**.** *For any random matrix $\mathbf{Z} \in \mathcal{M}_{p,q}$ and any real values function h defined on $\mathcal{M}_{p,q}$ we have that*

$$\begin{aligned} \mathbb{E}h(\mathbf{Z}) &\leq h(\mathbb{E}\mathbf{Z}) \quad \textit{if } h \textit{ is concave,} \\ \mathbb{E}h(\mathbf{Z}) &\geq h(\mathbb{E}\mathbf{Z}) \quad \textit{if } h \textit{ is convex.} \end{aligned}$$

Proof. We'll just prove it for the convex case, since the concave case can be obtained by change of symbol. From the section of subdifferential calculus of Section C.10, we know that the subdifferential of h is defined as

$$\partial h(\mathbf{B}) = \{\mathbf{G} : h(\mathbf{F}) \geq h(\mathbf{B}) + \langle \mathbf{G}, \mathbf{F} - \mathbf{B} \rangle, \forall \mathbf{F} \in \mathcal{M}_{p,q}\},$$

[1] The notation $\mathbf{Z}^{-}B$ in this context indicates the set of $\omega's$ such that $\mathbf{Z}(\omega) \in B$.

[2] We work with the metric space $(\mathcal{M}_{p,q}, ||| \cdot |||)$ so we can define open sets.

where $\langle \mathbf{B}_1, \mathbf{B}_2 \rangle = \text{tr}\ (\mathbf{B}_1^\intercal \mathbf{B}_2)$. Then, for any $\mathbf{G} \in \partial h(\mathbb{E}\mathbf{Z})$ we have that

$$h(\mathbf{Z}) \geq h(\mathbb{E}\mathbf{Z}) + \langle \mathbf{G}, \mathbf{Z} - \mathbb{E}\mathbf{Z} \rangle.$$

Taking expectations and noting that $\mathbb{E}\langle \mathbf{G}, \mathbf{Z} - \mathbb{E}\mathbf{Z} \rangle = \text{tr}\ (\mathbb{E}[\mathbf{G}^\intercal(\mathbf{Z} - \mathbb{E}\mathbf{Z})]) = 0$ ends the proof. □

C.4 Properties of covariance matrix estimator

Proposition C.4.1. *Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ be iid copies of the random vector $\boldsymbol{X} \in \mathbb{R}^p$ with $\mathbb{E}\boldsymbol{X} = \boldsymbol{\mu}$ and $\text{Cov}\boldsymbol{X} = \boldsymbol{\Sigma}$. Define the random matrix*

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{X}_j - \bar{\boldsymbol{X}})(\boldsymbol{X}_j - \bar{\boldsymbol{X}})^\intercal.$$

Then, $\frac{n}{n-1}\widehat{\boldsymbol{\Sigma}}$ is an unbiased estimator of $\boldsymbol{\Sigma}$, i.e., $\widehat{\boldsymbol{\Sigma}}$ is asymptotically unbiased.

Proof. First note that

$$(\boldsymbol{X}_j - \boldsymbol{\mu})(\boldsymbol{X}_j - \boldsymbol{\mu})^\intercal = \boldsymbol{X}_j\boldsymbol{X}_j^\intercal - \boldsymbol{X}_j\boldsymbol{\mu}^\intercal - \boldsymbol{\mu}\boldsymbol{X}_j^\intercal + \boldsymbol{\mu}\boldsymbol{\mu}^\intercal.$$

Then for every $j = 1, ..., n$

$$\mathbb{E}\boldsymbol{X}_j\boldsymbol{X}_j^\intercal = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\intercal.$$

Also, $\text{Cov}\bar{\boldsymbol{X}} = \frac{1}{n}\boldsymbol{\Sigma}$. This is true because

$$(\bar{\boldsymbol{X}} - \boldsymbol{\mu})(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^\intercal = \begin{pmatrix} (\bar{X}_1 - \mu_1)^2 & \cdots & (\bar{X}_1 - \mu_1)(\bar{X}_p - \mu_p) \\ \vdots & \ddots & \vdots \\ (\bar{X}_1 - \mu_1)(\bar{X}_p - \mu_p) & \cdots & (\bar{X}_p - \mu_p)^2 \end{pmatrix},$$

$\mathbb{E}(\bar{X}_1 - \mu_1)^2 = \sigma_1^2/n$ and for $i \neq j$,

$$\mathbb{E}(\bar{X}_i - \mu_i)(\bar{X}_j - \mu_j) = \frac{1}{n^2}\text{Cov}\left(\sum_{k=1}^{n} X_{ki}, \sum_{k=1}^{n} X_{kj}\right) = \frac{1}{n^2}n\text{Cov}(X_i, X_j) = \frac{1}{n}\text{Cov}(X_i, X_j).$$

Therefore,

$$
\begin{aligned}
\mathbb{E}\widehat{\boldsymbol{\Sigma}} &= \frac{1}{n}\mathbb{E}\sum_{j=1}^{n}\left[\boldsymbol{X}_j\boldsymbol{X}_j^\intercal - \boldsymbol{X}_j\bar{\boldsymbol{X}}^\intercal - \bar{\boldsymbol{X}}\boldsymbol{X}_j^\intercal + \bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^\intercal\right] \\
&= \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\boldsymbol{X}_j\boldsymbol{X}_j^\intercal - \mathbb{E}\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^\intercal \\
&= \frac{1}{n}\sum_{j=1}^{n}[\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\intercal] - \frac{1}{n}\boldsymbol{\Sigma} - \boldsymbol{\mu}\boldsymbol{\mu}^\intercal \\
&= \frac{n-1}{n}\boldsymbol{\Sigma}.
\end{aligned}
$$

This ends the proof. □

In the case $\boldsymbol{\mu} = \boldsymbol{0}$ the estimator

$$
\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{X}_j\boldsymbol{X}_j^\intercal
$$

is indeed unbiased because $\mathbb{E}\boldsymbol{X}_j\boldsymbol{X}_j^\intercal = \boldsymbol{\Sigma} + \boldsymbol{0}\boldsymbol{0}^\intercal = \boldsymbol{\Sigma}$.

