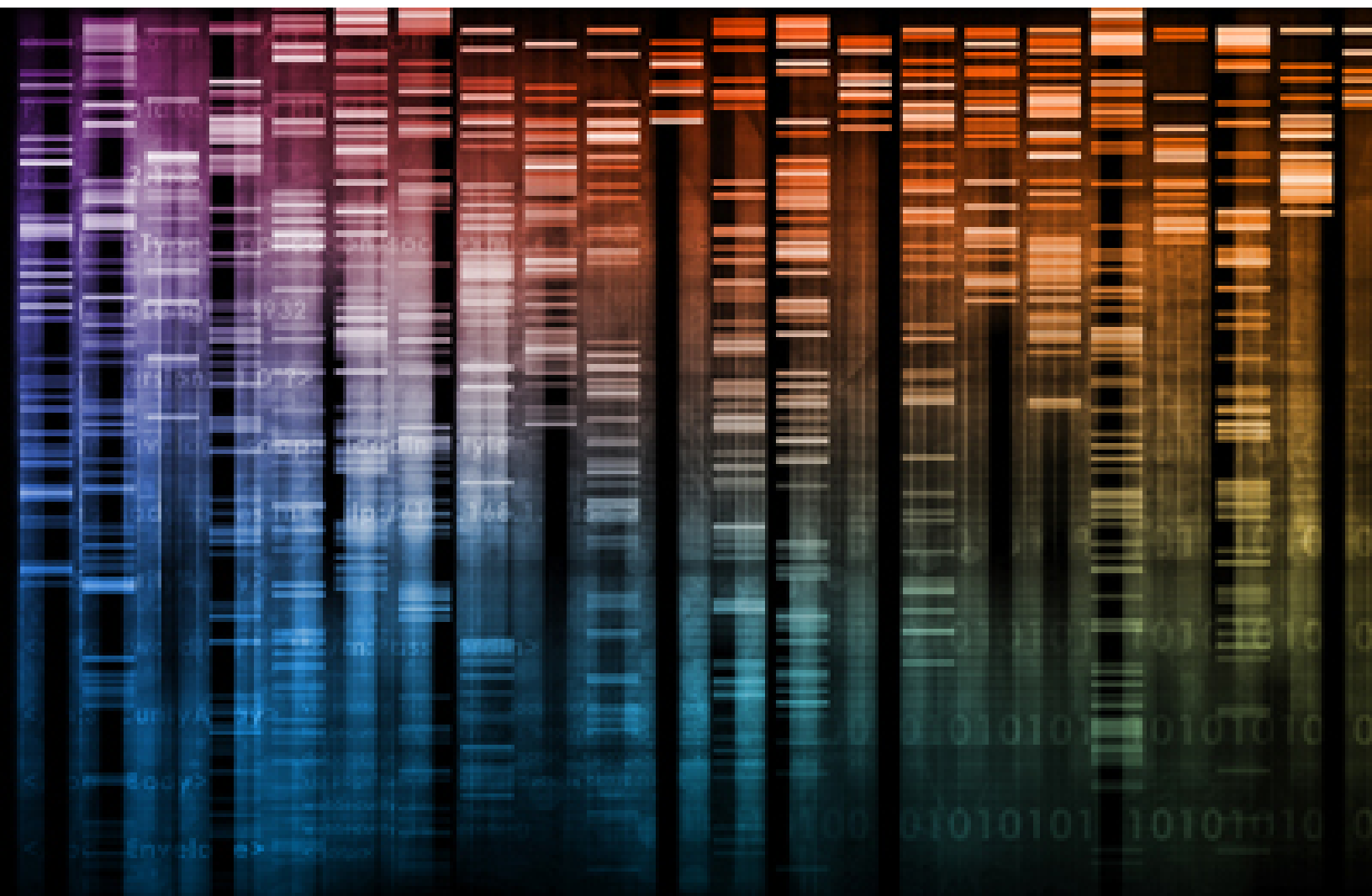


STATISTICS & LEARNING

1st Year, CentraleSupélec

2020-2021

Paul-Henry Cournède



'INGÉNIEUR' CURRICULUM, CENTRALESUPÉLEC

Official handbook of the Statistics and Learning course written by Paul-Henry Cournède, MICS laboratory & Mathematics department, CentraleSupélec. Please report any errors you may find at paul-henry.cournede@centralesupelec.fr.

First édition, April 2020

Contents

I	Elements of Mathematical Statistics	7
1	Data and Statistical Models	9
1.1	From data to statistical inference.	9
1.2	Probabilistic framework	9
1.2.1	Random Variables	9
1.2.2	Density	10
1.2.3	Conditional law and density function	11
1.3	Sampling theory	14
1.4	Empirical statistical method	15
1.4.1	Expectation and empirical mean	16
1.4.2	Empirical Variance	17
1.4.3	Empirical Distribution Function	17
1.4.4	Empirical Density	18
1.5	Statistical Model	20
1.5.1	Some examples of simple models	20
1.5.2	Regression Models	20

2	Parameter Estimation	23
2.1	Fisher Information	23
2.2	Point Estimation	27
2.2.1	Substitution method, method of moments	27
2.2.2	Maximum Likelihood	29
2.3	Properties of point estimators	31
2.3.1	Quadratic Risk and Bias	32
2.3.2	Efficiency of Unbiased Estimators	34
2.4	Limit theorems	37
2.5	Asymptotic properties of estimators	40
2.6	Asymptotic Properties of the Maximum Likelihood Estimator	43
2.7	Bayesian Estimation	45
2.7.1	Bayesian Model and Conditional Probability Distributions	45
2.7.2	Bayesian Method	46
2.7.3	Choice of the prior distribution	47
2.7.4	Point Bayesian Estimation	50
2.8	Confidence Region	52
2.8.1	Exact Confidence Region	52
2.8.2	Asymptotic confidence region	57
3	Statistical Hypothesis Testing	63
3.1	Problem Formulation	63
3.1.1	Statistical Hypothesis and Test	63
3.1.2	Risk of a test	64
3.2	Parametric Tests	67
3.2.1	Parametric Tests and Power Function	67
3.2.2	Method to Build Critical Regions	68
3.2.3	p -value for Parametric Tests	72
3.2.4	Pearson's χ^2 Test for the Multinomial Model	73

