STATISTICAL METHODS IN FINANCE

(Preliminary version)

Dr Antoine Jacquier

Department of Mathematics

Imperial College London

Autumn Term 2019-2020

MSc in Mathematics and Finance

This version: December 5, 2019

Contents

Notations and standard definitions

The notations below will be used throughout the notes. We also wish to emphasize some common notational mistakes.

$\mathbb{N}$	integer numbers $\{0, 1, 2, \ldots\}$ (including 0)
$\mathbb{N}^*$	non null integer numbers $\{1, 2, \ldots\}$
$\mathcal{M}_{m,n}(\mathbb{R})$	set of $m \times n$ matrices with real elements
$\mathcal{M}_{n}(\mathbb{R})$	set of $n \times n$ matrices with real elements
A^o	interior of a set A
$\overline{A}$	closure of a set A
$\mathcal{N}$	cumulative distribution function of the standard Gaussian distribution
$X = (X_t)_{t\geq 0} \neq X_t$	a process evolving in time, as opposed to X_t, which represents the (possibly random) value of the process X at time t
$f \neq f(x)$	f represents a function and $f(x)$ the value of the function f at the point x. Equivalently the function f can be written as $x \mapsto f(x)$
$\widehat{f}$	Fourier transform of a function f
$f(x) = \mathcal{O}(g(x))$ $(x \to \infty)$	there exist $M, x_0 > 0$ such that $\lvert f(x)\rvert \leq M\lvert g(x)\rvert$ for all $x > x_0$
$f(x) = \mathcal{O}(g(x))$ $(x \to a)$	there exist $M, \delta > 0$ such that $\lvert f(x)\rvert \leq M\lvert g(x)\rvert$ for all $\lvert x-a\rvert < \delta$
$f(x) = o(g(x))$ $(x \to a)$	$\lim_{x\to a} \frac{f(x)}{g(x)} = 0$, where $a \in \mathbb{R} \cup \{\pm\infty\}$
$\mathbb{1}_{\{x\in A\}}$	indicator function equal to 1 if $x \in A$ and zero otherwise
$x \wedge y$	$\min(x, y)$
a.s.	almost surely
$(x-y)_+$	$\max(0, x-y)$

Chapter 1

Descriptive Statistics and Python

1.1 Python for Statistics

1.1.1 A quick introduction to programming languages

Computing and programming are ubiquitous, in every area of every-day life, and are becoming increasingly important to deal with large flows of information. On financial markets, programming is fundamental to analyse time series of data, to evaluate financial derivatives, to run risk analyses, and to trade at high frequency, for example. Which programming language to use depends on one's needs, and the main factor is time: there are two types of times one should consider:

- Execution time is the time it takes to run the programme itself;
- Development time is the time it takes to write the code.

For ultra high-frequency trading, for instance, execution time is the most important, as the algorithm needs to make a decision very quickly. For long-term trading strategies, however, execution time is less important, and one might favour quicker development time.

1.1.2 Statically typed languages

For short execution time, lower level languages, which compile directly to machine code, are preferred. The often use static typing, namely data types have to be specified. C++ is the main example, and has been the main language used in quantitative finance; however, there is a non-negligible entry cost to it, understanding its underlying concepts such as memory allocation or pointers. Java is also a statically typed language, but automatically manages low-level memory allocation; that said, it does not compile directly to machine code, and needs a Java Virtual Machine to execute the Java bytecode generated by the programme. Historically slower than C++ (because of the virtual machine layer), recent advances have now made their speeds comparable.

1.1.3 Interpreted languages

When execution time is not the priority, and development time is preferred (for example for long-term strategies), one can instead use interpreted languages, such as Matlab, Python, or R. While the execution time is slower, these languages are dynamically typed, so that variables' types are automatically recognised by the programme and do not need to be specified by the user. Matlab has been a popular language in quantitative finance, and is still around because of its legacy code. However, in recent years, R (historically the preferred language for Statistics) and Python have been taking over, as they are open source and the range of available packages for applications has been growing exponentially. They are obviously slower than lower level languages, and some in-between languages have recently appeared, in particular Julia, which, when first run, generates machine code for execution.

1.1.4 Functional and query languages

1.2 Python

1.2.1 Python in Finance

A large number of financial companies, banks, hedge funds, asset managers,... have recently adopted Python. JP Morgan's Quartz, based on Python, is used for pricing and risk analyses; Bank of America has its own version called Athena. One major drawback of Python is its Global Interpreter Lock (GIL), which only allows one thread to execute at every point in time, making it difficult to parallelise code. Some Python libraries bypass the issue, for example the multiprocessing one, allowing the user to use multiple cores. Cython, on the other hand, is a static compiler for Python, and allows to convert some slow Python code (in particular loops) into much faster C versions.

1.2.2 General Python libraries

`Python` 3.5 is the default version of `Python`instead of 2.7. It is well supported by many packages to analyse data and perform statistical analysis.

- **NumPy** is the fundamental basic package for scientific computing with Python.
- **SciPy** supplements **NumPy**.
- **pandas** is a high-performance library for data analysis.
- **matplotlib** is the standard Python library for plots and graphs.

1.2.3 Python for Economics and Finance

- **quantdsl** is a functional programming language for financial derivatives.
- **statistics** is a built-in Python library for basic statistical computations.
- **ARCH**: tools for econometrics.
- **statsmodels** allows to explore data, estimate statistical models, and perform statistical tests.
- **QuantEcon**: library for economic modelling

1.2.4 Python libraries for plotting

- **matplotlib** is the standard Python library for plots and graphs. It is fairly basic but can basically, with enough commands, generate any graphs.
- **Seaborn** is a powerful plotting library built on top of matplotlib.

1.2.5 Python libraries for Machine learning

- **scikit-learn** adds to SciPy and NumPy common machine learning and data mining algorithms, such as clustering, regression, and classification.
- **Theano** has machine learning algorithms using the computer's GPU, and is hence extremely powerful for deep learning and heavy tasks.
- **TensorFlow** is Google-supported machine learning library based on a multi-layer architecture.

1.3 Online data sources

Python makes it straightforward to query online databases directly, without having to import data locally.

Economics database

An important database for economists is FRED, a vast collection of time series data maintained by the St. Louis Federal Reserve. For example, the entire series for the US civilian unemployment rate is available at https://research.stlouisfed.org/fred2/series/unrate/downloaddata/UNRATE.csv.

Another useful data for Economics data is the World Bank, which collects and organises data on a huge range of indicators.

Finance database

Yahoo Finance, Google Finance are publicly available. Options market data, though, are not, but can be accessed via WRDS/OptionMetrics.

Other interesting databases

- Google Trends Can you give some reasons explaining this graph and that one?
- Google Books
- Million Song Dataset
- Comprehensive list of available data

Chapter 2

Applied Multivariate Statistical Analysis

2.1 A short introduction to Matrix algebra

In this part, we recall some fundamental definitions, tools and properties of finite-dimensional algebra. Unless otherwise specified–chiefly because of the financial applications in mind–all quantities will be real valued.

2.1.1 Introductory tools

For m, n integers, we shall denote by $\mathcal{M}_{m,n}$ the space of matrices with real entries with m rows and n columns, endowed with the scalar product

$$\langle \mathbf{A}, \mathbf{B} \rangle := \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} b_{ij}, \qquad \text{for any } \mathbf{A}, \mathbf{B} \in \mathcal{M}_{m,n},$$

and the associated Euclidean norm $\|\mathbf{A}\| := \langle \mathbf{A}, \mathbf{A} \rangle^{1/2}$, where we use capital letters to denote matrices, and lower-case letters for its entries, such as $\mathbf{A} = (a_{i,j})_{1\leq i\leq m, 1\leq j\leq n}$, and we denote by $\mathbf{A}^\top$ the transpose of the matrix $\mathbf{A}$, i.e. $\mathbf{A}^\top = (a_{j,i})_{1\leq j\leq n, 1\leq i\leq m} \in \mathcal{M}_{n,m}$. For $\mathbf{A} \in \mathcal{M}_{m,n}$ and $\mathbf{B} \in \mathcal{M}_{n,p}$, the product $\mathbf{C} := \mathbf{AB}$ belongs to $\mathcal{M}_{m,p}$ and $c_{i,k} = \sum_{k=1}^{n} a_{ij} b_{jk}$. Whenever $m = n$, the space of square matrices (and corresponding indices) will be denoted by $\mathcal{M}_n$ for simplicity. The matrix $\mathbf{I}_n$ is the identity matrix in $\mathcal{M}_n$, and $\mathbf{O}_{m,n}$ the null matrix in $\mathcal{M}_{m,n}$.

Definition 2.1.1. Let $\mathbf{A} \in \mathcal{M}_n$.

- The matrix $\mathbf{A}$ is called orthogonal if $\mathbf{A}\mathbf{A}^\top = \mathbf{A}^\top\mathbf{A} = \mathbf{I}_n$;
- The rank of $\mathbf{A}$, denoted rank($\mathbf{A}$) is the maximum number of linearly independent rows;

- Trace: $\mathrm{Tr}(\mathbf{A}) := \sum_{i=1}^{n} a_{ii}$;
- Determinant:

$$\det(\mathbf{A}) := \sum (-1)^{|\tau|} a_{1,\tau_1} \cdots a_{n,\tau_n},$$

  over all permutations $\tau \in \{1, \cdots, n\}$, and $|\tau| = 0$ if the permutation is a product of an even number of transpositions, and $|\tau| = 1$ otherwise;
- if $\det(\mathbf{A}) \neq 0$ then the inverse matrix $\mathbf{A}^{-1}$ exists and $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$;

Exercise 1. Following the notations in Definition 2.1.1, let $\alpha \in \mathbb{R}$, prove the following identities:

(a) $\det(\alpha\mathbf{A}) = \alpha^n \det(\mathbf{A})$;

(b) $\det(\mathbf{AB}) = \det(\mathbf{BA}) = \det(\mathbf{A})\det(\mathbf{B})$;

(c) $\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$;

(d) $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{Tr}(\mathbf{A}^\top \mathbf{B}) = \mathrm{Tr}(\mathbf{A}\mathbf{B}^\top)$;

(e) $0 \leq \mathrm{rank}(\mathbf{A}) \leq m \wedge n$;

(f) $\mathrm{rank}(\mathbf{A}) = \mathrm{rank}(\mathbf{A}^\top) = \mathrm{rank}(\mathbf{A}\mathbf{A}^\top)$;

(g) if $\mathbf{A} \in \mathcal{M}_n$ and $\det(\mathbf{A}) \neq 0$,, then $\det(\mathbf{A})^{-1} = \det(\mathbf{A}^{-1})$;

(h) if $\mathbf{A}$ is orthogonal, then $|\det(\mathbf{A})| = 1$;

2.1.2 Spectral Theory for matrices

In this section, we consider a square real-valued matrix $\mathbf{A} \in \mathcal{M}_n$. The spectral theory for matrices is based on the following definition:

Definition 2.1.2. The characteristic polynomial $\mathcal{P}_{\mathbf{A}}$ of the matrix $\mathbf{A}$ is defined as

$$\mathcal{P}_{\mathbf{A}}(\lambda) := \det(\mathbf{A} - \lambda\mathbf{I}).$$

It is easy to see that $\deg(\mathcal{P}_{\mathbf{A}}) = n$, and its n (possibly complex) roots are called the eigenvalues of $\mathbf{A}$. A root with algebraic multiplicity equal to one is called a simple eigenvalue, and we usually denote the set of eigenvalues $\sigma(\mathbf{A})$. For any $\lambda \in \sigma(\mathbf{A})$, a non-null vector $\mathrm{u} \in \mathbb{R}^n$ satisfying $\mathbf{A}\mathrm{u} = \lambda\mathrm{u}$ is called the associated eigenvector. We shall further denote by $\rho(\mathbf{A}) := \max\{|\lambda| : \lambda \in \sigma(\mathbf{A})\}$ the spectral radius of $\mathbf{A}$.

Exercise 2.

- Show that the eigenvalues of a square symmetric matrix are real.
- Let P be a polynomial. Show that, for any $\lambda \in \sigma(\mathbf{A})$, then $P(\lambda) \in \sigma(P(\mathbf{A}))$.

- Show that $\mathcal{P}_{\mathbf{A}}(\mathbf{A}) = 0$.
- Show that $\det(\mathbf{A}) = \prod_{\lambda\in\sigma(\mathbf{A})} \lambda$, and that $\mathrm{Tr}(\mathbf{A}) = \sum_{\lambda\in\sigma(\mathbf{A})} \lambda$;

Theorem 2.1.3 (Jordan (spectral) decomposition)**.** *Any symmetric matrix $\mathbf{A} \in \mathcal{M}_n$ admits a decomposition of the form $\mathbf{A} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^\top$, where $\mathbf{\Lambda}$ is the diagonal matrix of all eigenvalues of $\mathbf{A}$, and $\mathbf{\Gamma}$ the orthogonal matrix consisting of the eigenvectors of $\mathbf{A}$.*

Since the matrix $\mathbf{\Lambda}$ is diagonal, we shall use the standard notation $\Lambda = \mathrm{Diag}(\lambda_1, \dots, \lambda_n)$.

Example 2.1.4. Consider the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}.$$

We can find the eigenvalues by solving $\|\mathbf{A} - \lambda\mathbf{I}\| = 0$, i.e.

$$\|\mathbf{A} - \lambda\mathbf{I}\| = \begin{pmatrix} 1-\lambda & 2 \\ 2 & 3-\lambda \end{pmatrix} = (1-\lambda)(3-\lambda) - 4 = 0,$$

so that $\lambda \in \{2-\sqrt{5}, 2+\sqrt{5}\}$, and corresponding eigenvectors

$$\gamma_1 = \begin{pmatrix} \frac{2}{1-\sqrt{5}} \\ 1 \end{pmatrix} \quad \text{and} \quad \gamma_2 = \begin{pmatrix} \frac{2}{1+\sqrt{5}} \\ 1 \end{pmatrix}.$$

Check that the matrix $\mathbf{\Gamma}$ formed by the two vectors γ_1 and γ_2 is indeed orthogonal, i.e. $\mathbf{\Gamma}^\top\mathbf{\Gamma} = \mathbf{I}$.

Exercise 3. Consider the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/2 & 1/2 \\ 0 & 1/2 & 1/2 \end{pmatrix}.$$

Show that $\mathbf{A}$ is idempotent, i.e. $\mathbf{A}\mathbf{A} = \mathbf{A}$. By computing its eigenvalues and eigenvectors, can you guess a property of such matrices?

The advantage of the Jordan decomposition is that it allows for quick computations of functions of matrices. Consider for example the function $P(x) \equiv x^\alpha$, for $\alpha \in \mathbb{R}$, applied to a symmetric matrix $\mathbf{A}$: $P(\mathbf{A}) = \mathbf{\Gamma}\Lambda^\alpha\mathbf{\Gamma}^\top$, where $\Lambda^\alpha = \mathrm{Diag}\left(\lambda_1^\alpha, \dots, \lambda_n^\alpha\right)$. This spectral decomposition theorem is fundamental, and will apply in particular to variance-covariance matrices, which, at least in theory, satisfy the required assumptions. However, many matrices (for example non-square matrices) cannot be handled by the decomposition, and the singular value decomposition generalises this. Before stating it, though, let us introduce some notations. Consider the matrix $\mathbf{A}$, and consider each row a_i as a point in $\mathbb{R}^n$. The problem is to determine the best (in the L^2 sense) subspace of dimension $k \leq n$:

Definition 2.1.5. Let $\mathbf{A} := (\mathrm{a}_1, \ldots, \mathrm{a}_m)$ be a set of points in $\mathbb{R}^n$. The best approximating k-dimensional linear subspace of $\mathbf{A}$ is the linear subspace $\mathbf{V} \in \mathbb{R}^k$ such that the distance from $\mathbf{A}$ to $\mathbf{V}$ is minimised.

Consider the case $k = 1$; we are looking for a line through the origin, closest to the cloud of points. By Pythagoras theorem, minimising the (square of the) distance of the cloud onto the line is equivalent to maximising the squared length of the projection onto the line. Let v be a unit vector along this line, and consider the projection of the point a_i onto v. The projection corresponds exactly to the vector $\langle \mathrm{a}_i, \mathrm{v}\rangle \mathrm{v}$, where the angle bracket here is nothing else than the dot product $\mathrm{a}_i^\top \mathrm{v}$. Since v is a unit vector, the length of this projection is equal to $\mathrm{a}_i^\top \mathrm{v}$. Therefore, by Pythagoras theorem, the distance from a_i to $\mathbf{V}$ is equal to

$$\mathrm{Dist}(\mathrm{a}_i, \mathbf{V})^2 = \mathrm{a}_i^\top \mathrm{a}_i - \left(\mathrm{a}_i^\top \mathrm{v}\right)^2,$$

and the distance between $\mathbf{A}$ and $\mathbf{V}$ is thus

$$\mathrm{Dist}(\mathbf{A}, \mathbf{V})^2 = \sum_{i=1}^{m} \left(\mathrm{a}_i^\top \mathrm{a}_i - \left(\mathrm{a}_i^\top \mathrm{v}\right)^2\right) = \|\mathbf{A}\|_F^2 - \|\mathbf{A}\mathrm{v}\|^2, \tag{2.1.1}$$

where $\|\cdot\|_F$ denotes the Frobenius norm $\|\mathbf{A}\|_F^2 := \mathrm{Tr}(\mathbf{A}^\top \mathbf{A})$. Since the first term is constant, minimising this distance is equivalent to maximising $\|\mathbf{A}\mathrm{v}\|$. The first singular vector v_1 is the best line fit (through the origin), defined as

$$\mathrm{v}_1 := \underset{\|\mathrm{v}\|=1}{\arg\max} \|\mathbf{A}\mathrm{v}\|, \tag{2.1.2}$$

and we call $\sigma_1(\mathbf{A}) := \|\mathbf{A}\mathrm{v}_1\|$ the first singular value. Note that, since (2.1.1) has to be non-negative, the maximum value for $\sigma_1(\mathbf{A})$ is the Frobenius norm of $\mathbf{A}$, which corresponds to all the points $\mathrm{a}_1, \ldots, \mathrm{a}_m$ lying on the same line. We can then iterate this procedure to define the second singular vector and value as

$$\mathrm{v}_2 := \underset{\mathrm{v}\perp\mathrm{v}_1, \|\mathrm{v}\|=1}{\arg\max} \|\mathbf{A}\mathrm{v}\| \qquad \text{and} \qquad \sigma_2(\mathbf{A}) := \|\mathbf{A}\mathrm{v}_2\|.$$

Clearly, the sequence of singular values is decreasing, and hence two cases can occur: either we reach n iterations, and there is hence no more vector v to choose, or we reach a level r such that $\sigma_{r+1}(\mathbf{A}) = 0$. In the latter case, this means that the data $\mathbf{A}$ lies fully in a r-dimensional subspace, spanned by the basis $\mathrm{v}_1, \ldots, \mathrm{v}_r$. The following result–stated without proof (not too hard, though)–justifies this algorithm:

Proposition 2.1.6 (Greedy Algorithm). *Let $\mathbf{A} \in \mathcal{M}_{m,n}$ and $\mathrm{v}_1, \ldots, \mathrm{v}_r$ its singular vectors constructed as above. For any $1 \leq k \leq r$, the subspace spanned by $\mathrm{v}_1, \ldots, \mathrm{v}_k$ is the best fit of dimension k for the matrix $\mathbf{A}$.*

Fix a row i. Since the vectors $\mathrm{v}_1, \ldots, \mathrm{v}_r$ span the space of all rows of the matrix $\mathbf{A}$, then clearly $\mathrm{a}_i^\top \mathrm{v} = 0$ for all v orthogonal to these vectors. Therefore, we can write $\sum_{j=1}^r \left(\mathrm{a}_i^\top \mathrm{v}_j\right)^2 = \|\mathrm{a}_i\|^2$ since v_i is a unit vector orthogonal to $(\mathrm{v}_1, \ldots, \mathrm{v}_{i-1}, \mathrm{v}_{i+1}, \ldots, \mathrm{v}_r)$, and hence

$$\sum_{i=1}^n \|\mathrm{a}_i\|^2 = \sum_{i=1}^n \sum_{j=1}^r \left(\mathrm{a}_i^\top \mathrm{v}_j\right)^2 = \sum_{j=1}^r \sum_{i=1}^n \left(\mathrm{a}_i^\top \mathrm{v}_j\right)^2 = \sum_{j=1}^r \|\mathbf{A}\mathrm{v}_j\|^2 = \sum_{j=1}^r \sigma_j^2(\mathbf{A}),$$

which in fact defines the so-called Frobenius norm.

Definition 2.1.7. The sequence of vectors $(\mathrm{u}_i)_{i=1,\ldots,n}$ defined by $\mathrm{u}_i := \dfrac{\mathbf{A}\mathrm{v}_i}{\sigma_i(\mathbf{A})}$ are called the left singular vectors of $\mathbf{A}$, and the v_i are the right singular vectors.

Theorem 2.1.8. *Both left and right singular vectors are orthogonal.*

The fact that the right singular vectors are orthogonal is trivial from their definition. We can prove by induction that so are the left singular vectors, but we shall omit the proof for sake of brevity here.

Theorem 2.1.9 (Singular Value decomposition). *Any matrix $\mathbf{A} \in \mathcal{M}_{m,n}$ with rank r admits a decomposition of the form $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}^\top$, where $\mathbf{\Lambda} = \mathrm{Diag}\left(\sigma_1(\mathbf{A}), \ldots, \sigma_r(\mathbf{A})\right)$, and $\mathbf{U} \in \mathcal{M}_{m,r}$ and $\mathbf{V} \in \mathcal{M}_{n,r}$ the matrices composed of the left and right singular vectors of $\mathbf{A}$.*

Proof. We want to show that the first r singular vectors form a linear subspace maximising the sum of squared projections of $\mathbf{A}$ onto it. It is trivial if $r = 1$, since the singular vector v_1 is the solution of the maximisation problem (2.1.2). Let now $\mathbf{W}$, with orthonormal basis $(\mathbf{w}_1, \mathbf{w}_2)$ be any best approximating two-dimensional linear subspace for $\mathbf{A}$. We are interested in maximising the quantity $\|\mathbf{A}\mathbf{w}_1\|^2 + \|\mathbf{A}\mathbf{w}_2\|^2$, and we pick the vector $\mathbf{w}_2 \perp \mathrm{v}_1$. Indeed, either v_1 is already orthogonal to $\mathbf{W}$ (trivial case), or it is not, and we let $\mathbf{w}_1$ be the orthogonal projection of v_1 onto $\mathbf{W}$, and take $\mathbf{w}_2$ as a unit vector orthogonal to $\mathbf{w}_1$. By construction, the vector v_1 maximises $\|\mathbf{A}\mathrm{v}\|$, so that $\|\mathbf{A}\mathrm{v}_1\| \geq \|\mathbf{A}\mathbf{w}_1\|$, and $\|\mathbf{A}\mathrm{v}_2\| \geq \|\mathbf{A}\mathbf{w}_2\|$ because $\mathbf{w}_2 \perp \mathrm{v}_1$. Therefore $\|\mathbf{A}\mathrm{v}_1\|^2 + \|\mathbf{A}\mathrm{v}_2\|^2 \geq \|\mathbf{A}\mathbf{w}_1\|^2 + \|\mathbf{A}\mathbf{w}_2\|^2$ as desired. The general r-dimensional case can be deduced by induction. □

The most useful application of Singular Value Decomposition in this course will be PCA, which we will see in full details later. It also has many applications, in particular to compute the so-called pseudo-inverse, and for image compression. To convince yourself, at least intuitively, consider the transmission of an image, where the matrix $\mathbf{A} \in \mathcal{M}_{n,n}$ represents the pixel description of the image. For large n, the transmission cost is of order n^2. Suppose that, instead of transmitting $\mathbf{A}$, we only transmit the first k singular values and left and right singular vectors; this would cost $\mathcal{O}(kn)$ operations. Of course, details are lost, and quality decreases, but, by picking the dimension k, one effectively chooses the size of the resolution. This can be formalised, and is, in fact, the content of the following theorem, which forms the basis of image reduction:

Theorem 2.1.10 (Eckart-Young-Mirsky Theorem). *Let $\mathbf{A} \in \mathcal{M}_{m,n}$ with rank r, and fix some $k \leq r$. The solution to the optimisation problem*

$$\min_{\widehat{\mathbf{A}}} \left\{ \left\| \mathbf{A} - \widehat{\mathbf{A}} \right\|_F : \mathrm{rank}(\widehat{\mathbf{A}}) \leq k \right\}$$

is given by $\widehat{\mathbf{A}} = \mathbf{U}_1 \mathbf{\Lambda}_1^{1/2} \mathbf{V}_1^\top$. Here, starting from the SVD of $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}^\top$, we write

$$\mathbf{U} = \begin{pmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{pmatrix}, \qquad \mathbf{\Lambda} = \begin{pmatrix} \mathbf{\Lambda}_1 & \mathbf{O}_{k,r-k} \\ \mathbf{O}_{r-k,k} & \mathbf{\Lambda}_2 \end{pmatrix}, \qquad \mathbf{V} = \begin{pmatrix} \mathbf{V}_1 & \mathbf{V}_2 \end{pmatrix},$$

with $\mathbf{U}_1 \in \mathcal{M}_{m,k}$, $\mathbf{V}_1 \in \mathcal{M}_{n,k}$, $\mathbf{\Lambda}_1 \in \mathcal{M}_{k,k}$

IPython notebook SVD.ipynb IPython notebook SVD_ImageCompression.ipynb

2.1.3 Quadratic forms

Let $\mathbf{A}$ be a square matrix in $\mathcal{M}_n$. We call $\mathcal{Q}_\mathbf{A}(\mathrm{x}) := \mathrm{x}^\top \mathbf{A} \mathrm{x}$ the quadratic form, from $\mathbb{R}^n$ to $\mathbb{R}$, associated to $\mathbf{A}$.

Definition 2.1.11. The matrix $\mathbf{A}$ is said to be positive semi-definite (resp. positive definite), and we write $\mathbf{A} \geq \mathbf{O}_n$ (resp. $\mathbf{A} > \mathbf{O}_n$), if $\mathcal{Q}_\mathbf{A}(\mathrm{x}) \geq 0$ (resp. $\mathcal{Q}_\mathbf{A}(\mathrm{x}) > 0$) for all non-zero vector $\mathrm{x} \in \mathbb{R}^n$.

We shall denote by $\mathcal{M}_n^+$ (resp. $\mathcal{M}_n^{++}$) the space of symmetric positive semi-definite (resp. positive definite) matrices in $\mathcal{M}_n$.

Example 2.1.12. The identity matrix $\mathbf{I}_n$ is positive definite.

Theorem 2.1.13. *If $\mathbf{A}$ is symmetric, then $\mathcal{Q}_\mathbf{A}(\mathrm{x}) = \mathrm{y}^\top \Lambda \mathrm{y}$ for any $\mathrm{x} \in \mathbb{R}^n$, with $\mathrm{y} := \mathbf{\Gamma}^\top \mathrm{x}$, where Λ and $\mathbf{\Gamma}$ arise from the Jordan decomposition of $\mathbf{A}$.*

Proof. Since $\mathbf{A} = \mathbf{\Gamma}\Lambda\mathbf{\Gamma}^\top$ from Theorem 2.1.3, then $\mathrm{x}^\top \mathbf{A} \mathrm{x} = \mathrm{x}^\top \mathbf{\Gamma}\Lambda\mathbf{\Gamma}^\top \mathrm{x} = \mathrm{y}^\top \Lambda \mathrm{y}$, with $\mathrm{y} = \mathbf{\Gamma}^\top \mathrm{x}$, and the theorem follows. □

Proposition 2.1.14. *Let $\mathbf{A} \in \mathcal{M}_n^+$. For any $N \in \mathbb{N}$, there exists a unique $\mathbf{B} \in \mathcal{M}_n^+$ such that $\mathbf{A} = \mathbf{B}^N$.*

Quadratic forms provide an easy way to check positivity of eigenvalues:

Proposition 2.1.15. *The matrix $\mathbf{A}$ is positive definite if and only if $\min_{\lambda \in \sigma(\mathbf{A})} \lambda > 0$.*

Proof. Since $\mathbf{A} > \mathbf{O}_n$, then $0 < \mathcal{Q}_\mathbf{A}(\mathrm{x}) = \mathrm{y}^\top \Lambda \mathrm{y}$ by Theorem 2.1.13, and the proposition follows. □

Corollary 2.1.16. *If $\mathbf{A}$ is positive definite, then $\mathbf{A}^{-1}$ exists and $\|\mathbf{A}\| > 0$.*

Exercise 4. Compute the quadratic forms of the identity matrix in $\mathcal{M}_n$ and of the matrices

$$\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

and determine their (absence of?) positivity.

Positive matrices appear very often in mathematical finance and in Statistics (in particular as covariance matrices), and admit a certain number of useful factorisations. Recall that a matrix $\mathbf{T}$ is called upper triangular if $t_{ij} = 0$ whenever $i > j$.

Proposition 2.1.17. *If $\mathbf{A} > \mathbf{O}_n$ (resp. $\mathbf{A} \geq \mathbf{O}_n$) then there exists a unique (resp. non-unique) upper triangular matrix $\mathbf{T} \in \mathcal{M}_n$ with strictly positive (resp. non-negative) diagonal elements such that $\mathbf{A} = \mathbf{T}^\top \mathbf{T}$.*

This apparently simple proposition allows, for example, to simulate general Gaussian processes simply from the knowledge of their covariance matrices. The following theorem will be fundamental when analysing reduction of variance in multivariate statistics, in particular for Principal Component Analysis:

Theorem 2.1.18. *Let $\mathbf{A}$ and $\mathbf{B}$ two symmetric matrices in $\mathcal{M}_n$ with $\mathbf{B} > \mathbf{O}_n$. Then*

$$\min_{\mathrm{x}:\mathcal{Q}_{\mathbf{B}}(\mathrm{x})=1} \mathcal{Q}_{\mathbf{A}}(\mathrm{x}) = \min\{\sigma(\mathbf{B}^{-1}\mathbf{A})\} \leq \max\{\sigma(\mathbf{B}^{-1}\mathbf{A})\} = \max_{\mathrm{x}:\mathcal{Q}_{\mathbf{B}}(\mathrm{x})=1} \mathcal{Q}_{\mathbf{A}}(\mathrm{x}).$$

Proof. Using the Jordan Decomposition, and writing the corresponding matrix as an index, we can write $\mathbf{B}^{1/2} = \mathbf{\Gamma}_{\mathbf{B}} \Lambda_{\mathbf{B}}^{1/2} \mathbf{\Gamma}_{\mathbf{B}}^\top$. Setting $\mathrm{y} := \mathbf{B}^{1/2}\mathrm{x}$, we can therefore write

$$\max_{\mathrm{x}:\mathcal{Q}_{\mathbf{B}}(\mathrm{x})=1} \mathcal{Q}_{\mathbf{A}}(\mathrm{x}) = \max_{\mathrm{y}:|\mathrm{y}\|=1} \mathcal{Q}_{\mathbf{A}}(\mathbf{B}^{-1/2}\mathrm{y}).$$

Using the Jordan Decomposition again, we can write $(\mathbf{B}^{-1/2})^\top \mathbf{A}\mathbf{B}^{-1/2} = \mathbf{\Gamma}\Lambda\mathbf{\Gamma}^\top$, so that, with $\mathrm{z} := \mathbf{\Gamma}^\top \mathrm{y}$, we have $\|\mathrm{z}\| = \|\mathbf{\Gamma}^\top \mathrm{y}\| = \|\mathrm{y}\|$ since $\mathbf{\Gamma}$ is orthogonal, and hence

$$\max_{\mathrm{y}:|\mathrm{y}\|=1} \mathcal{Q}_{\mathbf{A}}(\mathbf{B}^{-1/2}\mathrm{y}) = \max_{\mathrm{z}:\|\mathrm{z}\|=1} \mathrm{z}^\top \Lambda \mathrm{z} = \max_{\mathrm{z}:\|\mathrm{z}\|=1} \sum_{i=1}^n \lambda_i \mathrm{z}_i^2 \leq \left(\max_{\lambda\in\sigma(\mathbf{A})} \{\lambda\}\right)\left(\max_{\mathrm{z}:\|\mathrm{z}\|=1} \|\mathrm{z}\|\right) = \max\{\sigma(\mathbf{A})\}.$$

Clearly the maximum is attained at the point $\mathrm{z} = (1, 0, \ldots, 0)^\top$. Since the matrices $\mathbf{B}^{-1}\mathbf{A}$ and $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}$ have the same eigenvalues, the theorem follows. □

2.1.4 Derivatives

Let $\mathrm{x} \in \mathbb{R}^n$ and $\mathrm{y} \in \mathbb{R}^m$ related by $\mathrm{y} = \psi(\mathrm{x})$, where $\psi : \mathbb{R}^n \to \mathbb{R}^m$ is a smooth function. The Jacobian of the transformation is defined as

$$\nabla_{\mathbf{X}} \psi(\mathrm{x}) = \left(\partial_{x_j} y_i\right)_{1\leq i\leq m, 1\leq j\leq n} \in \mathcal{M}_{m,n}.$$

Example 2.1.19. Show that the following derivatives hold:

- $\nabla_{\mathrm{x}}(\mathbf{A}\mathrm{x}) = \mathbf{A}$, for any $\mathbf{A} \in \mathcal{M}_{m,n}$;
- $\nabla_{\mathrm{x}}\left(\mathrm{x}^\top \mathbf{A}\right) = \mathbf{A}^\top$, for any $\mathbf{A} \in \mathcal{M}_{n,m}$;
- $\nabla_{\mathbf{X}}(\mathcal{Q}_{\mathbf{A}}(\mathrm{x})) = \mathrm{x}^\top\left(\mathbf{A} + \mathbf{A}^\top\right)$, for any $\mathbf{A} \in \mathcal{M}_n$;
- $\nabla^2_{\mathbf{X}}(\mathcal{Q}_{\mathbf{A}}(\mathrm{x})) = \left(\mathbf{A} + \mathbf{A}^\top\right)$, for any $\mathbf{A} \in \mathcal{M}_n$.

2.1.5 Block matrices

For large matrices, it is sometimes convenient to decompose them into blocks of sub-matrices. Think for example of the covariance matrix of the S&P constituents, where one may be interested in sub-portfolios only. Let $\mathbf{A} \in \mathcal{M}_n$ be a square matrix with $n = p + q$ ($p, q \geq 1$), partitioned as

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix},$$

where $\mathbf{A}_{11} \in \mathcal{M}_p$, $\mathbf{A}_{22} \in \mathcal{M}_q$, $\mathbf{A}_{12} \in \mathcal{M}_{p,qq}$ and $\mathbf{A}_{21} \in \mathcal{M}_{q,p}$. Whenever it exists, the inverse matrix is denoted by

$$\mathbf{A}^{-1} = \begin{pmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{pmatrix},$$

and the blocks of the inverse are related to the original blocks through the following result, the proof of which is left as a simple yet tedious exercise:

Proposition 2.1.20. *Assuming all terms exist, the following identities hold:*

$$\begin{aligned} \mathbf{A}^{11} &= \left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)^{-1}, & \mathbf{A}^{12} &= -\mathbf{A}^{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1}, \\ \mathbf{A}^{22} &= \left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right)^{-1}, & \mathbf{A}^{21} &= -\mathbf{A}^{22}\mathbf{A}_{21}\mathbf{A}_{11}^{-1}. \end{aligned}$$

Proposition 2.1.21. *The following hold:*

$$\det\begin{pmatrix} \mathbf{A}_{11} & \mathbf{O}_{p,q} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} = \det(\mathbf{A}_{11})\det(\mathbf{A}_{22}) = \det\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{21} \\ \mathbf{O}_{q,p} & \mathbf{A}_{22} \end{pmatrix}$$

and, assuming they exist.

$$\det(\mathbf{A}) = \det(\mathbf{A}_{11})\det\left(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\right) = \det(\mathbf{A}_{22})\det\left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right).$$

2.2 Essentials of probability theory

We provide here a brief overview of standard results in probability theory and convergence of random variables needed in these lecture notes.

2.2.1 PDF, CDF and characteristic functions

In the following, $(\Omega, \mathcal{F}, \mathbb{P})$ shall denote a probability space and X a random variable defined on it. We define the cumulative distribution function $F : \mathbb{R} \to [0, 1]$ of X by

$$F(x) := \mathbb{P}(X \leq x), \qquad \text{for all } x \in \mathbb{R}.$$

The function F is increasing and right-continuous and satisfies the identities $\lim_{x\downarrow -\infty} F(x) = 0$ and $\lim_{x\uparrow\infty} F(x) = 1$. If the function F is absolutely continuous, then the random variable X has a probability density function $f : \mathbb{R} \to \mathbb{R}_+$ defined by $f(x) = F'(x)$, for all real number x. Note that this in particular implies the equality $F(x) = \int_{-\infty}^{x} f(u)\mathrm{d}u$. Recall that a function $F : \mathcal{D} \subset \mathbb{R} \to \mathbb{R}$ is said to be *absolutely continuous* if for any $\varepsilon > 0$, there exists $\delta > 0$ such that the implication

$$\sum_n |b_n - a_n| < \delta \qquad \Longrightarrow \qquad \sum_n |F(b_n) - F(a_n)| < \varepsilon$$

holds for any sequence of pairwise disjoint intervals $(a_n, b_n) \subset \mathcal{D}$. Define now the characteristic function $\phi : \mathbb{R} \to \mathbb{C}$ of the random variable X by

$$\phi(u) := \mathbb{E}\left(\mathrm{e}^{\mathrm{i}uX}\right).$$

Note that it is well defined for all real number u and the identity $|\phi(u)| \leq 1$ always holds on $\mathbb{R}$. Its extension to the complex plane ($u \in \mathbb{C}$) is more subtle; while it is fundamental for option pricing, it is less so for Statistics, and we shall leave it aside in these notes.