Proposition C.4.2. *Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ be iid copies of the random vector $\boldsymbol{X} \in \mathbb{R}^p$ with $\mathbb{E}\boldsymbol{X} = \boldsymbol{\mu}$ and $\operatorname{Cov}\boldsymbol{X} = \boldsymbol{\Sigma}$. Then, for any $i = 1, ..., n$,*

$$
\mathbb{E}\left[(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\intercal\right] = \frac{n-1}{n}\boldsymbol{\Sigma}.
$$

Proof. Since $\mathbb{E}[\boldsymbol{X}_i\boldsymbol{X}_i^\intercal] = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\intercal$ and $\mathbb{E}[\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^\intercal] = \frac{1}{n}\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\intercal$ (see the proof of the previous proposition) we get that

$$
\begin{aligned}
\mathbb{E}\left[(X_i - \bar{\boldsymbol{X}})(X_i - \bar{\boldsymbol{X}})^\intercal\right] &= \mathbb{E}\left[\boldsymbol{X}_i\boldsymbol{X}_i^\intercal - \bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^\intercal - \boldsymbol{X}_i\bar{\boldsymbol{X}}^\intercal - \bar{\boldsymbol{X}}\boldsymbol{X}_i^\intercal\right] \\
&= \frac{n+1}{n}\boldsymbol{\Sigma} + 2\boldsymbol{\mu}\boldsymbol{\mu}^\intercal - \mathbb{E}[\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^\intercal] - \mathbb{E}[\boldsymbol{X}_i\bar{\boldsymbol{X}}^\intercal - \bar{\boldsymbol{X}}\boldsymbol{X}_i^\intercal].
\end{aligned}
$$

On the other hand, for $\ell = 1, ..., n$ and $k = 1, ..., p$,

$$
\mathbb{E}[X_{i\ell}\bar{X}_k] = \frac{1}{n}\left(\mathbb{E}[X_{i\ell}X_{ik}] + \sum_{j \neq i}\mathbb{E}[X_{i\ell}X_{jk}]\right)
$$

$$= \frac{1}{n} \left(\sigma_{\ell k} + n \mu_\ell \mu_k \right).$$

So, $\mathbb{E}[\boldsymbol{X}_i \bar{\boldsymbol{X}}^\intercal] = \mathbb{E}[\bar{\boldsymbol{X}} \boldsymbol{X}_i] = \frac{1}{n}\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\intercal$. Therefore,

$$\mathbb{E}\left[(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^\intercal\right] = \frac{n+1}{n}\boldsymbol{\Sigma} + 2\boldsymbol{\mu}\boldsymbol{\mu}^\intercal - 2\frac{1}{n}\boldsymbol{\Sigma} + 2\boldsymbol{\mu}\boldsymbol{\mu}^\intercal = \frac{n-1}{n}\boldsymbol{\Sigma}.$$

□

C.5 Convergence in probability

We say that a sequence $X_0, X_1, ...$ of random variables converges in probability to a random variable X if for every $t > 0$

$$\mathbb{P}(|X_n - X| \geq t) \longrightarrow 0, \quad n \to \infty,$$

and we write $X_n \xrightarrow{\mathbb{P}} X$.

Proposition C.5.1. *Let $(X_n)_{n \geq 0}$ be a sequence of non-negative random variables. Define the sequence of functions $\epsilon_n : (0, \infty) \to (0, \infty)$ such that $\epsilon_n(\delta) \downarrow f(\delta)$ when $n \to \infty$, where f is a increasing function. If there exists a sequence of reals (b_n) such that $b_n \downarrow 0$ and*

$$\mathbb{P}\left(X_n \geq \epsilon_n(\delta)\right) \leq b_n, \quad \forall\, n \geq 0\, \delta > 0,$$

then $X_n \xrightarrow{\mathbb{P}} 0$.

Proof. For any $t > 0$, there exist $\delta_t > 0$ such that $t/2 > f(\delta_t)$. Since $\epsilon_n(\delta_t) - f(\delta_t) \downarrow 0$ when $n \to \infty$, for n sufficiently large

$$\epsilon_n(\delta_t) - f(\delta_t) \leq \frac{t}{2} \quad \text{and} \quad \epsilon_n(\delta_t) \leq t.$$

Therefore, for any $t > 0$,

$$\mathbb{P}(X_n \geq t) \leq \mathbb{P}(X_n \geq \epsilon_n(\delta_t)) \leq b_n \longrightarrow 0,\ n \to \infty.$$

Then, by definition, $X_n \xrightarrow{\mathbb{P}} 0$. □

C.6 Two unseful inequalities for Gaussian vectors

Proposition C.6.1. *Let $\boldsymbol{Y} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I})$. Define $\mathcal{D}$ as the set of all $p \times p$ positive definite and diagonal matrices such that $\sum_{j=1}^{p} \lambda_j(\mathbf{D}) = 1$ for all $\mathbf{D} \in \mathcal{D}$. Let $F : \mathcal{D} \to [0, \infty)$ be defined as*

$$F(\mathbf{D}) = \mathbb{E}\|\sqrt{\mathbf{D}}\boldsymbol{Y}\|_2.$$

Then, for any $\mathbf{D} \in \mathcal{D}$ we have that

$$F(\mathbf{D}) \leq \frac{1}{\sqrt{p}}\mathbb{E}\|\boldsymbol{Y}\|_2.$$

Proof. Define the probability simplex Λ as

$$\Lambda = \left\{ \boldsymbol{\lambda} = (\lambda_1, ...\lambda_p)^\intercal \in \mathbb{R}^p \ : \ \lambda_j \geq 0 \text{ for all } j \, , \sum_{j=1}^{p} \lambda_j = 1 \right\}.$$

We can express F as a function from Λ to $[0, \infty)$, namely,

$$F(\lambda_1, ..., \lambda_p) = \mathbb{E}\sqrt{\sum_{j=1}^{p} Y_j^2 \lambda_j}.$$

It is clear that F is continuous and permutation invariant, i.e., $F(\lambda_{\pi(1)}, ..., \lambda_{\pi(p)}) = F(\lambda_1, ..., \lambda_p)$ for each permutation π of the set $\{1, 2, ..., p\}$. Also F is concave since it is the expected value of a concave function[3]. Additionally, the probability simplex Λ defined is a compact set. Therefore, the function F reaches its maximum on Λ.

