Project report

Antieigenvalues and sample coviarance matrices

Fabio Ferrari

PoliMi - LiTH - - 2016 / 07 - SE / IT

Antieigenvalues and sample coviarance matrices

Mathematical Engineering, Politecnico di Milano, Linköpings Universitet

Fabio Ferrari

PoliMi - LiTH - - 2016 / 07 - SE / IT

Project report: **12 credits**

Level: **Master thesis**

Supervisor: **Martin Singull, Piercesare Secchi**,
Mathematical Engineering, Politecnico di Milano, Linköpings Universitet

Examiner: **Martin Singull, Piercesare Secchi**,
Mathematical Engineering, Politecnico di Milano, Linköpings Universitet

Linköping: **July 2016**

Abstract

In the last decades the necessity of finding statistical tools to analyze larger and larger data, has been an important topic for researchers. A huge theory about the limit distribution of the *eigenvalues* of sample covariance matrices, whose dimensions diverge under certain conditions, has been developed. Almost 50 years ago, Karl Gustafson introduced new quantities, called *antieigenvalues*, whose theory in a context of fixed matrices has been developing, with important applications in numerical analysis, wavelets, statistics, quantum mechanics, finance and optimization. In this report it is presented the basic theory concerning the limit spectral distribution of sample covariance matrices, by reporting the Wigner's law and the Marčenko-Pastur's law, and it is introduced to the so far known theory on the antieigenvalues. The real aim is to try to find a law that describes the limit distribution of the first antieigenvalue of a sample covariance matrix.

Keywords: Antieigenvalues, eigenvalues, sample covariance matrices, random matrices, Beta distribution, limit spectral distribution.

Contents

Chapter 1

Introduction

In the last decades, data sets have become large in both size and dimension. Hence, statisticians have been facing large dimensional data analysis in both theoretical investigation and real applications. The aim of this project is to investigate the distribution of eigenvalues and antieigenvalues of random matrices when their dimensions go to infinity.

In this chapter it will be provided a statistical background to introduce to the main character of this project: the estimator sample covariance matrix $\mathcal{S}$. The geometrical meaning of $\mathcal{S}$, of its eigenvalues and of its eigenvectors, will be presented.

In the second chapter, two important results about the limit distribution of eigenvalues of $\mathcal{S}$ will be reported: the Wigner semicircular law, where one dimension of the data is thought to be fixed, and the Marčenko-Pastur law, where both of the dimensions tend to infinity. Some simulations in R software will be shown to visualize some theoretical results.

In the third chapter, it will be given a brief introduction to the already known theory about antieigenvalues of operators and some simulations in R software will be shown to investigate their limit distribution, when the considered operator is $\mathcal{S}$, in a context of random matrices, which has not already been investigated.

1.1 Statistical background

1.1.1 Observations of statistical units

Let $\mathbf{x} = (x_1, \dots, x_p)^T$ be a *sample*, i.e. an observation of p variables. When we collect n observations, we can store the values in a matrix X, with n rows and p columns. $[X]_{ij} \in \mathbb{R}$ is the result of a measurement: the *j-th* variable of the *i-th* observation. The $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ observations are n points in $\mathbb{R}^p$, that is a vector space as $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ and $\forall \alpha, \beta \in \mathbb{R}$, $\alpha\mathbf{x} + \beta\mathbf{y} \in \mathbb{R}^p$.

A possible inner product to be associated to $\mathbb{R}^p$ is

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{j=1}^{p} x_j y_j = \mathbf{x}^T \mathbf{y} \quad (\textit{scalar product}).$$

An other possibility is to define the inner product as

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T Q \mathbf{y}, \tag{1.1}$$

for some $Q \in \mathbb{R}^{p \times p}$ positive definite.

Each observation $\mathbf{x}_i$ can be viewed as a realization of a random vector $\mathbf{X} = (X_1, \ldots, X_p)^T$, a so called *population* vector, whose distribution is $F(\boldsymbol{\mu}, \Sigma)$. Here $\boldsymbol{\mu}$ is the mean of the F population and it is defined as $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_p])^T$, while Σ is the covariance matrix.

Definition 1 *The **covariance matrix** of a p-variate population vector $\boldsymbol{X}$, with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T$, is defined as*

$$\Sigma = Cov(\mathbf{X}) = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T].$$

Each element of Σ is

$$[\Sigma]_{ij} = \sigma_{ij} = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] = cov(X_i, X_j),$$

the covariance between the components X_i and X_j. In particular, the elements on the diagonal of Σ are

$$[\Sigma]_{ii} = \sigma_{ii} = \mathbb{E}[(X_i - \mu_i)(X_i - \mu_i)] = var(X_i),$$

the variances of the components of the population vector $\mathbf{X}$.

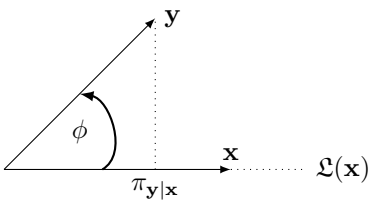
Let us now investigate the properties of Σ. It is easily notable that Σ is a squared matrix $\in \mathbb{R}^{p \times p}$ and that it is symmetric. This implies (Spectral theorem) that $\exists Q$ ortogonal matrix and $\exists D$ diagonal matrix, such that

$$\Sigma = Q^T D Q,$$

where Q is the matrix that stores the eigenvectors $(\mathbf{e}_1, \ldots, \mathbf{e}_p)$ of Σ, while the elements on the diagonal of D are its eigenvalues $(\lambda_1, \ldots, \lambda_p)$. In an other way, we can write

$$\Sigma = \sum_{j=1}^{p} \lambda_j \mathbf{e}_j \mathbf{e}_j^T.$$

Remark 1. $\frac{\boldsymbol{x}\boldsymbol{x}^T}{\boldsymbol{x}^T\boldsymbol{x}}$ *is the operator that projects orthogonally a vector in $\mathbb{R}^p$ on $\mathfrak{L}(\boldsymbol{x})$, the space generated by $\boldsymbol{x} \in \mathbb{R}^p$.*



Let $\pi_{\mathbf{y}|\mathbf{x}}$ be the orthogonal projection of a vector $\mathbf{y}$ on the space $\mathfrak{L}(\mathbf{x})$ generated by a vector $\mathbf{x}$. Let the cosine of the angle ϕ be defined as

$$cos(\phi) = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|\|\mathbf{y}\|},$$

then, it is easy to derive the expression for $\pi_{\mathbf{y}|\mathbf{x}}$ as

$$\pi_{\mathbf{y}|\mathbf{x}} = \|\mathbf{y}\| cos(\phi) \frac{\mathbf{x}}{\|\mathbf{x}\|} = \|\mathbf{y}\| \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|\|\mathbf{y}\|} \frac{\mathbf{x}}{\|\mathbf{x}\|} = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\|^2} \mathbf{x} = \frac{\mathbf{y}^T \mathbf{x}}{\mathbf{x}^T \mathbf{x}} \mathbf{x} = \frac{\mathbf{x}\mathbf{x}^T}{\mathbf{x}^T \mathbf{x}} \mathbf{y} = P\mathbf{y},$$

where $P = \frac{\mathbf{x}\mathbf{x}^T}{\mathbf{x}^T \mathbf{x}}$ is the symmetric and idempotent operator that projects a vector $\mathbf{y}$ orthogonally on $\mathfrak{L}(\mathbf{x})$.

Reminding that the norm of eigenvectors is equal to one, $\Sigma = \sum_{j=1}^{p} \lambda_j \mathbf{e}_j \mathbf{e}_j^T$ is the sum of orthogonal projections on the spaces generated by its eigenvectors.

Let $\mathbf{X} = (X_1, \ldots, X_p)^T$ be a vector of p random variables - with expected value $\boldsymbol{\mu}$ and covariance matrix Σ - and $\mathbf{c} = (c_1, \ldots, c_p)^T$ a vector of p real coefficients. Then, $\mathbf{c}^T \mathbf{X}$ is a linear combination of p random variables, with expected value $\mathbb{E}[\mathbf{c}^T \mathbf{X}] = \mathbf{c}^T \boldsymbol{\mu}$ and variance $Var[\mathbf{c}^T \boldsymbol{\mu}] = \mathbf{c}^T \Sigma \mathbf{c}$. As this holds $\forall \mathbf{c} \in \mathbb{R}^p$ and the variance is always non negative, we have that $\mathbf{c}^T \Sigma \mathbf{c} \geq 0 \ \forall \mathbf{c} \in \mathbb{R}^p$.

Defintion 2. *A $p \times p$ matrix A is called **positive semidefinite** if*

$$\mathbf{x}^T A \mathbf{x} \geq 0 \ \ \forall \mathbf{x} \in \mathbb{R}^p,$$

and we write $A \geq 0$.

This means that Σ is positive semidefinite. Hence, its eigenvalues $(\lambda_1, \ldots, \lambda_p)$ are all non negative. If all the eigenvalues are strictly positive, then Σ is said to be positive definite and we write $A > 0$.

1.1.2 Sample mean and sample covariance matrix

When we store n indipendent p-variate observations $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ in the data matrix

$$X = (\mathbf{x}_1 \ \ \mathbf{x}_2 \ \ \ldots \ \ \mathbf{x}_n)^T = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}$$

it can be viewed as a realization of the following random matrix

$$\mathcal{X} = (\boldsymbol{X_1} \ \ \boldsymbol{X_2} \ \ \ldots \ \ \boldsymbol{X_n})^T = \begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{np} \end{bmatrix}.$$

Definition 3. *Assuming that* $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ *are* n *independent realizations of* $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$*, random vectors with distribution* F*, mean* $\boldsymbol{\mu}$ *and covariance matrix* Σ*, then*

$$\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i = \frac{1}{n}\mathcal{X}^T \boldsymbol{1} = \begin{bmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \vdots \\ \overline{X}_p \end{bmatrix}, \quad \overline{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij} \;\; \forall j \in \{1, \ldots, p\},$$

is called the ***sample mean****. While, the matrix*

$$\mathcal{S} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^T, \quad \mathcal{S}_{ij} = \frac{1}{n}\sum_{k=1}^{n} (X_{ki} - \overline{X}_i)(X_{kj} - \overline{X}_j)$$

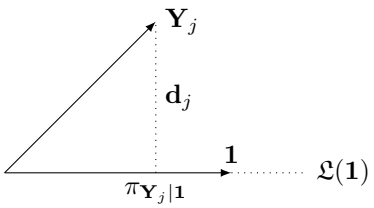
is called the ***sample covariance matrix****.*

The sample mean $\overline{\mathbf{X}}$ and the sample covariance matrix $\mathcal{S}$ are the *estimators* of $\boldsymbol{\mu}$ and Σ, while $\overline{\mathbf{x}}$ and S are the *estimates*, i.e., a realization of the estimators according to the observed data.

In order to highlight the marginal information about the variables that we are simultaneously observing, we can write $\mathcal{X} = [\boldsymbol{Y_1} \;\; \boldsymbol{Y_2} \;\; \ldots \;\; \boldsymbol{Y_p}]$. In a world without variability, all the observations are equal to each other, which means that $\mathbf{Y}_j = [k_j, k_j, \ldots, k_j]$, with $k_j \in \mathbb{R}$, $\forall j \in \{1, \ldots, p\}$, i.e. $\mathbf{Y}_j \in \mathfrak{L}(\mathbf{1}) \;\; \forall j \in \{1, \ldots, p\}$. If we project our data in this world without variability, we get, for each $j \in \{1, \ldots, p\}$:

$$\pi_{\mathbf{Y}_j|\mathbf{1}} = \frac{\mathbf{1}\mathbf{1}^T}{\mathbf{1}^T\mathbf{1}}\mathbf{Y}_j = \begin{bmatrix} 1/n & 1/n & \ldots & 1/n \\ 1/n & 1/n & \ldots & 1/n \\ \vdots & \vdots & \ddots & \vdots \\ 1/n & 1/n & \ldots & 1/n \end{bmatrix} \begin{bmatrix} X_{1j} \\ X_{2j} \\ \vdots \\ X_{nj} \end{bmatrix} = \begin{bmatrix} \frac{\sum_{i=1}^{n} X_{ij}}{n} \\ \frac{\sum_{i=1}^{n} X_{ij}}{n} \\ \vdots \\ \frac{\sum_{i=1}^{n} X_{ij}}{n} \end{bmatrix} = \begin{bmatrix} \overline{X}_j \\ \overline{X}_j \\ \vdots \\ \overline{X}_j \end{bmatrix}.$$

In this way we can notice that the *j-th* component of the sample mean $\overline{\mathbf{X}}$ is $\overline{X}_j$, the value of each component of the projection of the *j-th* variable $\mathbf{Y}_j$ on $\mathfrak{L}(\mathbf{1})$.



What is the error that we are committing by approximating the real $\mathbf{Y}_j$ with the projection $\pi_{\mathbf{Y}_j|\mathbf{1}}$?

$$\mathbf{d}_j = \mathbf{Y}_j - \pi_{\mathbf{Y}_j|\mathbf{1}} = \begin{bmatrix} X_{1j} - \overline{X}_j \\ X_{2j} - \overline{X}_j \\ \vdots \\ X_{nj} - \overline{X}_j \end{bmatrix}$$

In order to measure the error we can derive the expression of the squared euclidean norm of $\mathbf{d}_j$:

$$\|\mathbf{d}_j\|^2 = \langle \mathbf{Y}_j - \overline{X}_j\mathbf{1}, \mathbf{Y}_j - \overline{X}_j\mathbf{1}\rangle = \sum_{i=1}^{n}(X_{ij} - \overline{X}_j)^2 = nS_{jj}.$$

In this way we can see what is the relation between the sample variance $S_{jj} = var(\mathbf{Y}_j)$ and the euclidian distance between $\mathbf{Y}_j$ and $\mathfrak{L}(\mathbf{1})$.

Hence, each variable $\mathbf{Y}_j$ is associated with two values: $\overline{\mathbf{X}}_j$, the *j-th* component of the estimate of the sample mean, representing the orthogonal projection of $\mathbf{Y}_j$ on the space $\mathfrak{L}(\mathbf{1})$ generated by the vector $\mathbf{1}$, in a world without variability; and S_{jj}, the *j-th* element on the diagonal of the estimate of the sample covariance matrix, representing the distance between $\mathbf{Y}_j$ and $\mathfrak{L}(\mathbf{1})$, i.e., the error we commit when we approximate $\mathbf{Y}_j$ with its projection $\pi_{\mathbf{Y}_j|\mathbf{1}}$ on $\mathfrak{L}(\mathbf{1})$.

The correlation between a variable $\mathbf{Y}_j$ and $\mathbf{Y}_k$, in general for each $j, k \in \{1, \ldots, p\}$, is

$$r_{jk} = \frac{\sum_{i=1}^{n}(X_{ij} - \overline{\mathbf{X}}_j)(X_{ik} - \overline{\mathbf{X}}_k)}{\sqrt{\sum_{i=1}^{n}(X_{ij} - \overline{\mathbf{X}}_j)^2}\sqrt{\sum_{i=1}^{n}(X_{ik} - \overline{\mathbf{X}}_k)^2}} = \frac{S_{jk}}{\sqrt{S_{jj}}\sqrt{S_{kk}}}.$$

Hence, $S_{jk} = r_{jk}\sqrt{S_{jj}}\sqrt{S_{kk}}$ is a measure of the correlation between $\mathbf{Y}_j$ and $\mathbf{Y}_k$, also considering the distance between $\mathbf{Y}_j$ and $\mathcal{L}(\mathbf{1})$, and between $\mathbf{Y}_k$ and $\mathcal{L}(\mathbf{1})$.

Let us now investigate what is the relation between the estimators $\overline{\mathbf{X}}$, $\mathcal{S}$ and $\boldsymbol{\mu}$, Σ.

Proposition 1. Let $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ be i.i.d. $\sim F$, with mean $\boldsymbol{\mu}$ and covariance matrix Σ, then

1. $\mathbb{E}[\overline{\mathbf{X}}] = \boldsymbol{\mu}$ (unbiased estimator) and $cov(\overline{\mathbf{X}}) = \frac{1}{n}\Sigma$.
2. $\mathbb{E}[\mathcal{S}] = \frac{n-1}{n}\Sigma$ (biased estimator).