2.2.2 Some useful inequalities

We recall here a few inequalities that appear frequently in Probability and Statistics. We shall always consider random variables supported on the whole real line. The results below are not restricted to this case, though, but notations are simpler then.

Proposition 2.2.1 (Markov Inequality). *Let f be an increasing function and X a random variable such that $\mathbb{E}[f(X)]$ is finite. Then, for any $x \in \mathbb{R}$ such that $f(x) > 0$,*

$$\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}[f(X)]}{f(x)}.$$

Proof. Since f is increasing, then

$$\mathbb{P}(X \geq x) \leq \mathbb{P}(f(X) \geq f(x)) = \mathbb{E}\left(1\!\!1_{\{f(X)\geq f(x)\}}\right) \leq \mathbb{E}\left(\frac{f(X)}{f(x)} 1\!\!1_{\{f(X)\geq f(x)\}}\right) \leq \frac{\mathbb{E}\left(f(X)\right)}{f(x)}.$$

□

The following proposition is in fact an immediate corollary and is left as a simple exercise.

Proposition 2.2.2 (Chebychev Inequality). *If $X \in L^2(\mathbb{R})$, then, for any $x > 0$,*

$$\mathbb{P}(|X| \geq x) \leq \frac{\mathbb{E}(X^2)}{x^2} \qquad \textit{and} \qquad \mathbb{P}(|X - \mathbb{E}(X)| \geq x) \leq \frac{\mathbb{V}(X^2)}{x^2}.$$

Proposition 2.2.3 (Hölder Inequality). *Let $p \in (1,\infty)$ and q such that $p^{-1}+q^{-1}=1$. If X and Y are random variables such that $\mathbb{E}(|X|^p)$ and $\mathbb{E}(|Y|^q)$ are finite, then $\mathbb{E}(|XY|)$ is finite and*

$$\mathbb{E}(|XY|) \leq \mathbb{E}\left(|X|^p\right)^{1/p} \mathbb{E}\left(|Y|^q\right)^{1/q}.$$

Proof. Since the logarithm function is convex, the identity

$$\frac{\log(x)}{p} + \frac{\log(y)}{q} \leq \log\left(\frac{x}{p}+\frac{y}{q}\right)$$

holds for all $x, y > 0$. Taking exponential on both sides, this is obviously equivalent to $x^{1/p}y^{1/q} \leq \frac{x}{p}+\frac{y}{q}$. Setting $x = |X|^p/\mathbb{E}(|X|^p)$ and $y = |Y|^q/\mathbb{E}(|Y|^p)$ yields the result directly. □

The following inequality is a simple corollary, the proof of which is left as an exercise.

Proposition 2.2.4 (Lyapunov Inequality). *Let $0 < p < q$, and X a random variable such that $\mathbb{E}(|X|^q)$ is finite. Then*

$$\mathbb{E}(|X|^p)^{1/p} \leq \mathbb{E}(|X|^q)^{1/q}.$$

The kurtosis of a distribution X is defined as

$$\kappa := \frac{\mathbb{E}\left[(X-\mathbb{E}(X))^4\right]}{\mathbb{V}(X)^2},$$

and the excess kurtosis $\kappa_+ := \kappa - 3$.

Exercise 5. Using Lyapunov's Inequality, show that the excess kurtosis is always greater than -2. Show that this lower bound is attained for the Bernoulli distribution with equal chances.

Kurtosis measures the fatness of a distribution tails. Distributions can be classified as follows:

- Mesokurtic ($\kappa_+ = 0$): the Gaussian distribution for example;
- Leptokurtic ($\kappa_+ > 0$) distributions correspond to fat tails, and are of fundamental importance to describe returns of financial assets (in particular on Equity markets). The Student, Poisson, Laplace or Exponential distributions all belong to this category;
- Platykurtic ($\kappa_+ < 0$) correspond to thin-tail distributions, such as the uniform distribution.

Note that the Lyapunov Inequality in particular implies the sequence of inequalities for the moments of X,

$$\mathbb{E}(|X|) \leq \mathbb{E}(|X|^2)^{1/2} \leq \cdots \leq \mathbb{E}(|X|^q)^{1/q},$$

as long as the last one is finite for some integer q. This can be generalised as follows, the proof of which is left as an exercise:

Proposition 2.2.5 (Jensen Inequality). *Let f be a convex function and X a random variable such that $\mathbb{E}[f(X)]$ is finite. Then $f(\mathbb{E}(X)) \leq \mathbb{E}[f(X)]$.*

Proposition 2.2.6 (Cauchy-Schwarz Inequality)**.** *Let X and Y be two square-integrable random variables with $\mathbb{E}|XY|$ finite. Then $(\mathbb{E}(XY))^2 \leq (\mathbb{E}|XY|)^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$. The inequalities are equalities if X is almost surely a linear transformation of Y.*

Theorem 2.2.7 (Hoeffding Inequality)**.** *Let $X_1, \ldots, X_n$ be centered iid random variables with $a_i \leq X_i \leq b_i$. For any $\varepsilon > 0$, and any $z > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq \varepsilon\right) \leq \mathrm{e}^{-z\varepsilon} \prod_{i=1}^n \exp\left(\frac{z^2(b_i - a_i)^2}{8}\right).$$

Exercise 6. Let $X_1, \ldots, X_n$ be a sequence of iid Bernoulli(p) random variables. Using Theorem 2.2.7, show that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - p\right| > \varepsilon\right) \leq 2\mathrm{e}^{-2n\varepsilon^2}, \qquad \text{for any } \varepsilon > 0.$$

2.2.3 Gaussian distribution

A random variable X is said to have a Gaussian distribution (or Normal distribution) with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$, and we write $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ if and only if its density reads

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\mu)^2\right), \qquad \text{for all } x \in \mathbb{R}.$$

For such a random variable, the following identities are obvious:

$$\mathbb{E}\left(\mathrm{e}^{\mathtt{i}uX}\right) = \exp\left(\mathtt{i}\mu u - \frac{1}{2}u^2\sigma^2\right), \qquad \text{and} \qquad \mathbb{E}\left(\mathrm{e}^{uX}\right) = \exp\left(\mu u + \frac{1}{2}u^2\sigma^2\right),$$

for all $u \in \mathbb{R}$. The first quantity is the characteristic function whereas the second one is the Laplace transform or the random variable. If $X \in \mathcal{N}\left(\mu, \sigma^2\right)$, then the random variable $Y := \exp(X)$ is said to be lognormal and

$$\mathbb{E}(Y) = \exp\left(\mu + \frac{1}{2}\sigma^2\right) \qquad \text{and} \qquad \mathbb{E}\left(Y^2\right) = \exp\left(2\mu + 2\sigma^2\right).$$

2.2.4 Convergence of random variables

We recall here the different types of convergence for family of random variables $(X_n)_{n\geq 1}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We shall denote $F_n : \mathbb{R} \to [0,1]$ the corresponding cumulative distribution functions and $f_n : \mathbb{R} \to \mathbb{R}_+$ their densities whenever they exist. We start with a definition of convergence for functions, which we shall use repeatedly.

Definition 2.2.8. The family $(h_n)_{n\geq 1}$ of functions from $\mathbb{R}$ to $\mathbb{R}$ converge pointwise to a function $h : \mathbb{R} \to \mathbb{R}$ if and only if the equality $\lim_{n\uparrow\infty} h_n(x) = h(x)$ holds for all real number x.

This is a notoriously weak form of convergence, which, in particular does not preserve continuity (check for example the sequence $h_n(x) = x^n$ on $[0,1]$). It will however suffice here.

Convergence in distribution

This is the weakest form of convergence, and is the one appearing in the Central Limit Theorem.

Definition 2.2.9. The family $(X_n)_{n\geq 1}$ converges in distribution—or weakly or in law—to a random variable X if and only if $(F_n)_{n\geq 1}$ converges pointwise to a function $F : \mathbb{R} \to [0,1]$, i.e. if

$$\lim_{n\uparrow\infty} F_n(x) = F(x),$$

holds for all real number x where F is continuous. Furthermore, F is the CDF of X.

Example 2.2.10. Consider the family $(X_n)_{n\geq 1}$ such that each X_n is uniformly distributed on the interval $\left[0, n^{-1}\right]$. We then have $F_n(x) = nx\mathbb{1}_{\{x\in[0,1/n]\}} + \mathbb{1}_{\{x\geq 1/n\}}$. It is clear that the family of random variable converges weakly to the degenerate random variable $X = 0$. However, for any $n \geq 1$, we have $F_n(0) = 0$ and $F(0) = 1$.

Example 2.2.11. Weak convergence does not imply convergence of the densities, even when they exist. Consider the family such that $f_n(x) = \Big(1 - \cos(2\pi nx)\Big)\mathbb{1}_{\{x\in(0,1)\}}$.

Even though convergence in law is a weak form of convergence, it has a number of fundamental consequences for applications. We list them here without proof and refer the interested reader to [6] for details

Corollary 2.2.12. *Assume that the family $(X_n)_{n\geq 1}$ converges weakly to the random variable X. Then the following statements hold*

1. $\lim_{n\uparrow\infty} \mathbb{E}\left(h(X_n)\right) = \mathbb{E}\left(h(X)\right)$ *for all bounded and continuous function h.*

2. $\lim_{n\uparrow\infty} \mathbb{E}\left(h(X_n)\right) = \mathbb{E}\left(h(X)\right)$ *for all Lipschitz function h.*

3. $\lim \mathbb{P}\left(X_n \in A\right) = \mathbb{P}\left(X \in A\right)$ *for all continuity sets A of X.*

4. *(Continuous mapping theorem). The sequence $(h(X_n))_{n\geq 1}$ converges weakly to $h(X)$ for every continuous function h.*

The following theorem shall be of fundamental importance in many applications, and we therefore state it separately.

Theorem 2.2.13 (Lévy's continuity theorem). *The family $(X_n)_{n\geq 1}$ converges weakly to the random variable X if and only if the sequence of characteristic functions $(\phi_n)_{n\geq 1}$ converges pointwise to the characteristic function ϕ of X and ϕ is continuous at the origin.*

Exercise 7. Consider the sequence $(X_n)_{n\geq 0}$, where $X_n \sim \mathcal{N}(\mu_n, \sigma_n^2)$, and assume that $\lim_{n\uparrow\infty} \mu_n$ and $\lim_{n\uparrow\infty} \sigma_n^2$ exist. What can you conclude about the weak limit of the sequence $(X_n)_{n\geq 0}$?

Convergence in probability

Definition 2.2.14. The family $(X_n)_{n\geq 1}$ converges in probability to X if, for all $\varepsilon > 0$, we have

$$\lim_{n\uparrow\infty} \mathbb{P}\left(|X_n - X| \geq \varepsilon\right) = 0.$$

Remark 2.2.15. The continuous mapping theorem still holds under this form of convergence.

Almost sure convergence

This form of convergence is the strongest form of convergence and can be seen as an analogue for random variables of the pointwise convergence for functions.

Definition 2.2.16. The family $(X_n)_{n\geq 1}$ converges almost surely to the random variable X if

$$\mathbb{P}\left(\lim_{n\uparrow\infty} X_n = X\right) = 1.$$

Convergence in mean

Definition 2.2.17. Let $r \in \mathbb{N}^*$. The family $(X_n)_{n\geq 1}$ converges in the L^r norm to the random variable X if the r-th absolute moments of X_n and X exist for all $n \geq 1$ and if

$$\lim_{n\uparrow\infty} \mathbb{E}\left(|X_n - X|^r\right) = 0.$$

The following theorem makes the link between the different modes of convergence.

Theorem 2.2.18. *The following statements hold:*

- *Almost sure convergence implies convergence in probability.*
- *Convergence in probability implies weak convergence.*
- *Convergence in the L^r norm implies convergence in probability.*
- *For any $r \geq s \geq 1$, convergence in the L^r norm implies convergence in the L^s norm.*

2.2.5 Laws of large numbers and Central Limit Theorem

Consider an iid sequence $(X_1, \ldots, X_n)$ of random variables, with common finite mean μ and common variance σ^2, and define the arithmetic mean $\overline{X}_n := n^{-1}\sum_{i=1}^n X_i$. Direct computation yields $\mathbb{E}(\overline{X}_n) = \mu$ and $\mathbb{V}(\overline{X}_n) = \sigma^2/n$. The law of large numbers, presented below, is one of the fundamental results in probability, and is a key ingredient to prove convergence and bias of statistical estimators.

Theorem 2.2.19. *The weak law of large numbers state that the random variable $\overline{X}_n$ converges in probability to μ as n tends to infinity. The strong law of large numbers ensures that the convergence in fact holds almost surely.*

Note that, for the law of large numbers, weak or strong, to hold, we only require finiteness of the first moment, not of the second moment, although the proof when the latter is not finite is more involved. When the second moment is finite, we have the more precise formulation:

Theorem 2.2.20 (Central Limit Theorem). *If both μ and σ are finite, then the sequence $(\overline{X}_n - \mu)/(\sigma/\sqrt{n})$ converges in distribution to a centered Gaussian distribution with unit variance, or else*

$$\lim_{n\uparrow\infty} \mathbb{P}\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq x\right) = F_{\mathcal{N}(0,1)}(x), \qquad \textit{for all } x \in \mathbb{R}.$$

2.3 Introduction to statistical tools

In this section, we shall let $\mathbf{X} = (X_1, \ldots, X_n)$ denote a vector of size n (or equivalently $\mathbf{X} \in \mathcal{M}_{n,1}$) with random entries.

2.3.1 Joint distributions and change of variables

Let $\mathbf{X}$ denote a random vector taking values in $\mathbb{R}^n$. Its joint density distribution (whenever it exists) is the function $f : \mathbb{R}^n \to \mathbb{R}_+$ such that

$$\mathbf{F}_{\mathbf{X}}(\mathrm{x}) := \mathbb{P}(\mathbf{X} \leq \mathrm{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(y_1, \ldots, y_n)\mathrm{d}y_1 \cdots \mathrm{d}y_n, \qquad \text{for any } \mathrm{x} \in \mathbb{R}^n.$$

For any $i = 1, \ldots, n$, the marginal distribution of X_i is then given by

$$\mathbf{F}_{X_i}(x) = \lim_{(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)\uparrow\infty} \mathbf{F}_{\mathbf{X}}(\mathrm{x}).$$

In the continuous case, we shall always assume that $\mathbf{F}$ admits a non-negative density with respect to the Lebesgue measure on $\mathbb{R}^n$, so that

$$f_{\mathbf{X}}(\mathrm{x}) = \nabla_{\mathrm{x}} \mathbf{F}_{\mathbf{X}}(\mathrm{x})$$

is well defined for all $\mathrm{x} \in \mathbb{R}^n$ and satisfies $\int_{\mathbb{R}^n} f_{\mathbf{X}}(\mathrm{x})\mathrm{dx} = 1$. For each $i \in \{1, \ldots, n\}$, the marginal density function $f_i : \mathbb{R} \to \mathbb{R}_+$ is defined as

$$f_i(x_i) := \int_{\mathbb{R}^{n-1}} f(y_1, \ldots, y_{i-1}, x_i, y_{i+1}, \ldots, y_n)\mathrm{d}y_1 \cdots \mathrm{d}y_{i-1}\mathrm{d}y_{i+1} \cdots \mathrm{d}y_n.$$

We shall say that the random components of $\mathbf{X}$ are independent if

$$f_{\mathbf{X}}(\mathrm{x}) = \prod_{i=1}^{n} f_i(x_i), \qquad \text{for any } \mathrm{x} \in \mathbb{R}^n.$$

In the discrete case, the random vector $\mathbf{X}$ takes a finite number of values $(\mathrm{x}^1, \ldots, \mathrm{x}^m)$ for some integer m, and the marginal law of X_i is therefore given by

$$\mathbb{P}(X_i = \mathrm{x}_i^j) = \sum_{k_1,\ldots,k_{i-1},k_{i+1},\ldots,k_n} \mathbb{P}\left(X_1 = \mathrm{x}_1^{k_1}, \ldots, X_{i-1} = \mathrm{x}_{i-1}^{k_{i-1}}, X_i = \mathrm{x}_i^j, X_{i+1} = \mathrm{x}_{i+1}^{k_{i+1}}, \ldots, X_n = \mathrm{x}_n^{k_n}\right).$$

Remark 2.3.1. The marginal laws do not fully determine the joint law. Consider for example the following two functions:

$$f(x,y) := \frac{1}{2\pi}\exp\left(-\frac{x^2+y^2}{2}\right) \qquad \text{and} \qquad g(x,y) := \frac{1+xy\mathbb{1}_{[-1,1]}(x)\mathbb{1}_{[-1,1]}(y)}{2\pi}\exp\left(-\frac{x^2+y^2}{2}\right).$$

Show that they are both genuine two-dimensional density functions and that their marginals are all Gaussian.

Let now $\mathbf{X}$ and $\mathbf{Y}$ be two random vectors in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, admitting a joint marginal density $f_{\mathbf{X},\mathbf{Y}}$. The conditional density of $\mathbf{Y}$ with respect to $\mathbf{X}$ is defined as

$$f_{\mathbf{Y}|\mathbf{X}}(\mathrm{y}|\mathrm{x}) := \begin{cases} \dfrac{f_{\mathbf{Y},\mathbf{X}}(\mathrm{y},\mathrm{x})}{f_{\mathbf{X}}(\mathrm{x})} & \text{if } f_{\mathbf{X}}(\mathrm{x}) \neq 0, \\ f_{\mathbf{Y}}(\mathrm{y}) & \text{if } f_{\mathbf{X}}(\mathrm{x}) = 0. \end{cases}$$

Assume now that $\mathbf{X}$ admits a differentiable probability distribution function with density $f_{\mathbf{X}}$, and define $\mathbf{Y} := g(\mathbf{X})$ for some function $g \in \mathcal{C}^1(\mathbb{R}^n \to \mathbb{R}^n)$. Then $\mathbf{Y}$ admits a density function $f_{\mathbf{Y}}$ given by

$$f_{\mathbf{Y}}(\mathrm{y}) = f_{\mathbf{X}}\left(g^{-1}(\mathrm{y})\right)\left|\det\left(\nabla_{\mathrm{y}}(g^{-1}(\mathrm{y})\right)\right|,$$

where $\nabla_{\mathrm{y}}(h(\mathrm{y})) = (\partial_{y_j}h_i(\mathrm{y}))_{1\leq i,j\leq n}$ is the Jacobian matrix, and where the inverse function theorem gives $\partial_{\mathrm{y}}g^{-1}(\mathrm{y}) = 1/\partial_{\mathrm{x}}g(\mathrm{x})$.

Example 2.3.2. Given the random vector $\mathbf{X} \in \mathbb{R}^n$, which admits a smooth density, define $\mathbf{Y} := \mathbf{A}\mathbf{X} + \mathrm{b}$, where $\mathbf{A} \in \mathcal{M}_n$ is invertible, and $\mathrm{b} \in \mathbb{R}^n$. Then $\mathbf{Y}$ admits a density and

$$f_{\mathbf{X}}(\mathrm{x}) = \frac{f_{\mathbf{Y}}(\mathbf{A}^{-1}(\mathrm{x}-\mathrm{b}))}{|\det(\mathbf{A})|} \qquad \text{for all } \mathrm{x} \in \mathbb{R}^n.$$

2.3.2 Mean, covariance and correlation matrices

Whenever it exists the moment of order p is defined as

$$\mathbb{E}\left(\mathbf{X}^p\right) := \begin{pmatrix} \mathbb{E}(X_1^p) \\ \vdots \\ \mathbb{E}(X_n^p) \end{pmatrix} \in \mathbb{R}^n.$$

The second moment, whenever it exists, will also play a fundamental role later, and is defined as

$$\mathbb{E}\left(\mathbf{X}\mathbf{X}^\top\right) := \Big(\mathbb{E}(X_iX_j)\Big)_{1\leq i,j\leq n} \in \mathcal{M}_n.$$

Proposition 2.3.3. *Whenever it exists, the matrix $\mathbb{E}\left(\mathbf{X}\mathbf{X}^\top\right)$ is symmetric positive semi-definite.*

Proof. For any $\mathrm{u} \in \mathbb{R}^n$, we can write

$$\mathrm{u}^\top\mathbb{E}\left(\mathbf{X}\mathbf{X}^\top\right)\mathrm{u} = \mathbb{E}\left(\mathrm{u}^\top\mathbf{X}\mathbf{X}^\top\mathrm{u}\right) = \mathbb{E}\left(\|\mathbf{X}^\top\mathrm{u}\|^2\right) \geq 0.$$

Furthermore, $\mathbb{E}(\mathbf{X}\mathbf{X}^\top) > 0$ unless there exists $\mathrm{u} \in \mathbb{R}^n$ such that $\mathbb{P}(\mathrm{u}^\top\mathbf{X} = 0) = 1$. □

Definition 2.3.4. For any $\mathbf{X} \in \mathbb{R}^m$ and $\mathbf{Y} \in \mathbb{R}^n$, the matrix

$$\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) := \mathbb{E}\left((\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{Y} - \mathbb{E}(\mathbf{Y}))^\top\right) = \left(\mathrm{Cov}(X_i, Y_j)\right)_{1\leq i\leq m, 1\leq j\leq n} \in \mathcal{M}_{m,n}$$

is called the covariance matrix of $\mathbf{X}$ and $\mathbf{Y}$. The variance-covariance matrix of $\mathbf{X}$ is defined as $\mathbb{V}(\mathbf{X}) := \mathrm{Cov}(\mathbf{X}, \mathbf{X})$. Furthermore, the correlation matrix between $\mathbf{X}$ and $\mathbf{Y}$ is defined as

$$\mathrm{Corr}(\mathbf{X}, \mathbf{Y}) := \left(\frac{\mathrm{Cov}(X_i, Y_j)}{\sqrt{\mathbb{V}(X_i)\mathbb{V}(Y_j)}}\right)_{1\leq i\leq m, 1\leq j\leq n} \in \mathcal{M}_{m,n}.$$

The following properties are easy to prove and are left as an exercise:

Proposition 2.3.5. *Let $\mathbf{X} \in \mathbb{R}^m$, $\mathbf{Z} \in \mathbb{R}^m$, $\mathbf{Y} \in \mathbb{R}^n$ be random vectors. Show that*

- $\mathrm{Cov}(\mathbf{X}, \mathbf{X}) = \mathbb{V}(\mathbf{X})$*;*
- $\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = \mathrm{Cov}(\mathbf{Y}, \mathbf{X})^\top$*;*
- $\mathrm{Cov}(\mathbf{X} + \mathbf{Z}, \mathbf{Y}) = \mathrm{Cov}(\mathbf{X}, \mathbf{Y}) + \mathrm{Cov}(\mathbf{Z}, \mathbf{Y})$*;*
- $\mathbb{V}(\mathbf{X} + \mathbf{Z}) = \mathbb{V}(\mathbf{X}) + \mathbb{V}(\mathbf{Z}) + \mathrm{Cov}(\mathbf{X}, \mathbf{Z}) + \mathrm{Cov}(\mathbf{Z}, \mathbf{X})$*;*
- $\mathrm{Cov}(\mathbf{AX}, \mathbf{BY}) = \mathbf{A}\mathrm{Cov}(\mathbf{X}, \mathbf{Y})\mathbf{B}^\top$*, for any* $\mathbf{A} \in \mathcal{M}_{p,m}$*,* $\mathbf{B} \in \mathcal{M}_{q,n}$*.*

Two random vectors $\mathbf{X} \in \mathbb{R}^m$ and $\mathbf{Y} \in \mathbb{R}^n$ with finite second moments are said to be uncorrelated if $\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbf{O}_{m,n}$. If the two vectors are independent, then

$$\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbb{E}\left((\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{Y} - \mathbb{E}(\mathbf{Y}))^\top\right) = (\mathbb{E}(\mathbf{X}) - \mathbb{E}(\mathbf{X}))(\mathbb{E}(\mathbf{Y}) - \mathbb{E}(\mathbf{Y}))^\top = \mathbf{O}_{m,n},$$

hence they are uncorrelated. The converse is not necessarily true, however, as can be seen in Exercise 8 below. We finish this reminder on multivariate computations with the following simple statement:

Lemma 2.3.6. *If $\mathbf{X} \in \mathbb{R}^n$ is a random vector with finite second moment, then, for any* $\mathrm{u} \in \mathbb{R}^n$ *and* $\mathbf{A} \in \mathcal{M}_{m,n}$*, we have the identities*

$$\mathbb{E}(\mathrm{u} + \mathbf{AX}) = \mathrm{u} + \mathbf{A}\mathbb{E}(\mathbf{X}) \qquad \textit{and} \qquad \mathbb{V}(\mathrm{u} + \mathbf{AX}) = \mathbf{A}\mathbb{V}(\mathbf{X})\mathbf{A}^\top.$$

Exercise 8.

- Consider the one-dimensional case $X \sim \mathcal{N}(0, 1)$ and $Y := X^2$. Is the knowledge of the covariance enough to conclude about independence here?
- Prove that the correlation coefficient always lies in $[-1, 1]$;
- Prove the identity $\mathbb{V}(Y) = \mathbb{V}(\mathbb{E}(Y|X)) + \mathbb{E}(\mathbb{V}(Y|X))$;

2.3.3 Forecasting

The goal of this short section is not to have a full overview of forecasting, but only to show how conditional expectations enter as optimal (in some sense) forecasting tools. For two square integrable random vectors $\mathbf{X} \in \mathbb{R}^m$ and $\mathbf{Y} \in \mathbb{R}^n$, we understand $\mathbf{X}$ as observed data, and we wish to obtain some estimates for the unknown $\mathbf{Y}$.

Definition 2.3.7. The random vector $G(\mathbf{X})$, for some function $G : \mathbb{R}^m \to \mathbb{R}^n$ is called best forecast if

$$\mathbb{E}\left((\mathbf{Y} - G(\mathbf{X}))(\mathbf{Y} - G(\mathbf{X}))^\top\right) \le \mathbb{E}\left((\mathbf{Y} - H(\mathbf{X}))(\mathbf{Y} - H(\mathbf{X}))^\top\right),$$

holds for any function $H : \mathbb{R}^m \to \mathbb{R}^n$.

The following result is simple to prove, but provides a fundamental understanding of conditional expectation as an optimal projection operator.

Theorem 2.3.8. *If the joint law of $\mathbf{X}$ and $\mathbf{Y}$ admits a density, then $G(\mathbf{X}) = \mathbb{E}(\mathbf{Y}|\mathbf{X})$.*

It will often happen, at least as first approximations, that the random variables under consideration are Gaussian. We therefore need to be able to compute those conditional expectations and variances.

Theorem 2.3.9. *Let $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma} \in \mathcal{M}^+_{n,n}$, with the decomposition*

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

where $\mathbf{X}_1 \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$, $\mathbf{X}_2 \sim \mathcal{N}_q(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$ and $p + q = n$. With $\Theta := \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}$, the random variables $\mathbf{X}_1$ and $\mathbf{X}_2 - \Theta\mathbf{X}_1$ are independent, and, almost surely,

$$\mathbb{E}(\mathbf{X}_2|\mathbf{X}_1) = \boldsymbol{\mu}_2 + \Theta\left(\mathbf{X}_1 - \boldsymbol{\mu}_1\right) \qquad \text{and} \qquad \mathbb{V}(\mathbf{X}_2|\mathbf{X}_1) = \boldsymbol{\Sigma}_{22} - \Theta\boldsymbol{\Sigma}_{12}.$$

2.4 Multivariate distributions

2.4.1 A detailed example: the multinormal distribution

The Gaussian distribution is ubiquitous in Probability, Statistics and applications, and hence deserve a dedicated treatment. We start with the easy one-dimensional case, stating and proving a certain number of its properties, before delving into the multivariate case.

The univariate case

Definition 2.4.1. A real-valued random variable X is called standard Gaussian, and we write $X \sim \mathcal{N}(0,1)$ if its probability distribution reads, for all $x \in \mathbb{R}$,

$$\mathbb{P}(X \in \mathrm{d}x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x.$$

The following representation of its characteristic function is left as an exercise:

Proposition 2.4.2. *The characteristic function of $X \sim \mathcal{N}(0,1)$ is given by*

$$\phi_X(z) := \mathbb{E}\left(\mathrm{e}^{\mathtt{i}zX}\right) = \exp\left(-\frac{z^2}{2}\right), \qquad \textit{for all } z \in \mathbb{R}.$$

Proof. Since the density of X is known in closed form, we can write, for any $z \in \mathbb{R}$,

$$\phi_X(z) = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \exp\left\{\mathtt{i}uz - \frac{u^2}{2}\right\}\mathrm{d}u = \exp\left(-\frac{z^2}{2}\right)\left(\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}-\mathtt{i}z}\exp\left\{-\frac{u^2}{2}\right\}\mathrm{d}u\right).$$

Since the map $z \mapsto \exp(-z^2/2)$ is analytic on $\mathbb{R}$, Cauchy's theorem shows that

$$\left(\int_{-r-\mathtt{i}z}^{r-\mathtt{i}z} + \int_{r-\mathtt{i}z}^{r} + \int_{r}^{-r} + \int_{-r}^{-r-\mathtt{i}z}\right)\exp\left\{-\frac{z^2}{2}\right\}\mathrm{d}z = 0.$$

This identity allows us to write

$$\begin{aligned}
\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}-\mathtt{i}z}\exp\left\{-\frac{u^2}{2}\right\}\mathrm{d}u - 1 &= \frac{1}{\sqrt{2\pi}}\left[\left(\int_{\mathbb{R}-\mathtt{i}z} - \int_{\mathbb{R}}\right)\exp\left\{-\frac{u^2}{2}\right\}\mathrm{d}u\right]\\
&= \lim_{r\uparrow\infty}\left(\int_{-r-\mathtt{i}z}^{r-\mathtt{i}z} + \int_{r}^{-r}\right)\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{u^2}{2}\right\}\mathrm{d}u\\
&= \lim_{r\uparrow\infty}\left(\int_{r}^{r-\mathtt{i}z} + \int_{-r-\mathtt{i}z}^{-r}\right)\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{u^2}{2}\right\}\mathrm{d}u.
\end{aligned}$$

Since $|\exp(-z^2/2)| = \exp(-\Re(z^2)/2)$, then this limit is equal to zero, and the proposition follows. □

Proposition 2.4.3. *Let $X \sim \mathcal{N}(0,1)$. Then all moments exists,*

$$\mathbb{E}\left(X^p\right) = \begin{cases} 0 & \textit{if } p \textit{ is odd},\\ \dfrac{p!}{2^{p/2}(p/2)!} & \textit{if } p \textit{ is even},\end{cases}$$

and

$$\mathbb{E}\left(|X|^p\right) = 2^{p/2}\frac{\Gamma((p+1)/2)}{\Gamma(1/2)}, \qquad \textit{for all } p \geq 0.$$

Proof. Let p be even so that we can write $p = 2n$. Then

$$\begin{aligned}
\mathbb{E}(X^{2n}) &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} x^{2n}\exp\left(-\frac{x^2}{2}\right)\mathrm{d}x = \sqrt{\frac{2}{\pi}}\int_0^\infty x^{2n}\exp\left(-\frac{x^2}{2}\right)\mathrm{d}x\\
&= \frac{2^n}{\sqrt{\pi}}\int_{\mathbb{R}} z^{n-1/2}\mathrm{e}^{-z}\mathrm{d}z = \frac{2^n}{\sqrt{\pi}}\Gamma\left(n+\frac{1}{2}\right),
\end{aligned}$$

and the result follows from the fact that $\Gamma(1/2) = \sqrt{\pi}$ and the recursion $\Gamma(n+1) = n\Gamma(n)$. The proof for the absolute moments is similar and left as an exercise. □

Gaussian random variables satisfy the following useful property:

Proposition 2.4.4. *Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \in \mathbb{R}^n$ be Gaussian random vector with independent components. Then, for any $\mathrm{u} \in \mathbb{R}^n$, the sum $\boldsymbol{S} := \mathrm{u}^\top\mathbf{X}$ is also Gaussian with*

$$\mathbb{E}(\boldsymbol{S}) = \mathrm{u}^\top\boldsymbol{\mu} \qquad \textit{and} \qquad \mathbb{V}(\boldsymbol{S}) = \mathcal{Q}_\Sigma(\mathrm{u}).$$

The matrix case

We now extend the results from the univariate case above to the more interesting multi-dimensional case. We denote by $\mathcal{N}(\mathbf{O}_n, \mathbf{I}_n)$ the Gaussian random vector $\mathbf{Y} = (Y_1 \ldots, Y_n)$, where each Y_i is a univariate centered Gaussian random variable with unit variance. More generally, we define a Gaussian vector as follows:

Definition 2.4.5. Let $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\Sigma \in \mathcal{M}_n^+$ such that $\Sigma = \mathbf{T}^\top \mathbf{T}$ (by Proposition 2.1.17). The vector $\mathbf{X}$ is said to follow a Gaussian random distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix Σ, and we write $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ if the equality $\mathbf{X} = \boldsymbol{\mu} + \mathbf{T}^\top \mathcal{N}(\mathbf{O}_n, \mathbf{I}_n)$ holds in distribution.

A simple way to understand this is to start from a random vector $\mathbf{Z} = (Z_1, \ldots, Z_n)$, constituting an iid sequence of standard Gaussian distributions. Its joint density then reads

$$f_{\mathbf{Z}(\mathrm{z})} = \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{\|\mathrm{z}\|^2}{2}\right\}, \qquad \text{for any } \mathrm{z} \in \mathbb{R}^n.$$

Define now the random vector $\mathbf{X} := \boldsymbol{\mu} + \mathbf{T}^\top \mathbf{Z}$, where $\mu \in \mathbb{R}^n$, and $\mathbf{T} \in \mathcal{M}_n(\mathbb{R})$ a matrix of rank k. If $k < n$, then $\mathbf{Z}$ is said to have a singular multivariate Gaussian distribution. If $k = n$, then $\mathbf{T}$ has full rank and $\Sigma := \mathbf{T}^\top \mathbf{T}$ is positive definite and $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$.

Proposition 2.4.6. *Let* $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$*. Then* $\mathbb{E}(X) = \boldsymbol{\mu}$*,* $\mathbb{V}(X) = \Sigma$*, and*

$$\phi_X(\mathrm{u}) := \mathbb{E}\left(\mathrm{e}^{\mathrm{i}\mathrm{u}^\top X}\right) = \exp\left\{\mathtt{i}\mathrm{u}^\top \boldsymbol{\mu} - \frac{1}{2}\mathrm{u}^\top \Sigma \mathrm{u}\right\}, \qquad \textit{for all } \mathrm{u} \in \mathbb{R}^n.$$

If furthermore $\Sigma \in \mathcal{M}_n^{++}$*, then* X *admits a density which reads*

$$\mathbb{P}(X \in \mathrm{dx}) = \frac{\mathrm{dx}}{(2\pi)^{n/2} \det(\Sigma)^{1/2}} \exp\left\{-\frac{1}{2}(\mathrm{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathrm{x} - \boldsymbol{\mu})\right\}, \qquad \textit{for all } \mathrm{x} \in \mathbb{R}^n.$$

Exercise 9. Let $\mathbf{X} \in \mathcal{M}_n$. Prove the following properties:

- For any $\mathbf{A} \in \mathcal{M}_{m,n}$, $\mathrm{u} \in \mathbb{R}^m$, then $\mathbf{A}\mathcal{N}(\boldsymbol{\mu}, \Sigma) + \mathrm{u} \stackrel{\Delta}{=} \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathrm{u}, \mathbf{A}\Sigma\mathbf{A}^\top)$;
- for any orthogonal matrix $\mathbf{A} \in \mathcal{M}_n$, $\mathbf{A}\mathcal{N}(\mathbf{O}_n, \mathbf{I}_n) \stackrel{\Delta}{=} \mathcal{N}(\mathbf{O}_n, \mathbf{I}_n)$;

In the case of Gaussian random vectors, independence can be characterised simply through the variance-covariance matrix:

Proposition 2.4.7. *The components of* $\mathbf{X} \stackrel{\Delta}{=} \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ *are independent if and only if* Σ *is diagonal.*

Proof. The vectors $X_1, \ldots, X_n$ are independent if and only if the equality

$$\mathbb{E}\left(\mathrm{e}^{\mathrm{i}\mathrm{u}^\top \mathbf{X}}\right) = \prod_{i=1}^n \mathbb{E}\left(\mathrm{e}^{\mathtt{i}u_i X_i}\right)$$

holds for all $\mathrm{u} \in \mathbb{R}^n$. We leave it to the reader to check this is indeed the case. □

2.4.2 Other useful distributions

Chi Square Distribution

Definition 2.4.8. Let $X_1, \ldots, X_n$ for an iid sequence of centered Gaussian distributions with unit variance. Then the law of $S_n := \sum_{i=1}^n X_i^2$ is called the χ^2 distribution with n degrees of freedom, and we write $S_n \sim \chi_n^2$.