3.3	Goodness of Fit Tests	75
3.3.1	χ^2 -Test	75
3.3.2	Goodness of Fit Test of Kolmogorov-Smirnov	76
II	Statistical Learning	79
4	Supervised learning	81
4.1	Probabilistic Framework of Statistical Learning	81
4.1.1	Regression Model, Classification Model	82
4.1.2	Decision Theory	82
4.2	Models of Linear Regression	84
4.2.1	Definition	84
4.2.2	Estimation	85
4.3	Linear Models for Classification	95
4.3.1	Logistic Model for Classification	96
4.3.2	Performance criteria and classification decision	101
4.4	Regularization, Variable Selections, Model Selection	107
4.5	Basis Expansion for Non-linear Regression	113
4.5.1	Decision Trees	114
4.6	Artificial Neural Networks	117
4.6.1	Multilayer Perceptron for Regression and Classification	117
4.6.2	Activation Function	119
4.6.3	Training Neural Networks	120
5	Unsupervised Learning	123
5.1	Representation Learning	123
5.1.1	Principal Component Analysis	123
5.2	Clustering	131
5.2.1	Algorithme des K moyennes ou K -means	134
5.2.2	Choice of the Optimal Number of Clusters	135

A	Annexes	141
A.1	Lois usuelles	141
A.1.1	Lois discrètes	141
A.1.2	Lois continues	141
A.2	Propriétés des échantillons gaussiens	143

Part I

Elements of Mathematical Statistics

1. Data and Statistical Models

1.1 From data to statistical inference.

The following definition sets the basic framework for statistics.

Definition 1.1 — Data. A data is the observed realization of a random variable.

The random variable corresponds to a phenomenon that we want to study (size of an individual in a population, price of a financial product, remission rate of sick patients, yield of a crop ...). The law of probability P^* of the random variable is generally unknown, and the purpose of statistics is to provide the methodological framework for analysing the data in order to deduce knowledge about the underlying phenomenon.

In a simplified way, we can distinguish between **descriptive statistics**, which seeks to give meaning to the data by their description (what is the empirical mean observed, are there typical sub-populations ...), and **statistical inference**, which is interested in identifying the unknown law P^* , i.e. to infer the unknown characteristics of the studied phenomenon from the observations (e.g. the unknown characteristics of a population from a sample from this population).

In this course, we are primarily interested in statistical inference. Statistical modeling will provide the specific probabilistic assumptions that allow statistical inference to be translated into a formal framework. We will focus here on the mathematical treatment of two specific inferential approaches: parameter estimation and hypothesis testing.

1.2 Probabilistic framework

1.2.1 Random Variables

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, it corresponds to the set of elementary events. We consider a random variable X defined on $(\Omega, \mathcal{F}, \mathbb{P})$ with values in a measurable space $(\mathcal{X}, \mathcal{A})$.

We recall that the measurability condition for the random variable X ensures that the reciprocal image by X of any element A of the σ -algebra $\mathcal{A}$ has a probability and thus allows to define, on

$(\mathcal{X}, \mathcal{A})$, a probability measure, noted P_X , by:

$$P_X(A) = \mathbb{P}(X^{-1}(A)) = \mathbb{P}(X \in A).$$

The measure P_X is the image, by the application X , of the probability $\mathbb{P}$ defined on $(\Omega, \mathcal{F})$.

Definition 1.2 P_X is called the probability law of the random variable X .

In the remainder of this course, we will generally consider random variables with values in $(\mathcal{X}, \mathcal{A})$ where $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{X} \in \mathcal{B}(\mathbb{R}^d)$. For a continuous random variable, we'll generally take $\mathcal{A} = \mathcal{B}(\mathcal{X})$ the Borel σ -algebra over $\mathcal{X}$, and for a discrete value random variable, that is, if $\mathcal{X}$ is of finite cardinal or countable, $\mathcal{A} = \mathcal{P}(\mathcal{X})$, that is, all parts of $\mathcal{X}$.

1.2.2 Density

We recall below a remarkable consequence of the Radon-Nikodym theorem, based on the concept of measure dominance.

Definition 1.3 The σ -finite measure P is said **dominated** by a σ -finite measure μ if the sets that are negligible for μ are also negligible for P . P is said **absolutely continuous** with respect to μ .

Proposition 1.1 Let P be a probability measure. P admits a density p (or density function) which is the derivative of Radon-Nikodym $\frac{dP}{d\mu}$ with respect to P if and only if P is dominated by μ .

In this course, for continuous random variables the reference measure μ will be the Lebesgue measure and for discrete random variables the reference measure μ will be the count measure. This allows us to consider these two situations in the same theoretical framework. Unless otherwise stated, **all the laws of random variables considered in this course will admit a density with respect to either of these reference measures.**

■ **Remark 1.1** Note, however, the importance of the "unless otherwise stated" because there are relatively simple cases of random variables that do not admit a density. For example, if we consider two random variables X and Y , independent and uniform on $[0; 1]$, we can show that the pair $(X, \sup(X, Y))$ does not admit density with respect to λ_2 (Lebesgue's measure in dimension 2). ■

Thus, for a continuous random variable, p represents the density in the usual sense (defined almost everywhere as the derivative of the partition function), and:

$$\forall A \in \mathcal{B}(\mathbb{R}^d), P_X(A) = \int_A p(x) \lambda_d(dx)$$

where λ_d denotes Lebesgue's measure on $\mathbb{R}^d$.