Let $\boldsymbol{\lambda}^* \in \Lambda$ be the maximum of F and take and arbitrary $\boldsymbol{\lambda}$ in the convex hull of $\Pi(\boldsymbol{\lambda}^*)$, the set of all permutations of the entries of $\boldsymbol{\lambda}^*$. Then, since F is concave and permutation invariant, for some $\alpha_1, \alpha_2, ... \geq 0$ such that $\sum_i \alpha_i = 1$ we get that

$$\begin{aligned} F(\boldsymbol{\lambda}) &= F\left(\sum_{\boldsymbol{x}_i \in \Pi(\boldsymbol{\lambda}^*)} \alpha_i \boldsymbol{x}_i \right) \\ &\geq \sum_{\boldsymbol{x}_i \in \Pi(\boldsymbol{\lambda}^*)} \alpha_i F(\boldsymbol{x}_i) \end{aligned}$$

[3] It is not difficult to see that the hessian of $\boldsymbol{\lambda} \mapsto \sqrt{\sum_{j=1}^{p} \lambda_j Y_j^2}$ is negative definite.

$$
\begin{aligned}
&= F(\boldsymbol{\lambda}^*) \sum_{\boldsymbol{x}_i \in \Pi(\boldsymbol{\lambda}^*)} \alpha_i \\
&= F(\boldsymbol{\lambda}^*).
\end{aligned}
$$

Therefore, $F(\boldsymbol{\lambda}) = F(\boldsymbol{\lambda}^*)$ for any $\boldsymbol{\lambda}$ in the convex hull of $\Pi(\boldsymbol{\lambda}^*)$. Since the entries of $\boldsymbol{\lambda}^*$ sum up to one, it is easy to verify that the vector $(1/p, ..., 1/p)^\intercal$ belongs to the convex hull of $\Pi(\boldsymbol{\lambda}^*)$[4]. Therefore, for each $\mathbf{D} \in \mathcal{D}$,

$$
\frac{1}{\sqrt{p}} \mathbb{E}\|\boldsymbol{Y}\|_2 = F(1/p, ..., 1/p) \geq F(\lambda_1(\mathbf{D}), ..., \lambda_p(\mathbf{D})) = F(\mathbf{D}).
$$

□

Proposition C.6.2. *Let $\boldsymbol{X} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{Y} \sim \mathcal{N}_m(\mathbf{0}, \mathbf{I})$ where $n \geq m$ and $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent. Then,*

$$
-\mathbb{E}\|\boldsymbol{X}\|_2 + \mathbb{E}\|\boldsymbol{Y}\|_2 \leq -\sqrt{n} + \sqrt{m}.
$$

Proof. It will be sufficient to prove it for $n = m+1$ since the general case follows inductively. Let $W_1, W_2, ...$ be a sequence of iid standard Gaussian random variables and define $Z = \sum_{j=1}^m W_j^2$.

It's easy to see that the two functions

$$
x \mapsto \sqrt{c + x^2},\ c \geq 0, \quad \text{and} \quad x \mapsto \sqrt{x+1} - \sqrt{x}
$$

are convex in $[0, \infty)$[5]. Therefore, by Jensen's inequality and the independence of the variables $W_1, W_2, ...$,

$$
\begin{aligned}
\mathbb{E}\|\boldsymbol{X}\|_2 - \mathbb{E}\|\boldsymbol{Y}\|_2 &= \mathbb{E}\sqrt{Z + W_{m+1}^2} - \mathbb{E}\sqrt{Z} \\
&= \mathbb{E}\left[\mathbb{E}\left[\sqrt{Z + W_{m+1}^2} \,\middle|\, Z\right]\right] - \mathbb{E}\sqrt{Z} \\
&\geq \mathbb{E}\left[\sqrt{Z + \mathbb{E}[W_{m+1}^2 | Z]}\right] - \mathbb{E}\sqrt{Z} \\
&= \mathbb{E}\left[\sqrt{Z+1}\right] - \mathbb{E}\sqrt{Z} \\
&= \mathbb{E}\left[\sqrt{Z+1} - \sqrt{Z}\right] \\
&\geq \sqrt{\mathbb{E}Z + 1} - \sqrt{\mathbb{E}Z} \qquad = \sqrt{m+1} - \sqrt{m}.
\end{aligned}
$$

This ends the proof. □

[4] Take $\alpha_i = 1/p$ for exactly p indexes and $\alpha_i = 0$ for the rest.

[5] The second derivatives are $(c + x^2)^{-1/2}\left(\frac{-x^2}{x^2+c} + 1\right) \geq 0$ and $-\frac{1}{4}(x+1)^{-3/2} + \frac{1}{4}x^{-3/2} \geq 0$.

C.7 Catoni's influence function

Proposition C.7.1. *The functions $x \mapsto \log(1+x+x^2/2)$ and $x \mapsto -\log(1-x+x^2/2)$ are well defined and*

$$-\log(1-x+x^2/2) \leq \log(1+x+x^2/2), \quad \forall x \in \mathbb{R}.$$

Proof. Note that the function $x \mapsto 1+x+x^2/2$ is positive for all $x \in \mathbb{R}$ because it is positive at least for $x=0$ and it has no real roots. The same reasoning works for $x \mapsto 1-x+x^2/2$. In consequence the functions $x \mapsto \log(1+x+x^2/2)$ and $x \mapsto -\log(1-x+x^2/2)$ are well defined for all $x \in \mathbb{R}$. Also $x^4/4 \geq 0$ for all $x \in \mathbb{R}$ and

$$\begin{aligned} 1 &\leq 1+(x-x)+(x^2/2+x^2/2-x^2)+(x^3/2-x^3/2)+x^4/4 \\ &= (1+x+x^2/2)(1-x+x^2/2), \end{aligned}$$

so $1+x+x^2/2 \geq (1-x+x^2/2)^{-1}$ and $\log(1+x+x^2/2) \geq -\log(1-x+x^2/2)$ for all $x \in \mathbb{R}$. □