Proof:

1. $$\mathbb{E}[\overline{\mathbf{X}}] = \begin{bmatrix} \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_{i1}] \\ \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_{i2}] \\ \vdots \\ \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_{ip}] \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\mu}_1 \\ \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\mu}_2 \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\mu}_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \vdots \\ \boldsymbol{\mu}_p \end{bmatrix} = \boldsymbol{\mu}.$$

$$cov(\overline{\mathbf{X}}) = \mathbb{E}[(\overline{\mathbf{X}} - \boldsymbol{\mu})(\overline{\mathbf{X}} - \boldsymbol{\mu})^T] = \mathbb{E}[(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu}))(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu}))^T] =$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}[(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_j - \boldsymbol{\mu})^T].$$

Now, if $i \neq j$, then

$$\mathbb{E}[(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_j - \boldsymbol{\mu})^T] = \mathbb{E}[(\mathbf{X}_i - \boldsymbol{\mu})]\mathbb{E}[(\mathbf{X}_j - \boldsymbol{\mu})^T]$$

due to the independency between the random vectors. But $\mathbb{E}[(\mathbf{X}_i - \boldsymbol{\mu})] = 0$, for each $i = 1, \ldots, n$, so all the terms $i \neq j$ are equal to zero.
If $i = j$, then

$$\mathbb{E}[(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})^T] = \Sigma$$

Hence, $cov(\overline{\mathbf{X}}) = \frac{1}{n}\Sigma$.

2. $$\begin{aligned}
\mathbb{E}[\mathcal{S}] &= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^T] = \\
&= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}[(\mathbf{X}_i - \boldsymbol{\mu}) - (\overline{\mathbf{X}} - \boldsymbol{\mu})][(\mathbf{X}_i - \boldsymbol{\mu}) - (\overline{\mathbf{X}} - \boldsymbol{\mu})]^T] = \\
&= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})^T] - \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu})(\overline{\mathbf{X}} - \boldsymbol{\mu})^T] - \\
&- \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(\overline{\mathbf{X}} - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})^T] + \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(\overline{\mathbf{X}} - \boldsymbol{\mu})(\overline{\mathbf{X}} - \boldsymbol{\mu})^T] = \\
&= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})^T] - \mathbb{E}[(\overline{\mathbf{X}} - \boldsymbol{\mu})(\overline{\mathbf{X}} - \boldsymbol{\mu})^T] - \\
&- \mathbb{E}[(\overline{\mathbf{X}} - \boldsymbol{\mu})(\overline{\mathbf{X}} - \boldsymbol{\mu})^T] + \mathbb{E}[(\overline{\mathbf{X}} - \boldsymbol{\mu})(\overline{\mathbf{X}} - \boldsymbol{\mu})^T] = \\
&= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})^T] - \mathbb{E}[(\overline{\mathbf{X}} - \boldsymbol{\mu})(\overline{\mathbf{X}} - \boldsymbol{\mu})^T] = \\
&= \Sigma - \frac{\Sigma}{n} = \frac{n-1}{n}\Sigma.
\end{aligned}$$

■

$\overline{\mathbf{X}}$ is an unbiased estimator for $\boldsymbol{\mu}$ and its coviarance matrix converges to the null matrix when the number n of observations goes to infinity. $\mathcal{S}$ is a biased estimator for Σ, but when n goes to infinity, its expected value converges to Σ, which means that it is an asymptotically unbiased estimator.

Let D be the matrix storing the previuosly defined errors: $D = [\mathbf{d}_1, \ldots, \mathbf{d}_p]$. Then,

$$\mathcal{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^T = \frac{1}{n}\begin{bmatrix} \mathbf{d}_1^T\mathbf{d}_1 & \mathbf{d}_1^T\mathbf{d}_2 & \ldots & \mathbf{d}_1^T\mathbf{d}_p \\ \mathbf{d}_2^T\mathbf{d}_1 & \mathbf{d}_2^T\mathbf{d}_2 & \ldots & \mathbf{d}_2^T\mathbf{d}_p \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{d}_p^T\mathbf{d}_1 & \mathbf{d}_p^T\mathbf{d}_2 & \ldots & \mathbf{d}_p^T\mathbf{d}_p \end{bmatrix} = \frac{1}{n}D^TD.$$

This means that the estimator $\mathcal{S}$ is generated by the vectors $(\mathbf{d}_1, \ldots, \mathbf{d}_p)$, but $\forall j \in \{1, \ldots, p\}$, $\mathbf{d}_j = \mathbf{y}_j - \overline{x}_j\mathbf{1} \in \mathfrak{L}^{\perp}(\mathbf{1}) \subset \mathbb{R}^{n-1}$. Hence, $\mathcal{S}$ is generated by p elements living in $\mathbb{R}^{n-1}$. That is why $\mathcal{S}_{n-1} = \frac{1}{n-1}D^TD = \frac{n}{n-1}\mathcal{S}$ is the unbiased

estimator of Σ.

Let be noticed that $\mathbf{d}_j = \mathbf{y}_j - \overline{x}_j\mathbf{1} = \mathbf{y}_j - \frac{\mathbf{1}\mathbf{1}^T}{\mathbf{1}^T\mathbf{1}}\mathbf{y}_j = (\mathbb{1} - \frac{\mathbf{1}\mathbf{1}^T}{\mathbf{1}^T\mathbf{1}})\mathbf{y}_j$.
So, $D = (\mathbb{1} - \frac{\mathbf{1}\mathbf{1}^T}{\mathbf{1}^T\mathbf{1}})\mathcal{X} = H\mathcal{X}$, where the matrix $H = \mathbb{1} - \frac{\mathbf{1}\mathbf{1}^T}{\mathbf{1}^T\mathbf{1}}$ is the operator of orthogonal projection on the space $\mathfrak{L}^{\perp}(\mathbf{1})$, so it is symmetric and idempotent. In this way we can derive the following expression for $\mathcal{S}$:

$$\mathcal{S} = \frac{1}{n}D^T D = \frac{1}{n}\mathcal{X}^T H H^T \mathcal{X} = \frac{1}{n}\mathcal{X}^T H \mathcal{X}.$$

1.1.3 Eigenvalues and eigenvectors of the sample covariance matrix. Mahalanobis distance

In this subsection it will be given a geometrical and statistical interpretation of the eigenvalues and eigenvectors of a sample covariance matrix.

Let us ask the question "What is the direction in $\mathbb{R}^p$ with the highest variability among $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \sim F(\boldsymbol{\mu}, \Sigma)$?". Such a question can be expressed as the following optimization problem.

$$\arg\max_{\|\mathbf{a}\|=1} var(\mathbf{a}^T\mathbf{X}) = \arg\max \frac{\mathbf{a}^T\Sigma\mathbf{a}}{\mathbf{a}^T\mathbf{a}}$$

where $\mathbf{a} \in \mathbb{R}^p$.

Theorem 1. Let $B \in \mathbb{R}^{p\times p}$ be a postive definite symmetric matrix, then

$$1. \max_{\mathbf{a}\in\mathbb{R}^p} \frac{\mathbf{a}^T B\mathbf{a}}{\mathbf{a}^T\mathbf{a}} = \lambda_1$$
$$2. \arg\max_{\mathbf{a}\in\mathbb{R}^p} \frac{\mathbf{a}^T B\mathbf{a}}{\mathbf{a}^T\mathbf{a}} = \mathbf{e}_1,$$

where λ_1 and $\mathbf{e}_1$ are respectively the first eigenvalue and the first eigenvector of B.

Proof.
B is positive definite, which means that $B = \sum_{i=1}^p \lambda_i \mathbf{e}_i\mathbf{e}_i^T$, with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. Hence,

$$\frac{\mathbf{a}^T B\mathbf{a}}{\mathbf{a}^T\mathbf{a}} = \frac{\mathbf{a}^T\sum_{i=1}^p \lambda_i\mathbf{e}_i\mathbf{e}_i^T\mathbf{a}}{\mathbf{a}^T\mathbb{1}_p\mathbf{a}} = \frac{\sum_{i=1}^p \lambda_i(\mathbf{e}_i^T\mathbf{a})^2}{\mathbf{a}^T\sum_{i=1}^p \mathbf{e}_i\mathbf{e}_i^T\mathbf{a}} = \frac{\sum_{i=1}^p \lambda_i y_i^2}{\sum_{i=1}^p y_i^2},$$

where $y_i = \mathbf{e}_i^T\mathbf{a}$. Now, we know that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. Hence, $\forall \mathbf{a} \in \mathbb{R}^p$, it holds

$$\frac{\sum_{i=1}^p \lambda_i y_i^2}{\sum_{i=1}^p y_i^2} \leq \frac{\lambda_1\sum_{i=1}^p y_i^2}{\sum_{i=1}^p y_i^2} = \lambda_1.$$

Now, if we find a value $\overline{\boldsymbol{a}} \in \mathbb{R}^p$ such that this disequality is actually an equality, then λ_1 is the maximum of the quantity $\frac{\mathbf{a}^T B \mathbf{a}}{\mathbf{a}^T \mathbf{a}}$, and $\overline{\boldsymbol{a}}$ is the corresponding arg max.
Let us consider $\mathbf{a} = \mathbf{e}_1$. $y_i = \mathbf{e}_i^T \mathbf{a} = \mathbf{e}_i^T \mathbf{e}_1$. This means that $y_1 = 1$ and $y_i = 0$ for $i = \{2, \dots, n\}$. Thus, $\frac{\mathbf{a}^T B \mathbf{a}}{\mathbf{a}^T \mathbf{a}} = \frac{\lambda_1 \cdot 1}{1} = \lambda_1$. ■

Hence, the first eigenvector of the covariance matrix Σ is the direction of highest variability in the data. The measure of such variability is given by the first eigenvalue of Σ.
What about the other eigenvalues and eigenvectors? Through the same method we have used to prove Theorem 1, we can show that

$$\max_{\mathbf{a} \perp \mathbf{e}_1} \frac{\mathbf{a}^T B \mathbf{a}}{\mathbf{a}^T \mathbf{a}} = \lambda_2, \qquad \arg\max_{\mathbf{a} \perp \mathbf{e}_1} \frac{\mathbf{a}^T B \mathbf{a}}{\mathbf{a}^T \mathbf{a}} = \mathbf{e}_2,$$
$$\max_{\mathbf{a} \perp \{\mathbf{e}_1, \mathbf{e}_2\}} \frac{\mathbf{a}^T B \mathbf{a}}{\mathbf{a}^T \mathbf{a}} = \lambda_3, \qquad \arg\max_{\mathbf{a} \perp \{\mathbf{e}_1, \mathbf{e}_2\}} \frac{\mathbf{a}^T B \mathbf{a}}{\mathbf{a}^T \mathbf{a}} = \mathbf{e}_3,$$
$$\dots$$

Hence, the second eigenvector of Σ is the direction of highest variability in the data in the subspace of $\mathbb{R}^p$ orthogonal to the space generated by the first eigenvector. The measure of this variability is the second eigenvalue. The third eigenvector of Σ is the direction of highest variability in the data in the subspace of $\mathbb{R}^p$ orthogonal to the space generated by the first eigenvector and the second one, and so on.

In practice, we have our observations stored in a matrix X, realization of a random matrix $\mathcal{X}$, and the estimate S of the sample covariance matrix $\mathcal{S}$. We know that $S = PDP^T$, where P is the matrix storing the eigenvectors of S and D is the diagonal matrix storing its eigenvalues.
We have just seen that the columns of P, i.e., the eigenvectors of S, generate $\mathbb{R}^p$ and the corresponding referement system indicate the directions of highest variabilities of the data. $\widetilde{X} = XP$ contains the projections of our data on this new referement system. Let us calculate the sample coviarance matrix of these projected data.

$$\widetilde{S} = \frac{1}{n} \widetilde{X}^T H \widetilde{X} = \frac{1}{n} P^T X^T H X P = P^T S P = P^T P D P^T P = D.$$

The sample covariance matrix of the projected data is the diagonal matrix storing the eigenvalues of S! This means that the variables of the projected data are uncorrelated to each other (if the data are gaussian, their variables are also independent), and a measure of their variability is given by the eigenvalues of S.

In (1.1) we have defined a possible inner product in $\mathbb{R}^p$. We have shown that if $\mathcal{S}$ has no zero eigenvalues, then it is positive definite and so is its inverse. Thus, we can replace in (1.1) Q with $\mathcal{S}^{-1}$, in order to have $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{S}^{-1}} = \mathbf{x}^T \mathcal{S}^{-1} \mathbf{y}$.

Definition 4. *Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$, then*

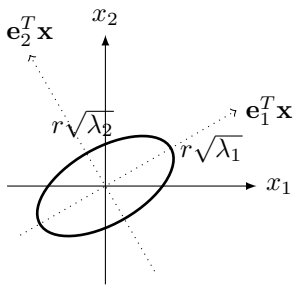
$$d^2_{\mathcal{S}^{-1}}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle_{\mathcal{S}^{-1}} = (\mathbf{x} - \mathbf{y})^T \mathcal{S}^{-1} (\mathbf{x} - \mathbf{y}) \tag{1.2}$$

is called the ***Mahalanobis distance****, between $\boldsymbol{x}$ and $\boldsymbol{y}$.*

In $\mathbb{R}^p$, $d^2_{\mathcal{S}^{-1}}(\mathbf{x}, \mathbf{x}) \leq r^2$, represents a sphere of radius r, according to the metrics defined by $d^2_{\mathcal{S}^{-1}}$. According to the Euclidean metrics, instead, it would simply be an ellipsoide.

Let us consider the case $p = 2$, and let us analyze the expression $d^2_{\mathcal{S}^{-1}}(\mathbf{x}, \mathbf{x}) = r^2$.

$$\begin{aligned} d^2_{\mathcal{S}^{-1}}(\mathbf{x}, \mathbf{x}) &= \mathbf{x}^T \mathcal{S}^{-1} \mathbf{x} = \mathbf{x}^T \left(\frac{1}{\lambda_1} \mathbf{e}_1 \mathbf{e}_1^T + \frac{1}{\lambda_2} \mathbf{e}_2 \mathbf{e}_2^T \right) \mathbf{x} = \\ &= \mathbf{x}^T \left(\frac{1}{\lambda_1} \mathbf{e}_1 \mathbf{e}_1^T \mathbf{x} + \frac{1}{\lambda_2} \mathbf{e}_2 \mathbf{e}_2^T \mathbf{x} \right) = \frac{1}{\lambda_1} \mathbf{x}^T \mathbf{e}_1 \mathbf{e}_1^T \mathbf{x} + \frac{1}{\lambda_2} \mathbf{x}^T \mathbf{e}_2 \mathbf{e}_2^T \mathbf{x} = \\ &= \left(\frac{\mathbf{e}_1^T \mathbf{x}}{\sqrt{\lambda_1}} \right)^2 + \left(\frac{\mathbf{e}_2^T \mathbf{x}}{\sqrt{\lambda_2}} \right)^2 = y_1^2 + y_2^2 = r^2. \end{aligned}$$



We have ended up with the equation of a circumference of radius r, in a new referement system defined by the new variables $y_1 = \frac{\mathbf{e}_1^T \mathbf{x}}{\sqrt{\lambda_1}}$ and $y_2 = \frac{\mathbf{e}_2^T \mathbf{x}}{\sqrt{\lambda_2}}$. In the original referement system we see an ellipse, rotated in direction of the two eigenvectors $\mathbf{e}_1$ and $\mathbf{e}_2$ of $\mathcal{S}$, orthogonal to each other. The first semi-axis has length equal to $r\sqrt{\lambda_1}$, while the second one has length equal to $r\sqrt{\lambda_2}$.

Summarizing this subsection, we have dealt with the problem of looking for the successively orthogonal directions that maximally explain the variability in the data. We have found that these directions are given by the eigenvectors of a population covariance matrix, and the variability in each direction is given by the corresponding eigenvalue. This problem is known as *Principal Component Analysis* (PCA). The directions given by the eigenvectors of Σ are called to be the *principal component directions*. We can project our data on this new referement system, where there is no correlation between the variables and where there is a sorting of the variability in the data ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$).

Since it may be not desirable to consider all the original p variables, expecially in high dimensions, PCA is one of the most important techniques for reduction of the data dimensionality: by projecting our data on the space generated by the first eigenvectors, we may explain most of the information, which is measured by the ratio between the sum of the corresponding eigenvalues and the sum of all of them. In presence of large data, an important problem is the choice of the number of the first eigenvectors to consider, dealing with the trade-

off between complexity of the projected data and their explained variability.

In practice, estimates of the sample covariance matrix and its eigenvalues are considered rather than the true objects. Thus, the study of the distribution of the eigenvalues of $\mathcal{S}$ becomes an important task that we are going to investigate in the following chapter.