For a discrete law with values in E , $p(x) = \mathbb{P}\{X = x\}$ denotes the point mass function, and we have:

$$\forall A \in \mathcal{P}(E), P_X(A) = \int_A p(x) K(dx) = \sum_{\alpha \in A} \mathbb{P}\{X = \alpha\}$$

where K is the counting measure on E .

Let us recall however that by using as a reference measure $\sum_{\alpha \in E} \delta_\alpha$, where E is a countable set containing at least all α values such that $P_X\{X = \alpha\} > 0$, we can also place ourselves on the Borel σ -algebra by considering:

$$\forall A \in \mathcal{B}(\mathbb{R}^d), P_X(A) = \int_A p(x) \sum_{\alpha \in E} \delta_\alpha(dx) = \sum_{\alpha \in A} \mathbb{P}\{X = \alpha\}.$$

Nous serons souvent amenés à réaliser des changements de variables aléatoires $Y = \varphi(X)$, qui se traduiront par des changements de densités, nous rappelons ci-dessous le théorème fondamental pour cela.

Theorem 1.2 — Change of variables in a probability density function. Let the random variable X with values in $\mathcal{S}_X \subset \mathbb{R}^d$ of density p_X with respect to Lebesgue's measure. Let $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}^d$ a transformation, such that φ is a $\mathcal{C}^1$ -diffeomorphism of $\mathcal{S}_X$ in $\mathcal{S}_Y := \varphi(\mathcal{S}_X)$. The Jacobian transformation matrix is given $\forall x \in \mathcal{S}_X$ by:

$$J_\varphi(x) = \left(\frac{\partial \varphi_i}{\partial x_j}(x) \right)_{1 \leq i, j \leq d}$$

The random variable $Y = \varphi(X)$ then admits for density function $\forall y \in \mathbb{R}^d$:

$$p_Y(y) = \frac{p_X(\varphi^{-1}(y))}{|J_\varphi(\varphi^{-1}(y))|} \mathbb{I}_{\mathcal{S}_Y}(y).$$

where $\mathbb{I}_{\mathcal{S}_Y}(y)$ is the indicator function, $\mathbb{I}_{\mathcal{S}_Y}(y) = 1$ if $y \in \mathcal{S}_Y$, 0 otherwise.

Remember that $\forall y \in \mathcal{S}_Y$:

$$|J_{\varphi^{-1}}(y)| = \frac{1}{|J_\varphi(\varphi^{-1}(y))|}$$

The proof derives from the formula of variable change for functions of several variables.

1.2.3 Conditional law and density function

The notion of conditioning is paramount in statistics. It expresses that the study of a random variable can generally benefit from a possible link between this variable and other auxiliary random variables (or events). The latter provide information to better determine the stochastic properties of the studied phenomenon.

Let two random variables X and Y with values respectively in $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$, two measurable spaces.

The course of probabilities (Billingsley, 1995) showed that it was easy to define the conditional law of Y given X when X is discrete, with values in $\mathcal{X}$. The conditional law of Y given X is the family of laws $\{P_{Y|X=x}, x \in \mathcal{A}\}$ on $\mathcal{B}$, such that if $P_X(\{x\}) \neq 0$,

$$P_{Y|X=x}(B) = \mathbb{P}(Y \in B | \{X = x\}) = \mathbb{P}(Y \in B, \{X = x\}) / P_X(\{x\}), \quad \forall B \in \mathcal{B} \quad (1.1)$$

and, $P_{Y|X=x}$ is any probability on $\mathcal{B}$ otherwise. This construction of conditional law does not extend to the general case, due to the impossibility of dividing by 0. The notion of a transitional kernel is introduced to get around this problem.

Definition 1.4 — Transition Kernel. A map $\nu: \mathcal{X}\mathcal{B} \rightarrow [0,1]$ is called a transition kernel if it satisfies the following properties:

- (i) $\forall x \in \mathcal{X}, \nu(x, \cdot)$ is a measure of probability on $\mathcal{B}$;
- (ii) $\forall B \in \mathcal{B}$, the map $\nu(\cdot, B)$ is $\mathcal{A}$ -measurable.

If μ is a probability measure on $\mathcal{A}$ and ν is a transition kernel on $\mathcal{B}$, then $\mu.\nu$ is the application defined on $\mathcal{A} \times \mathcal{B}$ by:

$$\mu.\nu(A \times B) = \int_A \nu(x, B) d\mu(x), \text{ for all } A \in \mathcal{A} \text{ and } B \in \mathcal{B}.$$

We can show that this application extends into a single probability on $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B})$ (see for example Billingsley (1995)). The extension is still denoted $\mu.\nu$. The application of Fubini's theorem is then direct and gives us the following properties, very useful in statistics.

Property 1.3 — Fubini's theorem. If μ is a measure of probability on $\mathcal{A}$ and ν a transition kernel on $\mathcal{X} \times \mathcal{B}$ and if $\phi: \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$ is a positive measurable function, then we can apply Fubini's theorem:

$$\int_{\mathcal{X} \times \mathcal{Y}} \phi(x, y) (\mu.\nu)(dx, dy) = \int_{\mathcal{X}} \left(\int_{\mathcal{Y}} \phi(x, y) \nu(x, dy) \right) \mu(dx) \quad (1.2)$$

Similarly, if ϕ is integrable on $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B}, \mu.\nu)$, then we have μ -almost everywhere, $y \rightarrow \phi(x, y)$ is $\nu(x, \cdot)$ -integrable, and the map defined μ -almost everywhere: $x \rightarrow \left(\int_{\mathcal{Y}} \phi(x, y) \nu(x, dy) \right)$ is μ -integrable. The above equality (1.2) is also true.

This notion of transitional kernel allows us to give a more general definition of a conditional law.

Definition 1.5 — Conditional law. We call the conditional law of Y given X a transition kernel ν on $\mathcal{X} \times \mathcal{B}$ such that $P_{(X,Y)} = P_X.\nu$.