Proposition C.7.2. *The truncation operator $\psi_1(x) = (|x| \wedge 1) sign(x)$ satisfies the inequality*

$$-\log\left(1-x+x^2\right) \leq \psi_1(x) \leq \log\left(1+x+x^2\right).$$

Proof. For $|x| \leq 1$ we want to prove that

$$\begin{aligned} e^x &\leq 1+x+x^2, \quad x \geq 0 \\ e^{-x} &\geq \frac{1}{1-x+x^2}, \quad x < 0. \end{aligned}$$

The second inequality can be stated as $e^x \leq 1-x+x^2$. Since $e^x \leq e^{-x}$ for $x<0$ and e^{-x} is an alternating sequence that converges we have that

$$e^x \leq e^{-x} \leq 1-x+\frac{x^2}{2} \leq 1-x+x^2.$$

The first inequality is proved in the following way: for $x \geq 0$,

$$e^x = 1+x+x^2\sum_{k=2}^{\infty}\frac{x^{k-2}}{k!}$$

$$
\begin{aligned}
&\leq 1 + x + x^2 \sum_{k=2}^{\infty} \frac{1}{k!} \\
&= 1 + x + x^2(e-2) \\
&\leq 1 + x + x^2.
\end{aligned}
$$

Whenever $|x| > 1$ we have that $1 + |x| + x^2 \geq e$ which proves that $-\log(1 - x + x^2) \leq -1$ for $x < -1$ and $\log(1 + x + x^2) \geq 1$ for $x > 1$. This ends the proof. □

C.8 Uniqueness of solution for equation (5.17)

Following the procedure of [49], we develop a sufficient condition to ensure that there exist a unique $\tau > 0$ such that equality (5.17) is true. This is a generalization of the procedure in [49], but we make some extra assumptions to get an straightforward result. Let $\mathbf{X} \in \mathcal{S}_p$ be a rank one random matrix with spectral decomposition

$$\mathbf{X} = \lambda \boldsymbol{v}\boldsymbol{v}^\intercal = \lambda \mathbf{V},$$

Where $\mathbf{V} = \boldsymbol{v}\boldsymbol{v}^\intercal$ and $\|\boldsymbol{v}\|_2 = 1$. We make the following two assumptions.

1. λ is a continuous random variable and $\boldsymbol{v}$ is a random vector with continuous coordinates.

2. For any $\tau \in (0, \infty)$, $\mathbb{P}(|\lambda| > \tau) < 1$.

The first assumption imply that $\mathbb{P}(|\lambda| > 0) = 1$, and for any non-zero vector $\boldsymbol{x} \in \mathbb{R}^p$ we have $\mathbb{P}(\boldsymbol{x}^\intercal \mathbf{V} \boldsymbol{x} = 0) = \mathbb{P}((\boldsymbol{x}^\intercal \boldsymbol{v})^2 = 0) = 0$.

For ease of development, lets define the next matrices:

$$
\begin{aligned}
\mathbf{G}(\tau) &= \mathbb{E}\left[\mathbf{1}(|\lambda| > \tau)\mathbf{V}\right] \\
\mathbf{P}(\tau) &= \mathbb{E}\left[\lambda^2 \mathbf{1}(|\lambda| \leq \tau)\mathbf{V}\right] \\
\mathbf{Q}(\tau) &= \mathbb{E}\left[\psi_\tau^2(\mathbf{X})\right] = \mathbb{E}\left[(|\lambda| \wedge \tau)^2 \mathbf{V}\right].
\end{aligned}
$$

One property of $\mathbf{P}(\tau)$ that will be useful is the following.

Lemma C.8.1. *The matrix $\mathbf{P}(\tau)$ is positive definite for any $\tau \in (0, \infty)$.*

Proof. Suppose that there exist a $\tau_0 \in (0, \infty)$ such that $\boldsymbol{x}^\intercal \mathbf{P}(\tau_0)\boldsymbol{x} = 0$ for some non-zero vector $\boldsymbol{x} \in \mathbb{R}^p$. Then,

$$\mathbb{E}\left[\boldsymbol{x}^\intercal \lambda^2 \mathbf{1}(|\lambda| \leq \tau_0)\mathbf{V}\boldsymbol{x}\right] = 0.$$

Since $\lambda^2 \mathbf{1}(|\lambda| \leq \tau_0)\mathbf{V} \succeq \mathbf{0}$, we conclude that $\lambda^2 \mathbf{1}(|\lambda| \leq \tau_0)\boldsymbol{x}^\intercal \mathbf{V}\boldsymbol{x} = \mathbf{0}$ almost surely. Since $\mathbb{P}(|\lambda| = 0) = \mathbb{P}(\boldsymbol{x}^\intercal \mathbf{V}\boldsymbol{x} = 0) = 0$,

$$1 = \mathbb{P}\left(\lambda^2 \mathbf{1}(|\lambda| \leq \tau_0)\boldsymbol{x}^\intercal \mathbf{V}\boldsymbol{x} = \mathbf{0}\right) = \mathbb{P}(|\lambda| > \tau_0)$$

This contradicts assumption 2, so we conclude that such τ_0 can not exist. □

Additionally, we define the matrices

$$\mathbf{p}(\tau) = \tau^{-2}\mathbf{P}(\tau) \quad \text{and} \quad \mathbf{q}(\tau) = \tau^{-2}\mathbf{Q}(\tau),$$

By the previous lemma we get that $\mathbf{p}(\tau) \succ \mathbf{0}$ for every $\tau \in (0, \infty)$.

For the moment, we just point out that we are interest in finding τ such that

$$|||\mathbf{q}(\tau)||| = z, \quad \text{for some } z > 0.$$

The next lemma gives conditions to ensure that the previous problem has a unique solution.

Lemma C.8.2.

(a) For any $\tau > 0$,

$$\mathbf{Q}(\tau) = 2\int_0^\tau y\mathbf{G}(y)\, dy, \quad \frac{d}{d\tau}\mathbf{q}(\tau) = -2\tau^{-1}\mathbf{p}(\tau)$$

where the integral and derivative is taken entry-wise, and

$$\mathbf{q}(\tau) = \mathbb{E}\mathbf{V} - 2\int_0^\tau y^{-1}\mathbf{p}(y)\, dy.$$

(b) For any $0 < r < s$, $\mathbf{q}(s) \prec \mathbf{q}(r)$ and $|||\mathbf{q}(s)||| < |||\mathbf{q}(r)|||$, and in consequence $|||\mathbf{q}(\tau)||| = z$, $z > 0$, has a unique solution whenever $|||z^{-1}\mathbb{E}\mathbf{V}||| > 1$.