1.1.4 Assumption of normality

In the last subsection we have introduced the Mahalanobis distance between two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, when the sample covariance matrix S is positive definite. Let us ask the question: "What is that distribution F, with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, such that the density function is centered in $\boldsymbol{\mu}$ and decreases exponentially when getting far from $\boldsymbol{\mu}$, with respect to the square distance $d_{\Sigma^{-1}}$?", i.e.,

$$f(\mathbf{x}) = k_1 \cdot exp\{-k_2 d^2_{\Sigma^{-1}}(\mathbf{x}, \boldsymbol{\mu})\} = k_1 \cdot exp\{-k_2 (\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})\},$$

where k_2 defines the rate of exponential velocity, while k_1 is to make f be a density function, such that the integral in $\mathbb{R}^p$ is equal to one.

Definition 5. *Fix $\boldsymbol{\mu} \in \mathbb{R}^p$ and a symmetric positive definite $p \times p$ matrix Σ. The random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^T$ is said to have a p-variate **Gaussian (normal)** distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ if its density function $f(\boldsymbol{x})$ is of the following form.*

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^p det\Sigma}} exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})\}.$$

This is denoted by $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$.

Data are quite often assumed to be realizations of independent multivariate *Gaussian (normal)* vectors. In the following chapters we will work under this assumption.

1.1.5 The Wishart distribution

In this subsection we are going to investigate the distribution of the sample covariance matrix $\mathcal{S}$.

Definition 7. Let $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ be i.i.d.$\sim N_p(\mathbf{0}, \Sigma)$, then

$$\sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i^T \sim W_p(\Sigma, n),$$

i.e., the Wishart distribution (W_p) with scale matrix $\Sigma \in \mathbb{R}^{p \times p}$ and n degrees of freedom is associated to $p \times p$ random matrices, that are sum of n i.i.d. random

vectors $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n \in \mathbb{R}^p$, normal distributed with $\mathbf{0}$ mean and Σ covariance matrix, multiplied by their transpose.

In general, let us consider the random matrix $\mathcal{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)^T$, where $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are i.i.d. random vectors from $N_p(\boldsymbol{\mu}, \Sigma)$. This is denoted by $\mathcal{X} \sim N_{n\times p}(\boldsymbol{\mu}, \Sigma)$. Let now $\mathcal{M}$ be $\mathcal{M} = \mathcal{X}^T\mathcal{X}$. Then, $\mathcal{M} \sim W_p(\Sigma, n)$.
The case when $\Sigma = \mathbb{1}_p$, where $\mathbb{1}_p$ represents the $p \times p$ identity identity matrix, is referred to as the *null* case, and the corresponding Wishart distribution is called *white Wishart distribution.*

The Wishart distribution has a density function only when $n \geq p$. Let A be from $W_p(\Sigma, n)$. Then, it has the following form (see [12]):

$$\frac{2^{-np/2}}{\Gamma_p(\frac{n}{2})(det\Sigma)^{n/2}} etr\left(-\frac{1}{2}\Sigma^{-1}A\right)(detA)^{(n-p-1)(2)},$$

where $etr(X) = e^{tr(X)}$.

In the univariate case $p = 1$, $z_1, z_2, \ldots, z_n$ i.i.d. $\sim N(0, \sigma^2)$, we have

$$\sum_{i=1}^{n} z_i z_i^T = \sum_{i=1}^{n} z_i^2 = \sigma^2 \sum_{i=1}^{n} \left(\frac{z_i}{\sigma}\right)^2 \sim \sigma^2\chi^2(n) \equiv W_1(\sigma^2, n).$$

Here are some properties of the Wishart distribution:

1. If $A \sim W_p(\Sigma, n_1) \perp B \sim W_p(\Sigma, n_2)$, then $A + B \sim W_p(\Sigma, n_1 + n_2)$.
2. If $C \in \mathbb{R}^{(q,p)}$ and $A \sim W_p(\Sigma, n)$, then $CAC^T \sim W_q(C\Sigma C^T, n)$.
3. From (2), we can derive that if $\mathbf{c} \in \mathbb{R}^p$ and $A \sim W_p(\Sigma, n)$, then $\mathbf{c}^T A\mathbf{c} \sim W_1(\mathbf{c}^T\Sigma\mathbf{c}, n) \equiv (\mathbf{c}^T\Sigma\mathbf{c})\chi^2(n)$.

It is finally time to introduce the distribution of the sample covariance matrix $\mathcal{S}$.

Theorem 2. *Let $\mathcal{X} \in \mathbb{R}^{n\times p}$ be a random matrix from $N_{n\times p}(\boldsymbol{\mu}, \Sigma)$. Then, the scaled sample covariance matrix $\mathcal{S}$ has the Wishart distribution:*

$$\mathcal{S} = \frac{1}{n}\mathcal{X}^T H\mathcal{X} \sim W_p\left(\frac{\Sigma}{n}, n-1\right). \tag{1.3}$$

1.2 Research questions

Suppose we have n p-variate observations, stored in a data matrix X, which can be viewed as a realization of a random matrix $\mathcal{X} \sim N_{n\times p}(\boldsymbol{\mu}, \Sigma)$. Then, we can compute the corresponding estimates $\overline{\boldsymbol{x}}$ and S of the sample mean $\overline{\boldsymbol{X}}$ and of

the sample covariance matrix $\mathcal{S}$ and we can project our data on the referement system centered in $\overline{\boldsymbol{x}}$ and whose axis-directions are given by the eigenvectors of S. The projected data have null sample mean and sample covariance matrix as diagonal matrix storing the eigenvalues of the sample covariance matrix of the original data.
Theoretically, we can think to translate the n rows of $\mathcal{X}$ by $\boldsymbol{\mu}$, such that the new data $\mathcal{Y}$ are from $N_{n\times p}(\mathbf{0}, \Sigma)$. Then, knowing that $\Sigma = \Gamma\Lambda\Gamma^T$, we can project our new data on the referement system whose directions are given by the eigenvectors of Σ, i.e., $\mathcal{Y}\Gamma \sim N_{n\times p}(\mathbf{0}, \Lambda)$.

Therefore, without loss of generality, we can consider a data matrix X as a realization of a random matrix $\mathcal{X} \sim N_{n\times p}(\mathbf{0}, diag(\lambda_i))$. Let S be the estimate of the corresponding sample covariance matrix, obtained from the observed X. Let $\bar{l}_1 \geq \bar{l}_2 \geq \cdots \geq \bar{l}_p \geq 0$ be the eigenvalues of S. They can be viewed as estimates of the eigenvalues $\hat{l}_1 \geq \hat{l}_2 \geq \cdots \geq \hat{l}_p \geq 0$ of $\mathcal{S}$, which are estimators of the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ of Σ.

In this settings, this thesis is going to investigate the distribution of the eigenvalues of $\mathcal{S}$, by reporting some of the already known theoretical results on this topics and by simulations in R to visualize such results.

Furthermore, the already known theory on the antieigenvalues, which are functions of the eigenvalues, of fixed matrices will be presented. Finally, an attempt to embark on a theory on antieigenvalues of random matrices ($\mathcal{S}$ is of our interest) by simulations in R, will be performed.

Chapter 2

Distribution of eigenvalues of sample covariance matrices

Since the second half of the XX century, the topic of the limiting spectral analysis of random matrices in large dimensions has been considered with interest from mathematicians, probabilists and statisticians. In this chapter, the asymptotic behaviour of the eigenvalues of the sample covariance matrix of a random matrix $\mathcal{X} \in \mathbb{R}^{n \times p}$ whose entries are i.i.d. normal distributed, will be investigated when n tends to infinity. Firstly p is thought to be fixed (Wigner's semicircular law) and then p is thought to tend to infinity as well, under the condition that the ratio $\frac{n}{p}$ tends to a positive costant $\gamma \in (0, \infty)$ (Marčenko-Pastur's law).

Some simulations in R will be performed in order to confirm the theoretical results.

2.1 Wigner matrices and Wishart ensemble

In this section Wigner matrices and the Wishart ensemble will be presented, in order to be ready in the next sections to discuss about the limit distribution of the eigenvalues of $\mathcal{S}$.

2.1.1 Haar measure

A *random matrix* can be defined as a probability space $(\Omega, \mathcal{F}, \mathcal{P})$, where Ω is the sample space of matrices of interest, $\mathcal{F}$ is a σ-algebra on Ω, i.e., a family of measurable subsets of Ω, $\mathcal{F} \subseteq 2^{\Omega}$, and $\mathcal{P}$ is a measure defined on Ω.
When a group Ω formed by an ensemble of random matrices is compact or locally compact, the *Haar measure* is a common used probability measure $\mathcal{P}_H$ on Ω and it is translation invariant. For any measurable set $B \in \mathcal{F}$ and any element $\mathcal{M} \in \Omega$.

$$\mathcal{P}_H(\mathcal{M}B) = \mathcal{P}_H(B), \text{ where } \mathcal{M}B = \{\mathcal{M}N | N \in B\}.$$

2.1.2 Wigner matrices

Wigner matrices were introduced by the Hungarian American theoretical physicist and mathematician Eugene Wigner in order to study the statistics of the excited energy levels of heavy nuclei.

A Wigner matrix is a square $n \times n$ symmetric (or Hermitian in the complex case, but this is not of our interest) random matrix $A = (A_{ij})$ with i.i.d. entries above the diagonal:

$$A_{ij} = A_{ji}, 1 \leq i \leq j \leq n, \{A_{ij}\}_{i \leq j} \text{ - i.i.d. real random variables.}$$

Let us assume that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. samples drawn from a p-dimensional multivariate normal population $N(\boldsymbol{\mu}, \mathbb{1}_p)$. When n tends to infinity, the sample covariance matrix $\mathcal{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^T \to \mathbb{1}_p$ and $\sqrt{n}(\mathcal{S} - \mathbb{1}_p) \to \sqrt{p}W_p$. It can be seen that the entries above the main diagonal of $\sqrt{p}W_p$ are i.i.d. $\mathcal{N}(0,1)$, while the entries on the diagonal are iid $\mathcal{N}(0,2)$. Such a matrix is called the (standard) Gaussian matrix or Wigner matrix.

2.1.3 Wishart ensemble

We already know the definition of the sample covariance matrix for a random matrix $\mathcal{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n]^T$, where $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^T$. In most cases of spectral analysis of large dimensional random matrices, though, the sample covariance matrix is simply defined as

$$\mathcal{S} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T = \frac{1}{n}\mathcal{X}^T\mathcal{X}.$$

Indeed, when we are in large dimensions, the projection of the vectors $\boldsymbol{Y_1}, \boldsymbol{Y_2}, \ldots, \boldsymbol{Y_p}$ on $\mathfrak{L}^{\perp}(\mathbf{1}) \subset \mathbb{R}^{n-1}$ almost coincides with the vectors theirselves in $\mathbb{R}^n$. Hence, $H\mathcal{X} \simeq \mathcal{X}$.

Suppose that $\{X_{ij}\}_{i,j=1,2,\ldots}$ is a double array of i.i.d. real random variables with zero mean and variance σ^2. The symmetrization of $\mathcal{X}$ by multiplying by its transpose $\mathcal{X}^T$, gives us the ensemble of the positive definite matrices $\mathcal{X}^T\mathcal{X}$, which coincides up to a constant with a family of sample covariance matrices. This ensemble was named after Wishart who first computed the joint element density of $\mathcal{X}^T\mathcal{X}$, which has a Wishart distribution, as already seen in the previous chapter.

It is in our interest to consider what in literature is known as the the *Wishart ensemble of real sample covariance matrices*. This is the case when $\mathcal{X}$ is a $n \times p$ real random matrix with np i.i.d. entries $\mathcal{N}(0,1)$.

$$X_{ij} \sim \mathcal{N}(0,1), 1 \leq i \leq n, 1 \leq j \leq p.$$

Let us now study the joint density of the eigenvalues of a Wishart matrix. The following theorem can be found in Muirhead [12] (1982, Th.3.2.17).

Theorem 3. *If $\mathcal{A}$ is a $p \times p$ positive definite random matrix with density function f($\mathcal{A}$) then the joint density function of the eigenvalues $l_1 > l_2 > \cdots > l_p > 0$ of $\mathcal{A}$ is*

$$\frac{\pi^{p^2/2}}{\Gamma_p(p/2)} \prod_{1 \le i \le j \le p} (l_i - l_j) \int_{\mathcal{O}_p} f(HLH^T)(dH).$$

Here dH stands for the Haar invariant probability measure on the orthogonal group $\mathcal{O}_p$, normalized so that

$$\int_{\mathcal{O}_p} (dH) = 1.$$

Let us now apply this general theorem to the Wishart case (Muirhead, Th.3.2.18).

Theorem 4. *If $\mathcal{A} \sim W_p(\Sigma, n)$ with $n > p-1$, the joint density of the eigenvalues $l_1, l_2, \ldots, l_p$ of $\mathcal{A}$ is*

$$\frac{\pi^{p^2/2} 2^{-np/2} (det\Sigma)^{-n/2}}{\Gamma_p(\frac{n}{2})\Gamma_p(\frac{p}{2})} \prod_{i=1}^{p} l_i^{(n-p-1)/2} \prod_{j>i}^{p} (l_i - l_j) \int_{\mathcal{O}_p} etr(-\frac{1}{2}\Sigma^{-1} HLH^T)(dH).$$

This formula can be obtained by substituing $f(\mathcal{A})$ in the previous theorem with the Wishart density function and by reminding that $det\mathcal{A} = \prod_{j=1}^{p} l_i$.

In the null case, when $\Sigma = \lambda \mathbb{1}$, we have that

$$\int_{\mathcal{O}_p} etr(-\frac{1}{2\lambda} HLH^T)(dH) = etr(-\frac{1}{2\lambda} L) \int_{\mathcal{O}_p} (dH) = e^{(-\frac{1}{2\lambda} \sum_{j=1}^{p} l_i)}.$$

Hence, the joint density distribution of the eigenvalues for the null Wishart matrix $\mathcal{A} \sim W_p(\lambda \mathbb{1}_p, n)$is

$$\frac{\pi^{p^2/2} 2^{-np/2} (det\Sigma)^{-n/2}}{\Gamma_p(\frac{n}{2})\Gamma_p(\frac{p}{2})} e^{(-\frac{1}{2\lambda} \sum_{1=1}^{p} l_i)} \prod_{i=1}^{p} l_i^{(n-p-1)/2} \prod_{j>i}^{p} (l_i - l_j).$$

2.2 Distribution of the largest eigenvalue: sample coviarance matrices

We already know that if $\mathcal{X} \sim N_{n \times p}(\boldsymbol{\mu}, \Sigma)$, the sample covariance matrix $\mathcal{S} = \frac{1}{n} \mathcal{X}^T H \mathcal{X}$ has the Wishart distribution $W_p(\frac{1}{n}\Sigma, n-1)$. Hence, we can just study the distribution of the eigenvalues of $\mathcal{A} = n\mathcal{S} \sim W_p(\Sigma, \widetilde{n})$, where $\widetilde{n} = n - 1$,

keeping in mind that its eigenvalues are just n times greater than those of $\mathcal{S}$.

In this context, let us formulate Theorem 4 in the following way:

Theorem 4. *The joint density function of the eigenvalues $l_1 > l_2 > \cdots > l_p > 0$ of a sample covariance matrix $\mathcal{S} \sim W_p(\Sigma, n)$ $(n > p)$ is of the following form*

$$\frac{\pi^{p^2/2}(det\Sigma)^{-\tilde{n}/2}}{\Gamma_p(\frac{\tilde{n}}{2})\Gamma_p(\frac{p}{2})}\frac{\tilde{n}^{-\tilde{n}p/2}}{2}\prod_{i=1}^{p} l_i^{(\tilde{n}-p-1)/2}\prod_{j>i}^{p}(l_i - l_j)\int_{\mathcal{O}_p} etr(-\frac{1}{2}n\Sigma^{-1}HLH^T)(dH).$$

This is the formula expressed in Theorem 4, with just a correction for n and the substitution of all l_i by $\tilde{n}l_i$.

Following Muirhead (Th. 7.2.12), by averaging over the group $\mathcal{O}_p$ we get the equality

$$etr(-\tfrac{1}{2}\tilde{n}\Sigma^{-1}HLH^T) = {}_0F_0(-\tfrac{1}{2}\tilde{n}L, \Sigma^{-1}),$$

where ${}_0F_0(\cdot,\cdot)$ is the *(two matrix) multivariate hypergeometric function.* Thus, the joint density function of the eigenvalues of $\mathcal{S}$ is

$$\frac{\pi^{p^2/2}(det\Sigma)^{-\tilde{n}/2}}{\Gamma_p(\frac{\tilde{n}}{2})\Gamma_p(\frac{p}{2})}\frac{\tilde{n}^{-\tilde{n}p/2}}{2}\prod_{i=1}^{p} l_i^{(\tilde{n}-p-1)/2}\prod_{j>i}^{p}(l_i - l_j){}_0F_0(-\frac{1}{2}\tilde{n}L, \Sigma^{-1}).$$

The distribution of the largest eigenvalue - which is measure of the variability in the data explained by the first principal component as we already know - can also be expressed in terms of the hypergeometric function of a matrix argument. The following theorem is formulated in Murihead as a corollary of a more general result regarding the positive definite matrices.