We show that, under fairly broad conditions (especially if $\mathcal{X} = \mathbb{R}^p$ and $\mathcal{Y} = \mathbb{R}^q$ and $\mathcal{A}, \mathcal{B}$ the associated Borel σ -agebras), the conditional law of Y given X exists (cf. Jirina's theorem (Billingsley, 1995)).

The ν transition kernel is then often denoted $P_{Y|X=}$. (or simply $P_{Y|X}$). Note, however, that a conditional law is really a family of laws of probability. The notation $P_{Y|X=x}$ thus represents $\nu(x, \cdot)$ for $x \in \mathcal{X}$, which is indeed a measure of probability on $\mathcal{B}$.

■ **Remark 1.2** In the case where X has discrete values, we find the conditional law defined by (1.1). Indeed, for any $A \subset \mathcal{X}$ and $B \in \mathcal{B}$, we have by definition the generalized conditional law:

$$P_{(X,Y)}(A \times B) = \left(P_X.P_{Y|X=} \right) (A \times B)$$

Using (1.2) and the count measure, we have:

$$P_{(X,Y)}(A \times B) = \int_A P_{Y|X=x}(B) P_X(dx) = \sum_{x \in D_0 \cap A} P_{Y|X=x}(B) P_X(\{x\}) .$$

where $D_0 = \{x \in \mathcal{X} \text{ tels que } \mathbb{P}(X = x) \neq 0\}$. ■

■ **Remark 1.3** Let us note in particular that if $P_{Y|X=}$ exists and if $\phi : \mathcal{Y} \rightarrow \mathbb{R}$ is P_Y -integrable, the random variable $\psi(X)$ defined for any $x \in \mathcal{X}$ by:

$$\psi(x) = \int_{\mathcal{Y}} \phi(y) P_{Y|X=x}(dy)$$

is a determination of the conditional expectation $\mathbb{E}(\phi(Y)|X)$. This gives meaning to the frequently used notation $\mathbb{E}(\phi(Y)|X = x)$, which is equal to $\psi(x)$. ■

Let X, Y, Z be three random variables such that there is a measurable function φ , such as $Z = \varphi(X, Y)$ and one wishes to study the conditional law of $Z|X$, then the application of the Transfer Theorem (Billingsley, 1995) to the conditional law allows to obtain a very useful result in practice in statistics.

Theorem 1.4 — Conditional Transfer. Let φ be a measurable function of $\mathcal{X} \times \mathcal{Y}$ such that $P_{\varphi(X,Y)|X=}$ and $P_{Y|X=}$ exist. Then, $\forall x \in \mathcal{X}$, $P_{\varphi(X,Y)|X=x} = P_{\varphi(x,Y)|X=x}$. In particular, if X and Y are independent, we have $\forall x \in \mathcal{X}$: $P_{\varphi(X,Y)|X=x} = P_{\varphi(x,Y)}$.

Corollary 1.5 Under the right conditions of integrability, the calculation of expectations of the type $\mathbb{E}[\varphi(X, Y)|X]$ is then facilitated:

$$\mathbb{E}[\varphi(X, Y)|X = x] = \mathbb{E}[\varphi(x, Y)|X = x] = \int_{\mathcal{Y}} \varphi(x, y) P_{Y|X=x}(dy) .$$

If, in addition, X and Y are independent, we have:

$$\mathbb{E}[\varphi(X, Y)|X = x] = \int_{\mathcal{Y}} \varphi(x, y) P_Y(dy).$$

These subtle concepts find a very simple expression when the pair of random variables (X, Y) admits a probability density function p .

■ **Definition 1.6 — Conditional Density.** If for P_X -almost all $x \in \mathbb{R}^p$, $P_{Y|X=x}$ has a density, it is called *conditional density* of Y given $X = x$.

Theorem 1.6 — Conditional Density. If (X, Y) admits a density p , then for all x such tha

$p_X(x) > 0$, the law $P_{Y|X=x}$ has a conditional density given by:

$$p_{Y|X=x}(y) = \frac{p(x, y)}{p_X(x)}.$$

Preuve : Let λ, μ be the reference measures on $(\mathcal{X}, \mathcal{A})$, $(\mathcal{Y}, \mathcal{B})$. Let ρ be any probability on $(\mathcal{Y}, \mathcal{B})$. For all $x \in \mathcal{X}$ and $B \in \mathcal{B}$, we define

$$\nu(x, B) = \int_B \frac{p(x, y)}{p_X(x)} \mu(dy),$$

if $p_X(x) > 0$, and $\nu(x, B) = \rho(B)$ if $p_X(x) = 0$. ν is then a conditional law of Y given X , since ν is a transition kernel on $\mathcal{X} \times \mathcal{Y}$ satisfying $P_{(X,Y)} = P_X \cdot \nu$.

Moreover, if $p_X(x) > 0$, the measure $\nu(x, \cdot)$ has the density $p_{(X,Y)}(x, \cdot)/p_X(x)$ with respect to μ . $\square$

■ **Remark 1.4** We will often use the simplified notation $p_{Y|X=x}(y) = p(y|x)$ if there is no ambiguity. ■

These preliminary remarks will later allow us to manipulate with a great deal of freedom conditional laws, whether discrete or continuous, or pairs of random variables mixing the two types.

1.3 Sampling theory

Let X be a random variable with values in $(\mathcal{X}, \mathcal{A})$, of unknown law of probability P_X , corresponding to our phenomenon of interest. We need data to make statistical inference on this unknown law. In general, a single data, i.e. a single observation of a realization of this random variable is not enough. Several observations of the random variable are then realized, which is formalized by the concept of sample.

Definition 1.7 — Sample of observations - Random sample. Let $(\Omega, \mathcal{F}, \mathbb{P})$, be a probability space over which is defined the random variable X .

The random variable defined on the product space $(\Omega^N, \mathcal{F}^{\otimes N})$ with values in $(\mathcal{X}^N, \mathcal{A}^{\otimes N})$:

$$(\omega_1, \omega_2, \dots, \omega_N) \mapsto (X(\omega_1), X(\omega_2), \dots, X(\omega_N))$$

is called a random sample of size N .