Proof. We can re express $(|\lambda| \wedge \tau)^2$ in the following way:

$$\begin{aligned}(|\lambda| \wedge \tau)^2 &= 2\int_0^\tau \mathbf{1}(|\lambda| > \tau) y \,\mathrm{d}y + 2\int_0^{|\lambda|} \mathbf{1}(|\lambda| \leq \tau) y \,\mathrm{d}y \\ &= 2\int_0^\tau \mathbf{1}(|\lambda| > \tau) y \,\mathrm{d}y + 2\int_0^\tau \mathbf{1}(|\lambda| > y)\mathbf{1}(|\lambda| \leq \tau) y \,\mathrm{d}y \\ &= 2\int_0^\tau \mathbf{1}(|\lambda| > y) y \,\mathrm{d}y.\end{aligned}$$

Therefore,

$$(|\lambda| \wedge \tau)^2 \mathbf{V} = 2\int_0^\tau \mathbf{1}(|\lambda| > y) y \mathbf{V} \,\mathrm{d}y,$$

and by Fubini's theorem, taking expectation on both sides, $\mathbf{Q}(\tau) = 2\int_0^\tau y\mathbf{G}(y)\,\mathrm{d}y$. From this relation it is clear that $\frac{d}{d\tau}\mathbf{Q}(\tau) = 2\tau\mathbf{G}(\tau)$. Now, note that

$$\mathbf{Q}(\tau) = \mathbb{E}\left[(|\lambda| \wedge \tau)^2 \mathbf{V}\left(\mathbf{1}(|\lambda| > \tau) + \mathbf{1}(|\lambda| \leq \tau)\right)\right] = \tau^2\mathbf{G}(\tau) + \mathbf{P}(\tau),$$

which also implies that

$$\mathbf{q}(\tau) = \tau^{-2}\mathbf{Q}(\tau) = \tau^{-2}\mathbf{P}(\tau) + \mathbf{G}(\tau) = \mathbf{p}(\tau) + \mathbf{G}(\tau).$$

And by definition of $\mathbf{q}(\tau)$ we have that $\frac{d}{d\tau}\mathbf{q}(\tau) = -2\tau^{-3}\mathbf{Q}(\tau) + \tau^{-2}\frac{d}{d\tau}\mathbf{Q}(\tau)$. Substituting $\mathbf{Q}(\tau)$ of the previous equality we get that

$$2\tau\mathbf{q}(\tau) + \tau^2\frac{d}{d\tau}\mathbf{q}(\tau) = 2\tau\mathbf{G}(\tau) = 2\tau(\mathbf{q}(\tau) - \mathbf{p}(\tau)),$$

so $\frac{d}{d\tau}\mathbf{q}(\tau) = -2\tau^{-1}\mathbf{p}(\tau)$.

To end part (a), observe that $\mathbf{q}(s) = \mathbf{q}(r) - 2\int_r^s \mathbf{p}(y)y^{-1}\,\mathrm{d}y$. Since

$$0 < \frac{(|\lambda| \wedge r)^2}{r^2} \leq 1, \quad \frac{(|\lambda| \wedge r)^2}{r^2} \longrightarrow 1,\ r \to 0,$$

by dominated convergence theorem[6],

$$\mathbf{q}(r) = \mathbb{E}\left[\frac{(|\lambda| \wedge r)^2}{r^2}\mathbf{V}\right] \longrightarrow \mathbb{E}\mathbf{V},\ r \to 0.$$

[6] Here the limit is taken entry-wise.

Then, by taking $r \to 0$ we get $\mathbf{q}(s) = \mathbb{E}\mathbf{V} - 2\int_0^s \mathbf{p}(y)y^{-1}\,\mathrm{d}y$.

For part (b) observe that for every $\boldsymbol{x} \in \mathbb{R}^p$ and $s > r > 0$,

$$\boldsymbol{x}^\intercal \int_0^r y^{-1}\mathbf{p}(y)\,\mathrm{d}y\boldsymbol{x} = \int_0^r \boldsymbol{x}^\intercal\mathbf{p}(y)y^{-1}\boldsymbol{x}\,\mathrm{d}y < \int_0^s \boldsymbol{x}^\intercal\mathbf{p}(y)y^{-1}\boldsymbol{x}\,\mathrm{d}y = \boldsymbol{x}^\intercal \int_0^s y^{-1}\mathbf{p}(y)\,\mathrm{d}y\boldsymbol{x},$$

so $\int_0^r y^{-1}\mathbf{p}(y)\,\mathrm{d}y \prec \int_0^s y^{-1}\mathbf{p}(y)\,\mathrm{d}y$. Since $\mathbf{p}(\tau) \succ \mathbf{0}$, we get that $\mathbf{q}(s) \prec \mathbf{q}(r)$ and that $|||\mathbf{q}(s)||| < |||\mathbf{q}(r)|||$ for $s > r > 0$. By the continuity of the norm and the integral representation of $\mathbf{q}(\tau)$, the function $|||\mathbf{q}(\tau)|||$ is continuous. Also,

$$0 < \frac{(|\lambda| \wedge \tau)^2}{\tau^2} \leq 1, \quad \frac{(|\lambda| \wedge \tau)^2}{\tau^2} \longrightarrow 0,\ \tau \to \infty,$$

and applying again dominated convergence theorem an the continuity of the norm, we get that $|||\mathbf{q}(\tau)||| \to 0$ as $\tau \to \infty$. Therefore, the function $|||\mathbf{q}(\tau)|||$ is continuous, monotone decreasing and $|||\mathbf{q}(0)||| = |||\mathbb{E}\mathbf{V}|||$. Then, the equation $|||\mathbf{q}(\tau)||| = z$, $z > 0$, has a unique solution whenever $z < |||\mathbb{E}\mathbf{V}|||$. □