It is easy to prove in particular that $\mathbb{E}(S_n) = n$ and $\mathbb{V}(S_n) = 2n$, that it admits a density

$$f_{S_n}(x) = \frac{x^{n/2-1}\mathrm{e}^{-x/2}}{2^{n/2}\Gamma(n/2)}, \qquad \text{for all } x \geq 0,$$

where $\Gamma(u) := \int_0^\infty z^{u-1}\mathrm{e}^{-z}\mathrm{d}z$ is the Gamma function, and its moment generating function reads

$$\mathbb{E}\left(\mathrm{e}^{uS_n}\right) = (1-2u)^{-n/2}, \qquad \text{for all } u < \frac{1}{2}.$$

Student Distribution

Definition 2.4.9. If $S_n \sim \chi_n^2$ for some integer n and $Z \in \mathcal{N}(0,1)$, then the ratio $T_n := \frac{Z}{\sqrt{S_n/n}}$ is called a Student distribution with n degrees of freedom, and we write $T_n \sim \mathcal{T}_n$.

One can show that its density reads

$$f_{T_n}(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}, \qquad \text{for all } x \in \mathbb{R}.$$

The expectation is finite if and only if $n > 1$, in which case $\mathbb{E}(T_n) = 0$. Likewise, the variance is finite if and only if $n > 2$, in which case $\mathbb{V}(T_n) = n/(n-2)$. The moment generating function, however, is always undefined.

Wishart distribution

We introduced above the χ^2 distribution, as a sum of squared iid Gaussian distributions. Its extension to the multivariate case is called the Wishart distribution, and will be fundamental in the study of estimators for covariance matrices.

Definition 2.4.10 (Wishart Distribution)**.** If $\mathbf{X}_1, \ldots, \mathbf{X}_n$ forms a sequence of $\mathbb{R}^p$-valued independent $\mathcal{N}(0, \mathbf{\Sigma})$ distributions, then the random matrix $\mathbf{W} := \sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top$ is called a Wishart distribution, denoted by $\mathbf{W}_p(\mathbf{\Sigma}, n)$.

We shall not dive into any details of this distribution here, but simply note that its density and characteristic function are available in closed form.

2.5 Application: Markowitz and CAPM

We now show how to apply these tools from multivariate analysis in order to solve the so-called Markowitz [1] efficient frontier problem. We consider n assets, and denote by $\mathbf{X} = (X_1, \ldots, X_n)$ the vector of returns over a given period. Following earlier notations, the mean and covariance read

$$\boldsymbol{\mu} := (\mu_1, \ldots, \mu_n) = (\mathbb{E}(X_i))_{i=1,\ldots,n} \qquad \text{and} \qquad \mathrm{Cov}(\mathbf{X}) =: \Sigma \in \mathcal{M}_{n,n}.$$

For a vector $\mathbf{w} \in \mathbb{R}^n$ of weights satisfying $\mathbf{w}^\top \mathbb{1}_n = 1$, we define the portfolio of returns $\Pi = \mathbf{w}^\top \mathbf{X}$, with mean $\mathbb{E}(\Pi) = \mathbf{w}^\top \boldsymbol{\mu}$, built by investing a share of $\mathbf{w}_i$ in asset i, for $i = 1, \ldots, n$. Markowitz' optimal portfolio is then defined as the solution to the following quadratic problem:

$$\min \left\{ \frac{1}{2} \mathcal{Q}_\Sigma(\mathbf{w}), \text{ such that } \mathbf{w}^\top \boldsymbol{\mu} = \widetilde{\mu}, \mathbf{w}^\top \mathbb{1}_n = 1 \right\}, \tag{2.5.1}$$

where $\mathcal{Q}$ is the quadratic form introduced in Section 2.1.3, and $\widetilde{\mu}$ some fixed target return. The coefficient $\frac{1}{2}$ is introduced here purely for technical reasons. Note that we did not impose that the weights should be non-negative, which is financially equivalent to allowing short-selling. This optimisation problem is quadratic, hence convex, and can be solved efficiently using convex optimisation tools. We adopt here a much simpler approach, based on the multivariate tools analysed above. The Lagrangian of the problem reads

$$\mathcal{L}(\mathbf{w}, \lambda_1, \lambda_2) := \frac{1}{2} \mathcal{Q}_\Sigma(\mathbf{w}) + \lambda_1 \left(\widetilde{\mu} - \mathbf{w}^\top \boldsymbol{\mu} \right) + \lambda_2 \left(1 - \mathbf{w}^\top \mathbb{1}_n \right).$$

The first-order conditions read

$$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \lambda_1, \lambda_2) = \Sigma^\top \mathbf{w} - \lambda_1 \boldsymbol{\mu} - \lambda_2 \mathbb{1}_n = 0,$$

since the covariance matrix Σ is symmetric. If it is also invertible, we can solve this equation as

$$\mathbf{w} = \Sigma^{-1} \left(\lambda_1 \boldsymbol{\mu} + \lambda_2 \mathbb{1}_n \right). \tag{2.5.2}$$

Recalling the constraints $\mathbb{1}_n^\top \mathbf{w} = 1$, we can pre-multiply the above by $\mathbb{1}_n^\top$ to obtain

$$\lambda_2 = \frac{1 - \lambda_1 \mathbb{1}_n^\top \Sigma^{-1} \boldsymbol{\mu}}{\mathbb{1}_n^\top \Sigma^{-1} \mathbb{1}_n}.$$

Plugging this optimal Lagrange multiplier into (2.5.2) yields

$$\mathbf{w}^* = \frac{\Sigma^{-1} \mathbb{1}_n}{\mathbb{1}_n^\top \Sigma^{-1} \mathbb{1}_n} + \lambda_1 \Sigma^{-1} \left(\boldsymbol{\mu} - \frac{\mathbb{1}_n^\top \Sigma^{-1} \boldsymbol{\mu}}{\mathbb{1}_n^\top \Sigma^{-1} \mathbb{1}_n} \mathbb{1}_n \right).$$

If we are only interested in variance efficient portfolio, then there is no constraint on target returns, i.e. $\lambda_1 = 0$, and hence

$$\mathbf{w}^* = \frac{\Sigma^{-1} \mathbb{1}_n}{\mathbb{1}_n^\top \Sigma^{-1} \mathbb{1}_n}.$$

[1] Harry Markowitz, born in 1927, won the Nobel Prize in Economics in 1990.

With this optimal weight, the portfolio has expectation and variance-covariance matrix

$$\mathbb{E}(\Pi) = (\mathbf{w}^*)^\top \boldsymbol{\mu} \qquad \text{and} \qquad \mathrm{Cov}(\Pi) = \mathcal{Q}_\Sigma(\mathbf{w}^*).$$

We allowed above for short-selling, so that the weights could be negative. If we impose positive of the weights, then the optimisation problem (2.5.1) transforms into

$$\min\left\{\frac{1}{2}\mathcal{Q}_\Sigma(\mathbf{w}), \text{ such that } \mathbf{w} \geq \mathbf{0}, \mathbf{w}^\top \boldsymbol{\mu} = \widetilde{\boldsymbol{\mu}}, \mathbf{w}^\top \mathbb{1}_n = 1\right\}.$$

Unfortunately, in this case, no closed-form solution exist, but the problem can easily be solved using quadratic programming principles.

IPython notebook Markowitz_Quadratic

The Capital Asset Pricing Model was introduced by Sharpe [32] and Lintner [28] on top of Markowitz' portfolio theory. Besides n risky assets available on the market, there exists a risk-free asset with lending and borrowing rate equal to r_f. The efficient frontier is defined as the line tangent to Markowitz' feasible region that goes through the point $(0, r_f)$. The one-fund theorem states that there exists only one contact point (σ_M, r_M) (called the market portfolio) between the efficient frontier and the Markowitz optima. Any point on the segment between $(0, r_f)$ and (σ_M, r_M) defines a portfolio consisting of the risk-free asset and the market portfolio. For a target expected return μ^*, the optimisation problem therefore reads

$$\min\left\{\frac{1}{2}\mathcal{Q}_\Sigma(\mathbf{w}), \text{ such that } \mathbf{w}^\top \boldsymbol{\mu} + (1 - \mathbf{w}^\top \mathbb{1}_n) r_f = \mu^*\right\}.$$

This is almost the same as (2.5.1), except that the weights do not have to sum up to one, since the remaining part not invested in the Markowitz portfolio can be invested in the risk-free asset. The Lagrangian reads

$$\mathcal{L}(\mathbf{w}, \lambda) := \frac{1}{2}\mathcal{Q}_\Sigma(\mathbf{w}) + \lambda\left(\mathbf{w}^\top \boldsymbol{\mu} + (1 - \mathbf{w}^\top \mathbb{1}_n) r_f - \mu^*\right).$$

The first-order conditions read

$$\begin{cases} \nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}, \lambda) & = \Sigma^\top \mathbf{w} + \lambda\left(\boldsymbol{\mu} - \mathbb{1}_n r_f\right) & = 0, \\ \partial_\lambda \mathcal{L}(\mathbf{w}, \lambda) & = \mathbf{w}^\top \boldsymbol{\mu} + (1 - \mathbf{w}^\top \mathbb{1}_n) r_f - \mu^* & = 0. \end{cases}$$

If the covariance matrix Σ is invertible, this equation can be solved as

$$\mathbf{w} = \frac{(\mu^* - r_f)\Sigma^{-1}\left(\boldsymbol{\mu} - r_f \mathbb{1}_n\right)}{\left(\boldsymbol{\mu} - r_f \mathbb{1}_n\right)^\top \Sigma^{-1}\left(\boldsymbol{\mu} - r_f \mathbb{1}_n\right)}.$$

Consider a portfolio whose returns have mean μ and variance σ^2. The capital market line joins $(0, r_f)$ to (σ, μ), and we can write it as

$$\mu = r_f + \frac{\mu_M - r_f}{\sigma_M}\sigma.$$

The coefficient $\frac{\mu_M - r_f}{\sigma_M}$ is called the Sharpe ratio and is the same for any efficient portfolio (in particular for the market portfolio).

Chapter 3

Statistical inference

In this part of the lectures, we will be interested in building tools to analyse data directly. The sequence $\mathcal{X}_n = (X_1, \ldots, X_n)$ is the sample data we observe, and which we want to explain. The fundamental hypothesis underlying statistical methods is that the observed sample $\mathcal{X}_n$ represents independent and identically distributed (iid) observations of some random variable X which we wish to describe.

Consider for example the evolution of the S&P500 between January 1st, 1986 and August 31st 2017, and let us call s_i its price on day i, for $i = 1, \ldots, n$, with n the number of trading days over the period; here $n = 7983$. Define the daily log-returns[1] $x_i := \log(s_i/s_{i-1})$, for $i = 1, \ldots, n-1$. Relabelling the index so that the sample reads $\mathcal{X}_n$ is now of size $n = 7982$, we can plot both the time series of the returns as well as their empirical distribution. Statistics' aim is to infer from these plots a distribution, or a model, describing the sample $\mathcal{X}_n$ of returns. If one assumes that the returns are Gaussian, i.e. $\mathcal{X}_n$ is the realisation of some $\mathcal{N}(\mu, \sigma^2)$ random variable, then the histogram should correspond (more or less) to the Gaussian density.

Exercise 10. There are two clear drops in the SPX evolution. What do they correspond to? What kind of observations can we make from the evolution of the returns?

3.1 Estimating statistical functionals

In Figure 3.1, we plotted the empirical distribution of the returns of the S&P500 over a given period. The first question one should ask is how it is in fact plotted; the second one, in order to be able to build some model, is to determine the shape/characteristics of this distribution.

[1] One may wonder why we consider logarithmic returns. Suppose that we were to consider returns of the form $\frac{s_i - s_{i-1}}{s_{i-1}} = \frac{s_i}{s_{i-1}} - 1$, which is equal—up to a second-order error—to the logarithmic returns.

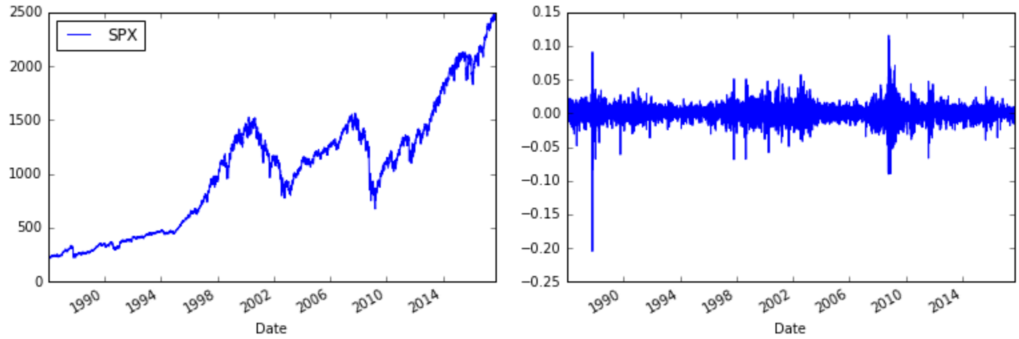

Figure 3.1: Time series of S&P 500 and its returns between 1/1/1986 and 31/12/2017.

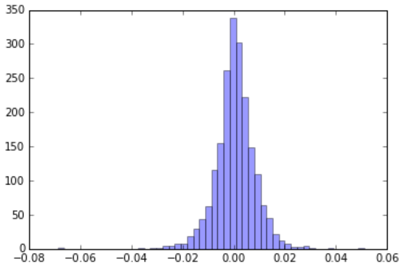

Figure 3.2: Empirical distribution of the SPX returns over the period from 1/1/1986 to 31/12/2017.

Definition 3.1.1. The empirical cumulative distribution function of the sample $\mathcal{X}_n$ is defined as

$$\widehat{F}_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \le x\}}, \qquad \text{for all } x \in \mathbb{R}.$$

Given a sample $\mathcal{X}_n$, the function $\widehat{F}_n$ is piecewise constant, right continuous, with jump sizes equal to $1/n$ and such that

$$\lim_{x \downarrow -\infty} \widehat{F}_n(x) = 0 \qquad \text{and} \qquad \lim_{x \uparrow \infty} \widehat{F}_n(x) = 1.$$

We expect that, as the sample grows larger, the empirical distribution becomes smoother (note that the size of the jumps become smaller), as Figure 3.3 shows. Now, for fixed x, $\widehat{F}_n(x)$ is a Binomial random variable, and hence, for any x and any n,

$$\mathbb{E}\left[\widehat{F}_n(x)\right] = F(x) \qquad \text{and} \qquad \mathbb{V}\left[\widehat{F}_n(x)\right] = \frac{F(x)(1-F(x))}{n},$$

and we have almost sure convergence by the strong law of large numbers (Theorem 2.2.19). A more precise version (uniform convergence) of this observation is the following:

Theorem 3.1.2 (Glivenko-Cantelli Theorem)**.** *If $X_1, \ldots, X_n$ are iid with common cdf F, then*

$$\lim_{n \uparrow \infty} \left\|\widehat{F}_n - F\right\|_\infty = \lim_{n \uparrow \infty} \sup_{x \in \mathbb{R}} \left|\widehat{F}_n(x) - F(x)\right| = 0 \quad \textit{almost surely.}$$

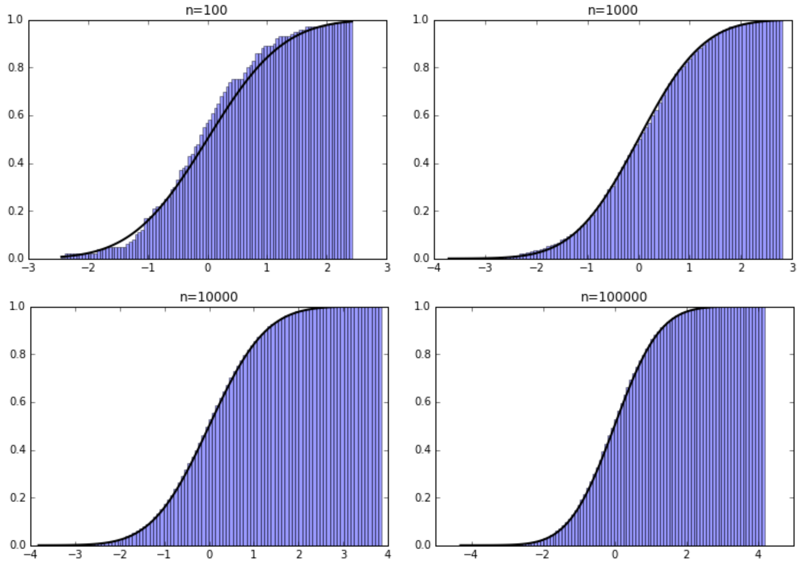

Figure 3.3: Empirical cdf of a Gaussian $\mathcal{N}(0,1)$ sample for different values of the sample size n, together with the exact Gaussian cdf (line).

Proof. Consider the simpler case where the function F is continuous. In that case, for any integer k, we can find a sequence $-\infty = x_0 < x_1 < \cdots < x_{k-1} < x_k = +\infty$ such that $F(x_i) = i/k$. Now, for any $x \in [x_{i-1}, x_i]$, the monotonicity of both $\widehat{F}_n$ and F imply

$$\widehat{F}_n(x_{i-1}) - F(x_{i-1}) - \frac{1}{k} = \widehat{F}_n(x_{i-1}) - F(x_i) \leq \widehat{F}_n(x) - F(x) \leq \widehat{F}_n(x_i) - F(x_{i-1}) = \widehat{F}_n(x_i) - F(x_i) + \frac{1}{k},$$

so that

$$\left|\widehat{F}_n(x) - F(x)\right| \leq \max_{i=1,\ldots,k-1} \left\{\left|\widehat{F}_n(x_i) - F(x_i)\right| + \frac{1}{k}\right\}.$$

Since $\widehat{F}_n(x)$ converges almost surely to $F(x)$ by the strong law of large numbers (Theorem 2.2.19 applied to a sequence of iid Bernoulli trials), then

$$\limsup_{n\uparrow\infty} \sup_{x\in\mathbb{R}} \left|\widehat{F}_n(x) - F(x)\right| \leq \frac{1}{k},$$

and the theorem follows by letting k tend to infnity. □

Figures 3.3 and 3.4 show convergence of the empirical densities and cumulative distribution functions. However, at first glance, it seems that the plots of the densities are more revealing than those of the cdfs, and one may wonder whether the Glivenko-Cantelli Theorem has an analogue for empirical densities. We first need to define properly what an empirical density is. We follow the intuition arising from the plots: let $\mathrm{h} = (h_1, \ldots, h_m)$ be an ordered series of bins containing the support of $\mathcal{X}_n$, for some integer m, with $\min_{i=1,\ldots,n} X_i \leq h_1 \leq \cdots \leq \max_{i=1,\ldots,n} X_i \leq h_m$, and such that the length of each interval $h_j - h_{j-1}$ is constant equal to h. For each $j = 2, \ldots, m$, we denote by $n_j := \sum_{i=1}^n \mathbb{1}_{\{X_i \in [h_{j-1}, h_j)\}}$ the number of elements from the sample falling into the bin $[h_{j-1}, h_j)$.

Definition 3.1.3. The empirical histogram of the sample $\mathcal{X}_n$ is defined as

$$\widehat{f}_n(x) := \sum_{j=1}^m \frac{n_j}{nh} \mathbb{1}_{\{x \in [h_{j-1}, h_j)\}}, \qquad \text{for all } x \in \mathbb{R}.$$

It is easy to see that $\widehat{f}$ is non-negative and integrates to one. This function is not continuous though, and one may want to smooth it in order to analyse it more in details. We could also use a different definition. Since the density is the derivative of the cfd, a first-order approximation yields, using the empirical cdf instead of the true, unknown cdf,

$$\widehat{f}_n(x) := \frac{\widehat{F}_n(x + h/2) - \widehat{F}_n(x - h/2)}{h} = \frac{1}{nh} \sum_{i=1}^n \mathbb{1}_{\{X_i \in B_x^h\}} = \frac{1}{nh} \sum_{i=1}^n K_0 \left(\frac{x - X_i}{h} \right),$$

where B_x^h is the half-open ball centred at x with radius $h/2$, and the kernel K_0 is naturally defined as $K_0(x) := \mathbb{1}_{\{-1/2 < x \leq 1/2\}}$. This is called a moving window estimator, but is still not continuous. However, from this representation, we see that the discontinuity comes from the kernel K_0. Kernel estimators are a natural generalisation of this, using a smooth kernel instead of an indicator function. The usual one is the Gaussian kernel, whereby the empirical density is defined as

$$\widehat{f}_n(x) := \frac{1}{nh} \sum_{i=1}^n K \left(\frac{x - X_i}{h} \right), \tag{3.1.1}$$

with $K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2} \right)$ is the Gaussian density.

Exercise 11. Show that the function $\widehat{f}_n$ in (3.1.1) is a valid density function.

IPython notebook GlivenkoCantelli.ipynb

Now that we have some sort of description of the data, we need to be able to analyse it. We therefore introduce statistical estimators, as Borel functions $S(\mathcal{X}_n)$ of the sample, for example in order to estimate some parameters. Consider for example the example in Figure 3.1. If we assume the returns of the S&P500 to be distributed as a Gaussian $\mathcal{N}(\mu, \sigma^2)$ random variable, we may want to use the fact that

$$\mu = \int_{\mathbb{R}} x f_{\mathcal{N}(\mu,\sigma^2)}(x) \mathrm{d}x = \int_{\mathbb{R}} x \mathrm{d}F_{\mathcal{N}(\mu,\sigma^2)}(x),$$

and introduce the statistical estimator $\mathcal{S}(\mathcal{X}_n) := \int_{\mathbb{R}} x \mathrm{d}\widehat{F}_n(x)$ for the true mean μ. Using Definition 3.1.1, we can therefore write

$$\mathcal{S}(\mathcal{X}_n) := \int_{\mathbb{R}} x \mathrm{d} \left(\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}} \right) = \frac{1}{n} \sum_{i=1}^n X_i =: \overline{X}, \tag{3.1.2}$$

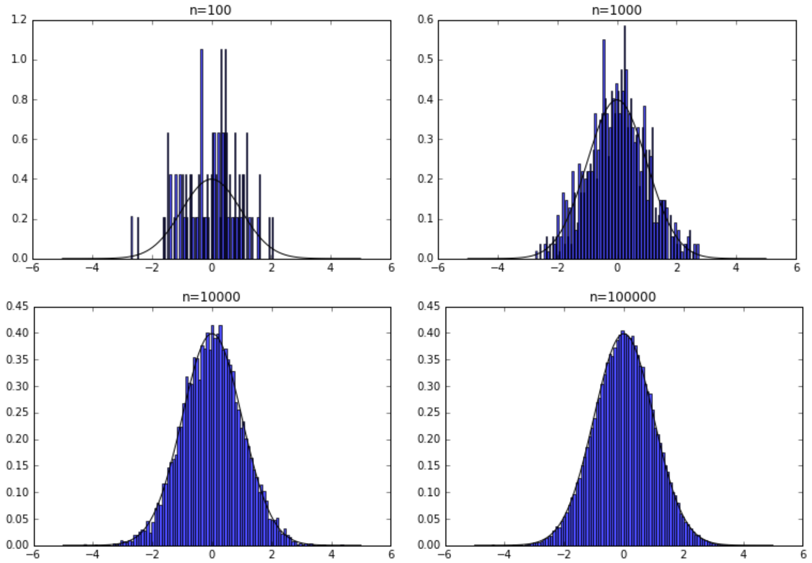

Figure 3.4: Empirical density of a Gaussian $\mathcal{N}(0,1)$ sample for different values of the sample size n.

which is nothing else than the arithmetic average. Under general assumptions on the function $\mathcal{S}$, one can prove that the Glivenko-Cantelli Theorem 3.1.2 yields convergence of $\mathcal{S}(\widehat{F}_n)$ to $\mathcal{S}(F)$ as the sample size n tends to infinity. With the function

$$\mathcal{S}(F) := \int x^2 F(\mathrm{d}x) - \left(\int xF(\mathrm{d}x)\right)^2,$$

the estimator is that of the variance, defined as

$$s_{\mathcal{X}}^2 := \frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}^2, \tag{3.1.3}$$

and $s_{\mathcal{X}}$ is called the standard deviation. Let us summarise a few properties of these two estimators:

Proposition 3.1.4. *Let $\mathcal{X} = (X_1,\ldots,X_n)$ be an iid sample with common distribution X satisfying $\mathbb{E}[X] = \mu$ and $\mathbb{V}[X^2] = \sigma^2 < \infty$, then*

$$\mathbb{E}\left[\overline{X}\right] = \mu, \qquad \mathbb{V}\left[\overline{X}\right] = \frac{\sigma^2}{n}, \qquad \mathbb{E}\left[s_{\mathcal{X}}^2\right] = \frac{n-1}{n}\sigma^2.$$

Furthermore, the sample mean $\overline{X}$ and the sample variance $s_{\mathcal{X}}^2$ are independent and converge almost surely to μ and σ^2 as n tends to infinity. Finally the random variable $ns_{\mathcal{X}}^2/\sigma^2$ is distributed as a Chi-Square distribution with $n-1$ degrees of freedom.

In order to prove this proposition, recall the following theorem, due to Cochran:

Theorem 3.1.5. *Let $\mathcal{X} := (X_1,\ldots,X_n)$ denote an iid sequence of $\mathcal{N}(0,\sigma^2)$ random variables, and assume that*

$$\sum_{i=1}^n X_i^2 = \sum_{i=i}^k Q_i,$$

where, for each $i = 1, \ldots, k$, Q_i is a positive semi-definite quadratic form in $\mathcal{X}$, i.e. $Q_i = \mathcal{X}^\top \mathbf{A}_i \mathcal{X}$ for some matrix $\mathbf{A}_i$. If $\sum_{i=1}^k \mathrm{rank}(\mathbf{A}_i) = n$ then all Q_i are independent and $Q_i \sim \sigma^2 \chi^2_{\mathrm{rank}(\mathbf{A}_i)}$.

Proof of Proposition 3.1.4. Before diving into the core of the proof, consider the following claims, for any constant α:

$$\mathbb{E}\left[(X-\mu)^2\right] = (\mathbb{E}[X]-\mu)^2 + \mathbb{V}[X], \tag{3.1.4}$$

$$\frac{1}{n}\sum_{i=1}^n (X_i-\alpha)^2 = \left(\overline{X}-\alpha\right)^2 + s^2_{\mathcal{X}}. \tag{3.1.5}$$

The first one is trivial. Regarding the second one, we can write, using (3.1.3),

$$\frac{1}{n}\sum_{i=1}^n (X_i-\alpha)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 + \alpha^2 - 2\alpha\overline{X} = \left(s^2_{\mathcal{X}} + \overline{X}^2\right) + \alpha^2 - 2\alpha\overline{X} = s^2_{\mathcal{X}} + \left(\overline{X}-\alpha\right)^2.$$

Using (3.1.5) with $\alpha = \mu$, we therefore have

$$\mathbb{E}[s^2_{\mathcal{X}}] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[(X_i-\mu)^2\right] - \mathbb{E}\left[\overline{X}-\mu\right]^2 = \mathbb{V}[X] - \mathbb{V}\left[\overline{X}\right] = \sigma^2 - \frac{\sigma^2}{n}.$$

The rest of the proof is slightly more involved. Since $\mathcal{X}^\top \sim \mathcal{N}_n(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, then the random variable $\eta := (\mathcal{X}^\top - \mathbb{E}[\mathcal{X}^\top]/\sigma$ is distributed as $\mathcal{N}_n(\mathbf{O}, \mathbf{I})$. It is easy to see that the matrix

$$\mathbf{A} := \frac{1}{n}\begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix} \in \mathcal{M}_{nn}$$

is idempotent ($\mathbf{A}^2 = \mathbf{A}$) and symmetric. Now,

$$\|\eta\|^2 = \eta^\top \mathbf{I}_n \eta = \eta^\top (\mathbf{I}-\mathbf{A})\,\eta + \eta^\top \mathbf{A}\eta =: Q_1 + Q_2.$$

Therefore

$$\eta_1 := \mathbf{A}\eta = \frac{1}{\sigma}\left(\overline{X}-\mu, \cdots, \overline{X}-\mu\right)^\top,$$

$$\eta_2 := (\mathbf{I}-\mathbf{A})\eta = \frac{1}{\sigma}(\mathbf{I}-\mathbf{A})(\mathcal{X}^\top - \mathbb{E}[\mathcal{X}^\top] = \frac{1}{\sigma}\begin{pmatrix} X_1-\mu \\ \vdots \\ X_n-\mu \end{pmatrix} - \frac{1}{\sigma}\begin{pmatrix} \overline{X}-\mu \\ \vdots \\ \overline{X}-\mu \end{pmatrix} = \frac{1}{\sigma}\begin{pmatrix} X_1-\overline{X} \\ \vdots \\ X_n-\overline{X}. \end{pmatrix}$$

Clearly $\mathrm{rank}(\mathbf{A}) = 1$ and $\mathrm{rank}(\mathbf{I}-\mathbf{A}) = n-1$, so that $\mathbf{A}$ and $\mathbf{I}-\mathbf{A}$ satisfy Cochran's hypotheses (Theorem 3.1.5) and hence η_1 and η_2 are independent. By construction, $\|\eta_2\|^2 \sim \chi^2_{n-1}$. Since

$$\|\eta_2\|^2 = \frac{1}{\sigma}\sum_{i=1}^n \left(X_i-\overline{X}\right)^2 = \frac{ns_{\mathcal{X}}}{\sigma^2},$$

the proposition follows. □

Corollary 3.1.6. *In the framework of Proposition 3.1.4, if $X \sim \mathcal{N}(\mu, \sigma^2)$, then the random variable $\sqrt{n-1}(\overline{X}-\mu)/s_{\mathcal{X}}$ is a Student distribution with $n-1$ degrees of freedom.*

Proof. Since $\sqrt{n}(\overline{X}-\mu)/\sigma$ is a centered Gaussian distribution with unit variance, then

$$\sqrt{n-1}\frac{\overline{X}-\mu}{s_{\mathcal{X}}} = \frac{\sqrt{n}\left(\overline{X}-\mu\right)}{\sigma}\sqrt{\frac{(n-1)\sigma^2}{ns_{\mathcal{X}}^2}} = \frac{\widetilde{n}}{\sqrt{\chi/(n-1)}},$$

where $\widetilde{n} \sim \mathcal{N}(0,1)$ and $\chi \sim \chi^2_{n-1}$. Since $\widetilde{n}$ and χ are independent by Proposition 3.1.4, the corollary follows immediately. □

Other examples are useful in mathematical finance, in particular the empirical quantile of order p is a key tool in risk management, to estimate portfolio losses, and abide by the Basel III regulatory commitments[2].

Definition 3.1.7. For a given random variable X with continuous and strictly increasing cdf F, the quantile of order $p \in (0,1)$ is the solution q_p to the equation $F(q_p) = p$.

However, if the cdf is not continuous or strictly increasing, this definition does not quite make sense, and should be refined in the following way:

$$q_p := \frac{1}{2}\left\{\inf_{F(q)>p} q + \sup_{F(q)<p} q\right\}. \tag{3.1.6}$$

Of particular interests are the following:

- the median corresponds to the quantile of order $p = 1/2$;
- the quartiles corresponds to the quantiles of order $p \in \{1/4, 3/4\}$;
- the difference $q_{3/4} - q_{1/4}$ is called the inter-quartile interval.

Note in passing that quantiles are always well defined, as opposed to the mean and the variance.

Exercise 12. From the definition (3.1.6) of the quantile, determine the values of q_p for p on each part of the following discontinuous cdf:

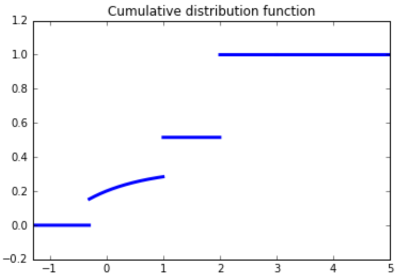

Figure 3.5: Determining quantiles

[2] seehttps://www.bis.org/bcbs/basel3.htm for details about the Basel III commitments

From (3.1.6), we can thus define the empirical quantile of order p for the sample $\mathcal{X}$ as

$$Q_{n,p} := \mathcal{S}(F) = \frac{1}{2}\left(\inf_{F(q)>p} q + \sup_{F(q)<p} q\right)$$

Exercise 13. By reordering the sample $\mathcal{X}$ in increasing order $X_{(1)} \leq \ldots \leq X_{(n)}$, show that

$$Q_{n,p} = \begin{cases} X_{(k)}, & \text{if } p \in \left(\frac{k-1}{n}, \frac{k}{n}\right), \\ \frac{1}{2}\left(X_{(k)} + X_{(k+1)}\right), & \text{if } p = \frac{k}{n} \text{ for some } k = 1, \ldots, n. \end{cases}$$

Suppose now that we observe two samples $\mathcal{X}_n$ and $\mathcal{Y}_n$ (say of two different indices, S&P500 and DAX). We can then define the empirical covariance $s_{\mathcal{X},\mathcal{Y}}$ and the empirical correlation $\rho_{\mathcal{X},\mathcal{Y}}$ between the two random vectors as

$$s_{\mathcal{X},\mathcal{Y}} := \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) \qquad \text{and} \qquad \rho_{\mathcal{X},\mathcal{Y}} := \frac{s_{\mathcal{X},\mathcal{Y}}}{s_{\mathcal{X}}s_{\mathcal{Y}}}.$$

It is easy to see that $|s_{\mathcal{X},\mathcal{Y}}| \leq 1$ always holds and that $s_{\mathcal{X},\mathcal{Y}} = 1$ if and only if there is a linear relationship between the two samples $\mathcal{X}$ and $\mathcal{Y}$. Note however, that a large value of $|s_{\mathcal{X},\mathcal{Y}}|$ does not imply that the two theoretical random variables are linearly related. An interesting and funny list of spurious relationships can be browsed through at http://www.tylervigen.com/spurious-correlations.

Exercise 14. Using the convergence results above as well as the strong law of large numbers, prove that the convergence of these two estimators to the true covariance and correlation.

Remark 3.1.8. Given some functional $\mathcal{S}(F)$ of an unknown cdf, the standard way to define an estimator thereof is to consider $\mathcal{S}(\widehat{F}_n)$.

Exercise 15. Write down an estimator for the skewness $s := \mathbb{E}[(X - \mathbb{E}[X])^3]/\mathbb{V}[X]^{3/2}$.

Exercise 16. In an IPython notebook, import two years of daily data of two stocks in the S&P500, compute the daily returns. From this, determine the values of the empirical mean, standard deviation, skewness and correlation.

Remark 3.1.9. One should be careful about drawing conclusions about the theoretical random variables from their empirical estimators. In particular very few occurrences of extreme observations can often lead to deceptive intuitions, for example, the 1987 crash. This requires tools from Extreme Value Theory, which are outside the scope of the present lectures.

IPython notebook SPXHistory.ipynb

3.2 Statistical inference

We considered so far the simple case of a sample of n observations of some random variable. We now look at the general case where the sample $\mathcal{X}_n = (\mathbf{X}_1, \dots, \mathbf{X}_n)$ consists of n iid random vectors, each real valued and of dimension p. Regarding the notations, given some (vector of) parameter(s) θ, and some random vector $\mathbf{X} \in \mathbb{R}^p$, we shall write f_θ the density of $\mathbf{X}$, F_θ its distribution, and correspondingly $\mathbb{E}_\theta$ the expectation.

3.2.1 Definition of estimators

By statistical model, we shall mean here a set of distribution (or density) functions, which can be either parameterised by some parameters or not. In the latter case, we speak of non-parametric estimation, and we consider general classes of functions for the distribution, as in the following examples:

Example 3.2.1 (Examples of non-parametric estimation)**.** We consider $X_1, \dots, X_n$ independent observations from an (unknown) cumulative distribution F, which we wish to estimate, given that it belongs to the family $\mathcal{F}$, the set of all cumulative distribution functions. Suppose now we wish to estimate its density $f = F'$. The set $\mathcal{F}$ is not valid as elements therein do not necessarily admit a density. A classical set to consider for $f \in \mathbf{f}$ is

$$\mathbf{f} := \mathbf{f}_0 \cap \mathbf{f}_{\mathrm{SOB}},$$

where $\mathbf{f}_0$ denotes the set of all densities and $\mathbf{f}_{\mathrm{SOB}}$ the Sobolev space

$$\mathbf{f}_{\mathrm{SOB}} := \left\{ f : \int \left(f''(x)\right)^2 \mathrm{d}x < \infty \right\},$$

which ensures that the class of densities is sufficiently smooth.

Example 3.2.2 (Further non-parametric statistical models)**.**

- The model $\mathcal{F} = \{\text{all distributions with finite first moments}\}$ is non parametric.
- Consider pairs of observations $((X_1, Y_1), \dots, (X_n, Y_n)))$, where X represents the predictor and Y the outcome. We wish to consider all the possible regression functions $r(x) := \mathbb{E}[Y|X = x]$. If the set $\mathcal{F}$ of such functions is finite (polynomials up to some degree), then $\mathcal{F}$ is parametric; on the other hand, if $\mathcal{F}$ is infinite dimensional, then it is non-parametric.

We now focus on parametric estimation, and will adopt as main assumption we the fact that the common distribution F is partially known, more specifically,

Assumption 3.2.3. The common distribution F belongs to some parametric family of distributions $\mathcal{F} = (\mathcal{F}_\theta)_{\theta \in \Theta}$.

Here, $\theta \in \Theta$ corresponds to the parameters of the distribution, for which we wish to determine some statistical estimator. It may be the case that we are only interested in some, but not all, parameters of a statistical model, for example if we already know the values of other parameters. Consider for example the model $\mathcal{F} = \{\mathcal{N}(\mu, \theta^2), \theta > 0\}$, where we assume that we already know the mean μ, and are only interested in estimating the volatility θ. Then $\Theta = \mathbb{R}_+^*$, and the parameter μ is called a nuisance parameter. We shall always assume the following:

Identifiability Hypothesis: two distributions in $\mathcal{F}$ are the same if and only if they have the same parameters.