$(X(\omega_1), X(\omega_2), \dots, X(\omega_N))$ is called a sample of observations. It is generally denoted by $(x_1, x_2, \dots, x_N)$.

■ **Remark 1.5** Note the difference between the notion of random sample and sample of observations. In one case, we consider the random variable and in the other its realization. However, we will not hesitate to speak simply of sample, in both situations, when there is no possible ambiguity. ■

An important case is when the random sample is of probability law $P_X^{\otimes N}$, i.e. the product law. It is equivalent to consider N independent realizations of a random variable of law P_X and the

realization of N independent random variables $X_1, X_2, \dots, X_N$ of the same law P_X . This is the framework of the renewable experiment. We will then speak of an independent identically distributed sample, which will often provide the basic framework for statistical inference on the unknown P_X law.

Definition 1.8 — Sample of independent and identically distributed random variables. Si les $(X_i)_{1 \leq i \leq N}$ sont des variables aléatoires dans $(\mathcal{X}, \mathcal{A})$, indépendantes de même loi P_X , alors le vecteur aléatoire $(X_1, X_2, \dots, X_N)$ dans $(\mathcal{X}^N, \mathcal{A}^{\otimes N})$ is said sample of independent and identically distributed random variables (i.i.d.) for P_X , or for short, i.i.d. sample. We will denote:

$$X_1, X_2, \dots, X_n \sim P_X$$

or if P_X has a density p_X :

$$X_1, X_2, \dots, X_n \sim p_X$$

The joint law of the random vector is the product law $P_{X_1} \otimes P_{X_2} \otimes \dots \otimes P_{X_N} = P_X^{\otimes N}$.

In practice, a sample is usually of finite size. However, the general framework of an infinite sampling model (in the countable sense) is indispensable for the theoretical study of asymptotic properties, which will call for the theory on the limits of sequences and series of random variables.

1.4 Empirical statistical method

Let us now give the mathematical definition of a statistic.

Definition 1.9 — Statistic. Let $(X_1, \dots, X_N)$ be a random sample of size N in $(\mathcal{X}^N, \mathcal{A}^{\otimes N})$. A statistic T is a measurable function on $(\mathcal{X}^N, \mathcal{A}^{\otimes N})$ with values in a measurable space $(\mathcal{Z}, \mathcal{C})$. By a slight misuse, we will also call statistic of the sample the function $T \circ (X_1, \dots, X_N)$ defined on $(\Omega^N, \mathcal{F}^{\otimes N})$ and denoted $T(X_1, \dots, X_N)$.

The $\mathcal{Z}$ space can be diverse. It can be $\mathcal{X}$, but it can also be a functional space. Below are some examples of common statistics.

■ Example 1.6

1. The empirical mean, $T : (\mathcal{X}^N, \mathcal{A}^{\otimes N}) \rightarrow (\mathcal{X}, \mathcal{A})$

$$T(X_1, \dots, X_N) = \frac{1}{N} \sum_{i=1}^N X_i$$

2. The maximum, $T : (\mathcal{X}^N, \mathcal{A}^{\otimes N}) \rightarrow (\mathcal{X}, \mathcal{A})$

$$T(X_1, \dots, X_N) = \max_{i=1, \dots, N} (X_i)$$

■

Subsequently, we will consider $(X_1, \dots, X_N)$ a N -sample of independent¹ and identically distributed random variables of law P_X . The statistics defined on this sample allow us to apprehend some characteristics of the unknown distribution, without any a priori presumption about it.

First of all, let us recall the definition of Dirac's measure.

Definition 1.10 — Dirac's Measure. Let $(\mathcal{X}, \mathcal{A})$ be a measurable space. Let $x \in \mathcal{X}$. The map δ_x defined on $\mathcal{A}$ such that for all $A \in \mathcal{A}$, $\delta_x(A) = 1$ if $x \in A$, and $\delta_x(A) = 0$ otherwise, is a measure called Dirac's measure associated to x .

We will also use the notation corresponding to the indicator function defined for all $A \in \mathcal{A}$ by

$$\mathbb{I}_A(x) = \begin{cases} 1 & \text{si } x \in A \\ 0 & \text{sinon} \end{cases}$$

Definition 1.11 — Empirical measure. For an i.i.d. sample $(X_1, \dots, X_N)$, the empirical measure on $(\mathcal{X}, \mathcal{A})$ is defined as the statistic $\hat{P}_X$ from $(\mathcal{X}^N, \mathcal{A}^{\otimes N})$ into the space of probability measures on $(\mathcal{X}, \mathcal{A})$ by:

$$\forall (x_1, \dots, x_N) \in \mathcal{X}^N, \hat{P}_X(x_1, \dots, x_N) = \frac{1}{N} \sum_{i=1}^N \delta_{x_i}.$$

The statistic will be denoted: $\hat{P}_X(X_1, \dots, X_N) = \frac{1}{N} \sum_{i=1}^N \delta_{X_i}$.

We can then approach the characteristics of the true P_X law from the empirical measure $\hat{P}_X$ as follows. If the characteristic is expressed as a mapping from $\mathcal{P}$, the set of probability measures on $(\mathcal{X}, \mathcal{A})$, into a measurable space $(\mathcal{Z}, \mathcal{C})$, $G : \mathcal{P} \rightarrow \mathcal{Z}$, then $G(P_X)$ is the true characteristic and $G(\hat{P}_X(x_1, \dots, x_N))$ is the empirical characteristic associated with the sample of observations $(x_1, \dots, x_N)$.

To determine whether an empirical characteristic converges to the true characteristic, limit theorems from probability theory must be used. Most useful theorems are recalled in the section 2.4. For characteristics that can be expressed as means, various versions of the law of large numbers are particularly useful to show the convergence.