Theorem 5. *If l_1 is the largest eigenvalue of $\mathcal{S}$, then the cumulative distribution function of l_1 can be expressed in the form*

$$\mathbb{P}(l_1 < x) = \frac{\Gamma_p(\frac{p+1}{2})}{\Gamma_p(\frac{n+p}{2})}det(\frac{\tilde{n}}{2}\Sigma^{-1})_1^{\tilde{n}/2}F_1(\frac{\tilde{n}}{2}; \frac{n+p}{2}; -\frac{n}{2}x\Sigma^{-1}).$$

2.3 Convergence of eigenvalues

It is finally time to analyze the limit distribution of the eigenvalues of $\mathcal{S}$.

2.3.1 The Wigner semicircular law

Dealing with eigenvalues' convergence in sample covariance matrices concerns not only the convergence's type (weak convergence, almost surely convergence, convergence in probability), but also the conditions imposed on the dimensions of the data matrix.

Following Anderson (2003) [1], in the case of a multinormal population, the asymptotic distribution of eigenvalues of a covariance matrix is expressed in the following theorem.

Theorem 6. *Suppose $\mathcal{S}$ is a $p \times p$ sample covariance matrix corresponding to a data matrix drawn from $N_{n\times p}(\boldsymbol{\mu}, \Sigma)$. Asymptotically, the eigenvalues $l_1, l_2, \ldots, l_p$ of $\mathcal{S}$ are distributed as follows:*

$$\sqrt{n}(l_j - \lambda_j) \xrightarrow{dist} N(0, 2\lambda_j^2), \quad for\ j = 1, \ldots, p,$$

where $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the (distinct) eigenvalues of the population covariance matrix Σ.

In this theorem, p is considered as a fixed value. The convergence is meant in the weak sense, i.e., the pointwise convergence of the cumulative distribution function of $\sqrt{n}(l_j - \lambda_j)$ to $\frac{1}{2\sqrt{\pi}\lambda_j} e^{-\frac{x^2}{4\lambda_j}}$ occurs.

Let $\#E$ be the cardinality of the set E, and $\mathbb{1}(E)$ be the event indicator function, such that

$$\mathbb{1}(E) = \begin{cases} 1, & \text{if } E \text{ is true} \\ 0, & \text{otherwise.} \end{cases}$$

Definition 8. Let A be a $p \times p$ matrix with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$. If all these eigenvalues are real, we can define the so called *empirical distribution function*

$$F^A(x) = \frac{1}{p}\#\{j \leq p \mid l_j \leq x\} = \frac{1}{p}\sum_{j=1}^{p} \mathbb{1}(l_j \leq x).$$

Given a sequence of random matrices $\{A_n\}$, it is interesting to investigate the convergence of the sequence of empirical spectral distributions $\{F^{A_n}\}$. The limit distribution F is called the *limiting spectral distribution* of the sequence $\{A_n\}$.

Reminding that the derivative of the event indicator function is the Dirac delta function $\delta(x)$, the probability density function of $F^A(x)$ is

$$f^A(x) = \frac{1}{p}\sum_{j=1}^{p}\delta(x - l_i).$$

Wigner developed the **semicircular law** by studying the spectral analysis of the large dimensional Wigner matrix. He proved that the limit of the expected empirical spectral distribution of a $n \times n$ Wigner matrix, normalized by $1/\sqrt{n}$, when n goes to infinity, is the semicircular law F, whose density is given by

$$F'(x) = \begin{cases} \frac{1}{2\pi}\sqrt{4 - x^2}, & \text{if } |x| \leq 2 \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

This is an interesting result concerning a limit distribution of eigenvalues of a sequence of $p \times p$ Wigner matrices, which can be seen as a sequence of $\sqrt{n}(\mathcal{S} - \mathbb{1}_p)$, with n (the number of rows of our data matrix $\mathcal{X}$) "fixed" to infinity, while p (the number of columns) tends to infinity.

Let us in the next subsection see what happens when both of the dimensions are allowed to tend to infinity.

2.3.2 Marčenko-Pastur distribution

In spectral analysis of large dimensional sample covariance matrices, it is usual to assume that the dimension n goes to infinity proportionally to p, i.e., $n = n(p) \to \infty$, s.t. $\frac{n}{p} \to \gamma \in (0, \infty)$.

In this context, the first success in finding the limiting spectral distribution of the large sample covariance matrix $\mathcal{S}$, was due to Marčenko and Pastur.

Theorem 7. (Marčenko-Pastur 1967) *Suppose that x_{ij} are i.i.d. real random variables with mean zero and variance σ^2. Also assume that $n/p \to \gamma \in (0, \infty)$. Then, with probability one, the empirical distribution function $F^{\mathcal{S}}$ of the eigenvalues $l_1^{(p)}, l_2^{(p)}, \ldots, l_p^{(p)}$ of the $p \times p$ sample covariance matrix $\mathcal{S}$ tends to the so called **Marčenko-Pastur law**, whose density is defined as follows.*

$$f_\gamma^{\mathcal{S}}(x) = \begin{cases} \frac{\gamma}{2\pi x\sigma^2}\sqrt{(b - x)(x - a)}, & \text{if } a \leq x \leq b, \\ 0, & \text{otherwise,} \end{cases} \tag{2.2}$$

and has a point mass $1 - \gamma$ at the origin if $\gamma \in (0, 1)$, where $a = \sigma^2(1 - \frac{1}{\sqrt{\gamma}})^2$ and $b = \sigma^2(1 + \frac{1}{\sqrt{\gamma}})^2$. Let be noticed that this theorem holds regardless of the choice of the entry distribution.
If $\sigma^2 = 1$, the M-P law is said to be the *standard M-P law.*

The investigation on limits of extreme eigenvalues is important in many areas, such as in signal processing, in pattern recognition and edge detection. It is known that

$$l_1^{(p)} \to b = \sigma^2 \left(1 + \frac{1}{\sqrt{\gamma}}\right)^2, \tag{2.3}$$

i.e., the largest eigenvalue $l_1^{(p)}$ of a sample covariance matrix tends almost surely to the right edge of the domain of the Marčenko-Pastur law, when $\gamma \in (0, \infty)$ under the assumption of the existence of the fourth moment of the underlying distribution. This condition is not only sufficient, but also necessary thanks to the work developed by Bai, Silverstein and Yin in [4] who proved that the limsup of the largest eigenvalue of a sample covariance matrix is infinity with probability 1 if the fourth moment of the underlying distribution is not finite.

Let us now ask what is the limit of the smallest eigenvalue of a large sample covariance matrix. The most current answer is due to Bai and Yin in [3], who proved that

$$l_{min\{n,p\}}^{(p)} \to a = \sigma^2 \left(1 - \frac{1}{\sqrt{\gamma}}\right)^2, \tag{2.4}$$

i.e., the smallest (non zero) eigenvalue $l_{min\{n,p\}}^{(p)}$ of a sample covariance matrix tends almost surely to the left edge of the domain of the Marčenko-Pastur law, when $\frac{n}{p} \to \gamma \in (0, \infty)$ under the existence of the fourth moment of the underlying distribution. Let be noticed that, if $n < p$, then $l_{n+1}^{(p)} = l_{n+2}^{(p)} = \cdots = l_p^{(p)} = 0$.

The *almost sure* convergence means here, that, for instance for the largest eigenvalue,

$$\mathbb{P}\left(\lim_{n/p \to \gamma} l_1^{(p)} = \sigma^2 (1 + \frac{1}{\sqrt{\gamma}})^2\right) = 1.$$

2.4 Results through simulations in R

So far, we have discussed about a few theoretical results about the limit distribution of the eigenvalues of a sample covariance matrix $\mathcal{S}$ related to a random matrix $\mathcal{X}$. In this section we are going to show some simulations in R to have a graphical representation of these theoretical results, assuming that $\mathcal{X}_{ij} \overset{iid}{\sim} N(0, \sigma^2 = 1)$.

2.4.1 M-P law: probability distribution function

Figure (2.1) represents the histograms of the eigenvalues $l_1, l_2, \ldots, l_p$ of sample covariance matrices when $p = 1000$ and respectively $n = \{300, 600, 900\}$. In each figure it is also drawn in blue color the Marčenko-Pastur theoretical density, with a mass point in $x = 0$, as $\gamma < 1$.

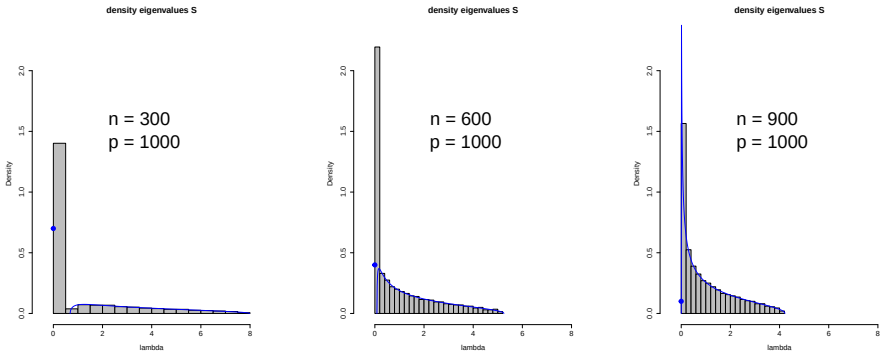


Figure 2.1: distribution of eigenvalues of $\mathcal{S}$ for $\gamma = \{0.3; 0.6; 0.9\}$.

It is noticeable that the results from the simulations follow the already known theoretical results.

Figure (2.2), the same content is presented, but when the number of columns of $\mathcal{X}$ is respectively $n = \{3000; 10000; 17000\}$. So, this time $\gamma > 1$ and there is no mass point in $x = 0$.

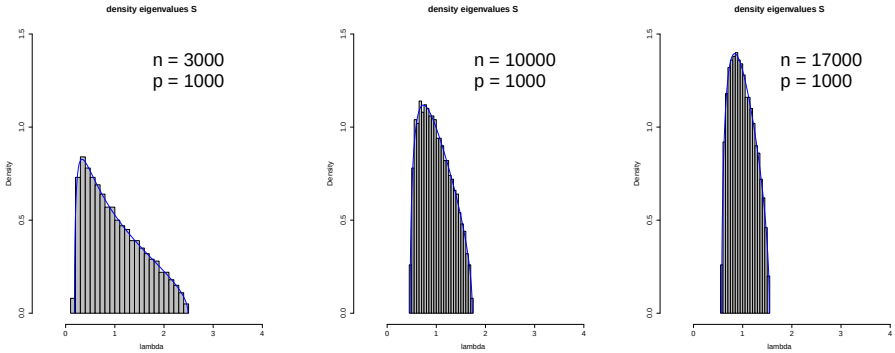


Figure 2.2: distribution of eigenvalues of $\mathcal{S}$ for $\gamma = \{3; 10; 17\}$.

Also in this case it is noticeable that the results from simulations fit pretty well the theory. Let be noticed that, the higher is γ and the more the eigenval-

ues distribute around the value $x = 1$. This confirms, indeed, our expectations, as we know that $\mathcal{S}$ converges to $\Sigma = \mathbb{1}_p$, whose all eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ are equal to one.

Figure (2.3) shows the density curves of the Marčenko-Pastur distribution when γ varies respectively from 0.1 to 1 and from 1 to 20.

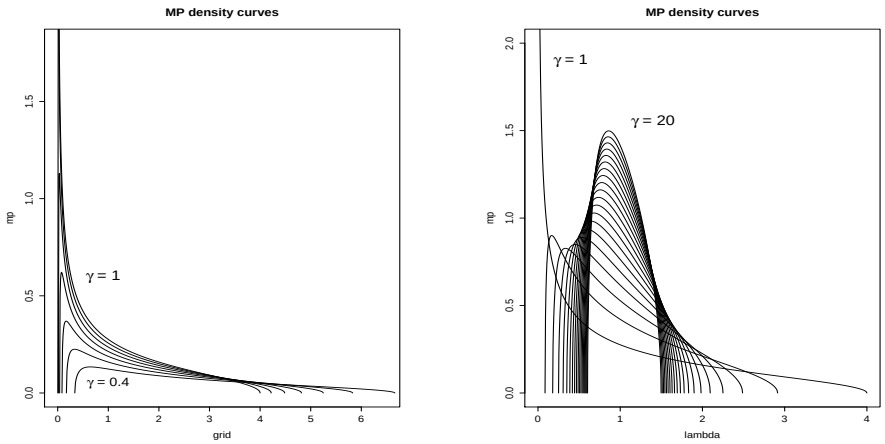


Figure 2.3: Marčenko-Pastur density curves depending on γ values.

2.4.2 Convergence of the largest eigenvalue and of the smallest eigenvalue

Figure (2.4), the convergence of the largest eigenvalue of $\mathcal{S}$ and of the smallest one is analyzed. In particular, it is studied the case $\gamma = \frac{n}{p} = 3$.

The simulation consists of 70 iterations. The first iteration starts with $p = 5$ and $n = \gamma \cdot p = 3 \cdot 5 = 15$. At the *i-th* iteration, the largest eigenvalue and the smallest one of the sample covariance matrix S of a random normal matrix with $n_i = n \cdot i$ rows and $p_i = p \cdot i$ columns, are computed, 100 times in the same way and in the end it is taken the mean for both the largest eigenvalue and the smallest one, out of the 100 samples. Hence, in the last iteration $n_{70} = 1050$ and $p_{70} = 350$.

In this algorithm we have set $\gamma = \frac{n}{p} = 3$; a sequence $\{p_i\}_{i=1}^{70}$, such that $\mathbf{p} = (5, 10, \ldots, 350)$; and a sequence $\{n_i\}_{i=1}^{70}$, $\mathbf{n} = (\gamma \cdot p_1, \gamma \cdot p_2, \ldots, \gamma \cdot p_{70}) = (15, 30, \ldots, 1050)$, in such a way that our limit conditions are $p \to$ "∞" and $\frac{n(p)}{p} = \frac{\gamma \cdot p}{p} = \gamma$ "$\to$" $\gamma = 3$. In this settings, we can see that $l_1^{(p)}$ converges to the right edge of the support of the Marčenko-Pastur law $b = (1 + \frac{1}{\sqrt{\gamma}})^2$; while $l_p^{(p)}$ converges to the left edge of the support of the Marčenko-Pastur law $a = (1 - \frac{1}{\sqrt{\gamma}})^2$, confirming the theoretical results (2.3) and (2.4).

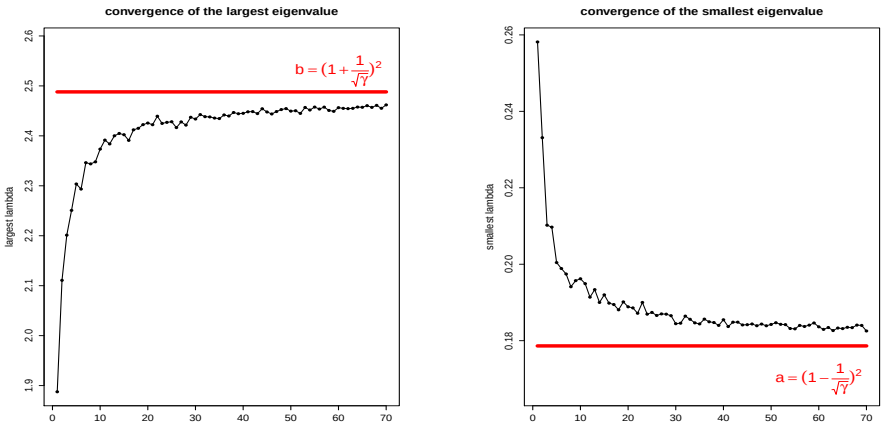


Figure 2.4: Convergence of the largest eigenvalue of $\mathcal{S}$ and of the smallest one.

Chapter 3

Distribution of antieigenvalues and sample covariance matrices

In this chapter, it will be presented the already known theory about the antieigenvalues, introduced almost 50 years ago by Karl Gustafson, now professor at the department of Mathematics at the University of Colorado Boulder. So far, antieigenvalues have been analyzed in a context of fixed matrices. Our purpose is to extend this theory in a context of random matrices, trying to set a rule for the limit distribution of the antieigenvalues (or at least for the first one), in analogy with what we have done in the previuos chapter for the eigenvalues.