Mathematically, we can restate this hypothesis in the following form:

Identifiability Hypothesis: the map $\theta \mapsto \mathbb{P}_\theta$ is injective.

Remark 3.2.4. The latter hypothesis may sound strange, but suppose that the statistical model $\mathcal{F}$ has the form

$$F_\theta(\mathrm{d}x) = \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{(x-\theta^2)^2}{2}\right\} \mathrm{d}x, \qquad \text{for all } x \in \mathbb{R},$$

for $\theta \in \Theta = \mathbb{R}$. Clearly, for any $\theta \in \Theta$, the laws F_θ and $F_{-\theta}$ are identical, and hence uniqueness of the parameter is compromised. This can here be circumvented by taking $\Theta = \mathbb{R}_+$ instead though.

Remark 3.2.5. If the observed sequence cannot be assumed to be iid, then one may have to consider the joint law of the vector $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ instead; the auto-regressive ARMA model is a classical example. We shall not consider this case in these lectures, though.

Example 3.2.6. The following examples are all parametric statistical models:

- $\mathcal{F} = \{\mathcal{N}(\theta, \sigma^2), \theta \in \mathbb{R}\}$ for some known $\sigma > 0$;
- $\mathcal{F} = \{\mathcal{N}(\theta_1, \theta_2^2), \theta_1 \in \mathbb{R}, \theta_2 > 0\}$;
- $\mathcal{F} = \{\mathcal{P}(\theta), \theta > 0\}$, the set of all Poisson distributions with parameter θ[3].

We shall from now on denote by $\widehat{\theta}_n$ an estimator of the real parameter θ, and, remembering that it is indeed a random variable, call it unbiased if $\mathbb{E}_\theta(\widehat{\theta}_n) = \theta$.

Definition 3.2.7. An estimator $\widehat{\theta}_n$ is said to be (respectively strongly) consistent if it converges to θ in probability (resp. almost surely) for all $\theta \in \Theta$, namely if

$$\lim_{n\uparrow\infty} \mathbb{P}_\theta\left(\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right) = 0, \qquad \text{for all } \varepsilon > 0, \theta \in \Theta.$$

Note that if $\widehat{\theta}_n$ is a consistent estimator of θ, then so is $\alpha_n \widehat{\theta}_n$ for any sequence (α_n) converging to 1, so that the notion of consistent estimator, though fundamental, is in fact rather weak.

[3] Remember that the Poisson distribution $X \sim \mathcal{P}(\theta)$ is characterised by $\mathbb{P}(X = n) = \theta^n \mathrm{e}^{-\theta}/(n!)$ for each $n = 0, 1, 2, \ldots$.

Definition 3.2.8. The quadratic error of the estimator $\widehat{\theta}_n$ of θ is defined as $R_n(\widehat{\theta}_n, \theta) := \mathbb{E}_\theta\left[\left(\widehat{\theta}_n - \theta\right)^2\right]$,

Proposition 3.2.9. *If $R_n(\widehat{\theta}_n, \theta)$ converges (pointwise in $\theta \in \Theta$) to zero as n tends to infinity, then $\widehat{\theta}_n$ is a consistent estimator of θ.*

Proof. Convergence of $R_n(\widehat{\theta}_n, \theta)$ is the same as L^2 convergence, and hence convergence in probability follows directly from Theorem 2.2.18. □

Remark 3.2.10. Alternatively, using Markov's inequality (Proposition 2.2.1), we can write, for any $a > 0$,

$$\mathbb{P}\left(\left|\widehat{\theta}_n - \theta\right| \geq a\right) \leq \frac{\mathbb{E}\left[\left(\widehat{\theta}_n - \theta\right)^2\right]}{a^2} = \frac{R_n(\widehat{\theta}_n, \theta)}{a^2},$$

and the corollary follows by taking limits.

Exercise 17. Let $\mathcal{X}_n$ denote n observations from a Bernoulli random variable with parameter $\theta \in [0, 1]$, and denote $\widehat{\theta}_n := n^{-1}\sum_{i=1}^n X_i$. Show that $\widehat{\theta}_n$ is a consistent estimator of θ.

Solution. *Recall that a Bernoulli random variable X with parameter $\theta \in [0, 1]$ takes value 1 with probability θ and zero with probability $1 - \theta$, and $\mathbb{E}[X] = \theta$ and $\mathbb{V}[X] = \theta(1 - \theta)$, so that $\mathbb{E}[X^2] = \mathbb{V}[X] + \mathbb{E}[X]^2 = \theta$. Therefore,*

$$\begin{aligned}
\mathbb{E}\left[\left(\widehat{\theta}_n - \theta\right)^2\right] = \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n X_i - \theta\right)^2\right] &= \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^n X_i\right)^2\right] - \frac{2\theta}{n}\mathbb{E}\left[\sum_{i=1}^n X_i\right] + \theta^2 \\
&= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^n X_i^2 + \sum_{i\neq j} X_i X_j\right] - 2\theta^2 + \theta^2 \\
&= \frac{1}{n^2}\left(\sum_{i=1}^n \mathbb{E}\left[X_i^2\right] + \sum_{i\neq j}\mathbb{E}[X_i X_j]\right) - \theta^2 \\
&= \frac{1}{n^2}\left(n\theta + n(n-1)\theta^2\right) - \theta^2,
\end{aligned}$$

which clearly converges to zero as n tends to infinity.

As mentioned above, there is no uniqueness of estimators, and one may ask how to choose between different consistent estimators. The following notions of efficient and admissible estimators clarify this.

Definition 3.2.11. Let $\widehat{\theta}_n^1$ and $\widehat{\theta}_n^2$ two estimators in the statistical model $(\mathcal{F}_\theta)_{\theta\in\Theta}$. If $R_n(\widehat{\theta}_n^1, \theta) \leq R_n(\widehat{\theta}_n^2, \theta)$, for all $\theta \in \Theta$ and $R_n(\widehat{\theta}_n^1, \theta) < R_n(\widehat{\theta}_n^2, \theta)$, for some $\theta \in \Theta$, then $\widehat{\theta}_n^1$ is called more efficient than $\widehat{\theta}_n^2$, which is then called inadmissible. The most efficient estimator is called admissible.

The quadratic risk of an estimator can easily be decomposed as follows:

$$R_n(\widehat{\theta}_n, \theta) := \mathbb{E}_\theta\left[\left(\widehat{\theta}_n - \theta\right)^2\right] = \left(\mathbb{E}_\theta(\widehat{\theta}_n) - \theta\right)^2 + \mathbb{E}_\theta\left[\left(\widehat{\theta}_n - \mathbb{E}_\theta(\widehat{\theta}_n)\right)^2\right] =: \beta_n^2(\widehat{\theta}_n, \theta) + \sigma_n^2(\widehat{\theta}_n, \theta), \tag{3.2.1}$$

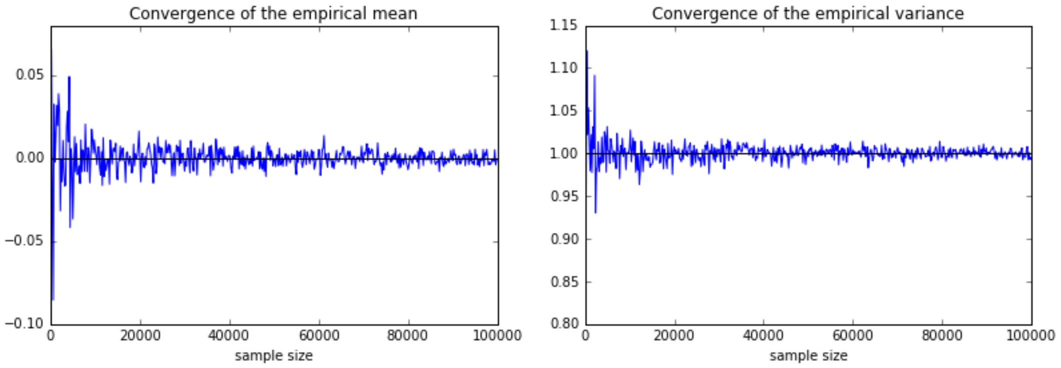

Figure 3.6: Convergence of the empirical mean and variances as estimators for the mean and the variance in the $\mathcal{N}(0,1)$ case.

where β_n is called the bias and σ_n^2 the variance of the estimator $\widehat{\theta}_n$, so that $\widehat{\theta}_n$ is unbiased if $\beta_n(\widehat{\theta}_n,\theta)=0$ for all $\theta\in\Theta$. The following classical exercise illustrates several far-reaching issues:

Exercise 18. Consider the statistical model $\mathcal{F}=\{\mathcal{N}(0,\sigma^2),\sigma\in\mathbb{R}_+^*\}$, and define the following two estimators of the true variance σ^2:

$$\widehat{\theta}_n^1 := \frac{1}{n}\sum_{i=1}^n\left(X_i-\overline{X}\right)^2 \qquad \text{and} \qquad \widehat{\theta}_n^2 := \frac{1}{n-1}\sum_{i=1}^n\left(X_i-\overline{X}\right)^2.$$

Show that $\widehat{\theta}_n^1$ is biased while $\widehat{\theta}_n^2$ is not, and compare their quadratic risks.

Solution. *From Proposition 3.1.4, noticing that $\widehat{\theta}_n^1$ is in fact the same as $s_{\mathcal{X}}^2$, we immediately obtain that it is biased, but that $\widehat{\theta}_n^2$ is not, and the bias of $\widehat{\theta}_n^1$ reads*

$$\mathbb{E}_\sigma(\widehat{\theta}_n)-\sigma^2=\frac{n-1}{n}\sigma^2-\sigma^2=-\frac{\sigma^2}{n}.$$

We can further compute the variances of the estimators as

$$\mathbb{E}_\sigma\left[\left(\widehat{\theta}_n^1-\mathbb{E}_\sigma(\widehat{\theta}_n^1)\right)^2\right]=\frac{2(n-1)\sigma^4}{n^2} \qquad \textit{and} \qquad \mathbb{E}_\sigma\left[\left(\widehat{\theta}_n^2-\mathbb{E}_\sigma(\widehat{\theta}_n^2)\right)^2\right]=\frac{2\sigma^4}{n-1},$$

and therefore, the quadratic risks read

$$R_n(\widehat{\theta}_n^1,\sigma^2)=\left(\frac{\sigma^2}{n}\right)^2+\frac{2(n-1)\sigma^4}{n^2}=\frac{2n-1}{n^2}\sigma^4 \qquad \textit{and} \qquad R_n(\widehat{\theta}_n^2,\sigma^2)=\frac{2\sigma^4}{n-1},$$

as well as the inequality $R_n(\widehat{\theta}_n^2,\sigma^2)>R_n(\widehat{\theta}_n^1,\sigma^2)$, meaning that the estimator $\widehat{\theta}_n^1$, despite being biased, is in fact more efficient while $\widehat{\theta}_n^2$ is inadmissible.

3.3 Parametric inference

3.3.1 The method of moments

We are interested here in estimating the moments of the law generating a given sample $X_1,\ldots,X_n$. Following the terminology above, the common law of the sample is drawn from the statistical model

$\mathcal{F} = (F_\theta)_{\theta \in \Theta}$. Define

$$\mu_r(\theta) := \mathbb{E}_\theta(X^r) = \int_{\mathbb{R}} x^r F_\theta(\mathrm{d}x), \tag{3.3.1}$$

the moment of order r for $\mathcal{F}_\theta$, which we assume to exist for all $r \leq q$, for some integer q. However, these moments are not known since they depend on the unknown parameter θ. We therefore consider their empirical estimators

$$m_r := \frac{1}{n}\sum_{i=1}^n X_i^r,$$

which we know converge to the true value as the sample size n tends to infinity. The following method is due to Pearson [4].

Definition 3.3.1. In the statistical model $\mathcal{F}$, the method of moments estimator $\widehat{\theta}_n^{\mathrm{MM}}$ for θ is the solution to the system

$$\mu_r\left(\widehat{\theta}_n^{\mathrm{MM}}\right) = m_r, \qquad \text{for } r = 1, \ldots, q.$$

Exercise 19.

- Let $\mathcal{F} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma \in \mathbb{R}_+^*\}$. Show that the estimators (3.1.2) and (3.1.3) correspond to method of moment estimators.
- Compute the method of moment estimator for the parameter of the exponential distribution using the first two moments.

3.3.2 The generalised method of moments

Instead of the moments μ_r, one could consider more general functions, rewriting (3.3.1) as

$$\mu_r(\theta) := \mathbb{E}_\theta(\phi_r(X)),$$

for a general family of functions $(\phi_r)_{r=1,\ldots,q}$, and we replace their estimators by

$$m_r := \frac{1}{n}\sum_{i=1}^n \phi_r(X_i).$$

We also denote by $\widehat{\theta}_n^{\mathrm{GM}}$ the corresponding generalised method of moment estimator. The reason for introducing this generalised method is that moments may not exist for some distribution, and functions other than polynomials may help circumventing this issue. Consider the Cauchy distribution, the density of which is given by

$$F_\theta(\mathrm{d}x) = \frac{1}{\pi}\frac{\mathrm{d}x}{1 + (x-\theta)^2}, \qquad \text{for all } x \in \mathbb{R}.$$

[4] Karl Pearson (1857-1936) was an English mathematician, and is one of the founders of mathematical statistics. His contributions are fundamental and widely spread, in particular the method of moments, Principal Component Analysis, the correlation coefficient, the histogram, and the p-value.

Exercise 20.

- Show that no moment exists.
- With the function $\phi_1(x) := \mathrm{sgn}(x) = \mathbb{1}_{\{x>0\}} - \mathbb{1}_{\{x\leq 0\}}$, show that the generalised method of moments estimator $\widehat{\theta}_n^{\mathrm{GM}}$, given a sample $\mathcal{X}_n$, is of the form
$$\widehat{\theta}_n^{\mathrm{GM}} = \tan\left(\frac{\pi}{2n}\sum_{i=1}^n \mathrm{sgn}(X_i)\right).$$
- Show that $\widehat{\theta}_n^{\mathrm{GM}}$ is a consistent estimator and that it is asymptotically Gaussian, i.e. that $\sqrt{n}\left(\widehat{\theta}_n^{\mathrm{GM}} - \theta^*\right)$ converges in distribution to a centered Gaussian random variable with variance to determine explicitly.

3.3.3 The Delta method

The Delta method is in fact another way of understanding the generalised method of moments: assume that, from the observations, one can construct an estimator of the form $\varphi(\theta)$. If the function φ is sufficiently smooth, then an estimator of θ can be obtain taking the reciprocal φ^{-1}. Let us state this more formally:

Theorem 3.3.2. *Consider the iid sample $(X_1, \ldots, X_n)$ with common law $\mathbb{P}_\theta$, for $\theta \in \Theta \subset \mathbb{R}$, and φ a $\mathcal{C}^1(\Theta \to \varphi(\Theta))$-diffeomorphism. If $\widehat{\varphi}_n = \widehat{\varphi}_n(X_1, \ldots, X_n)$ is a convergent estimator of $\varphi(\theta)$, and θ an interior point of Θ, then $\widehat{\theta}_n = \varphi^{-1}(\widehat{\varphi}_n)$ is defined almost surely as n tends to infinity and $\widehat{\theta}_n$ converges in probability to θ. Furthermore, if there exists a sequence $(\alpha)_n$ diverging to infinity and a random variable Z_θ such that*

$$\alpha_n\left(\widehat{\varphi}_n - \varphi(\theta)\right) \text{ converges in law to } Z_\theta,$$

then

$$\alpha_n\left(\widehat{\theta}_n - \theta\right) \text{ converges in law to } \frac{Z_\theta}{\varphi'(\theta)}.$$

Remark 3.3.3. The classical example is when $\alpha = n^{-1/2}$ and $Z_\theta \sim \mathcal{N}(0, \sigma_\theta^2)$.

Proof. First note that since the sequence $\left(\alpha_n\left(\widehat{\varphi}_n - \varphi(\theta)\right)\right)_n$ converges in law to Z_θ, then $(\widehat{\varphi}_n)_n$ converges in probability to $\varphi(\theta)$. Now, because θ belongs to the interior of Θ and φ is bijective and continuous, then $\varphi(\theta)$ belongs to the interior of $\varphi(\Theta)$. Furthermore, since $(\widehat{\varphi}_n)_n$ converges to $\varphi(\theta)$ in probability, then

$$\lim_{n\uparrow\infty} \mathbb{P}\left(\widehat{\varphi}_n \in \varphi(\Theta)\right) = 1.$$

Let $\psi \equiv \varphi^{-1}$ be the inverse function; by assumption, the Taylor expansion around the point $\varphi(\theta)$

$$\psi(x) = \psi(\varphi(\theta)) + (x - \varphi(\theta))\left[\psi'(\varphi(\theta)) + \varepsilon(x)\right]$$

holds, where $\varepsilon(x)$ tends to zero as x approaches $\varphi(\theta)$. Since $(\widehat{\varphi}_n)_n$ converges in probability, then, by continuity, so does $\psi(\widehat{\varphi}_n)$, and we can write the (random) expansion

$$\psi(\widehat{\varphi}_n) = \psi(\varphi(\theta)) + (\widehat{\varphi}_n - \varphi(\theta)) \left[\psi'(\varphi(\theta)) + \varepsilon(\widehat{\varphi}_n)\right],$$

where the sequence $(\psi'(\varphi(\theta)) + \varepsilon(\widehat{\varphi}_n))_n$ converges in probability to $\psi'(\varphi(\theta))$, and the result follows from Slutsky's theorem. We just recall in passing the useful identity:

$$\psi'(y) = \frac{1}{(\varphi' \circ \varphi)(y)} \Longrightarrow \psi'(\varphi(\theta)) = \frac{1}{\varphi'(\theta)}.$$

□

Example 3.3.4 (Sample variance)**.** Using previous notations, the sample variance of the iid sample $\mathcal{X} = (X_1, \ldots, X_n$ is defined in (3.1.3) as

$$s_{\mathcal{X}}^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}^2 =: \varphi\left(\overline{X}, \overline{X^2}\right),$$

where $\varphi(x, y) := y - x^2$. Assume that $s_{\mathcal{X}}^2$ is computed from a sample of a distribution with finite first four moments $\mu_1, \ldots, \mu_4$. The Central Limit Theorem implies that

$$\sqrt{n}\left[\begin{pmatrix}\overline{X}\\ \overline{X^2}\end{pmatrix} - \begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}\right] \text{ converges in distribution to } \mathbf{N} = \begin{pmatrix}\mathbf{N}_1\\ \mathbf{N}_2\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\mu_2 - \mu_1^2 & \mu_3 - \mu_1\mu_2\\ \mu_3 - \mu_1\mu_2 & \mu_4 - \mu_2^2\end{pmatrix}\right).$$

Clearly, the function φ is differentiable at $\theta = (\mu_1, \mu_2)^\top$ with gradient equal to $(-2\mu_1, 1)$. Therefore

$$\sqrt{n}\left(\varphi\left(\overline{X}, \overline{X^2}\right) - \varphi(\mu_1, \mu_2)\right) \qquad \text{converges in distribution to} \qquad -2\mu_1\mathbf{N}_1 + \mathbf{N}_2$$

3.3.4 Maximum likelihood method

We now move on to one of the most important and most widely used estimation method, namely Maximum Likelihood Estimation, popularised by Fisher[5] in the 1920s. In order to use this method, we need to assume that the statistical model $\mathcal{F}$ admits a density with respect to the Lebesgue measure, which we denote by f_θ for each $\theta \in \Theta$.

Definition 3.3.5. The maximum likelihood function is the map $\mathcal{L}_n$ from Θ to $\mathbb{R}$ defined as

$$\mathcal{L}_n(\theta) := \prod_{i=1}^n f_\theta(X_i), \qquad \text{for all } \theta \in \Theta,$$

and the function $l_n := -\frac{1}{n}\log \mathcal{L}_n$ is called the log-likelihood.

When the sequence of observations $(X_1, \ldots, X_n)$ is assumed to be iid, note that the maximum likelihood function is nothing else than the joint density of the sample.

[5] Ronald Fisher (1890-1972) was an English statistician and geneticist. He is vastly regarded as one of the founders of modern statistics, and of population genetics

Definition 3.3.6. The maximum likelihood estimator is defined as

$$\widehat{\theta}_n^{\mathrm{ML}} := \arg\max_{\theta\in\Theta} \mathcal{L}_n(\theta).$$

It is clear that we can also write $\widehat{\theta}_n^{\mathrm{ML}} := \arg\min_{\theta\in\Theta} l_n(\theta)$, and that a necessary condition is that the gradient of either l_n or $\mathcal{L}_n$ is null, and we call

$$\nabla l_n(\theta) = 0 \tag{3.3.2}$$

the likelihood equation (or likelihood system). Note that existence of a maximum likelihood estimator is not necessarily linked to that of a root to (3.3.2).

Example 3.3.7. [Maximum Likelihood Estimator for the Gaussian distribution] For the Gaussian statistical model $\mathcal{F} = \{\mathcal{N}(\mu,\sigma^2), \mu\in\mathbb{R}, \sigma>0\}$, we can write

$$\begin{aligned} \mathcal{L}_n(\theta) &= \left(\sigma\sqrt{2\pi}\right)^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\mu)^2\right), \\ l_n(\theta) &= \frac{\log(2\pi)}{2} + \log(\sigma) + \frac{1}{2n\sigma^2}\sum_{i=1}^n (X_i-\mu)^2. \end{aligned}$$

The likelihood system $\nabla l_n(\theta) = 0$ is therefore equivalent to

$$\begin{cases} \partial_\mu l_n(\theta) &= 0, \\ \partial_\sigma l_n(\theta) &= 0 \end{cases} \qquad \text{if and only if} \qquad \begin{cases} \displaystyle\sum_{i=1}^n (X_i-\mu) &= 0, \\ \displaystyle\sum_{i=1}^n (X_i-\mu)^2 &= \sigma^2 n, \end{cases}$$

so that

$$\mu = \overline{X} \qquad \text{and} \qquad \sigma = \left(\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2\right)^{1/2} = s_X.$$

Example 3.3.8 (Maximum Likelihood Estimator for the Uniform distribution)**.** Consider the Uniform distribution on the closed interval $[0,\theta]$, with density given by $f_\theta(x) = \theta^{-1}\mathbb{1}_{[0,\theta]}(x)$. The likelihood function is then given by

$$\mathcal{L}_n(\theta) := \prod_{i=1}^n f_\theta(X_i) = \frac{1}{\theta^n}\prod_{i=1}^n \mathbb{1}_{[0,\theta]}(X_i) = \frac{1}{\theta^n}\mathbb{1}_{[\max_{i=1,\ldots,n} X_i,\infty)}(\theta), \qquad \text{for all } \theta\in\Theta.$$

It is clear that the maximum of the function is then attained at the point $\widehat{\theta}_n^{\mathrm{ML}} = \max\limits_{i=1,\ldots,n} X_i$. This estimator has many immediate properties; in particular, $\mathbb{P}(\widehat{\theta}_n^{\mathrm{ML}} \leq t) = (t/\theta)^n$ for any $t\in[0,\theta]$ and its density and expectation therefore read

$$f_{\widehat{\theta}_n^{\mathrm{ML}}}(t) = \frac{n}{\theta^n} t^{n-1}\mathbb{1}_{[0,\theta]}(t) \qquad \text{and} \qquad \mathbb{E}\left[\widehat{\theta}_n^{\mathrm{ML}}\right] = \frac{n\theta}{n+1}.$$

It is biased, and we can compute directly its second moment and its quadratic risk as

$$\mathbb{E}\left[\left(\widehat{\theta}_n^{\mathrm{ML}}\right)^2\right] = \frac{n\theta^2}{n+2} \qquad \text{and} \qquad R\left(\widehat{\theta}_n^{\mathrm{ML}},\theta\right) = \frac{2\theta^2}{(n+1)(n+2)}.$$

Asymptotic behaviour of the log-likelihood function

In this section, we shall denote by θ^* the true value of the parameter, and will consider the following standing assumption:

Assumption 3.3.9. For any $\theta \in \Theta$, $\int |\log f_\theta(x)| F_{\theta^*}(\mathrm{d}x)$ is finite.

Under this assumption, it is easy to see that the parameterised sequence $(Z_i^\theta)_{i=1\ldots,n}$ defined by $Z_i^\theta := -\log f_\theta(X_i)$ is iid with

$$\mathbb{E}\left[Z_i^\theta\right] = -\int \log f_\theta(x) f_{\theta^*}(x)\mathrm{d}x =: J(\theta). \tag{3.3.3}$$

The function J is called the contrast (or divergence) function, and corresponds exactly, by the law of large numbers, to the limit in probability of the log-likelihood function. This link, together with the following lemma, justifies fully the maximum likelihood method.

Lemma 3.3.10. *Under Assumption 3.3.9, the inequality $J(\theta) \geq J(\theta^*)$ holds for all $\theta \in \Theta$. Furthermore, under the Identifiability Hypothesis, the inequality is strict whenever $\theta \neq \theta^*$.*

Proof. By convexity of the logarithm, the inequality $\log(1+z) - z \leq 0$ holds for all $z \geq -1$, and is an equality if and only if $z = 0$. Therefore, for any x, θ, we can write

$$\log\left(\frac{f_\theta(x)}{f_{\theta^*}(x)}\right) - \left(\frac{f_\theta(x)}{f_{\theta^*}(x)} - 1\right) = \log\left(1 + \left[\frac{f_\theta(x)}{f_{\theta^*}(x)} - 1\right]\right) - \left(\frac{f_\theta(x)}{f_{\theta^*}(x)} - 1\right) \leq 0.$$

Since

$$\int \left(\frac{f_\theta(x)}{f_{\theta^*}(x)} - 1\right) f_{\theta^*}(x)\mathrm{d}x = 0,$$

we therefore obtain

$$J(\theta) - J(\theta^*) = -\int f_{\theta^*}(x) \log\left(\frac{f_\theta(x)}{f_{\theta^*}(x)}\right)\mathrm{d}x = -\int f_{\theta^*}(x)\left\{\log\left(\frac{f_\theta(x)}{f_{\theta^*}(x)}\right) - \left(\frac{f_\theta(x)}{f_{\theta^*}(x)} - 1\right)\right\}\mathrm{d}x \geq 0.$$

Noting that the inner bracket is non-positive, it therefore has to be null on the set $\mathcal{A} := \{x : f_{\theta^*}(x) > 0\}$. By convexity of the logarithm, though, this is true if and only if $f_\theta(x)/f_{\theta^*}(x) = 1$ almost surely on $\mathcal{A}$, which implies $f_\theta(x) = f_{\theta^*}(x)$ almost surely for all x. Summarising, we have $J(\theta) = J(\theta^*)$ if and only if $f_\theta(x) = f_{\theta^*}(x)$ almost surely for all x, and the Identifiability Hypothesis yields $\theta = \theta^*$, and the lemma follows. □

Consistency of the maximum likelihood estimator

We saw above several types of estimators, which may or may not be consistent. In particular, in the case where $\mathcal{F} = \{\mathcal{N}(\theta, 1), \theta \in \mathbb{R}\}$, the maximum likelihood estimator $\widehat{\theta}_n^{\mathrm{ML}}$ corresponds exactly to the empirical average $\overline{X}$ whereas the real value is the theoretical mean which, in general, is different. The following theorem gathers a set of sufficient conditions ensuring that they are the same.

Theorem 3.3.11. *If the following conditions hold:*

1. *Θ is an open subset of $\mathbb{R}$,*

2. *the Identifiability Hypothesis holds,*

3. *for any $x \in \mathbb{R}$, the function $\theta \mapsto f_\theta(x)$ is continuous on Θ,*

4. *Assumption 3.3.9 holds,*

5. *$\widehat{\theta}_n^{\mathrm{ML}}$ exists for all n and the local minima of l_n forms a closed bounded set in the interior of Θ,*

then the maximum likelihood estimator is consistent.

3.3.5 Bayes estimators

We finish this review of estimators by Bayes estimators, which essentially relies on a generalised notion of the risk associated to an estimator (we so far only saw quadratic risk in Definition 3.2.8). We keep the same framework as before, and introduce a so-called loss function $\lambda : \Theta \times \Theta \to \mathbb{R}_+$ and the associated risk function

$$\mathcal{R}^\lambda(\theta, \widehat{\theta}) := \mathbb{E}_\theta\left[\lambda\left(\theta, \widehat{\theta}\right)\right] = \int_{\mathbb{R}^n} \lambda\left(\theta, \widehat{\theta}\right)\mathbb{P}_\theta(\mathrm{dx}), \qquad \text{for all } \theta \in \Theta,$$

where $\widehat{\theta}$ is a given estimator. The motivation for Bayes estimation is the following issue: consider the Gaussian statistical model $\mathcal{F} = \{\mathcal{N}(\theta, 1), \theta \in \mathbb{R}\}$, and, from a Gaussian $\mathcal{N}(\theta, 1)$ sample, we consider the Maximum Likelihood estimator $\widehat{\theta}_n^{\mathrm{ML}}$ as well as the constant estimator $\widetilde{\theta} = 0$. The quadratic risk, corresponding to the loss function $\lambda(x, y) = (x - y)^2$, can be computed as (we shall use the notation $\mathcal{R}^Q$ for the quadratic risk here)

$$\mathcal{R}^Q\left(\theta, \widehat{\theta}_n^{\mathrm{ML}}\right) = \mathbb{V}\left[\overline{\mathbf{X}}\right] = \frac{1}{n} \qquad \text{and} \qquad \mathcal{R}^Q\left(\theta, \widetilde{\theta}\right) = \theta^2. \qquad \text{for any } \theta \in \Theta.$$

Therefore, the maximum likelihood estimator is better only when $R^Q\left(\theta, \widehat{\theta}_n^{\mathrm{ML}}\right) < R^Q\left(\theta, \widetilde{\theta}\right)$, that is when $|\theta| > n^{-1/2}$, which is awkward since we do not actually know the value of θ. Bayes' paradigm is to endow θ with some distribution μ, supported on Θ, accounting for this uncertainty.

Definition 3.3.12. The Bayesian risk with the law μ, for some estimator $\widehat{\theta}$, is defined as

$$\mathcal{R}_B^\lambda\left(\mu, \widehat{\theta}\right) := \int_\Theta \mathcal{R}^\lambda\left(\theta, \widehat{\theta}\right)\mu(\mathrm{d}\theta) = \int_\Theta \left(\int_{\mathbb{R}^n} \lambda\left(\theta, \widehat{\theta}\right)\mathbb{P}_\theta(\mathrm{dx})\right)\mu(\mathrm{d}\theta)$$

Example 3.3.13. In the framework of the Gaussian estimation above, consider $\mu = \mathcal{N}(0, 1)$. Then

$$\begin{array}{rclclcl}
\mathcal{R}_B^Q\left(\mu, \widehat{\theta}_n^{\mathrm{ML}}\right) & := & \displaystyle\int_\Theta \mathcal{R}^Q\left(\theta, \widehat{\theta}_n^{\mathrm{ML}}\right)\mu(\mathrm{d}\theta) & = & \displaystyle\int_{\mathbb{R}} \frac{\mu(\mathrm{d}\theta)}{n} & = & \displaystyle\frac{1}{n}, \\
\mathcal{R}_B^Q\left(\mu, \widetilde{\theta}\right) & := & \displaystyle\int_\Theta \mathcal{R}^Q\left(\theta, \widetilde{\theta}\right)\mu(\mathrm{d}\theta) & = & \displaystyle\int_{\mathbb{R}} \theta^2\mu(\mathrm{d}\theta) & = & 1,
\end{array}$$

so that, in the Bayesian sense, the Maximum Likelihood estimator is better than the constant one. Note that, in fact, for any square integrable prior distribution μ, the Bayesian quadratic risk for the maximum likelihood estimator will always be equal to $\frac{1}{n}$, whereas it will be equal to some constant for the trivial estimator, and hence the MLE will be better for large enough sample size n.

The following definition is the Bayes' equivalent of the maximum likelihood estimator. Given the observation $\mathcal{X}$, we shall call $f(\theta|\mathcal{X})$ the posterior density, namely the density of θ conditional on the observation $\mathcal{X}$. Recall Bayes' formula:

$$f(\theta|\mathrm{x}) = \frac{f(\mathrm{x}|\theta)f(\theta)}{f(\mathrm{x})} = \frac{f(\mathrm{x}|\theta)f(\theta)}{\int_\Theta f(\mathrm{x}|\tau)f(\tau)\mathrm{d}\tau}. \tag{3.3.4}$$

Definition 3.3.14 (Bayes Estimator)**.** For a given prior distribution μ, the Bayes estimator $\widehat{\theta}^{\mathrm{B}}$ is defined as

$$\widehat{\theta}^{\mathrm{B}}(\mathrm{x}) := \operatorname*{arg\,min}_{\tau\in\Theta} \mathbb{E}\left[\lambda(\theta,\tau)|\mathcal{X}=\mathrm{x}\right] = \operatorname*{arg\,min}_{\tau\in\Theta} \int_\Theta \lambda(\theta,\tau)f(\theta|\mathrm{x})\mathrm{d}\theta.$$

Exercise 21. Show that, with the quadratic loss function $\lambda(x,y) = (x-y)^2$, the Bayes estimator is simply the conditional expectation $\widehat{\theta}^{\mathrm{B}}(\mathrm{x}) = \mathbb{E}\left[\theta|\mathcal{X}=\mathrm{x}\right]$.

Solution. *Let* $\Phi_{\mathrm{x}}(\tau) := \int_\Theta \lambda(\theta,\tau)f(\theta|\mathrm{x})\mathrm{d}\theta = \int_\Theta (\theta-\tau)^2 f(\theta|\mathrm{x})\mathrm{d}\theta$ *the function to be minimised, which does not depend on* θ*, with* x *as a parameter. Then, by Leibniz integral rule,*

$$\begin{aligned}
\partial_\tau\Phi_{\mathrm{x}}(\tau) &= \partial_\tau\int_\Theta (\theta-\tau)^2 f(\theta|\mathrm{x})\mathrm{d}\theta = \int_\Theta \partial_\tau\left[(\theta-\tau)^2 f(\theta|\mathrm{x})\right]\mathrm{d}\theta\\
&= -2\int_\Theta(\theta-\tau)f(\theta|\mathrm{x})\mathrm{d}\theta = -2\int_\Theta \theta f(\theta|\mathrm{x})\mathrm{d}\theta + 2\tau\int_\Theta f(\theta|\mathrm{x})\mathrm{d}\theta\\
&= -2\left(\mathbb{E}\left[\theta|\mathcal{X}=x\right]-\tau\right),
\end{aligned}$$

which is equal to zero if and only if $\tau = \mathbb{E}\left[\theta|\mathcal{X}=x\right]$*, and the result follows.*

Exercise 22. Compute the Bayes estimator for the loss function $\lambda(x,y) = |x-y|$.

Example 3.3.15. Consider $\mu \sim \mathcal{N}(0,1)$ and, knowing θ, the iid sample $\mathcal{X}$ is distributed according to $\mathcal{N}(\theta,1)$, that is

$$f(\mathrm{x}|\theta) = \frac{1}{(2\pi)^{n/2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^n (x_i-\theta)^2\right\}$$

Using Bayes' formula (3.3.4), the posterior density of θ reads

$$\begin{aligned}
f(\theta|\mathrm{x}) &= \frac{1}{f(\mathrm{x})}\frac{1}{(2\pi)^{n/2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^n (x_i-\theta)^2\right\}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{\theta^2}{2}\right\}\\
&\propto_{\mathrm{x}} \exp\left\{-\frac{n+1}{2}\left(\theta-\frac{n\overline{\mathrm{x}}}{n+1}\right)^2\right\},
\end{aligned}$$

where the symbol $\propto_{\mathrm{x}}$ means that the two sides are equal up to a constant multiple of x, independent of θ. Since a density has to integrate to unity, we do not need to compute the constant to conclude that the law of θ conditional on $\mathcal{X} = \mathrm{x}$ is Gaussian with mean $\frac{n\overline{\mathrm{x}}}{n+1}$ and variance $\frac{1}{n+1}$. The Bayes estimator for the quadratic risk is therefore $\frac{n}{n+1}\overline{X}_n$. It is biased and different from the maximum likelihood estimator $\overline{X}_n$, but the difference becomes negligible as the sample size increases. Since

$$\mathbb{V}\left[\widehat{\theta}^{\mathrm{B}}(\mathrm{x})\right] = \frac{n}{(n+1)^2} < \frac{1}{n} = \mathbb{V}\left[\widehat{\theta}_n^{\mathrm{ML}}\right],$$

the quadratic risks can be computed as

$$\mathcal{R}^Q\left(\theta, \widehat{\theta}^{\mathrm{B}}(\mathrm{x})\right) = \frac{n+\theta^2}{(n+1)^2} \qquad \text{and} \qquad \mathcal{R}^Q\left(\theta, \widehat{\theta}_n^{\mathrm{ML}}\right) = \frac{1}{n},$$

and their comparison again depends on the position of θ with respect to n. In the Bayesian framework, if we integrate with respect to the distribution of θ, we obtain

$$\mathcal{R}_B^Q\left(\theta, \widehat{\theta}^{\mathrm{B}}(\mathrm{x})\right) = \int_\Theta R\left(\theta, \widehat{\theta}^{\mathrm{B}}(\mathrm{x})\right)\mu(\mathrm{d}\theta) = \int_{\mathbb{R}} \frac{n+\theta^2}{(n+1)^2}\mu(\mathrm{d}\theta) = \frac{1}{n+1},$$

and

$$\mathcal{R}_B^Q\left(\theta, \widehat{\theta}_n^{\mathrm{ML}}\right) = \int_\Theta R\left(\theta, \widehat{\theta}_n^{\mathrm{ML}}\right)\mu(\mathrm{d}\theta) = \int_{\mathbb{R}} \frac{1}{n}\mu(\mathrm{d}\theta) = \frac{1}{n},$$

so the the Bayes estimator is slightly better.