1.4.1 Expectation and empirical mean

Let X be a random variable in $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, integrable, of law P_X . We recall that:

$$\mathbb{E}(X) = \int_{\mathbb{R}} x P_X(dx).$$

Let $(x_1, \dots, x_N)$ be a sample of observations for $(X_1, \dots, X_N)$, i.i.d. random variables of law P_X . Then the empirical mean denoted $\hat{\mathbb{E}}(x_1, \dots, x_N)$ associated to the empirical measure

¹Although it is beyond the scope of this course, it is worth keeping in mind that many results extend through the ergodic theorem for samples of non-independent random variables generated by a stationary Markov chain of distribution P , cf. Robert and Casella (1999), which is the basis of a paramount method in numerical statistics, the MCMC (Markov chain Monte-Carlo).

associated to $(x_1, \dots, x_N)$ is given by:

$$\hat{\mathbb{E}}(x_1, \dots, x_N) = \int_{\mathcal{X}} x \hat{P}_X(x_1, \dots, x_N)(dx) ,$$

that is to say:

$$\hat{\mathbb{E}}(x_1, \dots, x_N) = \int_{\mathbb{R}} x \left(\frac{1}{N} \sum_{i=1}^N \delta_{x_i} \right) (dx) = \frac{1}{N} \sum_{i=1}^N \int_{\mathcal{X}} x \delta_{x_i}(dx) = \frac{1}{N} \sum_{i=1}^N x_i .$$

We denote by $\bar{X}_N$ (or $\bar{X}$) the corresponding random variable:

$$\bar{X}_N = \frac{1}{N} \sum_{i=1}^N X_i .$$

Likewise, note that if X is a random variable in $(\mathcal{X}, \mathcal{A})$, and if $h : (\mathcal{X}, \mathcal{A}) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is P_X -integrable:

$$\mathbb{E}(h(X)) = \int_{\mathcal{X}} h(x) P_X(dx) ,$$

and for $(x_1, \dots, x_N)$ a sample of observations for $(X_1, \dots, X_N)$, i.i.d. random variables of law P_X , we obtain the corresponding empirical mean $\hat{E}_h$:

$$\hat{E}_h(x_1, \dots, x_N) = \int_{\mathcal{X}} x \hat{P}_X(x_1, \dots, x_N)(dx) = \frac{1}{N} \sum_{i=1}^N h(x_i) .$$

For a real-valued random variable, if we take $h(X) = X^k$, we thus obtain the empirical moment of order $k \geq 1$:

$$\hat{m}_k(x_1, \dots, x_N) = \frac{1}{N} \sum_{i=1}^N x_i^k .$$

1.4.2 Empirical Variance

Let X be a real-valued and square-integrable random variable, remembering that $\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$, we can deduce the empirical variance, given by the random variable:

$$S^2 = \frac{1}{N} \sum_{i=1}^N X_i^2 - \bar{X}_N^2$$

which is also written

$$S^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X}_N)^2 .$$

1.4.3 Empirical Distribution Function

Let X be a random variable, $X : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. We recall the definition of the distribution function on $\mathbb{R}$:

$$\forall x \in \mathbb{R}, F(x) = \mathbb{P}\{X \leq x\} = P_X([-\infty, x])$$

And for $(x_1, \dots, x_N)$, a sample of observations for $(X_1, \dots, X_N)$, i.i.d. random variables of law P_X , The empirical distribution function is given by:

$$\hat{F}(x_1, \dots, x_N)(x) = \left(\frac{1}{N} \sum_{i=1}^N \delta_{x_i} \right) ([-\infty, x]) = \frac{1}{N} \sum_{i=1}^N \mathbb{I}_{[-\infty, x]}(x_i)$$

1.4.4 Empirical Density

Let X be a random variable in $(bR, cB(bR))$, absolutely continuous with respect to Lebesgue's measure λ . It admits a density p_X . Let $(x_1, \dots, x_N)$ a sample of observations for $(X_1, \dots, X_N)$ an N -sample of i.i.d. random variables of law P_X . Then, the empirical measure $\hat{P}_X(x_1, \dots, x_N)$ is not absolutely continuous with respect to Lebesgue's measure: $\hat{P}_X(x_1, \dots, x_N)(\{x_i\}) = \frac{1}{N}$ whereas $\lambda(\{x_i\}) = 0$. Therefore, determining an empirical density is not possible.

However, there is a very useful way to approximate the density from the sample $(x_1, \dots, x_N)$, by histograms, which represent the number (or frequency) of observations at intervals. For a real-valued random variable, we choose $a_0 < a_1 < \dots < a_L$ in $\mathbb{R}$ such that

$$\{x_1, \dots, x_N\} \subset [a_0; a_L]$$

and we take:

$$\hat{f}(x_1, \dots, x_N)(x) = \sum_{j=1}^L \frac{\sum_{i=1}^N \mathbb{I}_{[a_{j-1}; a_j]}(x_i)}{N(a_j - a_{j-1})} \mathbb{I}_{[a_{j-1}; a_j]}(x)$$

Another way to approximate density from a sample is the kernel method.

Definition 1.12 The function $K : \mathbb{R} \rightarrow \mathbb{R}_+$ is called kernel if it is integrable, even and if:

$$\int_{\mathbb{R}} K(x) dx = 1 .$$

A kernel is thus a density of probability. Let us notice that if K is a kernel, then for any $h > 0$, $x \rightarrow \frac{1}{h} K\left(\frac{x}{h}\right)$ is also a kernel. So we can choose h as a scaling factor to fit the data.