Proof of Theorem 5.3.2. Under the assumptions of the theorem, the matrices $\boldsymbol{Y}_i\boldsymbol{Y}_i^\intercal/2$ satisfies the hypothesis of Lemma C.8.2. Define the random measure[7] $P_n = \frac{1}{N}\sum_{i=1}^N \delta(\boldsymbol{Y}_i\boldsymbol{Y}_i^\intercal/2)$. From Lemma C.8.2, taking expectation with respect to P_n, we can deduce that the equality[8]

$$\left|\left|\left|\frac{\mathbb{E}\left[\psi_\tau^2(\boldsymbol{Y}\boldsymbol{Y}^\intercal/2)\right]}{\tau^2}\right|\right|\right| = z, \quad z > 0,$$

is satisfied for a unique $\tau > 0$ as long as

$$\left|\left|\left|\mathbb{E}\left[\frac{\boldsymbol{Y}\boldsymbol{Y}^\intercal}{\|\boldsymbol{Y}\|_2^2}\right]\right|\right|\right| > z.$$

Taking $z = (\log(2p) + t)/m$ and observing that

$$\frac{\mathbb{E}\left[\psi_\tau^2(\boldsymbol{Y}\boldsymbol{Y}^\intercal/2)\right]}{\tau^2} = \frac{1}{\tau^2 N}\sum_{i=1}^N \psi_\tau^2(\boldsymbol{Y_i}\boldsymbol{Y}_i^\intercal/2),$$

and

$$\mathbb{E}\left[\frac{\boldsymbol{Y}\boldsymbol{Y}^\intercal}{\|\boldsymbol{Y}\|_2^2}\right] = \frac{1}{N}\sum_{i=1}^N \frac{\boldsymbol{Y}_i\boldsymbol{Y}_i^\intercal}{\|\boldsymbol{Y}_i\|_2^2}$$

ends the proof. □

[7] Here $\delta(\mathbf{A})$ assigns the value 1 to the matrix $\mathbf{A}$, i.e., is the Dirac measure in $\mathbf{A}$.

[8] $\boldsymbol{Y}\boldsymbol{Y}^\intercal/2$ is distributed according to P_n.

C.9 Covariance as minimization problem

Proposition C.9.1. *Let* $\mathbf{C} \in \mathcal{M}_p$ *be a fixed matrix. The function* $\mathbf{S} \mapsto |\!|\!|\mathbf{S} - \mathbf{H}|\!|\!|_2^2$ *is strictly convex in* $\mathcal{S}_p$*.*

Proof. Take $\mathbf{S}, \mathbf{R} \in \mathcal{S}_p$ with $\mathbf{S} \neq \mathbf{R}$ and $t \in (0,1)$, and define the inner product $\langle \mathbf{A}, \mathbf{B} \rangle = \operatorname{tr}(\mathbf{A}^\intercal \mathbf{B})$. Then,

$$\begin{aligned}
&t|\!|\!|\mathbf{C} - \mathbf{S}|\!|\!|_2^2 + (1-t)|\!|\!|\mathbf{C} - \mathbf{R}|\!|\!|_2^2 - |\!|\!|\mathbf{C} - t\mathbf{S} - (1-t)\mathbf{R}|\!|\!|_2^2 \\
&= t\left(|\!|\!|\mathbf{C}|\!|\!|_2^2 + |\!|\!|\mathbf{S}|\!|\!|_2^2 - 2\langle \mathbf{C}, \mathbf{S} \rangle\right) + (1-t)\left(|\!|\!|\mathbf{C}|\!|\!|_2^2 + |\!|\!|\mathbf{R}|\!|\!|_2^2 - 2\langle \mathbf{C}, \mathbf{R} \rangle\right) \\
&\quad - \left(|\!|\!|\mathbf{C}|\!|\!|_2^2 + t^2|\!|\!|\mathbf{S}|\!|\!|_2^2 + (1-t)^2|\!|\!|R|\!|\!|_2^2 - 2t\langle \mathbf{C}, \mathbf{S} \rangle - 2(1-t)\langle \mathbf{C}, \mathbf{R} \rangle + 2t(1-t)\langle \mathbf{R}, \mathbf{S} \rangle\right) \\
&= t(1-t)\left(|\!|\!|\mathbf{S}|\!|\!|_2^2 + |\!|\!|\mathbf{R}|\!|\!|_2^2 - 2\langle \mathbf{S}, \mathbf{R} \rangle\right) \\
&= t(1-t)|\!|\!|\mathbf{S} - \mathbf{R}|\!|\!|_2^2 \\
&> 0.
\end{aligned}$$

This ends the proof. □

Proof of Theorem 5.3.3. By the previous proposition, the function $\mathbf{A} \mapsto \mathbb{E}\|\boldsymbol{Z}\boldsymbol{Z}^\intercal - \mathbf{S}\|_2^2$ is strictly convex, so it has a unique minimum. Now, observe that

$$|\!|\!|\boldsymbol{Z}\boldsymbol{Z}^\intercal - \mathbf{S}|\!|\!|_2^2 = \operatorname{tr}(\mathbf{S}^2) - 2\operatorname{tr}(\mathbf{S}\boldsymbol{Z}\boldsymbol{Z}^\intercal) + \boldsymbol{Z}^\intercal\boldsymbol{Z}\operatorname{tr}(\boldsymbol{Z}\boldsymbol{Z}^\intercal).$$

Therefore,

$$\mathbb{E}|\!|\!|\boldsymbol{Z}\boldsymbol{Z}^\intercal - \mathbf{S}|\!|\!|_2^2 = \operatorname{tr}(\mathbf{S}^2) - 2\operatorname{tr}(\mathbf{S}\boldsymbol{\Sigma}).$$

From matrix derivation techniques presented in [36] one can prove that

$$\frac{\partial}{\partial \mathbf{S}}\mathbb{E}|\!|\!|\boldsymbol{Z}\boldsymbol{Z}^\intercal - \mathbf{S}|\!|\!|_2^2 = 4\mathbf{S} - 2\operatorname{diag}(\mathbf{S}) - 4\boldsymbol{\Sigma} + 2\operatorname{diag}(\boldsymbol{\Sigma}),$$

so $\mathbb{E}|\!|\!|\boldsymbol{Z}\boldsymbol{Z}^\intercal - \mathbf{S}|\!|\!|_2^2$ is minimized when $\mathbf{S} = \boldsymbol{\Sigma}$. □

C.10 Subdifferential calculus

Throughout this section $\mathcal{X}$ is a vector space with inner product $\langle \cdot, \cdot \rangle$. A function $F : \mathcal{X} \to \mathbb{R}$ is called *convex* if $F(t\boldsymbol{x} + (1-t)) \leq tF(\boldsymbol{x}) + (1+t)F(\boldsymbol{y})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ and $t \in [0,1]$. We call it *strictly convex* if $F(t\boldsymbol{x} + (1-t)) < tF(\boldsymbol{x}) + (1+t)F(\boldsymbol{y})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ and $t \in (0,1)$. An example of such function is given in the following theorem.