3.1 Introduction

In Karl Gustafson's work "The angle of an operator and positive operator products" (1968, [5]), he introduced two new quantities, namely *the angle of an operator* and *the minimum of a certain (norm) function related to tangent functionals.*

Definition 9. *Let $B : X \to X$ be an operator, where X is a complex Banach space, then*

$$\phi_R(B) = cos^{-1}\left(\inf_{\mathbf{x}\in D(B)} \frac{Re\langle B\mathbf{x}, \mathbf{x}\rangle}{\|B\mathbf{x}\|\|\mathbf{x}\|}\right), \tag{3.1}$$

is called the (real) ***angle of B****. It measures the maximum (real) turning effect of B.*

Definition 10. *Let $\epsilon \geq 0$, $g(\epsilon, B) = \|\epsilon B + \mathbb{1}\|$, then*

$$g_m(B) = \min_{\epsilon} g(\epsilon, B). \tag{3.2}$$

At the Los Angeles Inequalities Symposium in 1969 ([6]) Gustafson introduced the name *antieigenvalues* to indicate the following quantities.

Definition 11. *Let B be defined as above, then*

$$\mu_k(B) = \inf_{\mathbf{x}\perp\{\mathbf{x}_1,\dots,\mathbf{x}_{k-1}\}} \frac{Re\langle B\mathbf{x},\mathbf{x}\rangle}{\|B\mathbf{x}\|\|\mathbf{x}\|}, \quad \mathbf{x}\ s.t\ \mathbf{x}\in D(B)\setminus\{0\},\ B\mathbf{x}\neq 0. \tag{3.3}$$

are called the ***antieigenvalues*** *of B, while $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_k$ are the corresponding* ***antieigenvectors****.*

In particular, by looking at Definition 9, we can notice that the first (smallest) antieigenvalue

$$\mu_1(B) = \inf_{\mathbf{x}\in D(B)} \frac{Re\langle B\mathbf{x},\mathbf{x}\rangle}{\|B\mathbf{x}\|\|\mathbf{x}\|}. \tag{3.4}$$

is the cosine of the (real) angle of the operator B.

$$\mu_1(B) = cos\left(\phi_R(B)\right).$$

In short, we can write that the first antieigenvalue $\mu_1(B)$ of an operator B is its cosine, i.e., $\mu_1(B) = cosB$.

Coming back to Definition 10, the quantity $g_m(B) = \min_\epsilon \|\epsilon B + \mathbb{1}\|$ becomes more interesting when it is seen that

$$g_m(-B) = sin\left(\phi_R(B)\right) \quad (= sinB).$$

That is why it is introduced the quantity

$$\nu_1(B) = g_m(-B) = \min_\epsilon \|\epsilon B - \mathbb{1}\| = sinB. \tag{3.5}$$

Let B be a $p\times p$ symmetric positive definite matrix, then the following relation is satisfied.

$$\nu_1(B)^2 + \mu_1(B)^2 = sin^2(B) + cos^2(B) = 1.$$

In addition, the solution of (3.4) is

$$\mu_1 = \frac{2\sqrt{\lambda_1\lambda_p}}{\lambda_1+\lambda_p}, \tag{3.6}$$

and it is attained at the two maximally turned first antieigenvectors:

$$\mathbf{x}_{1,2} = \frac{\sqrt{\lambda_p}}{\sqrt{\lambda_1+\lambda_p}}\mathbf{e}_1 \pm \frac{\sqrt{\lambda_1}}{\sqrt{\lambda_1+\lambda_p}}\mathbf{e}_p,$$

where λ_1 and λ_p are respectively the largest and the smallest eigenvalue of B, while $\mathbf{e}_1$ and $\mathbf{e}_p$ are their corresponding eigenvectors.

Moreover, the solution of (3.3), is

$$\mu_i = \frac{2\sqrt{\lambda_i \lambda_{p-i+1}}}{\lambda_i + \lambda_{p-i+1}}, \quad i = \{1, \ldots, \lfloor \frac{p}{2} \rfloor\}. \tag{3.7}$$

If p is odd, then $\mu_{\frac{p+1}{2}} = 1$.

In [7], Gustafson gives an interesting motivation for the choice of the name "antieigenvalues". It comes from the fact that as much as eigenvalues are measures of the amount of dilation that a matrix B induces in those directions which are not turned, namely the eigenvectors, then the antieigenvalues are measures of the amount of rotation that a matrix B induces in those directions which are turned the most, namely the antieigenvectors.

In (1.1.3) we have expressed the meaning of the eigenvalues and eigenvectors of a matrix B, as the result of an optimization problem. Antieigenvalues and antieigenvectors are the result of an optimization problem as well: we ask the question what is the p-vector such that the (real) angle of B defined in (3.1) is maximum, or its cosine is minimum. The answer to this question is the first antieigenvector of B, and the minimum value of the cosine of the (real) angle B, namely the cosine of B, is the first antieigenvalue. Then we ask the same question, but in a subspace of $\mathbb{R}^p$ ortoghonal to the space generated by the first antieigenvector: the answer is the second antieigenvector and the minimim value of the new object function is the second antieigenvalue, and so on.

3.2 Chandler Davis' two contributions

In [7] Gustafson reports Chadler Davis' two principal contributions to the theory of the antieigenvalues. The first one is his relation of the antieigenvalues to the theory of the shell of an operator. The second one, perhaps more important, is his emphasis on estimating the antieigenvalues through pairs of eigenvalues.

According to Davis, the shell $s(B)$ of an operator B, defined in a Hilbert space, is the set of all values (ξ, η, h) obtained from all

$$\xi + i\eta = \frac{2\langle B\mathbf{x}, \mathbf{x}\rangle}{\|B\mathbf{x}\|^2 + \|\mathbf{x}\|^2}, \quad h = \frac{\|B\mathbf{x}\|^2 - \|\mathbf{x}\|^2}{\|B\mathbf{x}\|^2 + \|\mathbf{x}\|^2}, \quad \mathbf{x} \neq \mathbf{0}.$$

The set $s(B)$ is a convex subset of the unit ball $\xi^2 + \eta^2 + h^2 \leq 1$. In the finite dimensional case, the eigenvectors $\mathbf{x}$ of B are the intersection of $s(B)$ with the unit sphere $\xi^2 + \eta^2 + h^2 = 1$. Let us now investigate the connection between the shells and the antieigenvalues.

$$1 - h^2 = 1 - \left(\frac{\|B\mathbf{x}\|^2 - \|\mathbf{x}\|^2}{\|B\mathbf{x}\|^2 + \|\mathbf{x}\|^2} \right) = \frac{4\|B\mathbf{x}\|^2 \|\mathbf{x}\|^2}{(\|B\mathbf{x}\|^2 + \|\mathbf{x}\|^2)^2}$$

from which

$$\sqrt{1-h^2} = \frac{2\|B\mathbf{x}\|\|\mathbf{x}\|}{\|B\mathbf{x}\|^2+\|\mathbf{x}\|^2} = (\xi+i\eta)\frac{\|B\mathbf{x}\|\|\mathbf{x}\|}{\langle B\mathbf{x},\mathbf{x}\rangle}.$$

Thus,

$$\mu_1(B) = \inf_{\mathbf{x}\in D(B)} \frac{Re\langle B\mathbf{x},\mathbf{x}\rangle}{\|B\mathbf{x}\|\|\mathbf{x}\|} = \inf_{\mathbf{x}\in D(B)} \frac{Re(\xi+i\eta)}{\sqrt{1-h^2}} = \inf_{\mathbf{x}\in D(B)} \frac{\xi}{\sqrt{1-h^2}}.$$

Definition 12. *Let $m(B)$ be $m(B) = \inf Re\langle B\boldsymbol{x},\boldsymbol{x}\rangle$. Then, B is called **accretive** if $m(B)\geq 0$, **strongly accretive** if $m(B)>0$.*

If we consider finite dimensional strictly accretive operators B, the shell $s(B)$ is a convex polygon. The autoeigenvectors are the vectors that generate the corners of $s(B)$. Hence, upper bounds for the first antieigenvalue μ_1 could be given in terms of the sides of this polygon. For normal operators, the exact value of μ_1 depends on a pair of eigenvalues of B and all total antieigenvectors have only two nonzero components. Those componentes z_i and z_j satisy

$$|z_i|^2 = \frac{|\lambda_j|}{|\lambda_i|+|\lambda_j|}, \qquad |z_j|^2 = \frac{|\lambda_i|}{|\lambda_i|+|\lambda_j|},$$

where λ_i and λ_j are a pair of eigenvalues of B.

3.3 The variational characterization

In this section the Euler equation for the antieigenvalues will be presented.
For simplicity, let B be a strongly accretive bounded operator on a Hilbert space. Our antieigenvalues are of the following form:

$$\mu_k = \inf_H \mu(\mathbf{x}), \qquad \mu(\mathbf{x}) = Re\frac{\langle B\mathbf{x},\mathbf{x}\rangle}{\|B\mathbf{x}\|\|\mathbf{x}\|},$$

where $H = \mathcal{L}^{\perp}\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_{k-1}\}$, subset of the Hilbert space. This is an optimization problem: we have to minimize an object function within a certain space. To find the Euler equation, we consider the following differential quantity.

$$\left.\frac{\mathrm{d}\mu}{\mathrm{d}\mathbf{y}}\right|_{\epsilon=0} = \lim_{\epsilon\to 0}\frac{1}{\epsilon}\left(\frac{Re\langle B(\mathbf{x}+\epsilon\mathbf{y},\mathbf{x}+\epsilon\mathbf{y}\rangle)}{\langle B(\mathbf{x}+\epsilon\mathbf{y}),B(\mathbf{x}+\epsilon\mathbf{y})\rangle^{1/2}\langle\mathbf{x}+\epsilon\mathbf{y},\mathbf{x}+\epsilon\mathbf{y}\rangle^{1/2}} - \right.$$
$$\left. - \frac{Re\langle B\mathbf{x},\mathbf{x}\rangle}{\langle B\mathbf{x},B\mathbf{x}\rangle^{1/2}\langle\mathbf{x},\mathbf{x}\rangle^{1/2}}\right).$$

Following the steps in [7], we lead to the following Euler equation:

$$2\|B\mathbf{x}\|^2(ReB)\mathbf{x} - Re\langle B\mathbf{x},\mathbf{x}\rangle\left[B^*B+\|B\mathbf{x}\|^2\right]\mathbf{x} = 0, \qquad \|\mathbf{x}\| = 1.$$

In a "normalized" form, it may be written as

$$\frac{2(ReB)\mathbf{x}}{Re\langle B\mathbf{x},\mathbf{x}\rangle} - \frac{B^*B\mathbf{x}}{\|B\mathbf{x}\|^2} - \mathbf{x} = 0.$$

Definition 13. *An operator B is said to be* ***self-adjoint*** *if $\langle A\boldsymbol{x},\boldsymbol{y}\rangle = \langle \boldsymbol{x}, A\boldsymbol{y}\rangle$, $\forall \boldsymbol{x},\boldsymbol{y} \in X$.*

If B is a self-adjoint operator, the Euler equation becomes

$$\frac{B^2\mathbf{x}}{\langle B^2\mathbf{x},\mathbf{x}\rangle} - \frac{2B\mathbf{x}}{\langle B\mathbf{x},\mathbf{x}\rangle} + \mathbf{x} = 0.$$

Thus, we can conclude this section with the next theorem.

Theorem 8. *The Euler equation for the antieigenvalue functional*

$$\mu(\mathbf{x}) = \frac{Re\langle B\mathbf{x},\mathbf{x}\rangle}{\|B\mathbf{x}\|\|\mathbf{x}\|}$$

is

$$2\|B\mathbf{x}\|^2\|\mathbf{x}\|^2(ReB)\mathbf{x} - \|\mathbf{x}\|^2 Re\langle B\mathbf{x},\mathbf{x}\rangle B^*B\mathbf{x} - \|B\mathbf{x}\|^2 Re\langle B\mathbf{x},\mathbf{x}\rangle\mathbf{x} = 0.$$

3.4 Antieigenvalues and numerical analysis

In optimization theory, the method of the steepest descent is one of the basic methods, used to find the minimum of a given function, by performing the following basic algorithm.

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \left[\nabla f(\mathbf{x}_k)\right]^T.$$

Let us focus on the quadratic case, where the function f can be expressed with the following form.

$$f(\mathbf{x}) = \frac{\langle \mathbf{x}, A\mathbf{x}\rangle}{2} - \langle \mathbf{x},\mathbf{b}\rangle,$$

where A is a $p \times p$ symmetric positive definite matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$.

In this case, the point of minimum $\mathbf{x}^*$ solves the linear system

$$A\mathbf{x}^* = \mathbf{b}.$$

The algorithm for steepest descent is an iterative method, which is important for numerically solving large linear systems. In the case of quadratic minimization, this algorithm is the following.

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{\|\mathbf{y}_k\|^2 \mathbf{y}_k}{\langle A\mathbf{y}_k, \mathbf{y}_k \rangle},$$

where $\mathbf{y}_k = A\mathbf{x}_k - \mathbf{b}$ is called the *residual error*. The error $E_A(\mathbf{x})$ that we commit at each iteration k, whatever is the initial point $\mathbf{x}_0$, is the following.

$$\begin{aligned} E_A(\mathbf{x}_k) = f(\mathbf{x}_k) - f(\mathbf{x}^*) = f(\mathbf{x}_k) - \frac{\langle \mathbf{x}^*, A\mathbf{x}^* \rangle}{2} + \langle \mathbf{x}^*, \mathbf{b} \rangle = \\ = f(\mathbf{x}_k) - \frac{\langle \mathbf{x}^*, A\mathbf{x}^* \rangle}{2} + \langle \mathbf{x}^*, A\mathbf{x}^* \rangle = f(\mathbf{x}_k) + \frac{\langle \mathbf{x}^*, A\mathbf{x}^* \rangle}{2}. \end{aligned}$$

At each iteration, we have the following error bound:

$$E_A(\mathbf{x}_{k+1}) \leq \left(1 - \frac{4\lambda_1 \lambda_p}{(\lambda_1 + \lambda_p)^2}\right) E_A(\mathbf{x}_k). \tag{3.8}$$

In the right side of the inequality, the square of the first antieigenvalue of A appears. Hence, we can write

$$\begin{aligned} E_A(\mathbf{x}_{k+1}) \leq \left[1 - \mu_1^2(A)\right] E_A(\mathbf{x}_k) = \\ = (1 - cos^2(A)) E_A(\mathbf{x}_k) = \\ = (sin^2(A)) E_A(\mathbf{x}_k). \end{aligned}$$

This has an interesting geometrical interpretation. We know that the first antieigenvalue measures the maximum turning capability of A. The method of the steepest descent works fine, but the velocity of its convergence depends on the maximum turning capability of A, i.e., the maximal distance from a vector $\mathbf{x} \in \mathbb{R}^p$ to $A\mathbf{x}$, which is measured by $sin(A)$. The closer to zero $sin(A)$ is, i.e., the closer to 1 $\mu_1(A)$ is, and the faster the convergence of the steepest descent is.

This is an example in numerical analysis where the first antieigenvalue of an operator comes into play. We refer to [8] to see other applications to numerical analysis, wavelets, statistics, quantum mechanics, finance and optimization, analyzed from 1966 to 2010; and to [9] for new applications to continuum mechanics, economics, and number theory.

3.5 Random matrix

So far we have dealt with antieigenvalues within a context of fixed matrices. Let us now analyze the distribution of the antieigenvalues, in particular the first one,

in the same conditions we set while studying the distribution of the eigenvalues, so in a context of random matrices.

Why? A possible reason is explained in the next subsection.

3.5.1 Growth Curve Model

In this subsection The Growth Curve Model will be presented. It was introduced in 1964 by Potthoff and Roy ([13]), and it is defined as follows.

Definition *Let $X \in \mathbb{R}^{p\times N}$ and $\xi \in \mathbb{R}^{q\times m}$ be the observation and parameter matrices, respectively; and let $B \in \mathbb{R}^{p\times q}$ and $A \in \mathbb{R}^{m\times N}$ be the within and between individual design matrices, respectively. Suppose that $q \leq p$ and $p \leq N - r = n$, where $r = rank(A)$. Then, the Growth Curve model is given by*

$$X = B\xi A + \epsilon, \tag{3.9}$$

where the columns of ϵ are assumed to be independently p-variate normally distributed with mean zero and unknown positive definite covariance matrix Σ, i.e.,

$$\epsilon \sim N_{p,N}(\mathbf{0}, \Sigma, \mathbb{1}_N).$$

Here, X is thought to be a fixed matrix, representing p repeated measurements over time for N individuals; while m is the number of groups which the N individuals are divided in, and $q-1$ is the degree of the polynomial through which we want to express the m growth curves.