Among the most frequently used kernels, we find:

- the Gaussian kernel:

$$K(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} ,$$

- Epanechnikov's kernel:

$$K(x) = \frac{3}{4} (1 - x^2) \mathbb{I}_{\{|x| \leq 1\}}(x) ,$$

- or the triangle kernel:

$$K(x) = (1 - |x|) \mathbb{I}_{\{|x| \leq 1\}}(x) .$$

To approximate the probability density, we take:

$$\hat{f}_K(x_1, \dots, x_N) = \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x - x_i}{h}\right) ,$$

for a well-chosen h . Note that if K is the Gaussian kernel, then Silverman's rule suggests taking $h = 0.9 S N^{-0.2}$ where S is the empirical standard deviation on $(x_1, \dots, x_N)$, (Silverman, 1986). In Figure 1.1, we compare for different sizes ($N = 100, 1000, 10000$) of i.i.d. samples for the Gaussian distribution $\mathcal{N}(19.6^2)$ the histogram, the density reconstructed by the kernel method with a Gaussian kernel and Silverman's rule, and the true density. We find that for a sample of 100, if the histogram remains a very coarse representation of the density, the reconstruction by the kernel method is already very relevant. For a sample of 10000, both the histogram and the reconstructed density are excellent approximations of the true density.

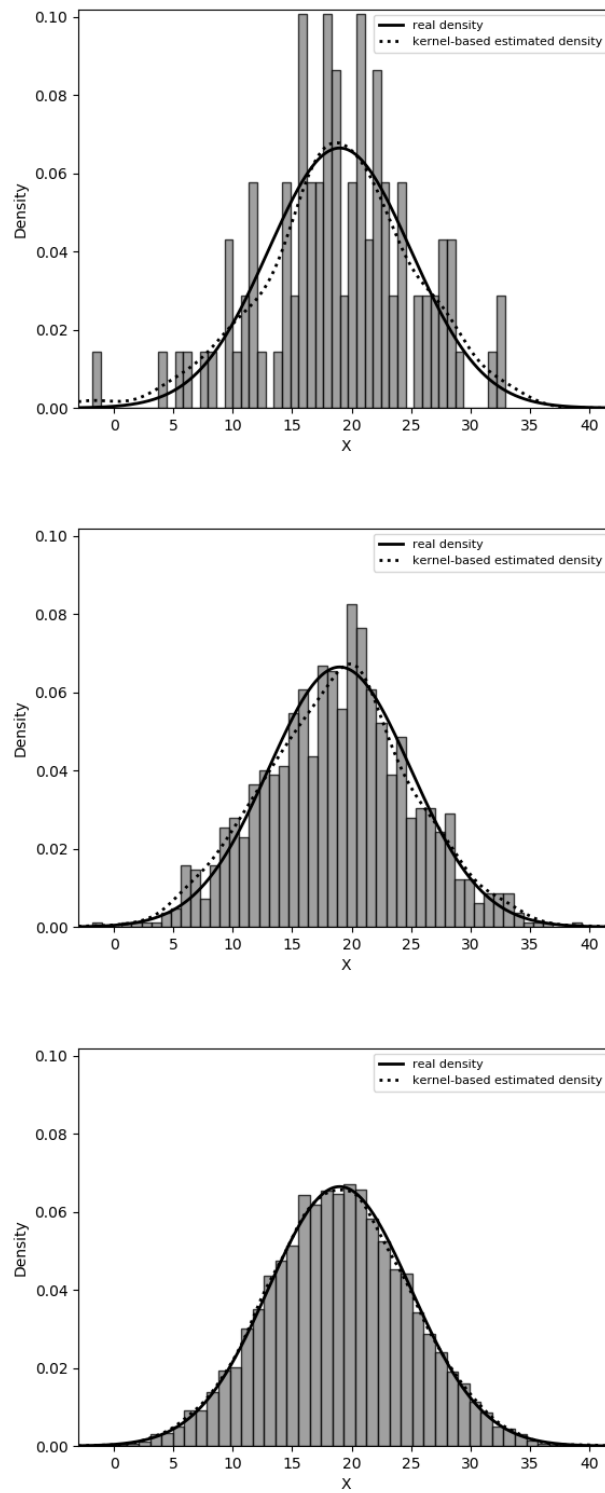


Figure 1.1: Histogram, density estimated by the kernel method and true density for sample sizes $N = 100, 1000, 10000$ respectively.

1.5 Statistical Model

For a random variable X in $(\mathcal{X}, \mathcal{A})$ with an unknown P_X distribution, the empirical method aimed to approximate some characteristics of P_X from observations of an i.i.d. sample, without a priori on the P_X distribution. We are now going to add a "modeling" step, i.e. a postulate on the family of laws to which P_X belongs. This is the principle of statistical modeling.

Definition 1.13 — Statistical Model. A $\mathcal{M}$ statistical model is a family of probability laws on $(\mathcal{X}, \mathcal{A})$. If this family can be parameterized by Θ , a finite dimensional space, we speak of a **parametric statistical model** and we denote: $\mathcal{M}_\Theta = \{P_\theta, \theta \in \Theta\}$.

The case where Θ is of infinite dimension (in a functional space for example) concerns the field of non-parametric statistics that will be seen later in the curriculum.

If the family of probability laws is dominated by a finite *sigma*-measure, then they admit densities p_θ and we get $\mathcal{M}_\Theta = \{p_\theta, \theta \in \Theta\}$.

If $X \sim p_\theta$, then we will also use the following notation: $p_\theta(x) = p(x|\theta)$, i.e. as a conditional density considering the parameter θ as a random variable. The reasons for this will become clear when we see the concepts of Bayesian statistics in part 2.7.1.

1.5.1 Some examples of simple models

We speak of a simple statistical model when the family of laws is given by a family of classical probability distributions. Below we consider various examples:

- Family of Poisson distributions (univariate, discrete), $\mathcal{M}_\Theta = \{P_\lambda, \lambda \in \mathbb{R}_+^*\}$

$$\mathbb{P}(X = k) = P_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

- Family of Gaussian distributions (univariate, continuous), $\mathcal{M}_\Theta = \{p_{(\mu, \sigma^2)}, (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*\}$

$$p_{(\mu, \sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Family of multivariate Gaussian distributions (multivariate, continuous),

$$\mathcal{M}_\Theta = \{p_{(\mu, \Sigma)}, (\mu, \Sigma) \in \mathbb{R}^d \times \mathcal{S}_d^+(\mathbb{R})\}, \text{ with } p_{(\mu, \Sigma)}(x) = (2\pi)^{-d/2} |\Sigma|^{-1/2} e^{-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)}$$

1.5.2 Regression Models

Let X and Y be two random variables with values in $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$ respectively, such that the pair (X, Y) admits a density with respect to the product measure on $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B})$.