Theorem C.10.1. *The function $F_\tau : \mathcal{M}_{n,p} \to \mathbb{R}$ defined as*

$$F_\tau(\mathbf{A}) = |||\mathbf{A} - \mathbf{X}|||_2^2 + \tau|||\mathbf{A}|||_1,$$

where $\mathbf{X} \in \mathcal{M}_{n,p}$ and $\tau > 0$, is strictly convex.

The proof is similar to the one from Promosition C.9.1 with the additional observation that the sum of a convex and strictly convex function is strictly convex.

We define the subdifferential of a convex function in the following way.

Definition C.10.2 (Subdifferential)**.** *Let $F : \mathcal{X} \to \mathbb{R}$ be a convex function. The subdifferential of F at $\boldsymbol{x} \in \mathcal{X}$ is the set*

$$\partial F(\boldsymbol{x}) = \{\boldsymbol{w} \in \mathcal{X} : F(\boldsymbol{y}) \geq F(\boldsymbol{x}) + \langle \boldsymbol{w}, \boldsymbol{y} - \boldsymbol{x} \rangle \quad \textit{for all } \boldsymbol{y} \in \mathcal{X}\}.$$

In [15, Appendix D] it is proved that a function $F : \mathcal{X} \to \mathbb{R}$ is convex if and only if the set $\partial F(\boldsymbol{x})$ is non-empty for all $\boldsymbol{x} \in \mathcal{X}$. One property that we are interested in is the monotonicity of the subdifferential.

Proposition C.10.3 (Monotonicity)**.** *If $F : \mathcal{X} \to \mathbb{R}$ is convex then*

$$\langle \boldsymbol{u} - \boldsymbol{v}, \boldsymbol{x} - \boldsymbol{y} \rangle \geq 0$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ and $\boldsymbol{u} \in \partial F(\boldsymbol{x})$, $\boldsymbol{v} \in \partial F(\boldsymbol{y})$.

Proof. By definition of subdifferential,

$$F(\boldsymbol{y}) \geq F(\boldsymbol{x}) + \langle \boldsymbol{u}, \boldsymbol{y} - \boldsymbol{x} \rangle \quad \text{and} \quad F(\boldsymbol{x}) \geq F(\boldsymbol{y}) + \langle \boldsymbol{v}, \boldsymbol{x} - \boldsymbol{y} \rangle.$$

Adding the two previous inequalities gives the desired result. □

An intuitive fact about subdifferentials is that for $F_1, F_2 : \mathcal{X} \to \mathbb{R}$ convex functions,

$$\partial(F_1 + F_2)(\boldsymbol{x}) = \partial F_1(\boldsymbol{x}) + \partial F_2(\boldsymbol{x}), \quad \forall \boldsymbol{x}. \tag{C.1}$$

The inclusion $\supset$ is immediate, but the other is more involved. One can see a proof of this fact in [46, Chapter 2].

Let $\mathcal{C} \subset \mathcal{X}$ be a convex set and define the indicator function $I_{\mathcal{C}} : \mathcal{X} \to \mathbb{R}$ as

$$I_{\mathcal{C}}(\boldsymbol{x}) = \begin{cases} 0, & \boldsymbol{x} \in \mathcal{C}, \\ \infty, & \boldsymbol{x} \notin \mathcal{C}. \end{cases}$$

Restricted to the set $\mathcal{C}$[9] we have that $\partial I_{\mathcal{C}}(\boldsymbol{x}) = N_{\mathcal{C}}(\boldsymbol{x})$, where $N_{\mathcal{C}}(\boldsymbol{x})$ is the *normal cone* of $\mathcal{C}$ at $\boldsymbol{x}$, i.e.,

$$N_{\mathcal{C}}(\boldsymbol{x}) = \{\boldsymbol{w} \in \mathcal{X} \,:\, \langle \boldsymbol{z}, \boldsymbol{x} - \boldsymbol{y} \rangle \geq \boldsymbol{0} \text{ for any } \boldsymbol{y} \in \mathcal{C}\}\,.$$

This follows immediately since $I_{\mathcal{C}}(\boldsymbol{x}) = I_{\mathcal{C}}(\boldsymbol{y}) = 0$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$.

The reason why we want this fact is that we can rewrite the restricted minimization problem

$$\min_{\boldsymbol{x} \in \mathcal{C}} F(\boldsymbol{x})$$

as the unrestricted problem

$$\min_{\boldsymbol{x} \in \mathcal{X}} \{F(\boldsymbol{x}) + I_{\mathcal{C}}(\boldsymbol{x})\}\,.$$

The usefulness of this representation is shown in the following proposition.

Proposition C.10.4. *Let $F : \mathcal{X} \to \mathbb{R}$ be a convex function and $\mathcal{C} \subset \mathcal{X}$ as convex set. Then,*

(a)

$$\boldsymbol{x}^* \in \min_{\boldsymbol{x} \in \mathcal{C}} F(\boldsymbol{x}) \Longleftrightarrow \boldsymbol{0} \in \partial F(\boldsymbol{x}^*);$$

(b)

$$\boldsymbol{0} \in \partial\left(F(\boldsymbol{x}) + I_{\mathcal{C}}(\boldsymbol{x})\right) \Longrightarrow \textit{there exists } \boldsymbol{w} \in \partial F(\boldsymbol{x}) \textit{ such that } \langle \boldsymbol{w}, \boldsymbol{y} - \boldsymbol{x} \rangle \geq 0 \,\forall \boldsymbol{y} \in \mathcal{C}.$$

Proof. The proof of (a) is immediate since $F(\boldsymbol{y}) \geq F(\boldsymbol{x}_*)$ for all $\boldsymbol{y} \in \mathcal{C}$ if and only if $F(\boldsymbol{y}) \geq F(\boldsymbol{x}_*) + \langle \boldsymbol{0}, \boldsymbol{y} - \boldsymbol{x}_* \rangle$.