Similarly to what the remark in (1.1.1) says, let $P_{A^T} = A^T(AA^T)^{-1}A$ be the projection on $\mathcal{L}(A^T)$, the space generated by A^T, and let $H = \mathbb{1}_N - P_{A^T}$ be the projection on $\mathcal{L}^{\perp}(A^T)$. Then we can define $V = (HX)^T(HX) = X^THX = X^THX$, N times the sample "covariance" matrix of X.
If B and A have full rank, the Maximum Likelihood Estimator for $\boldsymbol{\xi}$ is

$$\hat{\boldsymbol{\xi}}_{MLE} = (B^TV^{-1}B)^{-1}B^TV^{-1}XA^T(AA^T)^{-1}. \tag{3.10}$$

Mean and covariance for $\boldsymbol{\xi}_{MLE}$ are given in Kollo and von Rosen (2005, [10]) as

$$\begin{aligned}\mathbb{E}(\hat{\boldsymbol{\xi}}_{MLE}) &= \boldsymbol{\xi},\\ cov(\hat{\boldsymbol{\xi}}_{MLE}) &= \frac{n-1}{n-1-(p-q)}(AA^T)^{-1} \otimes (B^T\Sigma^{-1}B)^{-1},\end{aligned}$$

if $n-1-(p-q) > 0$, where $n = N - m$.

As we are interested in the case where $q \leq p \leq n$, then we have $\frac{n-1}{n-1-(p-q)} \geq 1$.

Let us now think X to be a random matrix, to discuss about to efficiency of $\hat{\boldsymbol{\xi}}_{MLE}$ when the number of individuals N is fixed, but the number of repeated measures p grows over time. This implies that the matrix V becomes closer to singularity. Thus, the variablity of $\hat{\boldsymbol{\xi}}_{MLE}$ grows and it becomes unstable.

In [14], an unweighted estimator of $\boldsymbol{\xi}$ is proposed, by considering the identity matrix instead of V, as follows.

$$\hat{\boldsymbol{\xi}} = (B^T B)^{-1} B^T X A^T (AA^T)^{-1}. \tag{3.11}$$

This estimator is simpler than the MLE, since we do not need to calculate the inverse of V. The distribution of this unweighted estimator is given by

$$\hat{\boldsymbol{\xi}} \sim N_{q,m}(\boldsymbol{\xi}, (B^T B)^{-1} B^T \Sigma B (B^T B)^{-1}, (AA^T)^{-1}),$$

i.e., we have

$$\begin{aligned} \mathbb{E}(\hat{\boldsymbol{\xi}}) &= \boldsymbol{\xi}, \\ cov(\hat{\boldsymbol{\xi}}) &= (AA^T)^{-1} \otimes (B^T B)^{-1} B^T \Sigma B (B^T B)^{-1}. \end{aligned}$$

After some simulations to compare the efficiency of the two estimators when p grows, it is possible to show that the MLE becomes more and more unstable, while the unweighted estimator is always close to the reality. Hence, a comparison between the two estimators seems reasonable to be done.

Both $\hat{\boldsymbol{\xi}}_{MLE}$ and $\hat{\boldsymbol{\xi}}$ are unbiased. Their covariances are given by

$$\begin{aligned} cov(\hat{\boldsymbol{\xi}}_{MLE}) &= \frac{n-1}{n-1-(p-q)} (AA^T)^{-1} \otimes (B^T \Sigma^{-1} B)^{-1}, \\ cov(\hat{\boldsymbol{\xi}}) &= (AA^T)^{-1} \otimes (B^T B)^{-1} B^T \Sigma B (B^T B)^{-1}. \end{aligned}$$

Hence, we want to compare $(B^T \Sigma^{-1} B)^{-1}$ and $(B^T B)^{-1} B^T \Sigma B (B^T B)^{-1}$.

It is possible to show that

$$(B^T \Sigma^{-1} B)^{-1} \leq (B^T B)^{-1} B^T \Sigma B (B^T B)^{-1}, \tag{3.12}$$

where the inequality is with respect to the Lowener partial ordening, i.e., $A \leq B$ if $B - A$ is positive semidefinite.
For large n, we can notice that the unweighted estimator of $\boldsymbol{\xi}$ has a larger covariance than the weighted one, as expected since this is the MLE.

When p is also large, though, and $q \leq p \leq n$ still holds, then

$$1 \ll \frac{n-1}{n-1-(p-q)},$$

which implies that

$$cov(\hat{\boldsymbol{\xi}}) < cov(\hat{\boldsymbol{\xi}}_{MLE}).$$

It is possible to show that the inequality (3.12) is actually an equality if and only if $\mathcal{L}(\Sigma^{-1}B) = \mathcal{L}(B)$, which happens for instance if $\Sigma = \sigma^2\mathbb{1}_p$ (sphericity assumption).
In this context, it is possible to show that

$$(B^TB)^{-1}B^T\Sigma B(B^TB)^{-1} \leq \frac{(\lambda_1+\lambda_p)^2}{4\lambda_1\lambda_p}(B^T\Sigma^{-1}B)^{-1}, \tag{3.13}$$

where λ_1 and λ_p are the largest and smallest eigenvalue of Σ.
Let us see again the fraction that appears in (3.13): this is $\frac{1}{\mu_1^2}$, where μ_1 is the first antieigenvalue of Σ!

Hence, if

$$\frac{1}{\mu_1^2} = \frac{(\lambda_1+\lambda_p)^2}{4\lambda_1\lambda_p} \leq \frac{n-1}{n-1-(p-q)}, \tag{3.14}$$

then, the unweighted estimator has *smaller* covariance matrix then the MLE. In other words, if the first antieigenvalue μ_1 of Σ, i.e., the maximum turning capability of Σ, is greater than a certain threshold, then the unweighted estimator is a better (more stable) estimator of $\boldsymbol{\xi}$ than the MLE.

In the next section, we are going to investigate the distribution of μ_1.

3.6 Simulations in R

The inequality (3.14), leads us to study the distribution of μ_1, the first antieigenvalue of the sample covariance matrix $\mathcal{S}$ of a random matrix $\mathcal{X} \in \mathbb{R}^{n\times p}$, whose np entries are i.i.d. $N(0,\sigma^2)$, as $\mathcal{S}$ is the estimator of $\Sigma = \sigma^2\mathbb{1}_p$, exactly as we did for the eigenvalues in Chapter 2.
We do not know any theoretical results about the distribution of μ_1. Hence, we want to perform some simulations in R and discuss the results. Without loss of generality we assume $\sigma^2 = 1$.

In analogy with the eigenvalues, we can plot histograms approximating the distributions of the antieigenvalues when $\gamma = n/p$ varies. This time, though, we are interested only in $\gamma > 1$, because our attention is mainly focused on the first antieigenvalue, which would be equal to zero in case $\gamma < 1$.

In Figure (3.1) the histograms of the antieigenvalues of $\mathcal{S}$ are shown, when $p = 1000$ and respectively $n = \{3000, 8000, 14000\}$. In analogy with the Marčenko-Pastur density functions for $\gamma < 1$, it seems that it could be possible to find a function, depending on γ that fits these histograms, with a mass

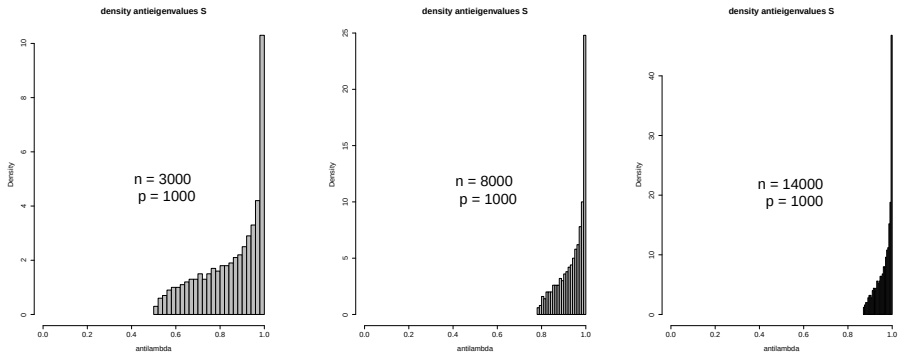


Figure 3.1: distribution of antieigenvalues of $\mathcal{S}$ for $\gamma = \{3; 8; 14\}$.

point in $x = 1$. A further research for this thesis could be trying to find the expression of such a function.

As already noticed for the eigenvalues, we know that $\mathcal{S}$ converges to $\Sigma = \mathbb{1}_p$, whose eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ are equal to one. By looking at the formulas for the antieiganvalues, this implies that all the antieigenvalues of Σ are equal to one. Indeed, the higher is γ and we can see that the more the antieigenvalues distribute next to the value $x = 1$.

3.6.1 Convergence of the first antieigenvalue

In the previous section, we have dealt with the limit distribution of the largest and of the smallest eigenvalues of $\mathcal{S}$, confirming the theoretical results by simulations.

By exploiting the results (2.3) and (2.4), i.e., $\lambda_1^{(p)} \to b$ and $\lambda_p^{(p)} \to a$ ($\gamma > 1 \Rightarrow p < n \Rightarrow min\{n, p\} = p$), and reminding that $\mu_1 = \frac{2\sqrt{\lambda_1 \lambda_p}}{\lambda_1 + \lambda_p}$, we expect that

$$\mu_1^{(p)} \to \frac{2\sqrt{ab}}{a+b} = \frac{\gamma - 1}{\gamma + 1}. \tag{3.15}$$

This is an important result, which says to us that when n and p go to infinite, and $\frac{n}{p}$ tends to a positive constant γ, if γ is almost one, i.e., $p \lessapprox n$, then $\mu_1^{(p)}$ tends to zero; if γ, instead, is really high, i.e., $p \ll n$, then $\mu_1^{(p)}$ tends to one.

Let us verify by simulations this result.

In Figure (3.2), the convergence of the first antieigenvalue of $\mathcal{S}$ is analyzed. In particular, it is studied the case $\gamma = \frac{n}{p} = 3$. This means that the limit value in (3.15) is equal to 0.5.

The simulation consists of 70 iterations and it is the same as the one we have

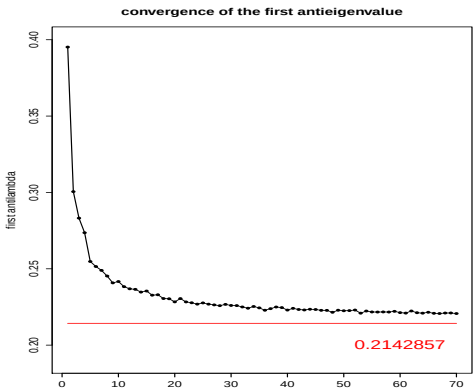


Figure 3.2: Convergence of the first antieigenvalue of $\mathcal{S}$.

seen to study the convergence of the largest and of the smallest eigenvalues, in the previous chapter. This time, though, we are interested in the first antieigenvalue.

We can see that $\mu_1^{(p)}$ seems to converge to the value 0.5.

3.6.2 The first antieigenvalue and the Beta distribution

In this subsection we are going to analyze the distribution of the first antieigenvalue. In Figure (3.3), the distribution of 1000 sampled μ_1's is shown, when $n = 100$ and respectively $p = \{5, 50, 80\}$.

By computing the same distributions for different values of p, it can be noticed

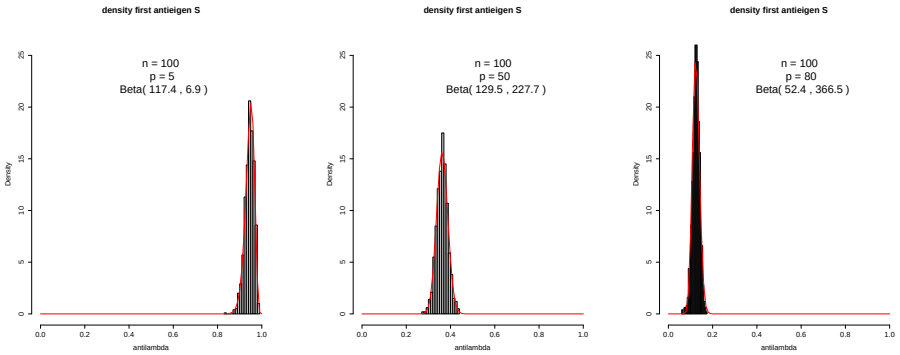


Figure 3.3: distribution of the first antieigenvalue of $\mathcal{S}$ for $n = 100$ and $p = \{5, 50, 80\}$ respectively.

that $\mu_1^{(p)}$ seems to follow a Beta distribution, whose parameters (α, β) depend on the values of p. In the three pictures, the histogram of the $\mu_1^{(p)}$'s is shown, but also the density function of a $Beta(\alpha^{(p)}, \beta^{(p)})$, where $\alpha^{(p)}$ and $\beta^{(p)}$ are calculated through the method of the moments.

It seems that when $p \ll n$, then $\alpha^{(p)} \gg \beta^{(p)}$; while when $p \lessapprox n$, then $\alpha^{(p)} \ll$

$\beta^{(p)}$. In the next subsection we are going to investigate the existence of a possible law that describes the relation between $\alpha^{(p)}$ and $\beta^{(p)}$.

3.6.3 The elliptical law

In Figure (3.4), it is presented the result of the following algorithm.

1. Set $n = 100$ and $p = \{2, 4, \ldots, 98\}$.

2. For each value of p, do 200 times the following procedure: sample 500 times the first antieigenvalue μ_1 of the sample covariance matrix $\mathcal{S}$ of the random matrix $\mathcal{X} \in \mathbb{R}^{n \times p}$, whose np entries are i.i.d. $N(0, 1)$, and estimate the parameters $\alpha^{(p)}$ and $\beta^{(p)}$ through the method of the moments.

3. Now, for each value of p, we have 200 values of the estimated $\alpha^{(p)}$ and $\beta^{(p)}$. Compute the mean, the empirical quantile .025 and the empirical quantile .975 of both of them.

4. Draw in the space (α, β), the mean values in red colour, the .025 quantile in blue colour, and the 0.975 quantile in purple colour, so that it is possible to see the 95% empirical interval of confidence for α and β for each p.

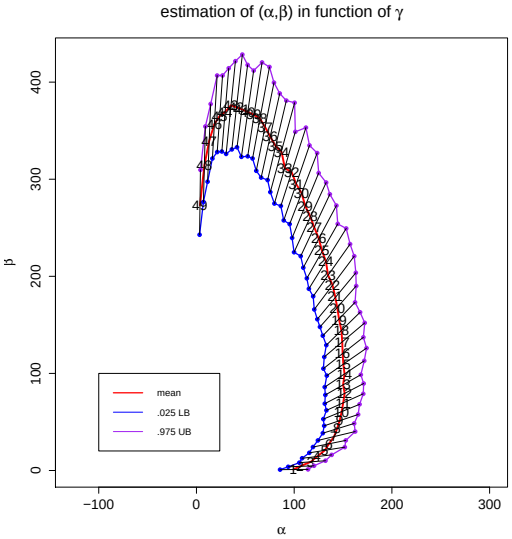


Figure 3.4: Estimation of $\alpha^{(p)}$ and $\beta^{(p)}$ thorugh the method of the moments, for $n = 100$ and $p = \{2, 4, \ldots, 98\}$.

The indexes $\{1, 2, \ldots, 49\}$ are the 49 iterations related to $\{p^{(1)} = 2, p^{(2)} = 4, \ldots, p^{(49)} = 98\}$.

It is interesting to notice that mean of the estimations of the points $(\alpha^{(p)}, \beta^{(p)})$ seems to follow an elliptical curve. Hence, let us try to express an elliptical formula that best fits the red line.
A parametric formula in the variable t to express an ellipse, in a xOy referment system, is

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} acos(t) \\ bsin(t) \end{pmatrix}, \qquad t \in [0, 2\pi],$$

where a is the length of the semi x-axis, while b is the length of the semi y-axis. If an ellipse is translated, i.e., its center is not in $O(0,0)$, but in $C(x_c, y_c)$, and if it is rotated, i.e., the angle between the referement system that is parallel to its axis and the original xOy referement system is equal to α, then the parametric equation for such an ellipse is the following.