3.3.6 Regular statistical models

We consider in this section only the one-dimensional case where Θ is a subset of the real line, i.e. we only estimate one parameter of the common distribution. We further assume that the density f_θ exists and is smooth, and define the function $l_\theta := \log f_\theta$.

Definition 3.3.16. The Fisher information is the function $I : \Theta \to \mathbb{R}_+$ defined as

$$I(\theta) := \int \left(\partial_\theta l_\theta(x)\right)^2 f_\theta(x)\mathrm{d}x = \int_{\{x: f_\theta(x)>0\}} \frac{\left(\partial_\theta f_\theta(x)\right)^2}{f_\theta(x)}\mathrm{d}x = \mathbb{E}_\theta\left[\left(\partial_\theta l_\theta(x)^2\right)\right]$$

In order to avoid degenerate situations, we shall work under the following set of assumptions:

Assumption 3.3.17 (Regularity Hypotheses)**.**

- Θ is an open subset of $\mathbb{R}$;
- for all x, the functions f_θ and l_θ are smooth;
- for any $\theta \in \Theta$, there exists a ball B_θ around θ and a function Λ such that, for all x,

$$\max\left\{\partial_{\theta\theta} l_\theta(x), \partial_\theta l_\theta(x), |\partial_\theta l_\theta(x)|^2\right\} \leq \Lambda(x) \qquad \text{and} \qquad \int \Lambda(x) \sup_{\theta \in B_\theta} f_\theta(x)\mathrm{d}x < \infty;$$

- $I(\theta) > 0$ for all $\theta \in \Theta$.

These regularity assumptions are not always met though, as the following examples show:

Example 3.3.18. Consider the exponential law with parameter $\theta \in \Theta = (0, \infty)$, for which the density reads $f_\theta(x) = \theta \exp(-\theta x) \mathbb{1}_{x>0}$, so that

- for any $x > 0$, the map $\theta \mapsto f_\theta(x)$ is smooth on Θ;
- for any $\theta \in \Theta$, the map

$$x \mapsto \frac{\left(\partial_\theta f_\theta(x)\right)^2}{f_\theta(x)} \mathbb{1}_{\{f_\theta(x)>0\}} = \frac{(1-\theta x)^2}{\theta} \mathrm{e}^{-\theta x}$$

is integrable on $\mathbb{R}_+$ and

$$I(\theta) = \frac{1}{\theta^2} \mathbb{E}_\theta \left[(1 - \theta X)^2\right] = \frac{1}{\theta^2}$$

is continuous on Θ.

Exercise 23.

- For the uniform distribution on $[0, \theta]$ (with $\theta > 0$), the function l_θ is not differentiable;
- Consider the statistical model $\{\mathcal{N}(\theta^2, 1), \theta \in \Theta\}$, with $\Theta = [0, \infty)$; Show that $I(0) = 0$, so that the model is not regular. Show that the model is regular, however, if $\Theta = (0, \infty)$.

The following lemma provides an alternative characterisation of the Fisher information. Its proof relies on integration by parts and the Fubini theorem, all justified by the regularity assumptions above. We recall that the divergence function $J(\cdot)$ was defined in (3.3.3).

Lemma 3.3.19. *For a regular model, the following equalities hold:*

$$\begin{aligned} I(\theta) &= -\int \partial_{\theta\theta} l_\theta(x) f_\theta(x) \mathrm{d}x, \qquad \textit{for all } \theta \in \Theta, \\ J(\theta^*) &= 0, \\ \partial_{\theta\theta} J(\theta^*) &= -\mathbb{E}_\theta \left[\partial_{\theta\theta} l_\theta(X, \theta^*)\right]. \end{aligned}$$

The lemma in particular implies that $I(\theta^*) = J''(\theta^*)$, which can be used as a geometric interpretation of the Fisher information based on the observed curvature of the function J around its minimum θ^*.

Definition 3.3.20. For any $\theta \in \Theta$, the quantity $K(\theta, \theta^*) := J(\theta) - J(\theta^*)$ is called the Kullback-Leibler divergence [27] (or relative entropy) between the two distributions F_θ and F_{θ^*}. Furthermore, the quantity $J(\theta^*)$ is called the Shannon[6] entropy, and is a key tool in information theory.

Theorem 3.3.21. *For a regular model, the sequence $\sqrt{n}\left(\widehat{\theta}_n^{\mathrm{ML}} - \theta^*\right)$ converges in distribution to $\mathcal{N}(0, 1/I(\theta^*))$ for any $\theta^* \in \Theta$ as n tends to infinity.*

[6] Claude Shannon (1916-2001) was an American mathematician and electrical engineer, and the father of information theory.

Proof. Since $\widehat{\theta}_n^{\mathrm{ML}}$ is a solution to the likelihood equation, we have $\partial_\theta l_\theta\left(\widehat{\theta}_n^{\mathrm{ML}}\right) = 0$, so that

$$-\partial_\theta l_\theta(\theta^*) = \partial_\theta l_\theta\left(\widehat{\theta}_n^{\mathrm{ML}}\right) - \partial_\theta l_\theta(\theta^*) = \left(\widehat{\theta}_n^{\mathrm{ML}} - \theta^*\right)\int_0^1 \partial_{\theta\theta}\left(u\widehat{\theta}_n^{\mathrm{ML}} + (1-u)\theta^*\right)\mathrm{d}u =: c_n\left(\widehat{\theta}_n^{\mathrm{ML}} - \theta^*\right).$$

Now,

$$\sqrt{n}\partial_\theta l_\theta(\theta^*) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \partial_\theta l_\theta\left(X_i, \theta^*\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i,$$

where the sequence $(Z_1, \ldots, Z_n)$ is iid with mean zero and variance $I(\theta^*)$. The Central Limit Theorem therefore implies that the sequence $(\sqrt{n}\partial_\theta l_\theta(\theta^*))_{n>0}$ converges in distribution to $\mathcal{N}(0, I(\theta^*))$ as n tends to infinity. We can further show–with more involved computations–that the sequence $(c_n)_{n>0}$ converges to $I(\theta^*)$ as n tends to infinity, but in probability, and the result follows by standard combinations of probabilistic limits. □

3.3.7 Comments about estimators and bias

We gather here some thoughts and comments, wrapping up some ideas from the previous sections. Note first that, for two given estimators $\widetilde{\theta}$ and $\widehat{\theta}$, we can have $R(\theta, \widehat{\theta}) < R(\theta, \widetilde{\theta})$ for some $\theta \in \Theta$, and the reverse inequality for other values of θ, which leaves some uncertainty about the choice of the estimator. The Bayesian framework endows θ with a distribution supported on Θ and allows, in some cases, to discard trivial estimators. However, the choice of the prior distribution might be debatable, and different choices may yield different answers. A general principle, however, relies on the decomposition (3.2.1), stating that a good estimator requires both small bias and small variance.

Regarding the bias, one should not be obsessively focussed on it, as, first a biased estimator may be better in the sense of quadratic risk, but also there may not exist any such estimators sometimes. Consider for example the case of an iid sample $\mathcal{X} = (X_1, \ldots, X_n)$ distributed according to a Binomial distribution $X \sim \mathcal{B}(n, \theta)$ for $\theta \in \Theta = (0, 1)$, and assume that there exists an unbiased estimator $\widehat{\theta}(\mathcal{X})$. Therefore, for any $\theta \in \Theta$, we can write

$$\frac{1}{\theta} = \mathbb{E}_\theta\left[\widehat{\theta}(\mathcal{X})\right] = \sum_{k=0}^n \widehat{\theta}(k)\mathbb{P}(X = k) = \sum_{k=0}^n \widehat{\theta}(k)\binom{n}{k}\theta^k(1-\theta)^{n-k}.$$

Since all the terms $\widehat{\theta}(k)$ are coefficients independent of θ, then we can rewrite this equality as

$$\frac{1}{\theta^{n+1}} - \sum_{k=0}^n \widehat{\theta}(k)\binom{n}{k}\left(\frac{1}{\theta} - 1\right)^{n-k} = 0.$$

This is a polynomial in $1/\theta$ of order $n + 1$, so only admits at most $n + 1$ roots. Since this equality has to be valid for all $\theta \in \Theta$, we therefore conclude that no unbiased estimator actually exists.

Stability

Furthermore, suppose a given unbiased estimator $\widehat{\theta}$ exists, and φ is a smooth strictly convex function. Then Jensen's inequality implies that the inequality

$$\mathbb{E}\left[\varphi\left(\widehat{\theta}\right)\right] > \varphi\left(\mathbb{E}\left[\widehat{\theta}\right]\right) = \varphi(\theta),$$

so that $\varphi(\widehat{\theta})$ is biased, even though $\widehat{\theta}$ is not.

Parallelisation

Now, usually, the bias (if any) is of order $\mathcal{O}(1/n)$, and so is the variance (you can check in all the previous examples), so that, in light of the decomposition (3.2.1) of the quadratic risk, the bias is in fact negligible. Suppose now that we have one unbiased estimator $\widetilde{\theta}$ and a biased one $\widehat{\theta}$, with bias $b_n := \mathbb{E}[\widehat{\theta}] - \theta = \mathcal{O}(1/n)$. The sample size n might be large, but we have in fact access to many computer cores or many computers. Let $N := \sqrt{n}$ and consider thus N cores, each treating a subset of the sample of size N. We denote by $\widehat{\theta}_N^{(1)}, \ldots, \widehat{\theta}_N^{(N)}$ the partial estimators (assumed iid), computed on each core, and define the final estimator as

$$\widehat{\tau}_N := \frac{1}{N}\sum_{i=1}^{N}\widehat{\theta}_N^{(i)}.$$

Then

$$\mathbb{E}\left[\widehat{\tau}_N\right] = b_N(\theta), \qquad \mathbb{V}\left[\widehat{\tau}_N\right] = \sigma_N^2(\theta), \qquad \mathcal{R}^Q\left(\theta, \widehat{\tau}_N\right) = b_N(\theta)^2 + \frac{\sigma_N^2(\theta)}{N}.$$

Likewise,

$$\mathbb{E}\left[\widetilde{\tau}_N\right] = 0, \qquad \mathbb{V}\left[\widetilde{\tau}_N\right] = s_N^2(\theta), \qquad \mathcal{R}^Q\left(\theta, \widetilde{\tau}_N\right) = \frac{s_N^2(\theta)}{N}.$$

If $b_N(\theta) = b(\theta)/N$, $\sigma_N^2(\theta) = \sigma^2(\theta)/N$ $s_N^2(\theta) = s^2(\theta)/N$, then

$$\mathcal{R}^Q\left(\theta, \widehat{\tau}_N\right) = \frac{b(\theta)^2 + \sigma^2(\theta)}{N} \qquad \text{and} \qquad \mathcal{R}^Q\left(\theta, \widetilde{\tau}_N\right) = \frac{s^2(\theta)}{N}.$$

Depending on the sign of $b(\theta)^2 + \sigma^2(\theta) - s^2(\theta)$, one estimator prevails above the other.

3.4 Hypothesis testing

We now wish to construct a methodology to differentiate two possible scenarios from the data. The standard set-up is to consider the null hypothesis $\mathcal{H}_0$ versus the alternative hypothesis $\mathcal{H}_1$, corresponding to two disjoint sets Θ_0 and Θ_1 of the parameter space:

$$\mathcal{H}_0 : \theta \in \Theta_0 \qquad \text{versus} \qquad \mathcal{H}_1 : \theta \in \Theta_1. \tag{3.4.1}$$

Starting from a given sample $\mathcal{X}_n$, the rejection region $\mathcal{R}$ allows to retain or reject the hypothesis based on the range of outcomes of $\mathcal{X}_n$, in the sense that, if $\mathcal{X}_n \in \mathcal{R}$, then $\mathcal{H}_0$ is rejected, otherwise $\mathcal{H}_0$

cannot be rejected. Regarding the terminology, a hypothesis of the form $\Theta_0 = \{\theta_0\}$ is called simple, and the corresponding test is one-sided; a hypothesis of the form $\Theta_0 = \{\theta > \theta_0\}$ is called composite, and the test is two-sided. There are two types of errors pertaining to hypothesis testing. Type I errors occur when the test rejects $\mathcal{H}_0$ while it is actually true; Type II errors occur when the test keep $\mathcal{H}_0$ while $\mathcal{H}_1$ is true.

Example 3.4.1. Consider the trivial, yet motivating, example of a statistical model $\mathcal{F} = \{\mathcal{N}(\theta, 1), \theta \in \Theta\}$, with $\Theta = \{0, 1\}$. Given a sample $\mathcal{X}_n = (X_1, \ldots, X_n)$, we consider the two sets $\Theta_0 = \{0\}$ and $\Theta_1 = \{1\}$, and the corresponding hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$ in (3.4.1). Since the empirical mean $\overline{X}_n$ is a good estimator of the true value for a large enough sample, we could consider the following test: reject $\mathcal{H}_0$ if $\overline{X}_n > 1/2$, so that the rejection region reads

$$\mathcal{R} = \left\{\mathcal{X}_n : \overline{X}_n > \frac{1}{2}\right\}.$$

Note that here, both the null and the alternative hypotheses are simple.

3.4.1 Simple tests

In simple hypothesis testing $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$, we can write these two types of errors as

$$\begin{aligned} &\text{Type I Error:} \qquad \mathbb{P}_{\theta_0}(\mathcal{X}_n \in \mathcal{R}), \\ &\text{Type II Error:} \qquad \mathbb{P}_{\theta_1}(\mathcal{X}_n \notin \mathcal{R}). \end{aligned} \tag{3.4.2}$$

The goal of any test is obviously to minimise the error. However, in order to minimise the Type I error, one needs to consider a small rejection region $\mathcal{R}$, which in turn is going to yield a large Type II error, so some balance needs to be set between the two. The idea is to set an acceptable threshold for the error, as follows:

Definition 3.4.2. The power function $\beta(\cdot)$ and the level $\alpha \in (0, 1)$ of a test with rejection region $\mathcal{R}$ are defined as

$$\beta(\mathcal{R}) := \mathbb{P}_{\theta_1}(\mathcal{X}_n \in \mathcal{R}) \qquad \text{and} \qquad \mathbb{P}_{\theta_0}(\mathcal{X}_n \in \mathcal{R}) \le \alpha.$$

If $\mathbb{P}_{\theta_0}(\mathcal{X}_n \in \mathcal{R}) = \alpha$, then α is called the size of the test.

In order to simplify the terminology, we shall call a test with rejection region $\mathcal{R}$ and $\mathcal{R}$-test from now on.

Definition 3.4.3. For a given level α, an $\mathcal{R}^*$-test with level α is called the most powerful test if $\beta(\mathcal{R}^*) \ge \beta(\mathcal{R})$ for any $\mathcal{R}$-test of level α.

The higher the power function the lower the Type II error, with a given bound on the Type I error. There are many such hypothesis tests in the literature, including the Wald test (which assumes that $\widehat{\theta}$ is asymptotically Gaussian), and the Neyman-Pearson test, which we present now. Define a rejection region of the form

$$\mathcal{R} := \{\mathcal{X}_n : \mathfrak{R}_n > c\}, \tag{3.4.3}$$

for some $c > 0$, where $\mathcal{L}_\theta(\mathcal{X}_n) := \prod_{i=1}^n f_\theta(X_i)$ denotes the likelihood function, and $\mathfrak{R}_n := \dfrac{\mathcal{L}_{\theta_1}(\mathcal{X}_n)}{\mathcal{L}_{\theta_0}(\mathcal{X}_n)}$ is a random variable called the likelihood ratio.

Theorem 3.4.4. *[Neyman-Pearson Lemma] Let $\alpha \in (0,1)$. If there exists $c^* > 0$ such that $\mathbb{P}_{\theta_0}(\mathfrak{R}_n > c^*) = \alpha$, then the test is the most powerful test of level α.*

Proof. Proving the theorem is equivalent to showing that for all rejection region $\mathcal{R}$ such that $\mathbb{P}_{\theta_0}(\mathcal{X}_n \in \mathcal{R}) \le \alpha$, then $\mathbb{P}_{\theta_1}(\mathcal{X}_n \notin \mathcal{R}) \ge \mathbb{P}_{\theta_1}(\mathcal{X}_n \notin \mathcal{R}^*)$, or else $\beta(\mathcal{R}) \le \beta(\mathcal{R}^*)$, where we denote $\mathcal{R}^*$ the optimal rejection region corresponding to the optimal value of c^* in the theorem. Now,

$$\mathbb{P}_{\theta_1}(\mathcal{X}_n \in \mathcal{R}^*) - \mathbb{P}_{\theta_1}(\mathcal{X}_n \in \mathcal{R}) = \int_{\mathcal{R}^*} \mathcal{L}_{\theta_1}(\mathrm{x})\mathrm{dx} - \int_{\mathcal{R}} \mathcal{L}_{\theta_1}(\mathrm{x})\mathrm{dx} = \left(\int_{\mathcal{R}^*\setminus\mathcal{R}} - \int_{\mathcal{R}\setminus\mathcal{R}^*}\right)\mathcal{L}_{\theta_1}(\mathrm{x})\mathrm{dx}.$$

Since $\mathcal{R}^* \setminus \mathcal{R} \subset \mathcal{R}^*$, then $\mathfrak{R}_n > c^*$ on this set, and obviously $\mathfrak{R}_n \le c^*$ on $\mathcal{R} \setminus \mathcal{R}^*$. Therefore,

$$\begin{aligned}
\left(\int_{\mathcal{R}^*\setminus\mathcal{R}} - \int_{\mathcal{R}\setminus\mathcal{R}^*}\right)\mathcal{L}_{\theta_1}(\mathrm{x})\mathrm{dx} &\ge c^*\left(\int_{\mathcal{R}^*\setminus\mathcal{R}} - \int_{\mathcal{R}\setminus\mathcal{R}^*}\right)\mathcal{L}_{\theta_0}(\mathrm{x})\mathrm{dx} \\
&= c^*\left(\int_{\mathcal{R}^*} - \int_{\mathcal{R}}\right)\mathcal{L}_{\theta_0}(\mathrm{x})\mathrm{dx} \\
&= c^*\left[\mathbb{P}_{\theta_0}(\mathcal{X}_n \in \mathcal{R}^*) - \mathbb{P}_{\theta_0}(\mathcal{X}_n \in \mathcal{R})\right].
\end{aligned}$$

By assumption, $\mathbb{P}_{\theta_0}(\mathcal{X}_n \in \mathcal{R}) \le \alpha$ and $\mathbb{P}_{\theta_0}(\mathcal{X}_n \in \mathcal{R}^*) = \alpha$, therefore, the right-hand side of the last inequality is non negative, and the theorem follows. □

We now introduce one of the key concepts in hypothesis testing, called the p-value, which we shall denote by π_0:

Definition 3.4.5. Consider a test of size $\alpha \in (0,1)$ and corresponding rejection region $\mathcal{R}_\alpha$. The p-value π_0 is defined as the smallest level at which the null hypothesis can be rejected, i.e.

$$\pi_0 := \inf\{\alpha : \mathcal{X}_n \in \mathcal{R}_\alpha\}.$$

Clearly the possible range of values is $(0,1)$. When the p-value is below 1%, then there is very strong evidence that the null hypothesis should be rejected; the range $(1\%, 5\%)$ represents strong evidence, $(5\%, 10\%)$ weak evidence, and when the p-value is greater than 10%, then the test is inconclusive, in the sense that we cannot decently reject the null hypothesis.

Example 3.4.6. We consider again the statistical model $\mathcal{F} = \{\mathcal{N}(\theta, \sigma^2), \theta \in \Theta\}$, where the variance σ^2 is known, and we consider

$$\mathcal{H}_0 : \theta \in \Theta_0 = \{0\} \qquad \text{versus} \qquad \mathcal{H}_1 : \theta \in \Theta_1 = \{1\}.$$

The likelihood function reads (see Example 3.3.7)

$$\mathcal{L}_\theta(\mathcal{X}_n) = \left(\sigma\sqrt{2\pi}\right)^n \prod_{i=1}^n \exp\left\{-\frac{(X_i-\theta)^2}{2\sigma^2}\right\},$$

and hence the likelihood ratio can be computed as

$$\mathfrak{R}_n = \frac{\mathcal{L}_{\theta_1}(\mathcal{X}_n)}{\mathcal{L}_{\theta_0}(\mathcal{X}_n)} = \exp\left\{\frac{n}{2\sigma^2}\left(2\overline{X}-1\right)\right\}.$$

The rejection region (3.4.3) therefore reads explicitly

$$\mathcal{R} = \left\{\exp\left\{\frac{n}{2\sigma^2}\left(2\overline{X}-1\right)\right\} \geq \widetilde{c}\right\} = \left\{\overline{X} > c\right\}, \tag{3.4.4}$$

for some $\widetilde{c} > 0$ with $c = \frac{1}{2} + \frac{\sigma^2}{n}\log(\widetilde{c})$. To choose c, we equate $\mathbb{P}_{\theta_0}(\mathcal{X}_n \in \mathcal{R}) = \alpha = \mathbb{P}_{\theta_0}(\overline{X} \geq c)$. Since the sample is Gaussian $\mathcal{N}(0,\sigma^2)$ under the null hypothesis, then $\overline{X} \sim \mathcal{N}(0, \frac{\sigma^2}{n})$, and therefore

$$c_\alpha = \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha) = \frac{\sigma}{\sqrt{n}}q_{1-\alpha}. \tag{3.4.5}$$

To compute the power of the test, we can write

$$\mathbb{P}_{\theta_1}\left(\mathcal{X}_n \in \mathcal{R}\right) = \mathbb{P}_{\theta_1}\left(\overline{X} > c\right) = \mathbb{P}_{\theta_1}\left(\mathcal{N}(0,1) > \frac{(c-\theta_1)\sqrt{n}}{\sigma}\right) = 1 - \Phi\left(\frac{(c-\theta_1)\sqrt{n}}{\sigma}\right).$$

Recall now that the p-value π_0 of a test is, for a given fixed sample, the largest value of α such that the null hypothesis $\mathcal{H}_0$ is not rejected. From (3.4.4), for a given level α, the rejection region is of the form $\mathcal{R} = \mathcal{R}_\alpha = \left\{\overline{X} > c_\alpha\right\}$, with c_α given in (3.4.5), or equivalently

$$1-\alpha = \Phi\left(\frac{\sqrt{n}c_\alpha}{\sigma}\right).$$

Therefore, for a given $\overline{X}$, the threshold from accepting $\mathcal{H}_0$ to rejecting it is $\alpha^*(\overline{X})$ such that $c_\alpha = \overline{X}$, i.e.

$$1-\alpha^*(\overline{X}) = \Phi\left(\frac{\sqrt{n}c_\alpha}{\sigma}\right) = \Phi\left(\frac{\sqrt{n}\overline{X}}{\sigma}\right),$$

and therefore the critical (p-value) threshold is equal to

$$\pi_0 = \alpha^*(\overline{X}) = 1 - \Phi\left(\frac{\overline{X}\sqrt{n}}{\sigma}\right).$$

3.4.2 Composite tests

We now consider composite tests, that is tests of the form $\Theta_0 = \{\theta > \theta_0\}$ versus $\Theta_1 = \{\theta \leq \theta_0\}$. We slightly modify the definitions of the error (3.4.2) in the following form:

$$\text{Type I Error:} \qquad \sup_{\theta\in\Theta_0} \mathbb{P}_\theta(\mathcal{X}_n \in \mathcal{R}). \tag{3.4.6}$$

Note that, in the composite case, we cannot make sense of the notion of Type II errors, but we shall use, similar to the simple case, the notions of level, size and power of a test:

Definition 3.4.7. The level $\alpha \in (0,1)$ of a test with rejection region $\mathcal{R}$ is such that

$$\sup_{\theta\in\Theta_0} \mathbb{P}_\theta\left(\mathcal{X}_n \in \mathcal{R}\right) \leq \alpha.$$

If the supremum is equal to α, then α is called the size of the test. The power function $\beta : \Theta \to [0,1]$ is defined as

$$\beta(\theta) := \mathbb{P}_\theta\left(\mathcal{X}_n \in \mathcal{R}\right).$$

This is a slight abuse of language as we previously defined the power function as a function of sets, but it should not create any confusion here. In order to extend the Neyman-Pearson lemma to the composite case, we need to introduce the following terminology:

Definition 3.4.8. A test $\mathcal{R}^*$ with level α is called Uniformly Most Powerful (UMP) if $\mathbb{P}_\theta\left(\mathcal{X}_n \in \mathcal{R}\right) \leq \mathbb{P}_\theta\left(\mathcal{X}_n \in \mathcal{R}^*\right)$ for all $\theta \in \Theta_1$ and any test $\mathcal{R}$ of level α.

A test is called consistent if $\beta(\theta)$ converges to 1 as the sample size tends to infinity, for any $\theta \in \Theta_1$, and is called unbiased if

$$\sup_{\theta\in\Theta_0} \beta(\theta) \leq \inf_{\theta\in\Theta_1} \beta(\theta).$$

Example 3.4.9. Consider $\mathcal{F} = \{\mathcal{N}(\theta,\sigma^2), \theta \in \mathbb{R}\}$, with $\sigma > 0$ known, and the hypotheses $\Theta_0 = (-\infty, 0]$ and $\Theta_1 = (0, +\infty)$. Consider the test with rejection region

$$\mathcal{R} := \{\overline{X} > c_\alpha\}, \qquad \text{with } c_\alpha := \frac{\sigma}{\sqrt{n}} q_{1-\alpha}. \tag{3.4.7}$$

Since, for any $\theta \in \mathbb{R}$, the random variable $\sqrt{n}\frac{\overline{X}-\theta}{\sigma} \sim \mathcal{N}(0,1)$, the power function reads

$$\begin{aligned}\beta(\theta) = \mathbb{P}(\overline{X} > c_\alpha) = \mathbb{P}_\theta\left(\frac{\sqrt{n}(\overline{X}-\theta)}{\sigma} > \frac{\sqrt{n}(c_\alpha-\theta)}{\sigma}\right) = 1 - \Phi\left(\frac{\sqrt{n}(c_\alpha-\theta)}{\sigma}\right) \\ = \Phi\left(\frac{\sqrt{n}(\theta-c_\alpha)}{\sigma}\right) = \Phi\left(\frac{\sqrt{n}\theta}{\sigma} - q_{1-\alpha}\right).\end{aligned}$$

Note further that $\beta(0) = \alpha$. Since the function Φ is monotone, then

$$\sup_{\theta\in\Theta_0} \mathbb{P}_\theta(\mathcal{X}_n \in \mathcal{R}) = \sup_{\theta\in\Theta_0} \beta(\theta) = \beta(0) = \Phi\left(q_{1-\alpha}\right) = \alpha.$$

Fix now some value $\theta' \in \Theta_1$, and consider the simple hypotheses

$$\widetilde{\mathcal{H}}_0 : \theta = 0 \qquad \text{vs} \qquad \widetilde{\mathcal{H}}_1 : \theta = \theta'.$$

By Neyman-Pearson's lemma (Theorem 3.4.4) and Example 3.4.6, the test $\mathcal{R}$ in (3.4.7) satisfies

$$\mathbb{P}_{\theta'}(\mathcal{X}_n \in \mathcal{R}) \geq \mathbb{P}_{\theta'}(\mathcal{X}_n \in \mathcal{R}_\alpha), \tag{3.4.8}$$

for any test $\mathcal{R}_\alpha$ of level α (e.g. such that $\mathbb{P}_0(\mathcal{X}_n \in \mathcal{R}_\alpha) \leq \alpha$). Since $0 \in \Theta_0$, any test satisfying

$$\sup_{\theta\in\Theta_0} \mathbb{P}_\theta(\mathcal{X}_n \in \mathcal{R}_\alpha) \leq \alpha$$

also satisfies $\mathbb{P}_0(\mathcal{X}_n \in \mathcal{R}_\alpha) \leq \alpha$, and therefore, for any test $\mathcal{R}_\alpha$ of level α for the null hypothesis $\mathcal{H}_0 : \theta \leq 0$ against $\mathcal{H}_1 : \theta > 0$, and for any $\theta' > 0$, the inequality (3.4.8) holds, so that $\mathcal{R}$ is UMP.

Exercise 24. Show that the test in Example 3.4.9 is consistent and unbiased.

Remark 3.4.10. In our recurring example above, we always assumed that the variance σ^2 of the Gaussian sample was known. In case it is not, we can however replace it by the unbiased estimator $ns_n^2/(n-1)$, where

$$s_n^2 := \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

In this case, the quantiles appearing in the rejection region will not be those of the Gaussian distribution any longer, but those of the Student distribution.

Exercise 25. Consider $\mathcal{F} = \{\mathcal{N}(\mu, \theta^2), \theta > 0\}$, with $\mu \in \mathbb{R}$ known, and the hypotheses $\Theta_0 = (0, \sigma_0]$ and $\Theta_1 = (\sigma_0, +\infty)$, for some $\sigma_0 > 0$. Analyse the test defined by

$$\mathcal{R} := \left\{\frac{\mathcal{L}_\theta(\mathcal{X}_n)}{\mathcal{L}_{\sigma_0}(\mathcal{X}_n)} > c\right\},$$

for some constant $c > 0$ to be determined, where $\mathcal{L}$ denote as usual the likelihood function.

Solution. *The log-likelihood ratio takes the form*

$$\frac{\mathcal{L}_\theta(\mathcal{X}_n)}{\mathcal{L}_{\sigma_0}(\mathcal{X}_n)} = \frac{\theta}{\sigma_0}\exp\left\{\left(\frac{1}{2\sigma_0^2} - \frac{1}{2\theta^2}\right)S_n\right\},$$

with $S_n := \sum_{i=1}^{n}(X_i - \mu)^2$, so that the rejection region can be written $\mathcal{R} = \{S_n > \tilde{c}\}$. As before, we choose the constant $\tilde{c}$ such that $\mathbb{P}_{\sigma_0}(\mathcal{X}_n \in \mathcal{R}) = \alpha$. Since the sample is assumed to be Gaussian, the random variable S_n/σ_0^2 follows, under $\mathbb{P}_{\sigma_0}$, a Chi-Squared χ_n^2 distribution, and hence

$$\mathbb{P}_{\sigma_0}(\mathcal{X}_n \in \mathcal{R}) = \mathbb{P}_{\sigma_0}(S_n > \tilde{c}) = \mathbb{P}_{\sigma_0}\left(\frac{S_n}{\sigma_0^2} > \frac{\tilde{c}}{\sigma_0^2}\right) = 1 - F_{\chi_n^2}\left(\frac{\tilde{c}}{\sigma_0^2}\right),$$

and therefore $\tilde{c} = q_{1-\alpha}^{\chi_n^2}\sigma_0^2$. The power of the test can then be computed as

$$\beta(\theta) = \mathbb{P}_\theta(\mathcal{X}_n \in \mathcal{R}) = \mathbb{P}_\theta\left(S_n > \tilde{c}\right) = \mathbb{P}_\theta\left(S_n > q_{1-\alpha}^{\chi_n^2}\sigma_0^2\right) = \mathbb{P}_\theta\left(\frac{S_n}{\theta^2} > \frac{\sigma_0^2}{\theta^2}q_{1-\alpha}^{\chi_n^2}\right) = 1 - F_{\chi_n^2}\left(\frac{\sigma_0^2}{\theta^2}q_{1-\alpha}^{\chi_n^2}\right),$$

and it is then easy to see that the test is of level α since

$$\beta(\theta) = 1 - F_{\chi_n^2}\left(\frac{\sigma_0^2}{\theta^2}q_{1-\alpha}^{\chi_n^2}\right) \leq 1 - F_{\chi_n^2}\left(q_{1-\alpha}^{\chi_n^2}\right) = \alpha.$$

3.4.3 Comparison of two Gaussian samples

We are interested here in comparing two samples $\mathcal{X}_{n_x}$ and $\mathcal{Y}_{n_y}$, respectively sampled from $\mathcal{N}(\mu_x, \sigma_x^2)$ and $\mathcal{N}(\mu_y, \sigma_y^2)$, and assumed to be independent. We only study here the simpler case of comparing the means knowing the variances, but of course a similar analysis can be performed for the variance as well.

Comparing the means

We assume that $\sigma_x = \sigma_y = \sigma > 0$ is known, and we wish to test the null hypothesis $\mathcal{H}_0 : \mu_x = \mu_y$ against the alternative $\mathcal{H}_1 : \mu_x \neq \mu_y$. Let

$$s_z^2 := \frac{1}{n_z}\sum_{i=1}^{n_z}\left(Z_i - \overline{Z}\right)^2, \qquad \text{for } (z, Z) \in \{(x, X), (y, Y)\}.$$

Let $n := n_x + n_y$. Since, for $z \in \{x, y\}$, $n_z s_z^2/\sigma_z^2$ follows a Chi Square distribution with $n_z - 1$ degrees of freedom, we deduce that

$$\Xi := \frac{n_x s_x^2 + n_y s_y^2}{\sigma^2} \sim \chi^2_{n-2}.$$

If the two means μ_x and μ_y are equal, then $\widetilde{Z} := \sqrt{n}\left(\overline{X} - \overline{Y}\right)/\sigma$ is centered Gaussian with variance equal to $\frac{n}{n_x} + \frac{n}{n_y}$. Let now $m := \left(\frac{n_x n_y (n-2)}{n(n_x s_x^2 + n_y s_y^2)}\right)^{1/2}$. Under the null hypothesis, we can write

$$\begin{aligned} m\left(\overline{X} - \overline{Y}\right) &= \left(\frac{n_x n_y (n-2)}{n(n_x s_x^2 + n_y s_y^2)}\right)^{1/2}\left(\overline{X} - \overline{Y}\right) \\ &= \left(\frac{n_x n_y (n-2)}{n(n_x s_x^2 + n_y s_y^2)}\right)^{1/2}\frac{\sigma\widetilde{Z}}{\sqrt{n}} \\ &= \left(\frac{n_x n_y (n-2)}{\Xi}\right)^{1/2}\frac{\widetilde{Z}}{n} = \frac{Z}{\sqrt{\Xi/(n-2)}}, \end{aligned}$$

where $Z \in \mathcal{N}(0, 1)$. In light of Definition 2.4.9, the random variable $m\left(\overline{X} - \overline{Y}\right)$ follows a Student-t_{n-2} distribution with $n - 2$ degrees of freedom, so that we can consider the rejection region $\mathcal{R} = \left\{\left|\overline{X} - \overline{Y}\right| > cm\right\}$. Picking $c = q^{t_{n-2}}_{1-\alpha/2}$, we obtain a test of size α.

3.4.4 Confidence intervals

Definition 3.4.11. Let $\alpha \in (0, 1)$. The $1 - \alpha$ confidence set $\mathcal{C}_n$ for θ, in general depending on the data, is such that

$$\mathbb{P}_\theta\left(\theta \in \mathcal{C}_n\right) \geq 1 - \alpha, \qquad \text{for all } \theta \in \Theta.$$

Example 3.4.12. Consider the statistical model $\mathcal{F} = \{\mathcal{N}(\theta, \sigma^2), \theta \in \mathbb{R}\}$ for $\sigma > 0$ known, and let $\alpha \in (0, 1)$. Consider now the (random) interval

$$\mathcal{C}_n := \left[\overline{X} - \frac{\sigma}{\sqrt{n}}q_{1-\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}}q_{1-\alpha/2}\right].$$

Then we can compute, for any $\theta \in \mathbb{R}$,

$$\mathbb{P}_\theta(\theta \in \mathcal{C}_n) = \mathbb{P}_\theta\left(\left|\overline{X} - \theta\right| \leq \frac{\sigma}{\sqrt{n}}q_{1-\alpha/2}\right) = \mathbb{P}_\theta\left(|Z| \leq q_{1-\alpha/2}\right) = 1 - \alpha,$$

where $Z \sim \mathcal{N}(0, 1)$, so that $\mathcal{C}_n$ is indeed a confidence interval of level $1 - \alpha$.

Exercise 26. For $\alpha \in (0,1)$ and the statistical model $\mathcal{F} = \{\mathcal{N}(\theta, \sigma^2), \theta \in \mathbb{R}\}$ for $\sigma > 0$ known, show that the (random) interval

$$\mathcal{C}_n := \left[\overline{X} - \frac{\sigma}{\sqrt{n}} q_{1-3\alpha/4}, \overline{X} + \frac{\sigma}{\sqrt{n}} q_{1-\alpha/4}\right].$$

is also a confidence interval of level $1-\alpha$. How does it compare to the one in Example 3.4.12?

Exercise 27 (Confidence interval for Bernoulli draws)**.** Consider the Bernoulli example above (Exercise 17), and denote $\mathcal{C}_n = (\widehat{\theta}_n - \widehat{\varepsilon}_n, \widehat{\theta}_n + \widehat{\varepsilon}_n)$. Show, using Theorem 2.2.7, that for any $\theta > 0$, $\mathbb{P}(\theta \in \mathcal{C}_n) \geq 1 - \alpha$ holds, where $2n\widehat{\varepsilon}_n = \log(2/\alpha)$.