To prove (b) note that from (C.1), $\boldsymbol{0} \in \partial\left(F(\boldsymbol{x}) + I_{\mathcal{C}}(\boldsymbol{x})\right)$ imply that there exists a $\boldsymbol{w} \in \mathcal{X}$ such that $\boldsymbol{w} \in \partial F(\boldsymbol{x})$ and $-\boldsymbol{w} \in \partial I_{\mathcal{C}}(\boldsymbol{x}) = N_{\mathcal{C}}(\boldsymbol{x})$. Then, $\langle -\boldsymbol{w}, \boldsymbol{x} - \boldsymbol{y} \rangle \geq 0$ for all $\boldsymbol{y} \in \mathcal{C}$, i.e., $\langle \boldsymbol{w}, \boldsymbol{y} - \boldsymbol{x} \rangle \geq 0$. □

In Chapter 6 we work with functions defined on a set of matrices and we want to optimize the Frobenius norm penalized with the nuclear norm. The next Theorem proved, for example, in [27] gives the form of the subdiferential of the function F_τ defined in Theorem C.10.1.

[9] i.e., $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$ in the subdifferential.

Theorem C.10.5. *For $\mathbf{A} \in \mathcal{M}_{n,p}$ of rank r, the subdifferential of $|||\mathbf{A}|||_1$ is the set*

$$\partial|||\mathbf{A}|||_1 = \left\{ \sum_{j=1}^{r} \boldsymbol{u}_j(\mathbf{A})\boldsymbol{v}_j(\mathbf{A})^\intercal + \mathbf{P}_{S_1^\perp}\mathbf{W}\mathbf{P}_{S_2^\perp} \,:\, |||\mathbf{W}||| \leq 1 \right\},$$

where $\mathbf{P}_{S_1^\perp} = \mathbf{I} - \mathbf{P}_{S_1}$ and $\mathbf{P}_{S_2^\perp} = \mathbf{I} - \mathbf{P}_{S_2}$ with $\mathbf{P}_{S_1}$ y $\mathbf{P}_{S_2}$ the orthogonal projectors in $S_1 = span\{\boldsymbol{u}_1(\mathbf{A}), ..., \boldsymbol{u}_r(\mathbf{A})\}$ and $S_2 = span\{\boldsymbol{v}_1(\mathbf{A}), ..., \boldsymbol{v}_r(\mathbf{A})\}$, respectively. The pair (S_1, S_2) is called the support of $\mathbf{A}$. Even more, the subdifferential of $F_\tau(\mathbf{A})$ defined in Theorem C.10.1 is the set

$$\partial F_\tau(\mathbf{A}) = \left\{ 2(\mathbf{X} - \mathbf{A}) + \tau \left(\sum_{j=1}^{r} \boldsymbol{u}_j(\mathbf{A})\boldsymbol{v}_j(\mathbf{A})^\intercal + \mathbf{P}_u^\perp\mathbf{W}\mathbf{P}_v^\perp \right) \,:\, |||\mathbf{W}||| \leq 1 \right\}.$$

C.11 Minimization of Frobenius norm

For two matrices $\mathbf{A}, \mathbf{B} \in \mathcal{M}_{n,m}$ define the Hadmard or entry-wise product $\mathbf{A} \circ \mathbf{B}$ as the matrix with entries $(\mathbf{A} \circ \mathbf{B})_{ij} = a_{ij}b_{ij}$.

Proposition C.11.1. *For fixed $\mathbf{B} \in \mathcal{M}_n$, the function $\mathbf{A} \mapsto |||\mathbf{A} - \mathbf{B}|||_2^2$ with $\mathbf{A} \in \mathcal{S}_n$ has a unique minimum in $\mathcal{S}_n$ given by*

$$\mathbf{A} = \left(\frac{1}{2}\mathbf{J} + \frac{1}{2}\mathbf{I} \right) \circ \left(\mathbf{B} + \mathbf{B}^\intercal + \mathbf{B} \circ \mathbf{I} \right),$$

where $\mathbf{J}$ is the $n \times n$ matrix of ones, i.e., $(J)_{ij} = 1$ for all i, j.

Proof. Up to a constant factor we have that

$$|||\mathbf{A} - \mathbf{B}|||_2^2 = \mathrm{tr}\,(\mathbf{A}^2) - 2\mathrm{tr}\,(\mathbf{A}^\intercal\mathbf{B}),$$

where

$$\mathrm{tr}\,(\mathbf{A}^2) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2 \quad \text{and} \quad \mathrm{tr}\,(\mathbf{A}^\intercal\mathbf{B}) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ji}b_{ji}.$$

Therefore, recalling that $\mathbf{A} \in \mathcal{S}_n$ is straightforward to get that

$$\frac{\partial}{\partial \mathbf{A}}\mathrm{tr}\,(\mathbf{A}^2) = 4\mathbf{A} - 2\mathbf{A} \circ \mathbf{I},$$

and

$$\frac{\partial}{\partial \mathbf{A}} \operatorname{tr}\left(\mathbf{A}^2\right) = \mathbf{B} + \mathbf{B}^\intercal - \mathbf{B} \circ \mathbf{I}.$$

With the same procedure of Proposition C.9.1 one can see that $\mathbf{A} \mapsto |||\mathbf{A} - \mathbf{B}|||_2^2$ is strictly convex, so the minimum is obtained by the equality

$$\mathbf{0} = \frac{\partial}{\partial \mathbf{A}} |||\mathbf{A} - \mathbf{B}|||_2^2 = 4\mathbf{A} - 2\mathbf{A} \circ \mathbf{I} - 2\left(\mathbf{B} + \mathbf{B}^\intercal - \mathbf{B} \circ \mathbf{I}\right).$$

Finally, by definition of Hadamard product we have that $\mathbf{X} \circ \mathbf{C} = \mathbf{D}$ is solved by $\mathbf{X} = (\mathbf{C})^- \circ \mathbf{D}$ where $(\mathbf{C})^- = (1/c_{ij})_{ij}$. This ends the proof. □