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix} = \begin{pmatrix} x_c \\ y_c \end{pmatrix} + \begin{pmatrix} cos(\alpha) & -sin(\alpha \\ sin(\alpha) & cos(\alpha) \end{pmatrix} \begin{pmatrix} acos(t) \\ bsin(t) \end{pmatrix}, \qquad t \in [0, 2\pi]. \tag{3.16}$$

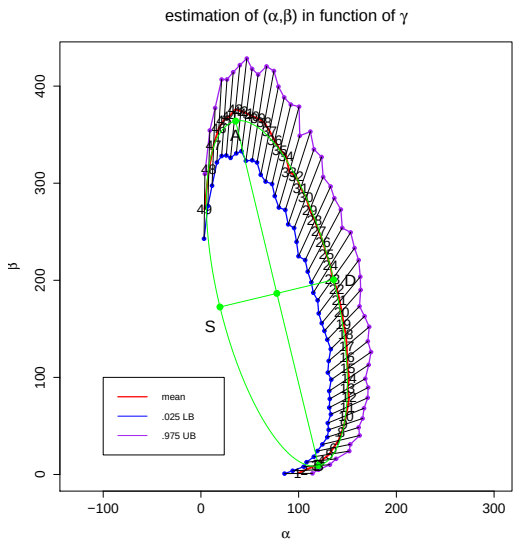


Figure 3.5: Estimation of $\alpha^{(p)}$ and $\beta^{(p)}$ thorugh the method of the moments, for $n = 100$ and $p = \{2, 4, \ldots, 98\}$. Parametric equation for the green ellipse.

In order to find our desired ellipse, we need x_c, y_c, α, a and b. In Figure (3.5), it is shown in green colour, the elliptical parametric equation estimated by following the next steps.

1. Set the values x_c and y_c, by looking at the graphics and figuring out where the real (x_c, y_c) could be approximately.

2. Let us now think to be in a referement system, whose center is $C(x_c, y_c)$, and rotated, such to see the ellipse parallel to the new axis: let us call

D,A,S,B the four points of intersection between the ellipse and our new axis, respectively at 0, $\pi/2$, π, and $3\pi/2$ degrees. Set the values x_d and y_d, x_a and y_a, by looking at the graphics and figuring out where the real D and A could be approximately.

3. Calculate the angle between the segments $\overline{CD}$ and $\overline{CA}$. If it is not very close to $\pi/2$, then repeat step 1,2.
4. Let us call P, the point $(x_c + 1, y_c)$. Calculate the angle between the segments $\overline{CD}$ and $\overline{CP}$: this is α.
5. Compute $a = \sqrt{(x_d - x_c)^2 + (y_d - y_c)^2}$ and $b = \sqrt{(x_a - x_c)^2 + (y_a - y_c)^2}$. Draw the ellipse by following the equation (3.16).
6. Repeat steps 1, 2, 3, 4, 5 by changing each value, in order to find an ellipse that better fits the red line.

The following values are the ones I got to draw the green ellipse in Figure (3.5):

$$\begin{aligned} x_c &= 77.7 \\ y_c &= 186.5 \\ x_d &= 135.9 \\ y_d &= 200.5 \\ x_a &= 35.5 \\ y_a &= 364 \\ angle(\overline{CA}, \overline{CD}) &= 89,84807 \\ \alpha &= 13,52552 \\ a &= 59,86017 \\ b &= 182.4475 \end{aligned}$$

Now, we have the parametric equation in t, that approximates pretty well the red line. Though, we are not interested in an equation in t. We are interested in having the equation of this ellipse in function of p. Hence, let us try to find a function $t = t(p)$.
Our idea is to associate to each value of $\overline{p}$, that value of $\overline{t} \in [0, 2\pi]$ such that the distance between the red point associated to $\overline{p}$ and the green point associated to $\overline{t}$ is the minimum one. The next table shows the results of this idea, where the first column represents the values of p, the second column represents the values of $t(p)$ and the third column represents the associated minimum Euclidean distance.

p	$t(p)$	dist	p	$t(p)$	dist	p	$t(p)$	dist	p	$t(p)$	dist
2	4.45	134	16	5.26	0.69	30	5.76	0.97	44	6.23	1.22
4	4.51	34.7	18	5.34	0.10	32	5.82	1.00	46	0.01	1.71
6	4.66	0.16	20	5.42	0.68	34	5.89	2.41	48	0.08	0.36
8	4.80	14.3	22	5.49	1.60	36	5.96	4.10	50	0.15	0.37
10	4.97	11.1	24	5.56	4.33	38	6.03	1.80	52	0.22	0.03
12	5.08	10.1	26	5.62	0.07	40	6.09	2.43	54	0.29	0.53
14	5.18	1.74	28	5.69	3.36	42	6.16	2.09	56	0.36	0.73

p	$t(p)$	dist	p	$t(p)$	dist	p	$t(p)$	dist
58	0.43	2.61	72	0.99	4.22	86	1.54	143.4
60	0.51	0.20	74	1.09	23.28	88	1.63	45.9
62	0.58	0.96	76	1.19	64.00	90	1.71	37.5
64	0.66	0.58	78	1.26	60.5	92	1.83	16.2
66	0.71	14.06	80	1.34	55.4	94	2.03	0.42
68	0.83	4.22	82	1.41	68.5	96	2.26	0.41
70	0.90	0.20	84	1.48	94.72	98	2.55	4.33

If the minimum distance is high, it means that the corresponding green and red points are far away from each other: for that value of p, the green ellipse is not a very good approximation of the red curve. The idea is to draw the green ellipse in such a way that the values of p in the middle are well fitted, without taking care so much of the values of p close to the border $\{0, 100\}$.

Now, the range of values $t \in [0, 2\pi]$ is just arbitrary: we could start from any point $\bar{t}$ and finish in $\bar{t} + 2\pi$. In our case we can start in $t(p = 2) - 2\pi = 4.45 - 2\pi = -1.84$ and finish in $t(p = 2) = 4.45$. Actually, we are not interested in all the green ellipse, but only in the part of it that fits the red line. Hence, we can stop at $t(p = 98) = 2.55$.

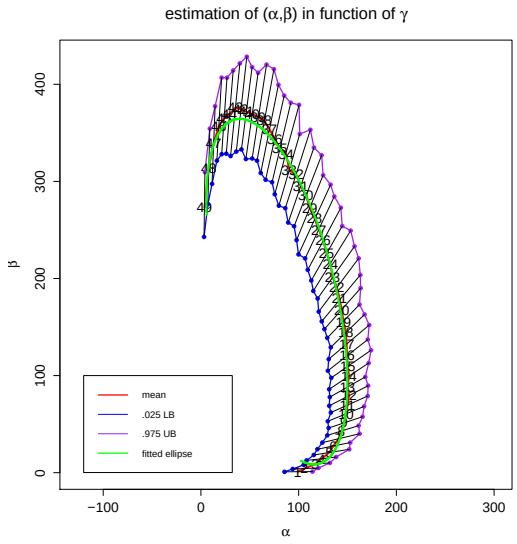


Figure 3.6: Estimation of $\alpha^{(p)}$ and $\beta^{(p)}$ thorugh the method of the moments, for $n = 100$ and $p = \{2, 4, \ldots, 98\}$. Parametric equation for the green ellipse.

In Figure (3.6) our green ellipse is shown for $t \in [-1.85, 2.60]$.

In Figure (3.7), the 49 points $(p, t(p))$ are shown. The aim is to find an explicit function for $t = t(p)$.

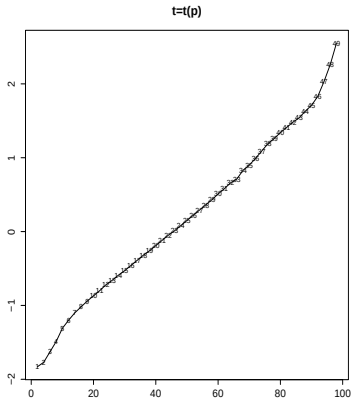


Figure 3.7: paramater t in function of p.

We can notice that, if we are not interested in border values of p, a good approximation of $t = t(p)$ could be a simple line. In general, a tangent curve seems to be the best option. Let us try with both of these ideas.

Linear method

In order to find the two parameters that define our desired line, we can exploit the command *lm()* in R, to perform a linear regression by considering only the values $\{p^{(7)} = 14, p^{(8)} = 16, \ldots, p^{(44)} = 88\}$. The output parameters are $\beta_0 = -1.634893$, $\beta_1 = 0.03651494$. Hence, our desired line could be $t = t(p) = \beta_1 p + \beta_0$, which is shown in figure (3.8).

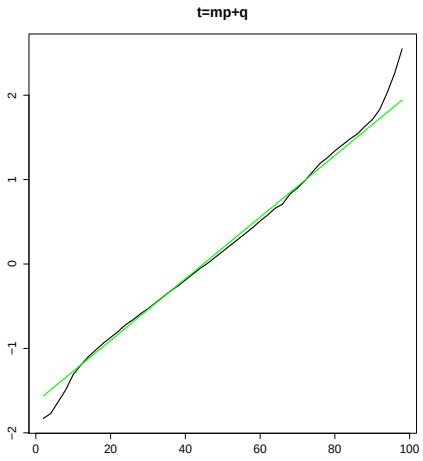


Figure 3.8: paramater t in function of p.

We have found a rule to estimate $(\alpha^{(p)}, \beta^{(p)})$, $\forall p \in (0, 100)$.

$$\begin{pmatrix} \hat{\alpha}(p) \\ \hat{\beta}(p) \end{pmatrix} = \begin{pmatrix} x_c \\ y_c \end{pmatrix} + \begin{pmatrix} cos(\alpha) & -sin(\alpha \\ sin(\alpha) & cos(\alpha) \end{pmatrix} \begin{pmatrix} a \cdot cos(mp + q) \\ b \cdot sin(mp + q) \end{pmatrix}, \tag{3.17}$$

where $x_c = 77.7$, $y_c = 186.5$, $\alpha = 13.52552°$, $a = 59.86017$, $b = 182.4475$, $m = 0.03651494$, $q = -1.634893$.

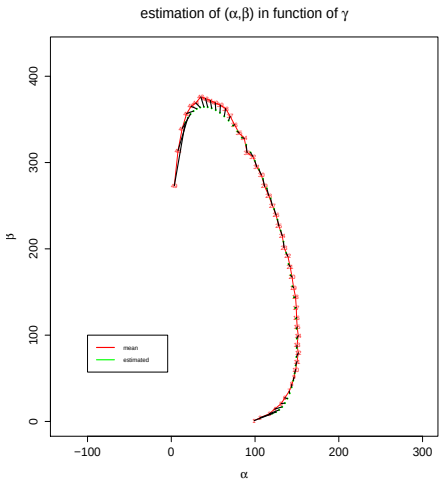


Figure 3.9: Distance between the estimated values of $(\alpha^{(p)}, \beta^{(p)})$ and the "real" ones. Linear method.

In order to estimate the efficiency of this rule, let us estimate $(\alpha^{(p)}, \beta^{(p)})$, for $p = \{2, 4, \ldots, 98\}$ and compare them to the "real" ones. In Figure (3.9), we can see the distance between the red points, corresponding to the "real" values, and the green points, corresponding to the estimated ones through the linear method. We can notice that for intermediate values of p, the estimation seems to be pretty good, while for values of p close to the border the estimation seems to be quite rough.

Tangent method

Let us try to approximate $t = t(p)$ with a tangent curve depending on three parameters: $t = t(p) = k_1 tan[k_2(p - k_3)]$. After some attempts we found that $k_1 = 1.62, k_2 = 0.0195$ and $k_3 = 45$ seem to be good values.

In Figure (3.10), such a curve is shown.

We have found a rule to estimate $(\alpha^{(p)}, \beta^{(p)})$, $\forall p \in (0, 100)$.

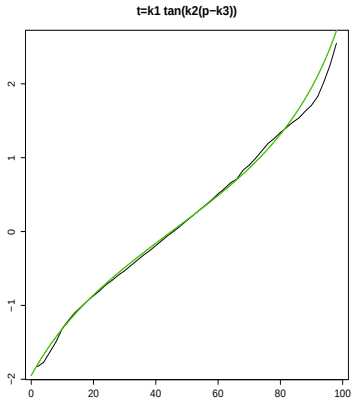


Figure 3.10: paramater t in function of p.

$$\begin{pmatrix}\hat{\alpha}(p)\\ \hat{\beta}(p)\end{pmatrix} = \begin{pmatrix}x_c\\ y_c\end{pmatrix} + \begin{pmatrix}cos(\alpha) & -sin(\alpha)\\ sin(\alpha) & cos(\alpha)\end{pmatrix} \begin{pmatrix}a \cdot cos(k_1 tan[k_2(p-k_3)])\\ b \cdot sin(k_1 tan[k_2(p-k_3)])\end{pmatrix}, \quad (3.18)$$

where $k_1 = 1.62$, $k_2 = 0.0195$, $k_3 = 45$.

In order to estimate the efficiency of this rule, let us estimate $(\alpha^{(p)}, \beta^{(p)})$, for $p = \{2, 4, \ldots, 98\}$ and compare them to the "real" ones.

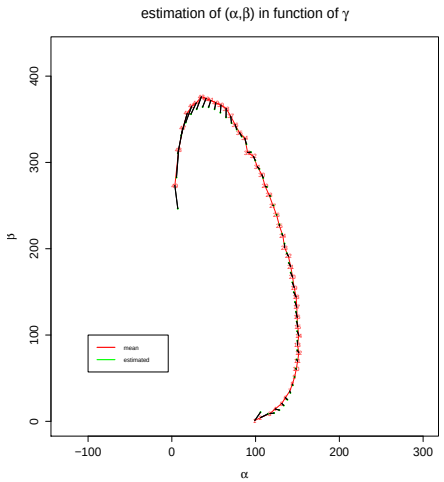


Figure 3.11: Distance between the estimated values of $(\alpha^{(p)}, \beta^{(p)})$ and the "real" ones.

In Figure (3.11), we can see the distance between the red points, corresponding to the "real" values, and the green points, corresponding to the estimated ones

through the tangent method. We can notice that for intermediate values of p, the estimation seems to be similar to the linear method, while for values of p close to the border the estimation seems to be much better.

Efficiency of the Elliptical Law

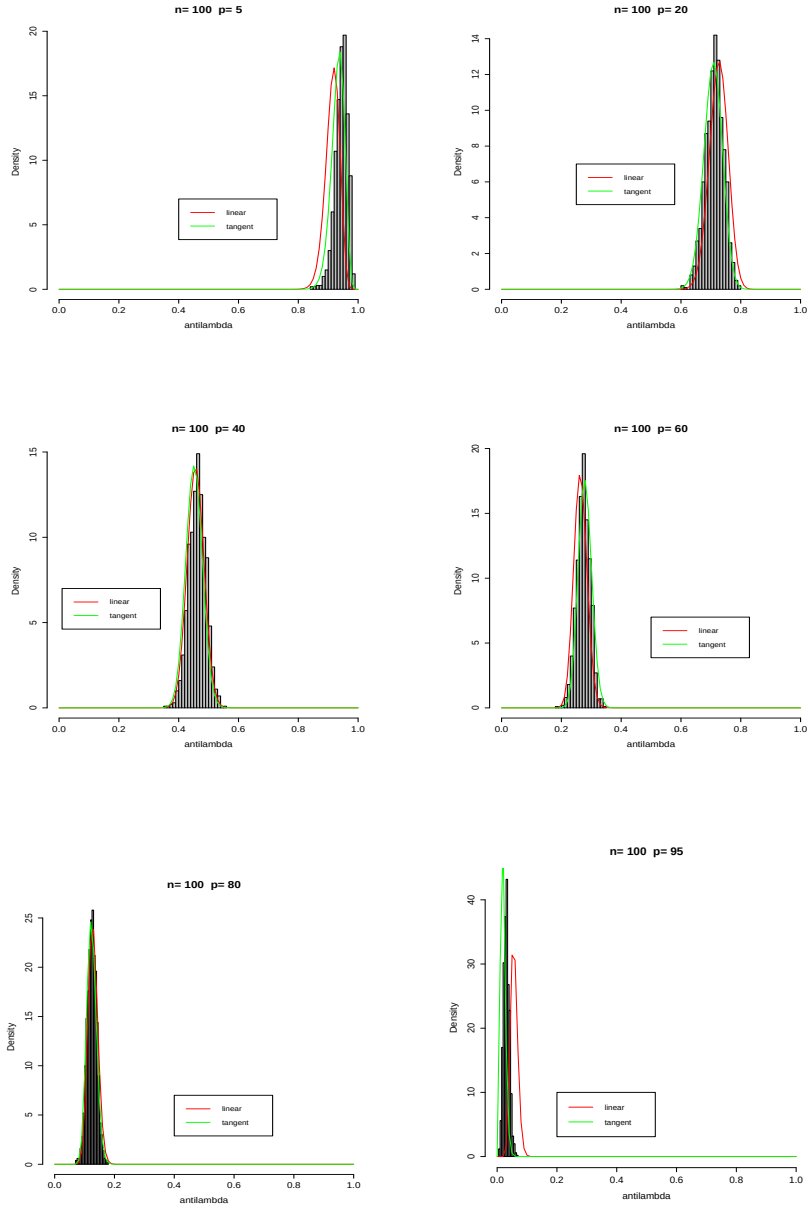


Figure 3.12: Estimation of $Beta(\alpha, \beta)$ through linear method and tangent method. Comparison to reality.