It often happens that we can construct confidence intervals based on the Gaussian distribution:

Theorem 3.4.13. *Assume that $\widehat{\theta}_n \sim \mathcal{N}(\theta, \widehat{\sigma}_n^2)$, and define the interval*

$$\mathcal{C}_n := \left(\widehat{\theta}_n - z_{\alpha/2}\widehat{\sigma}_n, \widehat{\theta}_n + z_{\alpha/2}\widehat{\sigma}_n\right),$$

where $z_{\alpha/2} := \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$, with Φ the Gaussian cdf. Then $\lim_{n\uparrow\infty} \mathbb{P}_\theta(\theta \in \mathcal{C}_n) = 1 - \alpha$.

Proof. Let $Z \in \mathcal{N}(0,1)$. Note first that the definition of $z_{\alpha/2}$ is equivalent to $\mathbb{P}(Z > z_{\alpha/2}) = \alpha/2$, or $\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$. Therefore,

$$\mathbb{P}_\theta(\theta \in \mathcal{C}_n) = \mathbb{P}_\theta\left(\widehat{\theta}_n - z_{\alpha/2}\widehat{\sigma}_n < \theta < \widehat{\theta}_n + z_{\alpha/2}\widehat{\sigma}_n\right) = \mathbb{P}_\theta\left(-z_{\alpha/2} < \frac{\theta - \widehat{\theta}_n}{\widehat{\sigma}_n} < z_{\alpha/2}\right).$$

Since the random sequence $\left(\frac{\theta - \widehat{\theta}_n}{\widehat{\sigma}_n}\right)_{n>0}$ converges in probability to $\mathcal{N}(0,1)$ as n tends to infinity, the theorem follows. □

Exercise 28. Construct such a confidence interval for Bernoulli draws and compare it with the one in Exercise 27.

Confidence interval for the cumulative distribution function

We built so far estimators and confidence intervals thereof; those were parametric in the sense that we assumed the true distribution to be known up to knowledge of its parameters. We may however challenge this and, getting back to the empirical cdf constructed at the very beginning of the chapter, try and determine whether the latter is in fact a good estimator for the true cdf. As proved in the discussion following Definition 3.1.1, we show that, pointwise for any $x \in \mathbb{R}$, the empirical cdf $\widehat{F}_n(x)$ is distributed as a Binomial distribution, so that the Central Limit Theorem implies that

$$\frac{\sqrt{n}\left(\widehat{F}_n(x) - F(x)\right)}{\sqrt{F(x)(1 - F(x))}}$$

converges in distribution to a centered Gaussian distribution with unit variance as n tends to infinity. Now, $F(x)$ is by definition unknown. However, since $\widehat{F}_n(x)$ converges in probability to $F(x)$, then Slutsky's theorem implies that

$$\frac{\sqrt{n}\left(\widehat{F}_n(x) - F(x)\right)}{\sqrt{\widehat{F}_n(x)(1-\widehat{F}_n(x))}}$$

still converges in distribution to a centered Gaussian distribution with unit variance. Therefore, for any $\alpha \in (0,1)$, a $(1-\alpha)$ confidence interval for $F(x)$ is given by

$$\mathcal{C}_n(\alpha) = \left[\widehat{F}_n(x) - q_{1-\alpha}\sqrt{\frac{\widehat{F}_n(x)\left(1-\widehat{F}_n(x)\right)}{n}}, \widehat{F}_n(x) + q_{1-\alpha}\sqrt{\frac{\widehat{F}_n(x)\left(1-\widehat{F}_n(x)\right)}{n}}\right].$$

In terms of hypothesis testing, pointwise again, it makes sense to consider the following test:

$$\mathcal{H}_0 : F(x) = F_0(x) \qquad \text{vs} \qquad \mathcal{H}_1 : F(x) \neq F_0(x), \tag{3.4.9}$$

for some given $F_0(x)$. Following similar steps to before, we can construct a test of level α based on a rejection region of the form

$$\mathcal{R} = \left\{ \frac{\left|\widehat{F}_n(x) - F_0(x)\right|}{\sqrt{F_0(x)\left(1-F_0(x)\right)}} > \frac{q_{\alpha/2}}{\sqrt{n}} \right\}.$$

Testing for the distribution

The previous paragraph focuses on estimator a cumulative distribution pointwise. We now wish to extend this to a uniform statement. Consider the so-called Kolmogorov-Smirnov statistic

$$D_n := \sup_{x\in\mathbb{R}} \left|\widehat{F}_n(x) - F(x)\right|.$$

Glivenko-Cantelli' result (Theorem 3.1.2) states that D_n converges almost surely to zero as n tends to infinity. This was refined by Kolmogorov as follows:

Theorem 3.4.14 (Kolmogorov-Smirnov). *For any $z \in \mathbb{R}$, the probability $\mathbb{P}\left(\sqrt{n}D_n \leq z\right)$ converges in distribution to $H(z)$ where the function H is the cumulative distribution function of the Kolmogorov-Smirnov distribution and is given explicitly by*

$$H(z) := 1 - 2\sum_{k\geq 1}(-1)^{k-1}\exp\left(-2k^2 z\right).$$

Remark 3.4.15. In fact, the distribution $H(\cdot)$ appearing in the theorem is exactly that of the supremum of the Brownian bridge on $[0,1]$, so that the theorem can equivalently be stated as the convergence in distribution of $\sqrt{n}D_n$ to $\sup\{|B(F_0(t)|, t \in [0,1]\}$, where B is a Brownian bridge and F_0 the hypothesized distribution.

Consider now the test

$$\mathcal{H}_0 : F = F_0 \qquad \text{vs} \qquad \mathcal{H}_1 : F \neq F_0,$$

for some given cdf F_0. Note that this represents a uniform version of the test (3.4.9). Now, under the null hypothesis $\mathcal{H}_0$, since F_0 is given a priori, the distribution of D_n, for any n fixed, can be tabulated. If $\mathcal{H}_0$ fails, however, calling F the true cdf, we know by the law of large numbers that $\widehat{F}_n$ converges to F, so that, for large enough n, $D_n > \delta$, for some $\delta > 0$, and hence $\sqrt{n}D_n > \delta\sqrt{n}$ and $\sqrt{n}D_n$ clearly tends to infinity as n becomes large. We can therefore construct a rejection region of the form

$$\mathcal{R} = \left\{\sqrt{n}D_n > c\right\}$$

Now, the Type-I error reads

$$\mathbb{P}_{\mathcal{H}_0}(\mathcal{X}_n \in \mathcal{R}) = \mathbb{P}_{\mathcal{H}_0}\left(\sqrt{n}D_n > c\right),$$

which converges to $1-H(c)$ as n tends to infinity. This corresponds to an asymptotic level $\alpha \in (0,1)$ if $1 - H(c) = \alpha$, and we can therefore determine the threshold c as

$$c = H^{-1}(1-\alpha).$$

3.4.5 Asymptotic tests

The tests discussed above are based on some knowledge of the distribution, which is rarely the case in practice. Suppose that the iid sequence $(X_1, \ldots, X_n)$ has constant mean $\theta \in \mathbb{R}$ and finite strictly positive variance. The Central Limit Theorem implies that $\sqrt{n}(\overline{X} - \theta)/\sigma(\theta)$ converges in distribution, under $\mathbb{P}_\theta$, to $\mathcal{N}(0,1)$ as n tends to infinity, where $\sigma^2(\theta) := \mathbb{V}_\theta(X_1)$. If the map $\sigma(\cdot)$ is continuous, using the fact that $\overline{X}$ converges to θ in probability, then Slutsky's theorem yields that $\sqrt{n}(\overline{X} - \theta)/\sigma(\overline{X})$ converges in distribution, under $\mathbb{P}_\theta$, to $\mathcal{N}(0,1)$ as n tends to infinity. Consider therefore the following test:

$$\mathcal{H}_0 : \theta = \theta_0 \qquad \text{vs} \qquad \theta > \theta_0,$$

with rejection region

$$\mathcal{R} := \left\{\overline{X} > \theta_0 + \frac{\sigma(\overline{X})}{\sqrt{n}} q_{1-\alpha}\right\}.$$

Then

$$\lim_{n\uparrow\infty} \mathbb{P}_\theta(\mathcal{X}_n \in \mathcal{R}) = \lim_{n\uparrow\infty} \mathbb{P}_\theta\left(\overline{X} > \theta_0 + \frac{\sigma(\overline{X})}{\sqrt{n}} q_{1-\alpha}\right) = \lim_{n\uparrow\infty} \mathbb{P}_\theta\left(\frac{\sqrt{n}(\overline{X} - \theta_0)}{\sigma(\overline{X})} > q_{1-\alpha}\right) = \alpha,$$

This leads us to the following definition:

Definition 3.4.16. A test $\mathcal{R}$ of the null hypothesis $\mathcal{H}_0 : \theta \in \Theta_0$ vs the alternative $\mathcal{H}_1 : \theta \in \Theta_1$ is called a test of asymptotic level α if

$$\sup_{\theta\in\Theta_0} \lim_{n\uparrow\infty} \mathbb{P}_\theta(\mathcal{X}_n \in \mathcal{R}) \leq \alpha.$$

Example 3.4.17. Consider the maximum likelihood estimator $\widehat{\theta}_n^{\mathrm{ML}}$ for some statistical model. Under the regularity hypotheses, we know that it converges in probability to θ and Theorem 3.3.21 implies that $\sqrt{nI(\theta)}\left(\widehat{\theta}_n^{\mathrm{ML}} - \theta\right)$ converges in distribution to $\mathcal{N}(0,1)$ as n tends to infinity. Assuming that the Fisher information $I(\cdot)$ is continuous, then, similarly to above, $I(\widehat{\theta}_n^{\mathrm{ML}})$ converges in probability to $I(\theta)$ and $\sqrt{nI(\widehat{\theta}_n^{\mathrm{ML}})}\left(\widehat{\theta}_n^{\mathrm{ML}} - \theta\right)$ converges in distribution to $\mathcal{N}(0,1)$ as n tends to infinity. This provides a natural test of asymptotic level α as

$$\mathcal{H}_0 : \theta = \theta_0 \qquad \text{vs} \qquad \theta \neq \theta_0,$$

with rejection region

$$\mathcal{R} := \left\{ \left|\widehat{\theta}_n^{\mathrm{ML}} - \theta_0\right| > \frac{q_{1-\alpha/2}}{\sqrt{nI(\widehat{\theta}_n^{\mathrm{ML}})}} \right\}.$$

3.5 Bootstrap

Let $\mathcal{S}_n := \mathcal{S}(\mathcal{X}_n)$ be a statistic, for which we need to compute the variance $\mathbb{V}[\mathcal{S}_n]$. Note here that the variance is computed from the unknown (parametric) distribution, which, for estimation, may or may not be parametric. We consider the non-parametric case, and use the empirical distribution function $\widehat{F}_n$ instead. Bootstrap then consists in, first approximating $\mathbb{V}[\mathcal{S}_n]$ by $\mathbb{V}_{\widehat{F}_n}[\mathcal{S}_n]$, then by approximating the latter by simulation. Consider the iid sample $\mathcal{X}_n = (X_1, \ldots, X_n)$ from a common (unknown) distribution F. The law of large numbers (Section 2.2.5) implies that $\overline{X} := n^{-1}\sum_{i=1}^n X_i$ converges in probability to $\mathbb{E}[X]$ as the number n of samples tends to infinity. By continuity, we also see that $n^{-1}\sum_{i=1}^n g(X_i)$ converges in probability to $\mathbb{E}[g(X)]$ for any smooth function $g()$, in particular

$$\lim_{n\uparrow\infty} \frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2 = \mathbb{V}[X] \text{ in probability.}$$

In the general case where we wish to compute $\mathbb{V}[\mathcal{S}_n]$, it is then clear that we need to compute

$$\frac{1}{m}\sum_{i=1}^m \left(\mathcal{S}_{n,i} - \frac{1}{m}\sum_{k=1}^m \mathcal{S}_{n,k}\right)^2,$$

for some m, where each $\mathcal{S}_{n,\cdot}$ is computed from sampling n values from $\widehat{F}_n$, repeated for $i = 1, \ldots, m$.

Exercise 29. Generate (in `Python`) $\mathcal{X}_n$, with $n = 10^6$ observations, from a common $\mathcal{N}(0,1)$ distribution, and compute the empirical distribution function $\widehat{F}_n$. Compute the median of the empirical distribution, and show the convergence of its variance as a function of m.

Bootstrap confidence interval

We consider a statistic $\widehat{\theta}_n = \mathcal{S}(\mathcal{X}_n)$ of the true value of some parameter θ, and try to determine confidence intervals at the level $1-\alpha$, for some $\alpha \in (0,1)$ (see Definition 3.4.11). The Normal interval is the simplest one and take the form (recall Theorem 3.4.13)

$$\mathcal{C}_n = \left(\widehat{\theta}_n - z_{\alpha/2}\widehat{\sigma}_n, \widehat{\theta}_n + z_{\alpha/2}\widehat{\sigma}_n\right),$$

where $\widehat{\sigma}_n$ is the bootstrap estimate of the standard error.

Another confidence is called the Pivotal interval, and works as follows: define the pivot $\pi_n := \widehat{\theta}_n - \theta$, and $\theta_{n,1}, \ldots, \theta_{n,m}$ bootstraps replications. We further let $F_{\pi_n}(x) := \mathbb{P}_F(\pi_n \leq x)$ denote the cumulative distribution of the pivot, in the true model F.

Lemma 3.5.1 (Pivot Confidence interval)**.** *The interval defined as*

$$\left(\widehat{\theta}_n - F_{\pi_n}^{-1}\left(1 - \frac{\alpha}{2}\right), \widehat{\theta}_n - F_{\pi_n}^{-1}\left(\frac{\alpha}{2}\right)\right)$$

is a $1-\alpha$ confidence interval for any $\alpha \in (0,1)$.

Proof. Let c_-, c_+ denote the lower and upper bound of the interval. The proof follows from the direct computation

$$\begin{aligned}
\mathbb{P}\left(c_- \leq \theta \leq c_+\right) &= \mathbb{P}\left(c_- - \widehat{\theta}_n \leq \theta - \widehat{\theta}_n \leq c_+ - \widehat{\theta}_n\right)\\
&= \mathbb{P}\left(\widehat{\theta}_n - c_+ \leq \pi_n \leq \widehat{\theta}_n - c_-\right)\\
&= F_{\pi_n}\left(\widehat{\theta}_n - c_-\right) - F_{\pi_n}\left(\widehat{\theta}_n - c_+\right)\\
&= F_{\pi_n}\left(F_{\pi_n}^{-1}\left(1 - \frac{\alpha}{2}\right)\right) - F_{\pi_n}\left(F_{\pi_n}^{-1}\left(\frac{\alpha}{2}\right)\right) = 1 - \alpha.
\end{aligned}$$

□

Note however than the pivot confidence interval depends on the true distribution F, but we can use a boostrap estimate of the form

$$\widehat{F}(x) := \frac{1}{m}\sum_{k=1}^{m} \mathbb{1}_{\{\pi_{n,k} \leq x\}},$$

with the bootstrap pivot $\pi_{n,k} := \theta_{n,k} - \widehat{\theta}_n$.

Chapter 4

Reducing Data Dimension

4.1 Principal Component Analysis

The starting point of Principal Component Analysis (PCA)–introduced in [31] in 1901!!!–is a random vector $\mathbf{X}$ of dimension n, representing, for example, the returns of n stocks. The only assumption we make here is that the first two moments of $\mathbf{X}$ exist, and we denote by Σ its variance-covariance matrix. The main goal of PCA is to reduce the dimension n of the problem in order to make it more tractable. The simplest solution would be to consider one single element of $\mathbf{X}$ as a simplified representation of the whole vector. However, the loss of information is potentially extreme, and it is furthermore not clear which element one should pick. Another approach could be to consider an average of all its values, but then all have the same weight, an assumption that is not realistic in practice.

4.1.1 Definitions and main properties

We consider data points in the form $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n) \in \mathcal{M}_{p,n}$, where each vector $\mathbf{X}_k$ is p-dimensional. For example, taking $p = 500$, each such vector may correspond to daily returns of the components of the S&P500, and the observations n are trading days. The goal of PCA is to describe / visualise this data in just a few dimensions. Intuitively, this could be done in two ways:

- Find a d-dimensional affine subspace on which the projected points are the best approximations of the original data;
- Find the projection preserving as much as possible the variance of the original data.

Recall that the sample mean $\boldsymbol{\mu}$ and sample covariance $\boldsymbol{\Sigma}$ are defined as

$$\boldsymbol{\mu} := \frac{1}{n}\sum_{k=1}^{n} \mathbf{X}_k = \frac{1}{n}\mathbf{X}\mathbb{1}_n \in \mathbb{R}^p \qquad \text{and} \qquad \boldsymbol{\Sigma} := \frac{1}{n-1}\sum_{k=1}^{n}(\mathbf{X}_k - \boldsymbol{\mu})(\mathbf{X}_k - \boldsymbol{\mu})^\top \in \mathcal{M}_{p,p}.$$

We consider the first approach first. For $1 \leq d \leq p$, let $\mathbf{E} = (\mathbf{E}_1, \ldots, \mathbf{E}_d) \in \mathcal{M}_{p,d}$ an orthonormal basis of the d-dimensional subspace we are interested in, and consider the affine fit

$$\mathbb{R}^p \ni \widehat{\mathbf{X}}_k^{\boldsymbol{\nu},\mathbf{B},\mathbf{E}} := \boldsymbol{\nu} + \sum_{i=1}^d \mathbf{B}_{k,i}\mathbf{E}_i = \boldsymbol{\nu} + \mathbf{E}\mathbf{B}_k, \qquad \text{for each } k = 1, \ldots, n,$$

where $\mathbf{B} \in \mathcal{M}_{n,d}$ is a vector of coefficients to estimate, and $\mathbf{B}_k \in \mathbb{R}^d$ its kth row. Note that, by construction, $\mathbf{E}^\top\mathbf{E} = \mathbf{I}_{d,d}$. A natural way to determine the optimal coefficients is to consider the least-square problem

$$\min_{\boldsymbol{\nu},\mathbf{B},\mathbf{E}:\|\mathbf{E}\|=1} \mathcal{D}(\boldsymbol{\nu}, \mathbf{B}, \mathbf{E}), \tag{4.1.1}$$

where $\mathcal{D}(\boldsymbol{\nu}, \mathbf{B}, \mathbf{E}) := \sum_{k=1}^n \left\|\mathbf{X}_k - \widehat{\mathbf{X}}_k^{\boldsymbol{\nu},\mathbf{B},\mathbf{E}}\right\|_2^2$. The first-order conditions in μ read

$$\nabla_{\boldsymbol{\nu}}\mathcal{D}(\boldsymbol{\nu}, \mathbf{B}, \mathbf{E}) = 0 \qquad \text{if and only if} \qquad \left(\mathbf{X} - \widehat{\mathbf{X}}^{\boldsymbol{\nu},\mathbf{B},\mathbf{E}}\right)\mathbb{1}_n = 0.$$

Since $\mathbb{1}_n^\top\mathbf{B} = \sum_{k=1}^n \mathbf{B}_k = \mathbf{O}$, we therefore deduce that the optimal $\boldsymbol{\nu}^*$ is given by $\boldsymbol{\nu}^* = \boldsymbol{\mu}$, the sample mean. The minimisation problem (4.1.1) therefore reduces to

$$\min_{\mathbf{E},\mathbf{B}:\|\mathbf{E}\|=1} \mathcal{D}(\boldsymbol{\mu}, \mathbf{B}, \mathbf{E}).$$

Focusing now on the matrix $\mathbf{B}$, since

$$\mathcal{D}(\boldsymbol{\mu}, \mathbf{B}, \mathbf{E}) = \sum_{k=1}^n \left\|\mathbf{X}_k - \widehat{\mathbf{X}}_k^{\boldsymbol{\mu},\mathbf{B},\mathbf{E}}\right\|_2^2 = \sum_{k=1}^n \|\mathbf{X}_k - (\boldsymbol{\mu} + \mathbf{E}\mathbf{B}_k)\|_2^2.$$

Since the minimisation problem decouples for each k, we can write

$$\nabla_{\mathbf{B}_k}\mathcal{D}(\boldsymbol{\mu}, \mathbf{B}, \mathbf{E}) = 0 \qquad \text{if and only if} \qquad \mathbf{X}_k - (\boldsymbol{\mu} + \mathbf{E}\mathbf{B}_k) = \mathbf{O},$$

or $\mathbf{B}_k^* = \mathbf{E}^\top(\mathbf{X}_k - \boldsymbol{\mu})$, since $\mathbf{E}^\top\mathbf{E} = \mathbf{I}$. The optimisation problem therefore reduces to the minimisation of $\mathcal{D}(\boldsymbol{\mu}, \mathbf{B}^*, \mathbf{E})$ subject to $\|\mathbf{E}\| = 1$. Note that now

$$\mathcal{D}(\boldsymbol{\mu}, \mathbf{B}^*, \mathbf{E}) = \sum_{k=1}^n \left\|(\mathbf{X}_k - \boldsymbol{\mu}) - \mathbf{E}\mathbf{E}^\top(\mathbf{X}_k - \boldsymbol{\mu})\right\|_2^2.$$

It is easy to see that, for each $k = 1, \ldots, n$, denoting $\widetilde{\mathbf{X}}_k := \mathbf{X}_k - \boldsymbol{\mu}$ for clarity:

$$\begin{aligned}\left\|\widetilde{\mathbf{X}}_k - \mathbf{E}\mathbf{E}^\top\widetilde{\mathbf{X}}_k\right\|_2^2 &= \widetilde{\mathbf{X}}_k^\top\widetilde{\mathbf{X}}_k - 2\widetilde{\mathbf{X}}_k^\top\mathbf{E}\mathbf{E}^\top\widetilde{\mathbf{X}}_k + \widetilde{\mathbf{X}}_k^\top\mathbf{E}(\mathbf{E}^\top\mathbf{E})\mathbf{E}^\top\widetilde{\mathbf{X}}_k \\ &= \widetilde{\mathbf{X}}_k^\top\widetilde{\mathbf{X}}_k - \widetilde{\mathbf{X}}_k^\top\mathbf{E}\mathbf{E}^\top\widetilde{\mathbf{X}}_k,\end{aligned}$$

so that minimising $\mathcal{D}(\boldsymbol{\mu}, \mathbf{B}^*, \mathbf{E})$ over $\mathbf{E}$ is equivalent to maximising $\widetilde{\mathbf{X}}_k^\top\mathbf{E}\mathbf{E}^\top\widetilde{\mathbf{X}}_k$ over $\mathbf{E}$. Now,

$$\begin{aligned}\sum_{k=1}^n \widetilde{\mathbf{X}}_k^\top\mathbf{E}\mathbf{E}^\top\widetilde{\mathbf{X}}_k &= \sum_{k=1}^n \mathrm{Tr}\left(\widetilde{\mathbf{X}}_k^\top\mathbf{E}\mathbf{E}^\top\widetilde{\mathbf{X}}_k\right) = \sum_{k=1}^n \mathrm{Tr}\left(\mathbf{E}^\top\widetilde{\mathbf{X}}_k\widetilde{\mathbf{X}}_k^\top\mathbf{E}\right) \\ &= \mathrm{Tr}\left(\mathbf{E}^\top\left\{\sum_{k=1}^n \widetilde{\mathbf{X}}_k\widetilde{\mathbf{X}}_k^\top\right\}\mathbf{E}\right) = (n-1)\mathrm{Tr}\left(\mathbf{E}^\top\boldsymbol{\Sigma}\mathbf{E}\right).\end{aligned} \tag{4.1.2}$$

We therefore deduce that the initial problem, namely projecting the vector of data onto a smaller subspace is equivalent to maximising the variance of the projection.

Now, consider the second approach, namely we want to find the projected points $\left(E_1^\top \mathbf{X}_k, \ldots, \mathrm{e}_d^\top \mathbf{X}_k\right)$ that have as much variance as possible. This is illustrated by the following figure, where a cloud of random points is generated. If we have to choose between the two axes (diagonal and anti-diagonal) for projection, it is clear that the anti-diagonal one is going to spread out the data as much as possible, and the two projections will clearly be separated, whereas all the points would be mixed up when projected onto the diagonal axis. More formally, we have to solve the following maximisation problem:

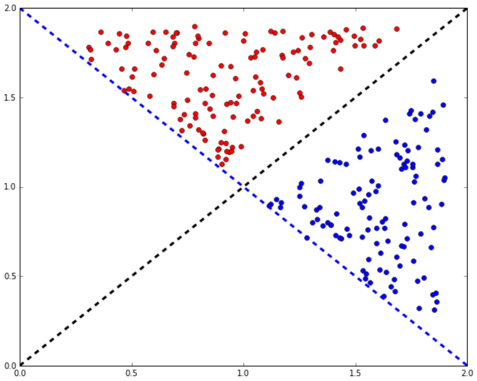

Figure 4.1: Cloud of random points

$$\max_{E:E^\top E=\mathbf{I}} \sum_{i=1}^{n} \left\| E^\top \mathbf{X}_k - \frac{1}{n}\sum_{j=1}^{n} E^\top \mathbf{X}_j \right\|^2 .$$

However,

$$\sum_{i=1}^{n} \left\| E^\top \mathbf{X}_k - \frac{1}{n}\sum_{j=1}^{n} E^\top \mathbf{X}_j \right\|^2 = \sum_{i=1}^{n} \left\| E^\top \left(\mathbf{X}_k - \frac{1}{n}\sum_{j=1}^{n} \mathbf{X}_j \right) \right\|^2 = \mathrm{Tr}\left(E^\top \mathbf{\Sigma} E\right),$$

which corresponds exactly to (4.1.2), so that the two approaches mentioned at the beginning of this section coincide.

PCA: theoretical setup

We consider some (theoretical) data $\mathbf{X}$ as above, and denote $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ its mean vector and covariance matrix, and write $\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^\top$ its spectral decomposition, where $\mathbf{\Gamma}$ is an orthogonal matrix in $\mathcal{M}_{p,p}$ and $\mathbf{\Lambda} = \mathrm{Diag}(\lambda_1, \ldots, \lambda_p)$ a diagonal matrix in $\mathcal{M}_{p,p}$. By rotation, we can assume without loss of generality that the eigenvalues are ordered, in the sense that $\lambda_1 \geq \ldots \geq \lambda_p$, and we denote by $\gamma_1, \ldots, \gamma_p$ the corresponding eigenvectors in $\mathbb{R}^p$, which satisfy

$$\|\gamma_i\|^2 = 1 \qquad \text{and} \qquad \gamma_i^\top \gamma_j = 0, \qquad \text{for all } i,j = 1,\ldots,p, \text{ with } i \neq j.$$

Definition 4.1.1. For any $i = 1, \dots, p$, the random variable $\boldsymbol{\eta}_i := \gamma_i(\mathbf{X} - \boldsymbol{\mu})$ is called the ith principal component of the random vector $\mathbf{X} \in \mathbb{R}^p$.

The following properties are simple to prove:

Proposition 4.1.2. *The random vector $\boldsymbol{\eta}$ is centered and its variance-covariance matrix satisfies $\mathbb{V}[\boldsymbol{\eta}_i] = \lambda_i$ and $\mathrm{Cov}[\boldsymbol{\eta}_i, \boldsymbol{\eta}_j] = 0$ for any $i \neq j$.*

Example 4.1.3. Let $\mathbf{X} \in \mathbb{R}^2$ be a random vector with

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

for some $\rho \in [0, 1]$. It is easy to show that the eigenvalues are $\lambda_1 = 1 + \rho$ and $\lambda_2 = 1 - \rho$, with corresponding eigenvectors

$$\widetilde{\gamma}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad \text{and} \qquad \widetilde{\gamma}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

Since $\|\widetilde{\gamma}_1\|^2 = \|\widetilde{\gamma}_2\|^2$, we need to normalise them, so that

$$\gamma_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad \text{and} \qquad \gamma_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

and hence we can deduce the first two principal components

$$\boldsymbol{\eta}_1 = \frac{X_1 + X_2}{\sqrt{2}} \qquad \text{and} \qquad \boldsymbol{\eta}_2 = \frac{X_1 - X_2}{\sqrt{2}}.$$

We can further compute the variances of the principal components:

$$\begin{aligned} \mathbb{V}[\boldsymbol{\eta}_1] &= \mathbb{V}\left[\frac{X_1 + X_2}{\sqrt{2}}\right] = \frac{\mathbb{V}[X_1 + X_2]}{2} = \frac{\mathbb{V}[X_1] + \mathbb{V}[X_2] + 2\mathrm{Cov}[X_1, X_2]}{2} = 1 + \rho = \lambda_1, \\ \mathbb{V}[\boldsymbol{\eta}_2] &= \mathbb{V}\left[\frac{X_1 - X_2}{\sqrt{2}}\right] = \frac{\mathbb{V}[X_1 - X_2]}{2} = \frac{\mathbb{V}[X_1] + \mathbb{V}[X_2] - 2\mathrm{Cov}[X_1, X_2]}{2} = 1 - \rho = \lambda_2, \end{aligned}$$

Theorem 4.1.4. *Let $\mathbf{X}$ be a random vector in $\mathbb{R}^p$ with finite second moment, mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and variance-covariance matrix $\boldsymbol{\Sigma}$. Then*

$$\underset{\mathrm{a} \in \mathbb{R}^p : \|\mathrm{a}\| = 1}{\arg\max}\ \mathbb{V}(\mathrm{a}^\top \mathbf{X}) = \underset{\mathrm{a} \in \mathbb{R}^p : \|\mathrm{a}\| = 1}{\arg\max}\ \mathbb{V}(\mathrm{a}^\top (\mathbf{X} - \boldsymbol{\mu})) = \gamma_1.$$

Proof. From the spectral decomposition $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^\top$, we can write

$$\mathbb{V}(\mathrm{a}^\top \mathbf{X}) = \sum_{i=1}^{p} \lambda_i \left(\mathrm{a}^\top \gamma_i\right) \left(\gamma_i^\top \mathrm{a}\right) =: \sum_{i=1}^{p} \lambda_i c_i^2,$$

where $c_i := \mathrm{a}^\top \gamma_i$ is the projection of the vector a in the direction γ_i. Since the vectors γ_i form an orthonormal basis, then $\sum_{i=1}^p c_i^2 = \|\mathrm{a}\|^2$. Furthermore,

$$\mathbb{V}\left(\mathrm{a}^\top \mathbf{X}\right) = \sum_{i=1}^{p} \lambda_i c_i^2 \leq \lambda_1 \sum_{i=1}^{p} c_i^2 = \lambda_1 \|\mathrm{a}\|^2 = \lambda_1.$$

With $\mathrm{a} = \gamma_1$, all the coefficients are therefore equal to zero except $c_1 = 1$; the maximum is clearly attained at this point since

$$\mathbb{V}\left(\gamma_1^\top \mathbf{X}\right) = \gamma_1^\top \mathbb{V}\left(\mathbf{X}\right)\gamma_1 = \gamma_1^\top \boldsymbol{\Sigma}\gamma_1 = \gamma_1^\top\left(\boldsymbol{\Sigma}\gamma_1\right) = \gamma_1^\top \lambda_1\gamma_1 = \lambda_1\|\gamma_1\|^2 = \lambda_1,$$

and the theorem follows. □

In the previous theorem, we showed that the first principal direction is in fact given by the eigenvector corresponding to the largest eigenvalue. It therefore sounds sensible to try and iterate the procedure. In order to find the second component, the proof of the theorem holds with almost no changes to prove that

$$\gamma_2 = \operatorname*{arg\,max}_{\mathrm{a}\in\mathcal{A}_2} \mathbb{V}(\mathrm{a}^\top \mathbf{X}) = \operatorname*{arg\,max}_{\mathrm{a}\in\mathbb{R}^p:\|\mathrm{a}\|=1} \mathbb{V}(\mathrm{a}^\top(\mathbf{X}-\boldsymbol{\mu})),$$

where

$$\mathcal{A}_2 := \left\{\mathrm{a}\in\mathbb{R}^p : \|\mathrm{a}\| = 1, \mathrm{a}\perp\gamma_1\right\}.$$

Indeed, repeating the arguments in the proof of the theorem, we can write, for any $\mathrm{a}\in\mathcal{A}_2$,

$$\mathbb{V}\left(\mathrm{a}^\top\mathbf{X}\right) = \sum_{i=2}^{p}\lambda_i c_i^2 \leq \lambda_2\sum_{i=1}^{p} c_i^2 = \lambda_2\|\mathrm{a}\|^2 = \lambda_2.$$

And we can iterate this procedure to any level up to p.

PCA: empirical analysis

We now consider a similar framework, but the the covariance matrix $\boldsymbol{\Sigma}$ replaced by the empirical covariance matrix $\boldsymbol{S}$. We can repeat the theoretical analysis, but now with empirical estimates.

Remark 4.1.5. Depending on the data, if several factors are expressed in different units, it might be more appropriate to perform PCA with the correlation matrix rather than the covariance matrix, as the former is unit-free.

Proposition 4.1.6 (Perron-Frobenius Theorem)**.** *Let $\mathbf{A}\in\mathcal{M}_{p,p}$ a symmetric matrix with all elements strictly positive. Then the coordinates of the first eigenvector all have the same sign.*

Remark 4.1.7. The procedure to perform PCA is clear and the steps are as follows:

(i) Compute the covariance matrix $\boldsymbol{\Sigma}$;

(ii) Apply the spectral decomposition theorem to $\boldsymbol{\Sigma}$;

From a computational point of view, however, this may be too expensive. From the definition of the covariance matrix, the cost of computing it is of order $\mathcal{O}(np^2)$, and it can be shown that the spectral decomposition has a cost of order $\mathcal{O}(p^3)$. Assuming that n is of the same order as p

(modulo some multiplicative constant), which is often the case in real data, the overall cost here is $\mathcal{O}(p^3)$. As an alternative, one could consider the following approach. Since

$$\mathbf{\Sigma} = \frac{1}{n-1}\left(\mathbf{X} - \boldsymbol{\mu}\mathbf{I}^\top\right)\left(\mathbf{X} - \boldsymbol{\mu}\mathbf{I}^\top\right)^\top, \tag{4.1.3}$$

we can first perform a Singular Value decomposition for $\mathbf{X} - \boldsymbol{\mu}\mathbf{I}^\top$ of the form

$$\mathbf{X} - \boldsymbol{\mu}\mathbf{I}^\top = \mathbf{UDV}^\top,$$

where $\mathbf{U} \in \mathcal{M}_p$ and $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$. Therefore, we can rewrite (4.1.3) as

$$\mathbf{\Sigma} = \frac{1}{n-1}\left(\mathbf{UDV}^\top\right)\left(\mathbf{UDV}^\top\right)^\top = \frac{1}{n-1}\mathbf{UDV}^\top\mathbf{VD}^\top\mathbf{U}^\top = \frac{1}{n-1}\mathbf{UD}^2\mathbf{U}^\top.$$

The SVD has a cost of order $\mathcal{O}(\inf\{n^2p, p^2n\})$, but in fact, if we wish to consider only the first d factors, the cost is reduced to $\mathcal{O}(dnp)$. With p and n of the same order, we deduce that this alternative approach is much more efficient than the brute-force spectral decomposition approach.

4.1.2 Examples

IPython notebook PCA.ipynb

4.2 Random matrix analysis in Finance

4.2.1 Definitions and properties

Consider a matrix $\mathbf{A} \in \mathcal{M}_{n,n}$ where all entries are independently distributed as centered Gaussian random variables with given unit variance, and define the symmetric matrix

$$\mathbf{X} := \frac{\mathbf{A} + \mathbf{A}^\top}{\sqrt{2n}}. \tag{4.2.1}$$

It is straightforward to see that $\mathbb{E}(\mathbf{X}) = \mathbf{O}$ and $\mathbb{V}(\mathbf{X}_{i,j}) = \frac{1}{n}\mathbb{1}_{\{i \neq j\}} + \frac{2}{n}\mathbb{1}_{\{i=j\}}$. Being real and symmetric, its spectrum consists of n distinct real eigenvalues $\sigma(\mathbf{X}) = (\lambda_1, \ldots, \lambda_n)$. The empirical distribution of the spectrum looks as follows:

We observe that, as the dimension n increases, the shape of the distribution of the spectrum approaches that of a semicircle. This is not a coincidence, and was proved in by Eugene Wigner [33, 34] (1902-1995), a Hungarian-American theoretical physicist who got awarded the Nobel Prize in Physics in 1963 '*for his contributions to the theory of the atomic nucleus and the elementary particles, particularly through the discovery and application of fundamental symmetry principles*': Before stating his main result, let us introduce a few notions, generalising and formalising the computations above.

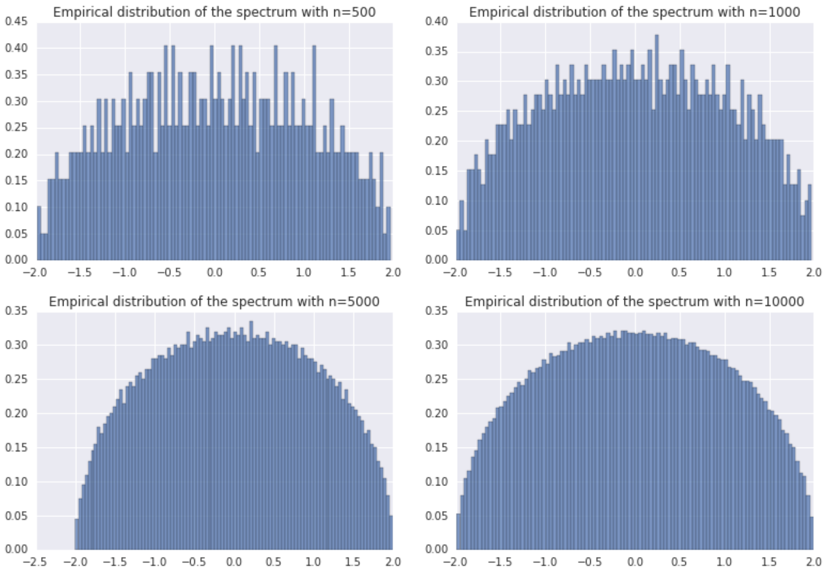

Figure 4.2: Empirical distribution of the eigenvalues of $\mathbf{X}$ in (4.2.1). See the IPython notebook RandomMatrix.ipynb.