It is now time to verify the goodness of fit of the elliptical law, for both the linear method and the tangent one. Let us fix $n = 100$ and sample 1000 times μ_1, for different values of p. In Figure (3.12), it is shown the histogram of the 1000 samples, the beta density function in red with parameters estimated through the linear method, and the green one through the tangent method, for some values of p.

We can notice that when p assumes intermediate values, then the estimations from both of the methods seem to be pretty good. When p assumes values close to the border, the linear method seems to be not so close to the reality, while the tangent method seems to be still good.

We have seen that the elliptical law estimates in a good way the parameters α and β of the Beta distribution where μ_1 seems to come from, when n is fixed and equal to 100. A natural question now is whether this elliptical law holds $\forall\gamma = \frac{n}{p}$, or not.

Let us try to sample 1000 times μ_1, from the sample covariance matrix $\mathcal{S}$ of a random matrix $X \in \mathbb{R}^{n\times p}$, whose np entries are i.i.d. $N(0,1)$, for instance when $n = 200$ and $p = 80$, or when $n = 500$ and $p = 200$, by using the estimated α and β for $n = 100$ and $p = 40$, i.e., for $\gamma = 2.5$.

We can notice from Figure (3.13) that the answer to our question is nega-

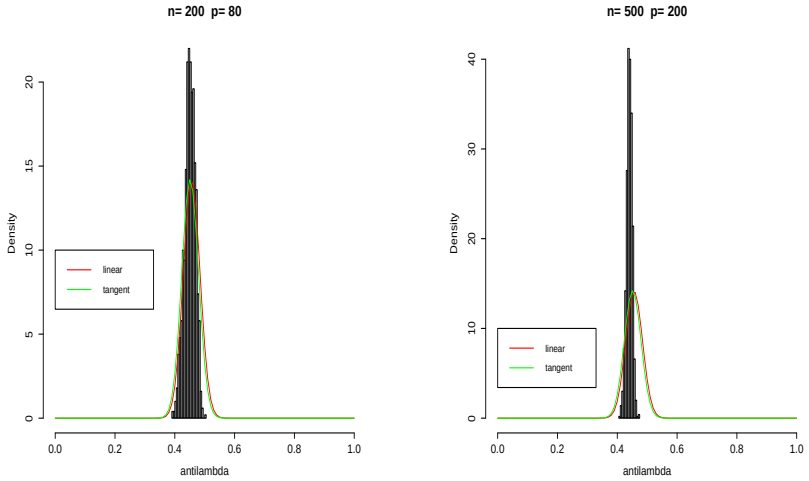


Figure 3.13: Estimation of $Beta(\alpha, \beta)$ when $n = 500$, $p = 200$. Comparison to reality.

tive. The elliptical law does not depend on only $\gamma = \frac{n}{p}$, but on both n and p. It seems that, if we think γ as a fixed value, the higher is n (and so $p = \frac{n}{\gamma}$), and the tighter is the μ_1 distribution (but still around the same values, for each γ). Thus, it could be the case that the mean value is the same for each n, but the variability decreases when n increases, which could mean that, for each $\gamma = \frac{n}{p}$, the parameters $\alpha^{(n+k)}$ and $\beta^{(n+k)}$, for some $k \in \mathbb{N}$, are proportional to $\alpha^{(n)}$ and

$\beta^{(n)}$, up to a constant $c_k > 1$.
Let us now try to do the same for $n = 200$ to answer to this question.

n = 200

Let us set $n = 200$, and $p = \{4, 8, \ldots, 196\}$. In Figure (3.14) it is shown in red colour the mean of the 100 (instead of 200) estimated parameters $\alpha^{(p)}$ and $\beta^{(p)}$ for each of the 49 values of p.

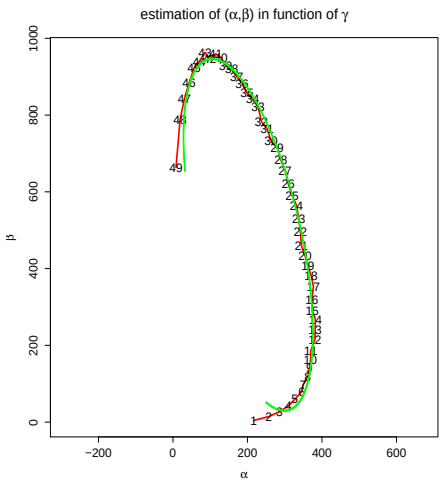


Figure 3.14: Estimation of $\alpha^{(p)}$ and $\beta^{(p)}$ thorugh the method of the moments, for $n = 200$ and $p = \{4, 8, \ldots, 196\}$. Parametric equation for the green ellipse.

In green colour it is shown the parametric equation for an ellipse drawn by following the same algorithm explained above, by attempts, and considering the range of values $t \in [-1.95, 2.70]$.

The following values represent the ratios between the 49 values of the estimated parameters α and β, in the case $n = 200$ and in the case $n = 100$:

p	$\frac{\alpha_{200}}{\alpha_{100}}$	$\frac{\beta_{200}}{\beta_{100}}$	p	$\frac{\alpha_{200}}{\alpha_{100}}$	$\frac{\beta_{200}}{\beta_{100}}$	p	$\frac{\alpha_{200}}{\alpha_{100}}$	$\frac{\beta_{200}}{\beta_{100}}$	p	$\frac{\alpha_{200}}{\alpha_{100}}$	$\frac{\beta_{200}}{\beta_{100}}$
2	2.19	4.30	16	2.52	2.77	30	2.49	2.65	44	2.46	2.59
4	2.43	3.33	18	2.50	2.73	32	2.50	2.65	46	2.51	2.63
6	2.44	3.03	20	2.49	2.70	34	2.53	2.68	48	2.50	2.62
8	2.51	2.97	22	2.46	2.66	36	2.50	2.64	50	2.49	2.61
10	2.49	2.87	24	2.51	2.70	38	2.48	2.62	52	2.48	2.60
12	2.55	2.88	26	2.54	2.72	40	2.46	2.59	54	2.50	2.62
14	2.48	2.76	28	2.52	2.69	42	2.43	2.56	56	2.49	2.61

p	$\frac{\alpha_{200}}{\alpha_{100}}$	$\frac{\beta_{200}}{\beta_{100}}$	p	$\frac{\alpha_{200}}{\alpha_{100}}$	$\frac{\beta_{200}}{\beta_{100}}$	p	$\frac{\alpha_{200}}{\alpha_{100}}$	$\frac{\beta_{200}}{\beta_{100}}$
58	2.51	2.62	72	2.46	2.57	86	2.44	2.56
60	2.45	2.56	74	2.44	2.55	88	2.43	2.55
62	2.48	2.59	76	2.43	2.54	90	2.40	2.53
64	2.44	2.55	78	2.42	2.53	92	2.35	2.48
66	2.53	2.64	80	2.46	2.57	94	2.35	2.48
68	2.46	2.57	82	2.47	2.58	96	2.37	2.51
70	2.46	2.57	84	2.42	2.53	98	2.28	2.43

The ratio values related to α have 2.459269 mean, and 0.06561574 standard deviation. While the ratio values related to β have 2.678803 mean, and 0.2827515 standard deviation. We can notice that the standard deviation corresponding to β seems to be quite high. Apparently, the answer to our previous question is negative. A further research for this thesis could be investigating more about this question.

In the same way as for the case $n = 100$, m and q can be estimated for the linear method, and k_1, k_2 and k_3 for the tangent method. Let us analyze the efficiency of these two methods, by sampling 1000 times μ_1 for different values of p, as we did for $n = 100$.

By looking at Figure (3.15), also in this case the linear method seems to be very good in some cases, and not very close to the reality in other cases. The tangent method instead, seems to be always better and close to the reality.

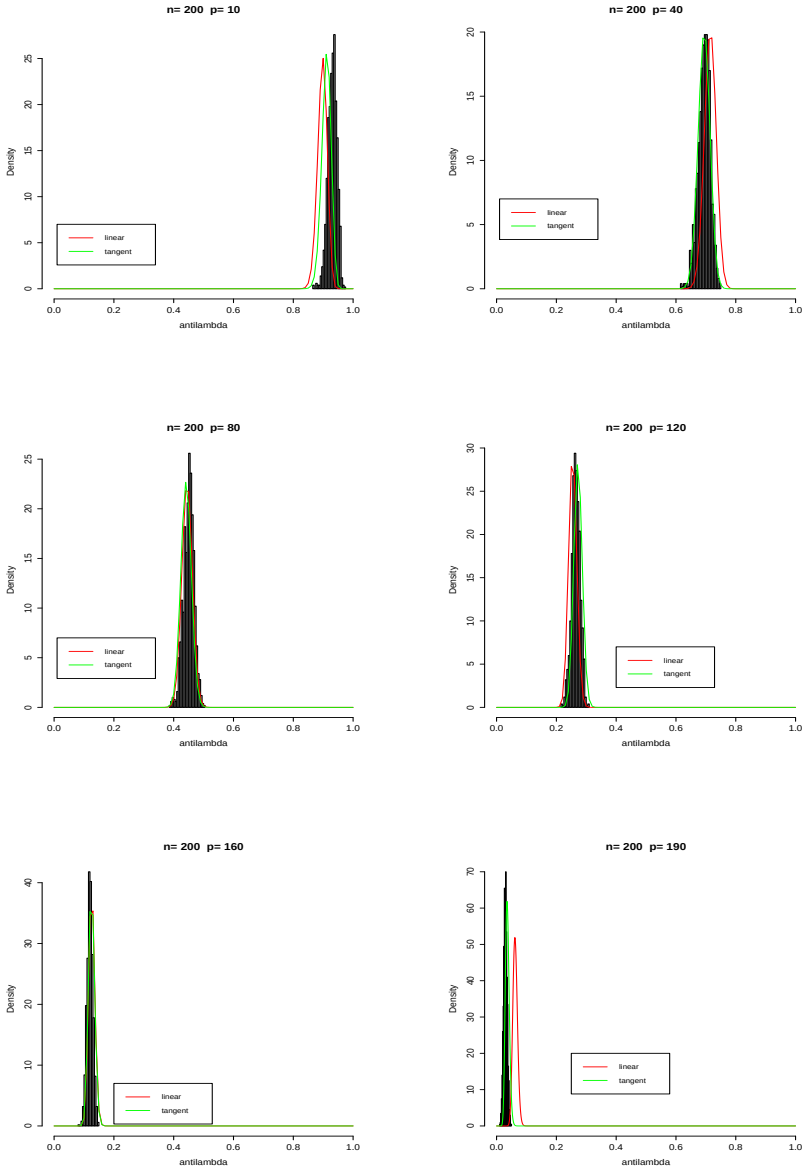


Figure 3.15: Estimation of $Beta(\alpha, \beta)$ through linear method and tangent method. Comparison to reality.

Chapter 4

Conclusions

After reporting some already known results about the distribution of eigenvalues of sample covariance matrices and some already known information about antieigenvalues, in particular of $p \times p$ symmetric positive definite operators, through simulations in R software we have started investigating the distribution of antieigenvalues of the sample covariance matrix $\mathcal{S}$ of a random matrix $\mathcal{X} \in \mathbb{R}^{n \times p}$, whose entries are i.i.d. $N(0,1)$.

We have analyzed the distribution of the antieigenvalues of $\mathcal{S}$ when $\gamma = \frac{n}{p}$ varies in $(1, +\infty)$, noticing that the distribution spreads in a tighter way towards the value 1, the higher is γ, as expected since all the antieigenvalues converge to the value 1 when n tends to infinity and p is fixed.

We have investigated the convergence property of the first antieigenvalue of $\mathcal{S}$ $\mu_1(\mathcal{S})$, obtaining a theoretical result by exploiting the already known results for the largest and smallest eigenvalues of $\mathcal{S}$. This result says that if $p \ll n$, then $\mu_1^{(p)}(\mathcal{S})$ converges to a value close to 1. While, if $p \lessapprox n$, then $\mu_1^{(p)}(\mathcal{S})$ converges to a value close to 0.

We have noticed that the first antieigenvalue of $\mathcal{S}$ seems to follow a Beta distribution, whose parameters α and β respect the following behaviour: if $p \ll n$, then $\alpha \gg \beta$; while, if $p \lessapprox n$, then $\alpha \ll \beta$.

We have tried to set a rule to estimate the parameters α and β of the Beta distribution where $\mu_1(\mathcal{S})$ comes from, in function of γ.
We have set $n = 100$ and expressed an elliptical law to estimate α and β in function of p, through both a linear method and a tangent method. By analyzing the performance of these two methods we have noticed that in general the tangent method is better and works fine.
Then, we have set $n = 200$ to see if such estimations keep holding, but we were contraddicted by the results. So, we asked the question if the real values of α and β grow linearly in function of n (with the same linear coefficient for each γ), but we have seen that the ratio of the estimated values for α and β when $n = 200$ and when $n = 100$, varies in a relevant way with γ for β.
Hence, it is probably the case that our elliptical law needs to be found for all the values of n that we are interested in, as we did for $n = 100$ and $n = 200$,

unless a relation for the real values of α and β in function of n, different from the linear one and holding for each value of γ, is found. In this case the elliptical law obtained by setting $n = 100$ should just be modified by introducing γ and such relation.

Further Research

Along the chapters we have already mentioned possible further researches. Firstly, we have mentioned the possibility of trying to find a function, in analogy to the Marčenko-Pastur law, that best fits the histograms of the antieigenvalues of $\mathcal{S}$. Then, a really important further research could be trying to find a relation for the values of the parameters α and β of the beta distribution where $\mu_1(\mathcal{S})$ comes from, in function of n and which, for each n, is the same for all the values of $\gamma \in (1, n)$. In addition, a further research could be trying to find theoretical results that confirm the results we have got through simulations in R software.

Bibliography

[1] Anderson, T. W.
An Introduction to Multivariate Stattistical Analysis (third edition). New York: Wiley (2003).

[2] Bai, Z. D. and Silverstein J. W.
Spectral Analysis of Large Dimensional Random Matrices. Springer (2010).

[3] Bai, Z. D. and Yin, Y. Q.
Limit of the smallest eigenvalue of large dimensional covariance matrix. The Annals of Probability, Vol.21, No.3 (1993).

[4] Bai, Z. D., Yin, Y. Q. and Silverstein, J. W.
A Note on the Largest Eigenvalue of a Large Dimensional Sample Covariance Matrix. Journal of Multivariate Analysis (1988).

[5] Gustafson K.
The angle of an operator and positive operator products. Bulletin of the American Mathematical Society (1968).

[6] Gustafson K.
Antieigenvalue inequalities in operator theory. Inequalities III, Los Angeles (1969).

[7] Gustafson K.
Antieigenvalues. Department of Mathematics, University of Colorado (1994).

[8] Gustafson K.
Antieigenvalue Analysis, with Applications to Numerical Analysis, Wavelets, Statistics, Quantum Mechanics, Finance and Optimization. World-Scientific, Singapore (2012).

[9] Gustafson K.
Antieigenvalue Analysis, New Applications: Continuum Mechanics, Economics, Number Theory. Department of Mathematics, University of Colorado (2015).

[10] Kollo, T. and von Rosen, D.
Advanced Multivariate Statistics with Matrices. Springer, Dordrecht, The Netherlands (2005).

[11] Marčenko V. A. and Pastur L. A.
Distribution of eigenvalues for some sets of random matrices. Math. USSR-Sbornik, No. 4 Vol. 1 (1967).

[12] Muirhead, R. J.
Aspects of Multivariate Statistical Theory. Wiley, New York (1982).

[13] Potthoff R. F. and Roy S. N.
A generalized multivariate analysis of variance model useful especially for growth curve problems. University of North Carolina (1964).

[14] Srivastava M. S. and Singull, M.
Test for the mean matrix in a Growth Curve model for high dimensions. Accepted for pubblication in Communication in Statistics - Theory and Methods (2016).

POLITECNICO DI MILANO

LINKÖPINGS UNIVERSITET