Definition 4.2.1. A matrix $\mathbf{X} \in \mathcal{M}_{n,n}$ is called a Wigner matrix if it is real, symmetric, and can be written as

$$\mathbf{X}_{i,j} = \mathbf{X}_{j,i} := \begin{cases} \frac{1}{\sqrt{n}}\mathbf{Z}_{i,j} & \text{if } i < j, \\ \frac{1}{\sqrt{n}}\mathbf{Y}_i & \text{if } i = j, \end{cases} \qquad \text{for } 1 \leq i, j \leq n,$$

where $(\mathbf{Z}_{i,j})_{1\leq i<j\leq n}$ and $(\mathbf{Y}_i)_{1\leq i\leq n}$ are independent and identically distributed centered, real-valued random variables with finite moments. If those are Gaussian, then $\mathbf{X}$ is called a Gaussian Wigner matrix.

For a given Wigner matrix $\mathbf{X}$, we denote its spectrum by $\sigma(\mathbf{X}) = \{\lambda_1, \ldots, \lambda_n\}$, and by L_n the empirical distribution of its eigenvalues

$$L_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i},$$

where δ denotes the Dirac mass. Note that L_n is a random probability measure on the real line. Introduce finally the semicircle distribution

$$\mu(\mathrm{d}x) = \frac{1}{2\pi}\sqrt{4-x^2}\mathbb{1}_{\{|x|\leq 2\}}\mathrm{d}x.$$

The following lemma is key in the proof of Wigner's semicircle law (Theorem 4.2.3). We shall not prove the theorem here, but prove the expressions for the moments of μ out of sheer mathematical interest.

Lemma 4.2.2. *For any integer k, we have*

$$\int_{\mathbb{R}} x^k \mu(\mathrm{d}x) = \begin{cases} 0, & \text{if } k \text{ is odd,} \\ \mathfrak{c}_p = \dfrac{(2p)!}{(p+1)!p!}, & \text{if } k \text{ is even with } k = 2p, \end{cases}$$

where the $\mathfrak{c}_p$ are called the Catalan numbers.

Proof. If k is odd, the result is trivial by symmetry. Otherwise, we can write, with $k = 2p$ and the change of variable $x = 2\sin(\theta)$,

$$\begin{aligned} \mathfrak{m}_{2p} := \int_{\mathbb{R}} x^{2p} \mu(\mathrm{d}x) &= \frac{1}{2\pi} \int_{-2}^{2} x^{2p} \sqrt{4 - x^2} \mathrm{d}x \\ &= \frac{2^{2p}}{\pi} \int_{-\pi/2}^{\pi/2} \sin(\theta)^{2p} \sqrt{4 - 4\sin(\theta)^2} \cos(\theta) \mathrm{d}\theta \\ &= \frac{2^{2p+1}}{\pi} \int_{-\pi/2}^{\pi/2} \sin(\theta)^{2p} \cos(\theta)^2 \mathrm{d}\theta \end{aligned}$$

Now, using by integration by parts, we can easily show that

$$\begin{aligned} \mathcal{I}_{2p+2} := \int_{-\pi/2}^{\pi/2} \sin(\theta)^{2p+2} \mathrm{d}\theta &= \int_{-\pi/2}^{\pi/2} \sin(\theta)^{2p+1} \sin(\theta) \mathrm{d}\theta \\ &= -\Big[\sin(\theta)^{2p+1} \cos(\theta)\Big]_{-\pi/2}^{\pi/2} + (2p+1) \int_{-\pi/2}^{\pi/2} \sin(\theta)^{2p} \cos(\theta)^2 \mathrm{d}\theta \end{aligned}$$

Since the bracket is null, we deduce the identity

$$\mathfrak{m}_{2p} := \int_{\mathbb{R}} x^{2p} \mu(\mathrm{d}x) = \frac{2^{2p+1}}{\pi} \int_{-\pi/2}^{\pi/2} \sin(\theta)^{2p} \cos(\theta)^2 \mathrm{d}\theta = \frac{2^{2p+1}}{(2p+1)\pi} \mathcal{I}_{2p+2},$$

and the result follows by recursion. □

Theorem 4.2.3 (Wigner semicircle law)**.** *The empirical measure L_n of a Wigner matrix converges weakly, in probability, to the standard semicircle distribution μ as n tends to infinity.*

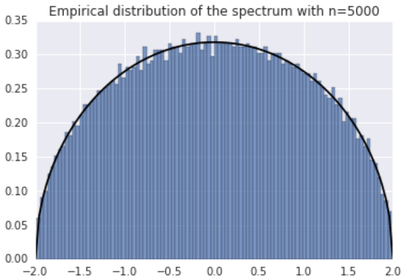

Figure 4.3: Empirical distribution of the eigenvalues of $\mathbf{X}$ with $n = 5000$ and the limit density of the Wigner semicircle law.

Consider now the observation of the matrix $\mathbf{X} \in \mathcal{M}_{n,p}$ where each $\mathbf{X}_{i,j}$ represents the returns of some stock j at time i. We assume that the matrix $\mathbf{X}$ is already centered and normalised, i.e.

$\mathbb{E}(\mathbf{X}_{i,j}) = 0$ and $\mathbb{V}(\mathbf{X}_{i,j}) = 1$. The empirical covariance matrix $\mathbf{S}_n$ is the $p \times p$ symmetric matrix defined as

$$\mathbf{S}_n := \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top, \tag{4.2.2}$$

where $\mathbf{X}_i \in \mathbb{R}^p$ denotes the ith observation (or equivalently row i).

Definition 4.2.4. A matrix $\mathbf{W} \in \mathcal{M}_{p,p}$ is said to have a Wishart distribution with scale matrix $\mathbf{\Sigma}$ and n degrees of freedom, and we write $\mathbf{W} \sim W_p(n, \mathbf{\Sigma})$, if $\mathbf{W} = \mathbf{X}^\top \mathbf{X}$ with $\mathbf{X} \sim \mathcal{N}_{n,p}(0, \mathbf{\Sigma})$.

In the case where the scale matrix $\mathbf{\Sigma}$ is the identity matrix, $\mathbf{W}$ is called a white Wishart matrix. A general $W_p(n, \mathbf{\Sigma})$ Wishart distribution admits a density (available in closed form) only when $n \geq p$. We leave the following lemma regarding properties of Wishart distributions as an exercise:

Lemma 4.2.5. *Let $\mathbf{W}_1 \sim W_p(n_1, \mathbf{\Sigma})$ and $\mathbf{W}_2 \sim W_p(n_2, \mathbf{\Sigma})$ be two independent Wishart matrices and $\mathbf{A} \in \mathcal{M}_{p,q}$. Then $\mathbf{A}^\top \mathbf{W}_1 \mathbf{A} \sim W_q(n_1, \mathbf{A}^\top \mathbf{\Sigma} \mathbf{A})$ and $\mathbf{W}_1 + \mathbf{W}_2 \in W_p(n_1 + n_2, \mathbf{\Sigma})$.*

The following theorem explains why Wishart matrices are key for covariance estimation:

Theorem 4.2.6. *The empirical covariance matrix in (4.2.2) satisfies $\mathbf{S}_n \sim W_p(n-1, n^{-1}\mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is the true covariance matrix. Furthermore, the unbiased empirical covariance matrix $\frac{n}{n-1}\mathbf{S}_n$ is distributed as $W_p(n-1, \frac{1}{n-1}\mathbf{\Sigma})$.*

Remark 4.2.7. We shall not prove the theorem here, but it is easy to see that the empirical covariance matrix $\mathbf{S}_n$ in (4.2.2) can be written as

$$\mathbf{S}_n := \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^\top = \frac{1}{n}\mathbf{X}^\top \mathbf{H} \mathbf{X},$$

where the so-called centering matrix $\mathbf{H}$ defined as $\mathbf{H} := \mathbf{I} - n^{-1}\mathbb{1}\mathbb{1}^\top$. It is easy to see that is a projector, i.e. $\mathbf{H}^2 = \mathbf{H}$ and $\mathbf{H}^\top = \mathbf{H}$.

The following result is classical and provides the asymptotic distribution of eigenvalues of sample covariance matrices, when the sample size grows large, but for a fixed number of factors:

Theorem 4.2.8. *Let $\mathbf{S}_n \in \mathcal{M}_{p,p}$ be a sample covariance matrix drawn from $\mathcal{N}_{n,p}(\mu, \mathbf{\Sigma})$ distributions, and denote by $l_1, \ldots, l_p$ its eigenvalues. Then, for any $i = 1, \ldots, p$, as n tends to infinity,*

$$\sqrt{n}\,(l_i - \lambda_i) \text{ converges in distribution to } \mathcal{N}(0, 2\lambda_i^2),$$

where $\lambda_1, \ldots, \lambda_p$ denote the eigenvalue of the true covariance matrix.

As mentioned before, however, p is in general of the same order as n, and the theorem does not hold any longer. The following theorem is a key result in this framework, and was proved in [30].

Theorem 4.2.9 (Marčenko-Pastur Theorem)**.** *In the previous framework, let $n = pq$ for some fixed integer q, and assume that $\mathbf{\Sigma} = \mathrm{Diag}(\sigma^2, \ldots, \sigma^2)$. As p tends to infinity, the density of the spectrum of $\mathbf{S}_n$ converges to the Marčenko-Pastur density*

$$\rho(\mathrm{d}\lambda) = \frac{q}{2\lambda\pi\sigma^2}\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}\mathrm{d}\lambda + (1-q)\mathbb{1}_{\{q<1\}}(\mathrm{d}\lambda),$$

where $\lambda_\pm := \sigma\left(1 \pm q^{-1/2}\right)^2$.

In particular, the above theorem implies that the top and bottom eigenvalues $\lambda_{\min\{p,qp\}}$ and λ_1 converge almost surely to the edge of the support of the Marčenko-Pastur density, and in particular, if $q < 1$, then $\lambda_{n+1}, \ldots, \lambda_p$ are null. It also means that we have

$$\mathbb{P}\left(\lim_{m\uparrow\infty} \lambda_1 = \left(1 + q^{-1/2}\right)^2\right) = 1.$$

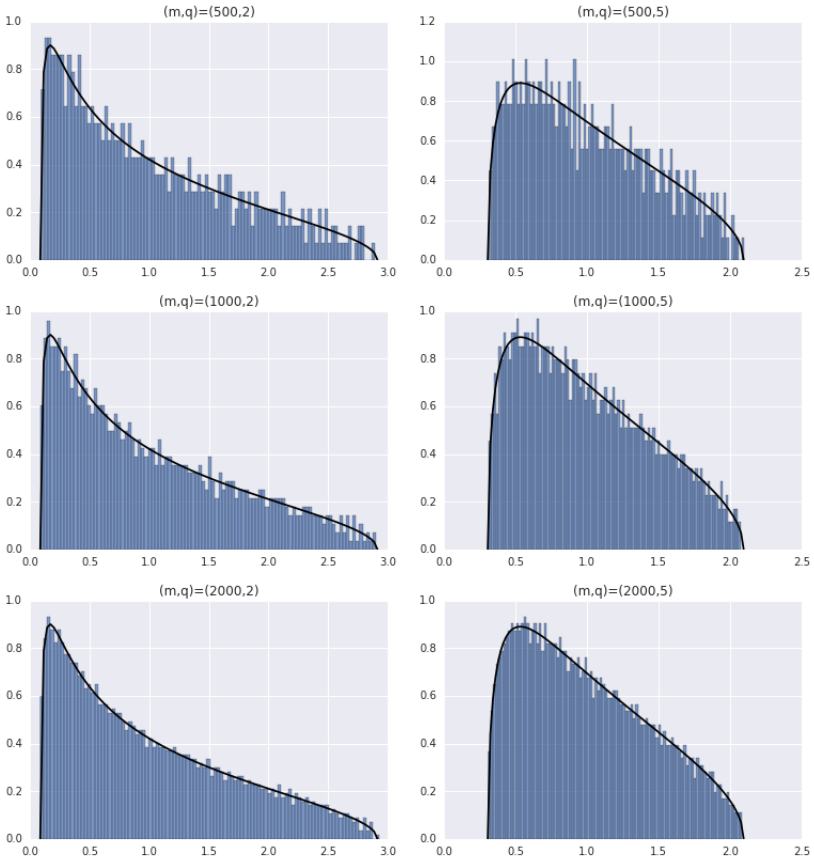

Figure 4.4: Empirical distribution of the eigenvalues of $\mathbf{E}^n$ (from Theorem 4.2.9) for different values of m and q, where the initial random matrix $\mathbf{X}$ is iid $\mathcal{N}(0,1)$.

Consider now the largest eigenvalue of the spectrum. The following result is due to Johnstone [26]:

Proposition 4.2.10. *Let $\mathbf{W} \sim W_p(n, \mathbf{I})$, and denote by λ_1 its largest eigenvalue. Then $(\lambda_1 - \mu)/\sigma$ converges in distribution to F_1, where F_1 is called the Tracy-Widom distribution of order one, and*

$$\mu := \left(\sqrt{n-1} + \sqrt{m}\right)^2 \qquad \textit{and} \qquad \sigma := \mu \left((n-1)^{-1/2} + m^{-1/2}\right)^{1/3}.$$

4.2.2 Application: cleaning correlation matrices

IPython notebook Random Matrix.ipynb

Chapter 5

Regression Methods

5.1 Regression methods

5.1.1 Linear regression: the simple case

We start this chapter with the simple case of data $(x_i, y_i)_{i=1,\ldots,n}$, from which some dependency can be observed. We would like to determine a relation of the form $y_i \approx f(x_i)$ for some function $f : \mathbb{R} \to \mathbb{R}$. A general formulation can be stated as

$$\min_{f \in \mathbf{F}} \sum_{i=1}^{n} \Phi \left[y_i - f(x_i) \right],$$

for some given cost function Φ, where $\mathbf{F}$ is a class of functions of interest.

Simple linear regression

This is the simplest case, where we assume a dependence of the form

$$y_i = f(x_i) + \varepsilon_i, \qquad \text{for } i = 1, \ldots, n$$

where $f(x) \equiv \alpha + \beta x$ is a linear function, and the sequence $(\varepsilon_i)_{i=1,\ldots,n}$ are centered independent random noises with constant variance σ^2. We can rewrite this relation as $\mathbf{Y} = \alpha \mathbb{1} + \beta \mathbf{X} + \boldsymbol{\varepsilon}$, with $\mathbf{X} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$, $\mathbf{Y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$.

Least-squares and properties

Since α and β are the only two parameters here, the infinite-dimensional minimisation problem (over $\mathbf{F}$) reduces to one over $\mathbb{R}^2$. In the least-square minimisation problem, we consider the following loss / cost / error function:

$$\mathfrak{L}(\alpha, \beta) := \|\mathbf{Y} - (\alpha \mathbb{1} + \beta \mathbf{X})\|_2^2 = \sum_{i=1}^{n} \left(y_i - \alpha - \beta x_i \right)^2.$$

Definition 5.1.1. The least-square estimator $(\widehat{\alpha}, \widehat{\beta})$ is the solution to the minimisation problem

$$(\widehat{\alpha}, \widehat{\beta}) := \operatorname*{arg\,min}_{(\alpha,\beta)} \mathfrak{L}(\alpha, \beta). \tag{5.1.1}$$

Proposition 5.1.2. *The solution to* (5.1.1) *reads*

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x} \qquad \textit{and} \qquad \widehat{\beta} = \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} \sum_{i=1}^{n} (x_i - \overline{x})\, y_i.$$

Note that the computation of these estimators is purely deterministic and do not rely on the iid assumption made about the errors.

Proof. Clearly, the function $\mathfrak{L}$ is smooth, convex and hence admits a unique minimum, and

$$\nabla\mathfrak{L}(\alpha, \beta) = \begin{pmatrix} \partial_\alpha \mathfrak{L}(\alpha, \beta) \\ \partial_\beta \mathfrak{L}(\alpha, \beta) \end{pmatrix} = -2 \begin{pmatrix} \displaystyle\sum_{i=1}^{n} (y_i - \alpha - \beta x_i) \\ \displaystyle\sum_{i=1}^{n} x_i\, (y_i - \alpha - \beta x_i) \end{pmatrix}.$$

The first element can be written as

$$\sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = n\left(\overline{y} - \alpha - \beta\overline{x}\right),$$

which is equal to zero if and only if $\alpha = \overline{y} - \beta\overline{x}$. Regarding the gradient with respect to β, we can write, plugging this optimal α,

$$\sum_{i=1}^{n} x_i\,(y_i - \alpha - \beta x_i) = \sum_{i=1}^{n} x_i \left[y_i - (\overline{y} - \beta\overline{x}) - \beta x_i\right] = \sum_{i=1}^{n} x_i y_i - \overline{y}\sum_{i=1}^{n} x_i + \beta\overline{x}\sum_{i=1}^{n} x_i - \beta\sum_{i=1}^{n} x_i^2,$$

which is equal to zero if and only if

$$\beta = \frac{\sum_{i=1}^{n} x_i\,(y_i - \overline{y})}{\sum_{i=1}^{n} x_i\,(x_i - \overline{x})},$$

and the proposition follows. □

Remark 5.1.3. Note that we can alternatively write the numerator and denominator in the expression for $\widehat{\beta}$ as follows:

$$\sum_{i=1}^{n} x_i\,(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \overline{y}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i - n\overline{y}\,\overline{x} = \sum_{i=1}^{n} x_i y_i - \overline{x}\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} y_i\,(x_i - \overline{x}),$$

and

$$\begin{aligned} \|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2 &= \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} \left(x_i^2 + \overline{x}^2 - 2x_i\overline{x}\right) = \sum_{i=1}^{n} x_i^2 + n\overline{x}^2 - 2\overline{x}\sum_{i=1}^{n} x_i \\ &= \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 = \sum_{i=1}^{n} x_i\,(x_i - \overline{x}). \end{aligned} \tag{5.1.2}$$

Let us prove some properties of these estimators:

Theorem 5.1.4. *The LSE are unbiased and*

$$\mathbb{V}\left[\widehat{\beta}\right] = \frac{\sigma^2}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2}, \qquad \mathbb{V}\left[\widehat{\alpha}\right] = \frac{\sigma^2 \|\mathbf{X}\|_2^2}{n\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2}, \qquad \mathrm{Cov}\left(\widehat{\alpha}, \widehat{\beta}\right) = -\frac{\sigma^2 \overline{x}}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2}.$$

Proof. Let us first show the following alternative representation for $\widehat{\beta}$:

$$\widehat{\beta} = \beta + \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} \sum_{i=1}^n (x_i - \overline{x})\,\varepsilon_i. \tag{5.1.3}$$

Combining the expression for $\widehat{\beta}$ in Proposition 5.1.2 and the definition of the linear model $\mathbf{Y} = \alpha\mathbb{1} + \beta\mathbf{X} + \boldsymbol{\varepsilon}$, we can write

$$\begin{aligned}\widehat{\beta} = \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} \sum_{i=1}^n (x_i - \overline{x})\, y_i &= \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} \sum_{i=1}^n (x_i - \overline{x}) \left[\alpha + \beta x_i + \varepsilon_i\right] \\ &= \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} \sum_{i=1}^n (x_i - \overline{x}) \left[\alpha + \beta x_i\right] + \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} \sum_{i=1}^n (x_i - \overline{x})\,\varepsilon_i.\end{aligned}$$

Regarding the first term, we can write

$$\frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} \sum_{i=1}^n (x_i - \overline{x}) \left[\alpha + \beta x_i\right] = \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} \left[\alpha \sum_{i=1}^n (x_i - \overline{x}) + \beta \sum_{i=1}^n x_i (x_i - \overline{x})\right] = \beta,$$

since $\sum_{i=1}^n (x_i - \overline{x}) = 0$ and using the identity (5.1.2), so that (5.1.3) holds. Since the sequence $(\varepsilon_1, \ldots, \varepsilon_n)$ is iid, then clearly $\mathbb{E}[\widehat{\beta}] = \beta$, and furthermore

$$\mathbb{E}[\widehat{\alpha}] = \mathbb{E}\left[\overline{y} - \widehat{\beta}\overline{x}\right] = \overline{y} - \beta\overline{x} = \alpha.$$

It is also easy from (5.1.3) to show that

$$\begin{aligned}\mathbb{V}\left[\widehat{\beta}\right] = \mathbb{V}\left[\beta + \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} \sum_{i=1}^n (x_i - \overline{x})\,\varepsilon_i\right] &= \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^4} \sum_{i=1}^n (x_i - \overline{x})^2\, \mathbb{V}[\varepsilon_i] \\ &= \frac{\sigma^2}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^4} \sum_{i=1}^n (x_i - \overline{x})^2 \\ &= \frac{\sigma^2}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2},\end{aligned}$$

using (5.1.2). Now, since $\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x}$ by Proposition 5.1.2, we can write

$$\begin{aligned}\mathbb{V}\left[\widehat{\alpha}\right] = \mathbb{V}\left[\overline{y} - \widehat{\beta}\overline{x}\right] &= \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^n y_i - \widehat{\beta}\overline{x}\right] \\ &= \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^n y_i\right] + \mathbb{V}\left[\widehat{\beta}\overline{x}\right] - \frac{2\overline{x}}{n}\sum_{i=1}^n \mathrm{Cov}\left(y_i, \widehat{\beta}\right) \\ &= \frac{\sigma^2}{n} + \overline{x}^2\mathbb{V}\left[\widehat{\beta}\right] - \frac{2\overline{x}}{n}\sum_{i=1}^n \mathrm{Cov}\left(y_i, \widehat{\beta}\right).\end{aligned} \tag{5.1.4}$$

Now, for any $i = 1, \ldots, n$, we have, using (5.1.3),

$$\begin{aligned}\mathrm{Cov}\left(y_i, \widehat{\beta}\right) &= \mathrm{Cov}\left(y_i, \beta + \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2}\sum_{k=1}^{n}\left(x_k - \overline{x}\right)\varepsilon_k\right) = \mathrm{Cov}\left(y_i, \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2}\sum_{k=1}^{n}\left(x_k - \overline{x}\right)\varepsilon_k\right)\\ &= \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2}\sum_{k=1}^{n}\left(x_k - \overline{x}\right)\mathrm{Cov}\left(y_i, \varepsilon_k\right) = \frac{\sigma^2}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2}\left(x_i - \overline{x}\right),\end{aligned}$$

since $\mathrm{Cov}\left(y_i, \varepsilon_k\right) = 0$ for all $k \neq i$ and is equal to σ^2 when $k = i$. Therefore $\mathrm{Cov}\left(\overline{y}, \widehat{\beta}\right) = 0$. Therefore, from (5.1.4), we obtain the desired variance of $\widehat{\alpha}$. Finally,

$$\mathrm{Cov}\left(\widehat{\alpha}, \widehat{\beta}\right) = \mathrm{Cov}\left(\overline{y} - \widehat{\beta}\overline{x}, \widehat{\beta}\right) = \mathrm{Cov}\left(\overline{y}, \widehat{\beta}\right) - \overline{x}\mathbb{V}\left[\widehat{\beta}\right] = -\overline{x}\mathbb{V}\left[\widehat{\beta}\right] = -\frac{\overline{x}\sigma^2}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2}.$$

□

Theorem 5.1.5 (Gauss Markov). *The LSE is optimal, in the sense that, among all possible unbiased estimators linear in* $\mathbf{Y}$*, it achieves minimal variance.*

Proof. The estimator for β, which we can write as $\widehat{\beta} = \sum_{i=1}^{n} w_i y_i = \mathbf{w}^\top \mathbf{Y}$, is clearly linear in $\mathbf{Y}$. Consider another unbiased estimator linear in $\mathbf{Y}$: $\widetilde{\beta} := \mathbf{p}^\top \mathbf{Y}$. We can then write

$$\mathbb{E}\left[\widetilde{\beta}\right] = \mathbf{p}^\top \mathbb{E}\left[\alpha\mathbb{1} + \beta\mathbf{X} + \boldsymbol{\varepsilon}\right] = \alpha\mathbf{p}^\top\mathbb{1} + \beta\mathbf{p}^\top\mathbf{X}$$

Since this new estimator is assumed unbiased and this relation holds for any β, we have $\alpha\mathbf{p}^\top\mathbb{1} = 0$ and $\mathbf{p}^\top\mathbf{X} = 1$. Now,

$$\mathbb{V}\left[\widetilde{\beta}\right] = \mathbb{V}\left[\widetilde{\beta} - \widehat{\beta} + \widehat{\beta}\right] = \mathbb{V}\left[\widetilde{\beta} - \widehat{\beta}\right] + \mathbb{V}\left[\widehat{\beta}\right] + 2\mathrm{Cov}\left(\widetilde{\beta} - \widehat{\beta}, \widehat{\beta}\right).$$

Now,

$$\mathrm{Cov}\left(\widetilde{\beta} - \widehat{\beta}, \widehat{\beta}\right) = \mathrm{Cov}\left(\widetilde{\beta}, \widehat{\beta}\right) - \mathbb{V}\left[\widehat{\beta}\right] = 0,$$

so that we deduce $\mathbb{V}\left[\widetilde{\beta}\right] \geq \mathbb{V}\left[\widehat{\beta}\right]$, and the theorem follows. □

Note that in the results presented so far, the variances and covariance of the estimators depend on the constant variance of the noise $\boldsymbol{\varepsilon}$, which is actually unknown. We can however provide an estimator for it. Define $\widehat{\mathbf{Y}} := \widehat{\alpha} + \widehat{\beta}\mathbf{X}$ and $\widehat{\boldsymbol{\varepsilon}} := \mathbf{Y} - \widehat{\mathbf{Y}}$.

Proposition 5.1.6. *The statistics* $\widehat{\sigma}^2 := \frac{1}{n-2}\left\|\widehat{\boldsymbol{\varepsilon}}\right\|_2^2$ *is an unbiased estimator of* σ^2.

Proof. For any $i = 1, \ldots, n$, since $y_i = \alpha + \beta x_i + \varepsilon_i$, then, summing over i yields $\overline{y} = \alpha + \beta\overline{x} + \overline{\varepsilon}$. Therefore, for any $i = 1, \ldots, n$, since $\overline{y} = \widehat{\alpha} + \widehat{\beta}\overline{x}$, we can write

$$\begin{aligned}\widehat{\varepsilon}_i := y_i - \widehat{y}_i &= \alpha + \beta x_i + \varepsilon_i - \left(\widehat{\alpha} + \widehat{\beta}x_i\right)\\ &= \left(\overline{y} - \beta\overline{x} - \overline{\varepsilon}\right) + \beta x_i + \varepsilon_i - \left(\overline{y} - \widehat{\beta}\overline{x}\right) - \widehat{\beta}x_i\\ &= \left(\beta - \widehat{\beta}\right)\left(x_i - \overline{x}\right) + \left(\varepsilon_i - \overline{\varepsilon}\right).\end{aligned}$$

Therefore, using (5.1.3) (in the second line),

$$\begin{aligned}\|\widehat{\boldsymbol{\varepsilon}}\|_2^2 = \sum_{i=1}^n \widehat{\varepsilon}_i^2 &= \left(\beta - \widehat{\beta}\right)^2 \sum_{i=1}^n (x_i - \overline{x})^2 + \sum_{i=1}^n (\varepsilon_i - \overline{\varepsilon})^2 + 2\left(\beta - \widehat{\beta}\right) \sum_{i=1}^n (x_i - \overline{x})(\varepsilon_i - \overline{\varepsilon}) \\ &= \left(\beta - \widehat{\beta}\right)^2 \sum_{i=1}^n (x_i - \overline{x})^2 + \sum_{i=1}^n (\varepsilon_i - \overline{\varepsilon})^2 - 2\left(\beta - \widehat{\beta}\right)^2 \sum_{i=1}^n (x_i - \overline{x})^2 \\ &= \sum_{i=1}^n (\varepsilon_i - \overline{\varepsilon})^2 - \left(\beta - \widehat{\beta}\right)^2 \|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2,\end{aligned}$$

and hence

$$\begin{aligned}\mathbb{E}\left[\|\widehat{\boldsymbol{\varepsilon}}\|_2^2\right] = \mathbb{E}\left[\sum_{i=1}^n (\varepsilon_i - \overline{\varepsilon})^2 - \left(\beta - \widehat{\beta}\right)^2 \|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2\right] &= \sum_{i=1}^n \mathbb{E}\left[(\varepsilon_i - \overline{\varepsilon})^2\right] - \|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2 \mathbb{V}\left[\widehat{\beta}\right] \\ &= (n-1)\sigma^2 - \sigma^2 = (n-2)\sigma^2.\end{aligned}$$

Indeed,

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}\left[(\varepsilon_i - \overline{\varepsilon})^2\right] &= \sum_{i=1}^n \mathbb{E}\left[\varepsilon_i^2 + \frac{1}{n^2}\left(\sum_{i=1}^n \varepsilon_i\right)^2 - 2\overline{\varepsilon}\sum_{i=1}^n \varepsilon_i\right] \\ &= \sum_{i=1}^n \left[\mathbb{E}\left[\varepsilon_i^2\right] + \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^n \varepsilon_i^2\right] - 2\overline{\varepsilon}\mathbb{E}\left[\sum_{i=1}^n \varepsilon_i\right]\right] \\ &= \sum_{i=1}^n \left[\sigma^2 + \frac{\sigma^2}{n} - 2\frac{\sigma^2}{n}\right] = (n-2)\sigma^2.\end{aligned}$$

The result then follows from Theorem 5.1.4. □

One goal of regression methods is to predict the behaviour of variables. Suppose then that we observe a new value, say x_{n+1}, and we want to be able to predict the unknown value of y_{n+1}. A natural candidate is to consider $\widehat{y}_{n+1} = \widehat{\alpha} + \widehat{\beta} x_{n+1}$. However, this yields the following error:

Proposition 5.1.7. *The forecasting error* $\widehat{\varepsilon}_{n+1} := y_{n+1} - \widehat{y}_{n+1}$ *satisfies*

$$\mathbb{E}\left[\widehat{\varepsilon}_{n+1}\right] = 0 \qquad \textit{and} \qquad \mathbb{V}\left[\widehat{\varepsilon}_{n+1}\right] = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2}\right).$$

Proof. Since ε_{n+1} is centered and the estimators $\widehat{\alpha}$ and $\widehat{\beta}$ are unbiased, then

$$\mathbb{E}\left[\widehat{\varepsilon}_{n+1}\right] = \mathbb{E}[\alpha - \widehat{\alpha}] + \mathbb{E}[\beta - \widehat{\beta}]x_{n+1} + \mathbb{E}[\varepsilon_{n+1}] = 0.$$

Furthermore,

$$\mathbb{V}\left[\widehat{\varepsilon}_{n+1}\right] = \mathbb{V}[y_{n+1} - \widehat{y}_{n+1}] = \mathbb{V}[y_{n+1}] + \mathbb{V}[\widehat{y}_{n+1}] = \sigma^2 + \mathbb{V}[\widehat{y}_{n+1}].$$

The second term can be computed explicitly as

$$\begin{aligned}\mathbb{V}[\widehat{y}_{n+1}] = \mathbb{V}\left[\widehat{\alpha}+\widehat{\beta}x_{n+1}\right] &= \mathbb{V}\left[\widehat{\alpha}\right] + x_{n+1}^2\mathbb{V}\left[\widehat{\beta}\right] + 2x_{n+1}\mathrm{Cov}\left(\widehat{\alpha},\widehat{\beta}\right)\\ &= \frac{\sigma^2}{\|\mathbf{X}-\overline{x}\mathbb{1}\|_2^2}\left(\frac{1}{n}\sum_{i=1}^n x_i^2 + x_{n+1}^2 - 2x_{n+1}\overline{x}\right)\\ &= \frac{\sigma^2}{\|\mathbf{X}-\overline{x}\mathbb{1}\|_2^2}\left(\frac{1}{n}\sum_{i=1}^n (x_i-\overline{x})^2 + \overline{x}^2 + x_{n+1}^2 - 2x_{n+1}\overline{x}\right)\\ &= \sigma^2\left(\frac{1}{n}+\frac{(x_{n+1}-\overline{x})^2}{\|\mathbf{X}-\overline{x}\mathbb{1}\|_2^2}\right),\end{aligned}$$

and the proposition follows. □

A useful quantity when performing linear regression is the so-called coefficient of determination $\mathfrak{R}^2$. Note first that, by Pythagoras' Theorem,

$$\mathrm{TSS} := \|\mathbf{Y}-\overline{y}\mathbb{1}\|_2^2 = \left\|\widehat{\mathbf{Y}}-\overline{y}\mathbb{1}\right\|_2^2 + \|\widehat{\boldsymbol{\varepsilon}}\|_2^2 =: \mathrm{ESS}+\mathrm{RSS},$$

so that the total sum of squares is equal to that explained by the model and the residual one.

Definition 5.1.8. The coefficient of determination $\mathfrak{R}^2$ is defined as $\mathfrak{R}^2 := \frac{\mathrm{ESS}}{\mathrm{TSS}}$.

If $\mathfrak{R}^2 = 1$, then the model fully explains the relationship, i.e. there is a perfectly linear relation between $\mathbf{X}$ and $\mathbf{Y}$. If $\mathfrak{R}^2 = 0$, then $\|\widehat{\mathbf{Y}}-\overline{y}\mathbb{1}\| = 0$, or $\widehat{y}_i = \overline{y}$ for all $i = 1,\ldots,n$, and the model is completely inadequate.

Exercise 30. Show that $\mathfrak{R}^2$ can also be understood as the square of the empirical correlation coefficient between the two vectors $\mathbf{X}$ and $\mathbf{Y}$, i.e. that

$$\mathfrak{R}^2 = \left(\frac{\sum_{i=1}^n (x_i-\overline{x})(y_i-\overline{y})}{\|\mathbf{X}-\overline{x}\|\|\mathbf{Y}-\overline{y}\|}\right)^2.$$

5.1.2 Study and estimation of the Gaussian linear regression model

We have so far assumed very little about the errors $\boldsymbol{\varepsilon}$. In order to refine the analysis above, we now assume that the sequence of errors $(\varepsilon_1,\ldots,\varepsilon_n)$ is iid Gaussian with constant variance σ^2. The linear regression, given the vector $\mathbf{X}$ known (observed), becomes a simple assumption on the Gaussian character of the output vector $\mathbf{Y} \sim \mathcal{N}(\alpha+\beta\mathbf{X},\sigma^2)$. We can first compute the likelihood of the sample as

$$\mathcal{L}_n(\alpha,\beta,\sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\alpha-\beta x_i)^2\right\},$$

so that the log-likelihood to be minimised reads

$$l_n(\alpha,\beta,\sigma^2) := -\frac{1}{n}\log\mathcal{L}_n(\alpha,\beta,\sigma^2) = \frac{n}{2}\log\left(2\pi\sigma^2\right) + \frac{\|\mathbf{Y}-\alpha\mathbb{1}-\beta\mathbf{X}\|_2^2}{2\sigma^2}.$$

Minimising over α and β is straightforward and yields exactly the least-square estimates computed above. Now,

$$\partial_{\sigma^2} l_n(\widehat{\alpha}, \widehat{\beta}, \sigma^2) = \frac{n}{2\sigma^2} - \frac{\left\| \mathbf{Y} - \widehat{\alpha}\mathbb{1} - \widehat{\beta}\mathbf{X} \right\|_2^2}{2\sigma^4} = \frac{n}{2\sigma^2} - \frac{\|\widehat{\varepsilon}\|_2^2}{2\sigma^4},$$

which is equal to zero if and only if $\sigma^2 = \widehat{\sigma}_L^2 := \frac{1}{n}\|\widehat{\varepsilon}\|_2^2$. It is easy to compute that

$$\mathbb{E}\left[\widehat{\sigma}_L^2\right] = \mathbb{E}\left[\frac{\|\widehat{\varepsilon}\|_2^2}{n}\right] = \frac{n-2}{n}\sigma^2,$$

so that the maximum likelihood estimator of the variance is biased. Let us now look at the fine properties of the estimators.

Theorem 5.1.9. *Both $\widehat{\alpha}$ and $\widehat{\beta}$ are Gaussian with means α and β and covariance matrix*

$$\mathrm{Cov}(\widehat{\alpha}, \widehat{\beta}) = \frac{1}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2} \begin{pmatrix} \frac{\|\mathbf{X}\|_2^2}{n} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}.$$

Furthermore, $\frac{(n-2)}{\sigma^2}\widehat{\sigma}^2 \sim \chi^2_{n-2}$, and $\widehat{\beta}$ and $\widehat{\sigma}^2$ are independent.

As before, one issue here is that the variance σ^2 of the errors is actually unknown, and we can replace it by its estimator, leading to more involved distributions for the above estimators. The following proposition allows us to build confidence intervals for the estimators.

Proposition 5.1.10. *Let t_{n-2} denote the Student law with $n-2$ degrees of freedom, and define*

$$\sigma_\alpha^2 := \frac{\sigma^2}{n\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2}\sum_{i=1}^n x_i^2 \qquad \text{and} \qquad \sigma_\beta^2 := \frac{\sigma^2}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2}$$

the variance of the least square estimators. Then

- $\frac{\widehat{\alpha} - \alpha}{\sigma_\alpha} \sim t_{n-2}$ *and* $\frac{\widehat{\beta} - \beta}{\sigma_\beta} \sim t_{n-2}$;
- *The intervals*

$$\left[\widehat{\alpha} - t_{n-2}^{1-\eta/2}\widehat{\sigma}_\alpha, \widehat{\alpha} + t_{n-2}^{1-\eta/2}\widehat{\sigma}_\alpha\right] \qquad \text{and} \qquad \left[\widehat{\beta} - t_{n-2}^{1-\eta/2}\widehat{\sigma}_\beta, \widehat{\beta} + t_{n-2}^{1-\eta/2}\widehat{\sigma}_\beta\right]$$

  are the confidence intervals for α and β with level $1 - \frac{\eta}{2}$.

Forecasting

Regarding forecasting, the general results above (without the Gaussian assumption for the errors) still hold for the mean and the variance, so that Proposition 5.1.7 remains unchanged. However, since σ^2 is again unknown, we need to replace its value by the estimator $\widehat{\sigma}^2$, and we obtain

Proposition 5.1.11.

$$\frac{\varepsilon_{n+1}}{\widehat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\|\mathbf{X} - \overline{x}\mathbb{1}\|_2^2}}} \sim t_{n-2},$$

and the corresponding confidence interval for y_{n+1} reads

$$\left[\widehat{y}_{n+1} - t_{n-2}^{1-\eta/2}\widehat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_{n+1}-\overline{x})^2}{\|\mathbf{X}-\overline{x}\mathbb{1}\|_2^2}}, \widehat{y}_{n+1} + t_{n-2}^{1-\eta/2}\widehat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_{n+1}-\overline{x})^2}{\|\mathbf{X}-\overline{x}\mathbb{1}\|_2^2}}\right].$$

5.1.3 Linear regression: the multidimensional case

We now consider a multidimensional version of the previous linear regression, so that we are interested in the following problem:

$$\min_{f\in\mathbf{F}}\sum_{i=1}^{n}\Phi\left[y_i - f(\mathrm{x}_i)\right],$$

for some given cost function Φ, where $\mathbf{F}$ is a class of functions of interest. The main difference here is that, for each $i = 1, \ldots, n$, x_i is a vector in $\mathbb{R}^p$, and the set $\mathbf{F}$ is composed of functions from $\mathbb{R}^p$ to $\mathbb{R}$. A multidimensional linear regression model is therefore a representation of the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{5.1.5}$$

where $\mathbf{Y} \in \mathbb{R}^n$ is the response/measured/endogenous variable, $\mathbf{X} \in \mathcal{M}_{n,p}(\mathbb{R})$ the exogenous or explanatory variable, $\boldsymbol{\beta} \in \mathbb{R}^p$ the vector of parameters to estimate, and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ the noise vector We shall always assume that the following conditions hold:

Assumption 5.1.12.

$$\mathrm{rank}(\mathbf{X}) = p \qquad \text{and} \qquad \mathbb{E}[\boldsymbol{\varepsilon}] = 0 \qquad \text{and} \qquad \mathbb{V}[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{I}_n.$$

Exercise 31. Consider the one-dimensional model

$$y_i = \alpha + \sum_{k=1}^{p}\beta_k x_i^k + \varepsilon_i,$$

for some integer p and all $i = 1, \ldots, n$. Write this problem in a linear form.

Exercise 32. Consider now a function f of the form $f(\mathrm{x}) = a\exp\left(\gamma^\top \mathrm{x}\right)$. Find a transform to reduce the model to a linear form.

Least-square estimators

Definition 5.1.13. The least-square estimator is defined as

$$\widehat{\boldsymbol{\beta}} = \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2.$$

We gather in the following proposition several results about $\widehat{\boldsymbol{\beta}}$. Note that the optimality statement is nothing else than a multi-dimensional version of the Gauss-Markov theorem (Theorem 5.1.5).

Proposition 5.1.14. *The optimal least-square estimator is* $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y}$ *and*

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta} \qquad \textit{and} \qquad \mathbb{V}[\widehat{\boldsymbol{\beta}}] = \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}.$$

Note that the matrix $\mathbf{X}^\top\mathbf{X}$ is invertible by Assumption 5.1.12.

Proof. The proof is a straightforward minimisation problem:

$$\nabla_{\boldsymbol{\beta}}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 = \nabla_{\boldsymbol{\beta}}\Big(\boldsymbol{\beta}^\top\left(\mathbf{X}^\top\mathbf{X}\right)\boldsymbol{\beta} - 2\mathbf{Y}^\top\mathbf{X}\boldsymbol{\beta} + \|\mathbf{Y}\|_2^2\Big) = 2\boldsymbol{\beta}\mathbf{X}^\top\mathbf{X} - 2\mathbf{Y}^\top\mathbf{X}.$$

Since the matrix $\mathbf{X}^\top\mathbf{X}$ is invertible, the result for $\widehat{\boldsymbol{\beta}}$ follows directly. Now,

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}] = \mathbb{E}\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y}\right] = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbb{E}\left[\mathbf{Y}\right] = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbb{E}\left[\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\right] = \boldsymbol{\beta}.$$

Likewise,

$$\begin{aligned}\mathbb{V}[\widehat{\boldsymbol{\beta}}] = \mathbb{V}\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y}\right] &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbb{V}\left[\mathbf{Y}\right]\left\{\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\right\}^\top \\ &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbb{V}\left[\mathbf{Y}\right]\mathbf{X}\left\{\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right\}^\top \\ &= \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\left\{\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right\}^\top,\end{aligned}$$

since $\mathbb{V}[\mathbf{Y}] = \mathbb{V}[\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}] = \mathbb{V}[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{I}_n$, and the proposition follows. □

Exactly like in the one-dimensional case, we can show that the least-square estimator is the optimal one, in the following sense:

Theorem 5.1.15. *Among all unbiased estimators of* $\boldsymbol{\beta}$ *linear in* $\mathbf{Y}$*, the least-square estimator* $\widehat{\boldsymbol{\beta}}$ *has minimal variance.*

Proof. Let $\widetilde{\boldsymbol{\beta}}$ be a linear estimator of $\boldsymbol{\beta}$. Then

$$\mathbb{V}\left[\widetilde{\boldsymbol{\beta}}\right] = \mathbb{V}\left[\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}+\widehat{\boldsymbol{\beta}}\right] = \mathbb{V}\left[\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right] + \mathbb{V}\left[\widehat{\boldsymbol{\beta}}\right] + \mathrm{Cov}\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\beta}}\right) + \mathrm{Cov}\left(\widehat{\boldsymbol{\beta}},\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}\right).$$

Since $\widetilde{\boldsymbol{\beta}}$ is linear in $\mathbf{Y}$, then we can write it $\widetilde{\boldsymbol{\beta}} = \mathbf{A}\mathbf{Y}$ for some matrix $\mathbf{A} \in \mathcal{M}_{p,n}$. Being unbiased, it further satisfies

$$\mathbb{E}\left[\widetilde{\boldsymbol{\beta}}\right] = \boldsymbol{\beta} = \mathbf{A}\mathbb{E}\left[\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\right] = \mathbf{A}\mathbf{X}\boldsymbol{\beta} + \mathbf{A}\mathbb{E}\left[\boldsymbol{\varepsilon}\right] = \mathbf{A}\mathbf{X}\boldsymbol{\beta},$$

so that $\mathbf{A}\mathbf{X} = \mathbf{I}$. The covariance term then reads (recall Proposition 2.3.5 for such computations)

$$\begin{aligned}\mathrm{Cov}\left(\widetilde{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\beta}}\right) &= \mathrm{Cov}\left(\mathbf{A}\mathbf{Y}-\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\beta}}\right) \\ &= \mathrm{Cov}\left(\mathbf{A}\mathbf{Y},\widehat{\boldsymbol{\beta}}\right) - \mathbb{V}\left[\widehat{\boldsymbol{\beta}}\right] \\ &= \mathrm{Cov}\left(\mathbf{A}\mathbf{Y},\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y}\right) - \mathbb{V}\left[\widehat{\boldsymbol{\beta}}\right] \\ &= \mathbf{A}\mathbb{V}\left[\mathbf{Y}\right]\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} - \mathbb{V}\left[\widehat{\boldsymbol{\beta}}\right] \\ &= \sigma^2\mathbf{A}\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} - \mathbb{V}\left[\widehat{\boldsymbol{\beta}}\right],\end{aligned}$$

which is equal to zero since $\widetilde{\boldsymbol{\beta}}$ is unbiased, and the theorem follows. □

From a geometric point of view, least-square minimisation is equivalent to determining the projection (in the Euclidean sense) of the random vector $\mathbf{Y}$ onto $\mathcal{M}_{\mathbf{X}}$, the subspace of $\mathbb{R}^n$ generated by the column vectors of the matrix $\mathbf{X}$. More specifically,

$$\mathcal{M}_{\mathbf{X}} = \left\{ \mathbf{X}\mathbf{w} = \sum_{i=1}^{p} w_1 \mathbf{X}_1, \mathbf{w} = (w_1, \dots, w_p) \in \mathbb{R}^p \right\},$$

where we denote by $\mathbf{X}_i$ the i-th column of $\mathbf{X}$. Since by assumption, $\mathbf{X}$ has rank p, then $\mathcal{M}_{\mathbf{X}}$ is of dimension p. By Proposition 5.1.14, the projection therefore reads $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} =: \mathbf{P}_{\mathbf{X}}\mathbf{Y}$, where $\mathbf{P}_{\mathbf{X}} := \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top$ is called the orthogonal projection matrix on $\mathcal{M}_{\mathbf{X}}$. (as a projection matrix, check that $\mathbf{P}_{\mathbf{X}}^2 = \mathbf{P}_{\mathbf{X}}$). Now, the residuals of the least square estimation read

$$\widehat{\boldsymbol{\varepsilon}} := \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{P}_{\mathbf{X}}\mathbf{Y} = \left(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}}\right)\mathbf{Y} =: \mathbf{P}_{\mathbf{X}^\top}\mathbf{Y} = \mathbf{P}_{\mathbf{X}^\top}\boldsymbol{\varepsilon}.$$

The matrix $\mathbf{P}_{\mathbf{X}^\perp}$ is the orthogonal projection matrix onto $\mathcal{M}_{\mathbf{X}^\perp}$. We can thus prove (exercise) the following properties:

Proposition 5.1.16. *The residuals are centered with* $\mathbb{V}\left[\widehat{\boldsymbol{\varepsilon}}\right] = \sigma^2\mathbf{P}_{\mathbf{X}^\perp}$ *and*

$$\mathbb{E}\left[\widehat{\mathbf{Y}}\right] = \mathbf{X}\boldsymbol{\beta}, \qquad \mathbb{V}\left[\widehat{\mathbf{Y}}\right] = \sigma^2\mathbf{P}_{\mathbf{X}}, \qquad \mathrm{Cov}\left(\widehat{\boldsymbol{\varepsilon}}, \widehat{\mathbf{Y}}\right) = 0.$$

Proposition 5.1.17. *The estimator* $\widehat{\sigma}^2 := \dfrac{\|\widehat{\boldsymbol{\varepsilon}}\|_2^2}{n-p}$ *is an unbiased estimator of the variance* σ^2.

Proof. First, note that

$$\mathbb{E}\left[\|\widehat{\boldsymbol{\varepsilon}}\|_2^2\right] = \mathbb{E}\left[\mathrm{Tr}\left(\|\widehat{\boldsymbol{\varepsilon}}\|_2^2\right)\right] = \mathbb{E}\left[\mathrm{Tr}\left(\widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}}\right)\right] = \mathbb{E}\left[\mathrm{Tr}\left(\widehat{\boldsymbol{\varepsilon}}\widehat{\boldsymbol{\varepsilon}}^\top\right)\right] = \mathrm{Tr}\left(\mathbb{E}\left[\widehat{\boldsymbol{\varepsilon}}\widehat{\boldsymbol{\varepsilon}}^\top\right]\right) = \mathrm{Tr}\left(\mathbb{V}\left[\widehat{\boldsymbol{\varepsilon}}\right]\right) = \mathrm{Tr}\left(\sigma^2\mathbf{P}_{\mathbf{X}^\perp}\right).$$

Finally, since

$$\mathrm{Tr}\left(\mathbf{P}_{\mathbf{X}^\perp}\right) = \mathrm{Tr}\left(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}}\right) = \mathrm{Tr}\left(\mathbf{I}_n\right) - \mathrm{Tr}\left(\mathbf{P}_{\mathbf{X}}\right) = \mathrm{Tr}\left(\mathbf{I}_n\right) - \mathrm{Tr}\left(\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\perp\right) = n - p,$$

the proposition follows. □

This in turn, combined with Proposition 5.1.14, directly gives an estimator of the variance of $\widehat{\boldsymbol{\beta}}$. Regarding forecast, we again mimic the one-dimensional case. Consider a new observation vector x_{n+1}, with new response variable $y_{n+1} = \mathrm{x}_{n+1}^\top\boldsymbol{\beta} + \varepsilon_{n+1}$, with $\mathbb{E}[\varepsilon_{n+1}] = 0$, $\mathbb{V}[\varepsilon_{n+1}] = \sigma^2$ and $\mathrm{Cov}(\varepsilon_{n+1}, \varepsilon_i) = 0$ for any $i = 1, \dots, n$. We define the new prevision as $\widehat{y}_{n+1} := \mathrm{x}_{n+1}\boldsymbol{\beta}$, and the prevision error $\widehat{\varepsilon}_{n+1} := y_{n+1} - \widehat{y}_{n+1}$.

Proposition 5.1.18. *The following identities hold*

$$\mathbb{E}\left[\widehat{\varepsilon}_{n+1}\right] = 0 \qquad \textit{and} \qquad \mathbb{V}\left[\widehat{\varepsilon}_{n+1}\right] = \sigma^2\left(1 + \mathrm{x}_{n+1}^\top\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathrm{x}_{n+1}\right).$$

Proof. Note first that

$$\widehat{\varepsilon}_{n+1} := y_{n+1} - \widehat{y}_{n+1} = \mathrm{x}_{n+1}^\top\boldsymbol{\beta} + \varepsilon_{n+1} - \mathrm{x}_{n+1}^\top\widehat{\boldsymbol{\beta}} = \mathrm{x}_{n+1}^\top\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right) + \varepsilon_{n+1},$$

so that, since ε_{n+1} is centered and $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$, we can write

$$\mathbb{E}\left[\widehat{\varepsilon}_{n+1}\right] = \mathbb{E}\left[\mathrm{x}_{n+1}^\top\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right) + \varepsilon_{n+1}\right] = \mathrm{x}_{n+1}^\top\left(\boldsymbol{\beta} - \mathbb{E}\left[\widehat{\boldsymbol{\beta}}\right]\right) = 0.$$

Now,

$$\mathbb{V}\left[\widehat{\varepsilon}_{n+1}\right] = \mathbb{V}\left[\mathrm{x}_{n+1}^\top\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right) + \varepsilon_{n+1}\right] = \mathbb{V}\left[\mathrm{x}_{n+1}^\top\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right)\right] + \mathbb{V}\left[\varepsilon_{n+1}\right] = \mathrm{x}_{n+1}^\top\mathbb{V}\left[\widehat{\boldsymbol{\beta}}\right]\mathrm{x}_{n+1} + \sigma^2$$

since ε_{n+1} is uncorrelated with $(\varepsilon_i)_{i=1,\ldots,n}$ and $\widehat{\boldsymbol{\beta}}$ only depends on the latter sequence. The result follows using Proposition 5.1.14. □

5.2 Departure from classical assumptions

5.2.1 Overfitting and regularisation

When performing regression analysis, overfitting occurs when noise or errors are described, rather than the actual relationships between the variables. This can essentially be due to two factors:

- the model may be too complicated (too many parameters compared to the number of observations). From Proposition 5.1.14, the matrix to invert is of size $p\times p$, but has rank $\min(p, n)$. Therefore is $n < p$, it is not invertible any longer;
- the data may contain some collinearity, when several explanatory variables are in fact highly (linearly) dependent.

We shall see below how to regularise this issue, but let us discuss multicollinearity first.

Multicollinearity

Ridge regression

With the classical linear regression problem in mind (Definition 5.1.13), consider the loss function

$$\mathfrak{L}(\boldsymbol{\beta}) := \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \tag{5.2.1}$$

which we aim to minimise. The ridge regression considers the minimisation, not of this loss function, but of the alternative, L^2-penalised version

$$\mathfrak{L}^R(\boldsymbol{\beta}) := \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2,$$

Following an analysis similar to the standard linear regression, we obtain that the optimal coefficient is given by

$$\widehat{\boldsymbol{\beta}}^R = \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{Y}.$$

The advantage of this penalisation is clear: even if the original matrix $\mathbf{X}^\top\mathbf{X}$ is not invertible, the new version is, for $\lambda > 0$. Furthermore, as the parameter λ increases, the coefficients of $\boldsymbol{\beta}$ decrease, making solutions with large coefficients less attractive.

Meaning of the ridge regression

Consider the Singular Value Decomposition for the matrix $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$. For the optimal linear regression coefficient vector, we can then write

$$\begin{aligned}\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y} &= \mathbf{U}\mathbf{D}\mathbf{V}^\top\left(\left[\mathbf{U}\mathbf{D}\mathbf{V}^\top\right]^\top\mathbf{U}\mathbf{D}\mathbf{V}^\top\right)^{-1}[\mathbf{U}\mathbf{D}\mathbf{V}^\top]^\top\mathbf{Y}\\ &= \mathbf{U}\mathbf{D}\mathbf{V}^\top\left(\mathbf{V}\mathbf{D}^2\mathbf{V}^\top\right)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}\mathbf{Y} = \mathbf{U}\mathbf{U}^\top\mathbf{Y},\end{aligned}$$

and in the case of ridge regression,

$$\mathbf{X}\widehat{\boldsymbol{\beta}}^R = \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{Y} = \mathbf{U}\mathbf{D}\left(\mathbf{D}^2 + \lambda\mathbf{I}\right)^{-1}\mathbf{D}\mathbf{U}^\top\mathbf{Y} = \sum_{j=1}^{p}\mathrm{u}_j\frac{d_j^2}{d_j^2+\lambda}\mathrm{u}_j^\top\mathbf{Y},$$

where the u_j represent the columns of the matrix $\mathbf{U}$. The implication is that the ridge regression applies more shrinkage when d_j^2 is small; since $d_1, \ldots, d_p$ represent the eigenvalues of the matrix $\mathbf{X}^\top\mathbf{X}$, ridge regression shrinks the coefficient $\boldsymbol{\beta}$ in the directions of small explained variances.

LASSO regression

This regression consists in penalising the loss function, not in the L^2 sense as in the ridge regression, but in the L^1 sense, as

$$\mathfrak{L}^L(\boldsymbol{\beta}) := \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1.$$

Unfortunately, here there is no closed-form expression for the solution vector $\widehat{\boldsymbol{\beta}}$, but this quadratic programming problem is easy to solve numerically.

5.2.2 Underfitting

If overfitting is an issue, underfitting is also one, when explanatory variables are forgotten. Assume that the true model is actually of the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

where the error term $\boldsymbol{\varepsilon}$ satisfies the classical assumptions, and the columns of $\mathbf{Z}$ are linearly independent of the columns of $\mathbf{X}$. Then the estimator $\widehat{\boldsymbol{\beta}}$ (Proposition 5.1.14) satisfies

$$\begin{aligned}\mathbb{E}\left[\widehat{\boldsymbol{\beta}}\right] = \mathbb{E}\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y}\right] &= \mathbb{E}\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\left(\mathbf{X}\beta + \mathbf{Z}\gamma\right)\right]\\ &= \mathbb{E}\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\beta\right] + \mathbb{E}\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Z}\gamma\right]\\ &= \boldsymbol{\beta} + \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Z}\gamma =: \boldsymbol{\beta} + \mathbf{L}\gamma,\end{aligned}$$

where $\mathbf{L}$ is the matrix of regression coefficients omitted in the regression considering only the $\mathbf{X}$ component. The estimator $\widehat{\boldsymbol{\beta}}$ is therefore biased. Note however that if the matrix $\mathbf{Z}$ is chosen such that its columns are orthogonal to those of $\mathbf{X}$, then $\mathbf{X}^\top\mathbf{Z} = \mathbf{O}$ and hence $\mathbf{L} = \mathbf{O}$, so that $\widehat{\boldsymbol{\beta}}$

becomes unbiased. Now, it is easy to see that the variance of the estimator remains unchanged (see Proposition 5.1.14):

$$\begin{aligned}
\mathbb{V}\left[\widehat{\boldsymbol{\beta}}\right] = \mathbb{V}\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y}\right] &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbb{V}\left[\mathbf{Y}\right]\left(\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\right)^\top \\
&= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbb{V}\left[\mathbf{Y}\right]\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} \\
&= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\left[\sigma^2\mathbf{I}_n\right]\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} \\
&= \sigma^2\mathbf{X}^\top\mathbf{X}.
\end{aligned}$$

However, concerning the estimator of the variance, similarly to Proposition 5.1.17, we can write

$$\begin{aligned}
\widehat{\sigma}^2 := \frac{\|\widehat{\varepsilon}\|^2}{n-p} = \frac{\left(\mathbf{Y}-\widehat{\mathbf{Y}}\right)^\top\left(\mathbf{Y}-\widehat{\mathbf{Y}}\right)}{n-p} &= \frac{\left([\mathbf{I}_n-\mathbf{P}_\mathbf{X}]\mathbf{Y}\right)^\top\left([\mathbf{I}_n-\mathbf{P}_\mathbf{X}]\mathbf{Y}\right)}{n-p} \\
&= \frac{\mathbf{Y}^\top\left(\mathbf{I}_n-\mathbf{P}_\mathbf{X}\right)^\top\left(\mathbf{I}_n-\mathbf{P}_\mathbf{X}\right)\mathbf{Y}}{n-p} \\
&= \frac{\mathbf{Y}^\top\left(\mathbf{I}_n-\mathbf{P}_\mathbf{X}\right)\mathbf{Y}}{n-p},
\end{aligned}$$

since the matrix $\mathbf{I}_n - \mathbf{P}_\mathbf{X}$ is a projection matrix (onto $\mathcal{M}_{\mathbf{X}^\perp}$). Therefore,

$$\begin{aligned}
(n-p)\mathbb{E}\left[\widehat{\sigma}^2\right] &= \mathbb{E}\left[\mathbf{Y}^\top\left(\mathbf{I}_n-\mathbf{P}_\mathbf{X}\right)\mathbf{Y}\right] \\
&= \mathrm{Tr}\left(\left(\mathbf{I}_n-\mathbf{P}_\mathbf{X}\right)\mathbb{V}\left[\mathbf{Y}\right]\right) + \mathbb{E}[\mathbf{Y}]^\top\left(\mathbf{I}_n-\mathbf{P}_\mathbf{X}\right)\mathbb{E}[\mathbf{Y}] \\
&= (n-p)\sigma^2 + (\mathbf{Z}\gamma)^\top\left(\mathbf{I}_n-\mathbf{P}_\mathbf{X}\right)\mathbf{Z}\gamma,
\end{aligned}$$

where we used the fact that the matrix $\mathbf{I}_n - \mathbf{P}_\mathbf{X}$ annihilates all linear dependence on $\mathbf{X}$ and so does not annihilate the $\mathbf{Z}$ part by assumption on the columns of $\mathbf{Z}$. Therefore $\widehat{\sigma}^2$ in general overestimates the variance σ^2. We also made use of the following simple result:

Lemma 5.2.1. *Let* $\mathbf{Y}$ *be an* $\mathbb{R}^n$*-valued random vector with second moment and* $\mathbf{A} \in \mathcal{M}_{nn}$ *a symmetric matrix. Then* $\mathbb{E}\left[\mathbf{Y}^\top\mathbf{A}\mathbf{Y}\right] = \mathrm{Tr}\left(\mathbf{A}\mathbb{V}[\mathbf{Y}]\right) + \mathbb{E}[\mathbf{Y}]^\top\mathbf{A}\mathbb{E}[\mathbf{Y}]$.

Proof. Since $\mathbf{Y}^\top\mathbf{A}\mathbf{Y}$ is of dimension one, it is equal to its trace, and hence

$$\begin{aligned}
\mathbb{E}\left[\mathbf{Y}^\top\mathbf{A}\mathbf{Y}\right] &= \mathrm{Tr}\left(\mathbb{E}\left[\mathbf{Y}^\top\mathbf{A}\mathbf{Y}\right]\right) = \mathbb{E}\left[\mathrm{Tr}\left(\mathbf{Y}^\top\mathbf{A}\mathbf{Y}\right)\right] = \mathbb{E}\left[\mathrm{Tr}\left(\mathbf{A}\mathbf{Y}\mathbf{Y}^\top\right)\right] \\
&= \mathrm{Tr}\left(\mathbb{E}\left[\mathbf{A}\mathbf{Y}\mathbf{Y}^\top\right]\right) = \mathrm{Tr}\left(\mathbf{A}\mathbb{E}\left[\mathbf{Y}\mathbf{Y}^\top\right]\right) = \mathrm{Tr}\left(\mathbf{A}\left\{\mathbb{V}[\mathbf{Y}] + \mathbb{E}[\mathbf{Y}]\mathbb{E}[\mathbf{Y}]^\top\right\}\right) \\
&= \mathrm{Tr}\left(\mathbf{A}\mathbb{V}[\mathbf{Y}]\right) + \mathbb{E}[\mathbf{Y}]^\top\mathbf{A}\mathbb{E}[\mathbf{Y}].
\end{aligned}$$

□

5.2.3 Incorrect variance matrix

Suppose that, instead of $\sigma^2\mathbf{I}_n$, the variance of the error is in fact equal to $\sigma^2\mathbf{V}$, for some matrix $\mathbf{V} \in \mathcal{M}_{nn}$. Then, it is straightforward to see that the estimator $\widehat{\boldsymbol{\beta}}$ remains unbiased. However,

$$\mathbb{V}\left[\widehat{\beta}\right] = \mathbb{V}\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y}\right] = \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{V}\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1},$$

which is in general different from the variance of $\widehat{\boldsymbol{\beta}}$ under the classical assumption. Furthermore, similar computations to above show that the estimator $\widehat{\sigma}^2$ for the variance is in general biased, as

$$\mathbb{E}\left[\widehat{\sigma}^2\right] := \frac{\mathbb{E}\left[\mathbf{Y}^\top(\mathbf{I}_n - \mathbf{P}_\mathbf{X})\mathbf{Y}\right]}{n-p} = \frac{\sigma^2}{n-p}\mathrm{Tr}\left((\mathbf{I}_n - \mathbf{P}_\mathbf{X})\mathbb{V}[\mathbf{Y}]\right) = \sigma^2\mathrm{Tr}(V).$$

5.2.4 Stochastic regressors

So far, one of the main assumptions of the model was that the regressor (or the vector of regressors) was deterministic. This is not so realistic in practice, and we hence need to extend the results to the case of random (or stochastic) regressors. We therefore rewrite the model (5.1.5) as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{5.2.2}$$

together with Assumption 5.1.12. Regarding the variable $\mathbf{X}$, we shall assume that the sequence $(\varepsilon_i, X_i)_{i=1,\ldots,n}$ is jointly iid. We now consider several dependence assumptions, each having particular consequences on the regression analysis:

- Independent $\mathbf{X}$: this is the strongest assumption possible, meaning that the distributions of $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ are unrelated.
- Conditional zero mean: $\mathbb{E}[\mathbf{X}|\varepsilon] = 0$.
- Uncorrelated $\mathbf{X}$: $\mathrm{Cov}(\mathbf{X}_i, \varepsilon_i) = 0$ for any $i = 1, \ldots, n$.
- Corelated $\mathbf{X}$: this is the most general case, where $\mathrm{Cov}(\mathbf{X}_i, \varepsilon_i) \neq 0$ for any $i = 1, \ldots, n$.

It is easy to see that the above conditions are ordered from the strongest to the weakest, each implying the weaker.

Remark 5.2.2. Suppose that we do not have accurate access to the data $\mathbf{X}$, but only to a noisy version $\widetilde{\mathbf{X}}$, say due to measurement errors (liquidity issues for some option prices, errors in the discounting curve used...), of the form

$$\mathbf{X} = \widetilde{\mathbf{X}} + \boldsymbol{\eta},$$

where $\boldsymbol{\eta}$ represents the error. The linear regression model (5.2.2) therefore reads

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \left(\widetilde{\mathbf{X}} + \boldsymbol{\eta}\right)\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \widetilde{\mathbf{X}}\boldsymbol{\beta} + (\boldsymbol{\eta}\boldsymbol{\beta} + \boldsymbol{\varepsilon}).$$

Now, even if the measurement error $\boldsymbol{\eta}$ is independent of the original error $\boldsymbol{\varepsilon}$, the actual regressor $\widetilde{\mathbf{X}}$ will be correlated with the new error $(\boldsymbol{\eta}\boldsymbol{\beta} + \boldsymbol{\varepsilon})$.

We now analyse the consequences of the randomness of the regressor on the linear regression. Recall from Proposition 5.1.14 that the optimal least-square estimator is

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y} \\
&= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\left(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right) \\
&= \boldsymbol{\beta} + \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}
\end{aligned}$$

For $i = 1, \ldots, p$, we can then write

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_i &= \boldsymbol{\beta}_i + \sum_{k=1}^{p} \left(\left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \right)_{ik} \left(\mathbf{X}^\top \boldsymbol{\varepsilon} \right)_k \\
&= \boldsymbol{\beta}_i + \sum_{k=1}^{p} \left(\left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \right)_{ik} \left(\sum_{j=1}^{n} \mathbf{X}_{jk} \boldsymbol{\varepsilon}_k \right) \\
&= \boldsymbol{\beta}_i + \sum_{j=1}^{n} \left[\sum_{k=1}^{p} \left(\left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \right)_{ik} \mathbf{X}_{jk} \right] \boldsymbol{\varepsilon}_j =: \boldsymbol{\beta}_i + \sum_{j=1}^{n} \mathbf{w}_i \boldsymbol{\varepsilon}_j
\end{aligned}$$

If $\boldsymbol{\varepsilon}_j$ is uncorrelated with $\mathbf{X}_k$ for $j \neq k$, it might not be so, in general, with the non-linear (in $\mathbf{X}$) term $\mathbf{w}$, so that, in general, we will have

$$\mathbb{E}\left[\widehat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta} + \mathbb{E}\left[\left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{X}^\top \boldsymbol{\varepsilon} \right] \neq \boldsymbol{\beta}.$$

Appendix A

Useful tools in probability theory and analysis

A.1 Useful tools in linear algebra

Let $n \in \mathbb{N}$ and consider a matrix $A = (a_{ij})_{1 \leq i,j \leq n} \in \mathcal{M}_n(\mathbb{R})$.

Definition A.1.1. The matrix A is said to be *positive definite* (respectively *positive semi-definite*) if $\mathrm{x}^t \mathrm{A} \mathrm{x} > 0$ (resp ≥ 0) for all non null vector $\mathrm{x} \in \mathbb{R}^n$.

For a matrix $\mathrm{A} \in \mathcal{M}_n(\mathbb{R})$, we define its *principal minors* as

$$\Delta_1 := a_{11}, \qquad \Delta_2 := \det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \qquad \ldots, \qquad \Delta_n := \det(\mathrm{A}).$$

Proposition A.1.2. *The following statements are equivalent:*

(i) A *is positive definite;*

(ii) all the eigenvalues of A *are positive;*

(iii) all leading principal minors of A *are positive.*

Exercise 33. Let $S \subset \mathbb{R}^n$ be a convex and open set. Let f be a continuously differentiable function on S. Recall that the function f is convex in S if for any two points x and y in S, the following inequality holds:

$$f(\alpha \mathrm{x} + (1-\alpha)\mathrm{y}) \leq \alpha f(\mathrm{x}) + (1-\alpha) f(\mathrm{y}), \quad \text{for any } \alpha \in [0,1].$$

Show the following:

(i) f is convex if and only if $f(\mathrm{y}) \geq f(x) + \nabla f(\mathrm{x})^T \cdot (\mathrm{y} - \mathrm{x})$ for all $(\mathrm{x}, \mathrm{y}) \in S \times S$;

(ii) if f is twice continuously differentiable on S, then f is convex if and only if the matrix $\nabla^2 f(\mathrm{x})$ is positive semi-definite for all $\mathrm{x} \in S$.

Definition A.1.3. The *spectral radius* ρ of a matrix $\mathrm{A} \in \mathcal{M}_n(\mathbb{R})$ is defined by $\rho(\mathrm{A}) := \max_{1 \leq i \leq n} \lambda_i$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A.

Bibliography

[1] M. Abramowitz and I. Stegun. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. New York: Dover Publications, 1972.

[2] L.B.G. Andersen, P. Jäckel and C. Kahl. Simulation of Square-Root Processes. Encyclopedia of Quantitative Finance, 2010.

[3] K.E. Atkinson. An introduction to numerical analysis, Second Edition. Wiley, 1989.

[4] D. H. Bailey and P. N. Swarztrauber. The fractional Fourier transform and applications. *SIAM Review*,33: 389-404, 1991.

[5] A.C. Berry. The Accuracy of the Gaussian Approximation to the Sum of Independent Variates. *Transactions of the American Mathematical Society*, 49 (1): 122-136, 1941.

[6] P. Billingsley. Convergence of probability measures (2nd ed.). John Wiley & Sons, 1999.

[7] F. Black and M. Scholes. The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, 81 (3): 637-654, 1973.

[8] P. Boyle. Option Valuation Using a Three-Jump Process. *International Options Journal* 3, 7-12, 1986.

[9] P. Carr, H. Geman, D. Madan and M. Yor. Stochastic volatility for Lévy processes. *Mathematical Finance*, 13(3): 345-382, 2003.

[10] P. Carr and D. Madan. Option valuation using the fast Fourier transform. *Journal of Computational Finance*, 2 (4): 61-73, 1999.

[11] P. Carr and D. Madan. Saddlepoint Methods for Option Pricing. *Journal of Computational Finance*, 13 (1): 4961, 2009.

[12] A.L. Cauchy. Cours d'analyse de l'Ecole Royale Polytechnique. Imprimerie royale, 1821. Reissued by Cambridge University Press, 2009.

[13] R. Cont and E. Voltchkova. A finite difference scheme for option pricing in jump diffusion and exponential Lévy models. *SIAM Journal On Numerical Analysis*, 43(4): 1596-1626, 2005.

[14] J.W. Cooley and J.W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19: 297-301, 1965.

[15] M. Cooney. Report on the accuracy and efficiency of the fitted methods for solving the Black-Scholes equation for European and American options. Working report, Datasim Education Ltd, Dublin, 2000.

[16] R. Courant, K. Friedrichs and H. Lewy. Über die partiellen Differenzengleichungen der mathematischen Physik. *Mathematische Annalen*, 100 (1): 32-74, 1928.

[17] J.C. Cox, J.E. Ingersoll and S.A. Ross. A Theory of the Term Structure of Interest Rates. *Econometrica*, 53: 385-407.

[18] J.C. Cox, S.A. Ross and M. Rubinstein. Option Pricing: A Simplified Approach. *Journal of Financial Economics*, 7: 229-263, 1979.

[19] J.Crank and P. Nicolson. A Practical Method for Numerical Evaluation of Solutions of Partial Differential Equations of Heat Conduction Type. *Proceedings of the Cambridge Philosophical Society* 43: 50-67, 1947.

[20] F. Delbaen and W. Schachermayer. A general version of the fundamental theorem of asset pricing. *Math Ann*, 300(1): 463-520, 1994.

[21] F. Delbaen and W. Schachermayer. The Mathematics of arbitrage. Springer, 2008.

[22] J. Douglas and H.H. Rachford. On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the AMS*, 82:421-439, 1956.

[23] D. Duffie, D. Filipovic, and W. Schachermayer. Affine processes and applications in finance. *The Annals of Applied Probability*, 13(3): 984-1053, 2003.

[24] C.-G. Esseen. On the Liapunoff limit of error in the theory of probability. *Arkiv f́ur Matematik, Astronomi och Fysik*, A28: 1-19, 1942.

[25] B. Flannery, W.H. Press, S. Teukolsky and W. Vetterling. Numerical Recipes. Cambridge University Press, Third Edition, 2007.

[26] I.M. Johnstone. On the distribution of the largest eigenvalue in principal component analysis. *Annals of Statistics*, 29(2): 295-327, 2001.

[27] S. Kullback and R.A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79-86, 1951.

[28] J. Lintner. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics*, 47(1): 13-37, 1965.

[29] R. Lugannani and S.O. Rice. Saddlepoint approximations for the distribution of the sum of independent random variables. *Advances in Applied Probability*, 12: 475-490, 1980.

[30] V.A. Marchenko and L.A. Pastur. Distribution of eigenvalues for some sets of random matrices. Mat. Sb. N.S. (in Russian). 72 (114:4): 507-536, 1967.

[31] K. Pearson. On lines and planes of closest

to systems of points in space. *Philosophical Magazine*, Series 6, 2(11):559-572, 1901.

[32] W.F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, 19 (3): 425-442, 1964.

[33] E. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, 62: 548-564, 1955.

[34] E. Wigner. On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics*, 67: 325-328, 1958.

[35] A.T.A Wood, J.G. Booth and R.W. Butler. Saddlepoint approximations with nonnormal limit distributions. *Journal of the American Statistical Association*, 88: 680-686, 1993.