We will generally consider cases where $(\mathcal{X}, \mathcal{A}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, we will speak of quantitative variables, or $(\mathcal{X}, \mathcal{A}) = (\mathbb{N}^m, \mathcal{P}(\mathbb{N}^m))$, we will speak of qualitative or categorical variables. We can also find a combination of the two types of variables $(\mathcal{X}, \mathcal{A}) = (\mathbb{R}^d \times \mathbb{N}^m, \mathcal{B}(\mathbb{R}^d) \otimes \mathcal{P}(\mathbb{N}^m))$. The same will be true for $(\mathcal{Y}, \mathcal{B})$, the two variables X and Y not necessarily being of the same type.

We are interested in a fundamental problem in statistics (and more specifically in the field of statistical learning), that of predicting (or estimating) Y from X . The random variable Y will

then be said to be an explained variable and X an explanatory variable (we will also speak of covariates or *features* or *inputs* for all the components of X).

For all $x \in \mathcal{X}$ such that $p_X(x) \neq 0$, we recall that we can define the conditional density $p_{Y|X=x}(y) = p(x, y)/p_X(x)$ (see Section 1.2.3).

Definition 1.14 — Modèle statistique de régression. Under the above assumptions, we call a statistical regression model a family of conditional probability laws given X . Especially if the family is parameterized and the conditional laws admit densities, we will have:

$$\mathcal{M}_\Theta = \{p_\theta(y|x), \theta \in \Theta, x \in \mathcal{X}\}$$

using the notation $p_{Y|X=x}(y) = p(y|x)$.

In practice, we will observe pairs of explanatory variables/explained variables (X_i, Y_i) and use them for inference.

■ **Example 1.7 — Linear Regression Model.**

Let X be a random variable with values in $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$, Y, ξ random variables in $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, and $\beta_0 \in \mathbb{R}$, $\beta \in \mathbb{R}^p$. If $\xi \sim \mathcal{N}(0, \sigma^2)$, then the equation:

$$Y = \beta_0 + \beta^t X + \xi$$

defines the statistical linear regression model:

$$\mathcal{M}_\Theta = \{p_{(\beta_0, \beta, \sigma^2)}(y|x), (\beta_0, \beta, \sigma^2) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}_+^*, x \in \mathbb{R}^p\}$$

with:

$$p_{(\beta_0, \beta, \sigma^2)}(y|x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y - \beta_0 - \beta^t x}{\sigma}\right)^2}$$

■

■ **Example 1.8 — Time dependent random variable.** By analogy, we can consider in the same framework the case where we have a stochastic process $(Y_x)_{x \in \Omega}$ indexed by Ω , a measurable set of $(\mathcal{X}, \mathcal{A})$. The random variable Y_x is then considered as the conditional random variable $Y|X = x$, and the same regression framework can be applied.

Let $Y(t)$, a time-indexed random variable with values in $\mathbb{R}$, and let $\xi \sim \mathcal{N}(0, \sigma^2)$. Let's consider the logistic model:

$$Y(t) = \frac{\theta_1}{1 + \theta_2 e^{-\theta_3 t}} (1 + \xi) .$$

Then:

$$\mathcal{M}_\Theta = \{p_{(\theta, \sigma^2)}(y|t), (\theta, \sigma^2) \in \mathbb{R}^3 \times \mathbb{R}_+^*\}$$

with $p_{(\theta, \sigma^2)}(y|t)$ la densité de $\mathcal{N}\left(\frac{\theta_1}{1 + \theta_2 e^{-\theta_3 t}}, \sigma^2 \left(\frac{\theta_1}{1 + \theta_2 e^{-\theta_3 t}}\right)^2\right)$

■

Assuming that the unknown law P^* of the random phenomenon studied belongs to a family of parametric laws $\mathcal{M}_\Theta$ raises two main questions that will be dealt with in the rest of this course:

- On one hand, based on the data, that is, the observations, how can we determine $\hat{\theta}$ such that $P_{\hat{\theta}} \in \mathcal{M}_{\Theta}$ is as close as possible to P^* ? That's the problem of **parameter estimation**. Not only will it be necessary to determine $\hat{\theta}$ from the data, but also to qualify the quality of the approximation, it will be characterized in particular by the uncertainty or confidence in $\hat{\theta}$.
- On the other hand, we can question the quality of our statistical model, i.e. our starting hypothesis. This can concern different situations, for example to verify if the postulated law P_{θ_0} is indeed equal to the unknown law P^* , or even to test if P^* really belongs to the parametric family of laws $\mathcal{M}_{\Theta}$. We talk of **hypothesis testing**.

2. Parameter Estimation

Let X be a random variable in $(\mathcal{X}, \mathcal{A})$, of unknown law P^* . Our objective is to identify P^* from observations of X or from an i.i.d. sample for the P^* law, $(X_1, \dots, X_N)$. This approach is greatly simplified if we assume that there is a statistical model $\mathcal{M}_\Theta = \{P_\theta, \theta \in \Theta\}$ such as $P^* \in \mathcal{M}_\Theta$. In such case, there exists $\theta^* \in \Theta$ (unknown), such that $P^* = P_{\theta^*}$. θ^* is said to be the true value of the parameter.

Note that we will generally expect θ^* to be unique, which will be ensured as soon as the model is identifiable.

Definition 2.1 — Identifiability. A parametric model $\mathcal{M}_\Theta = \{P_\theta, \theta \in \Theta\}$ is said identifiable if the map: $\theta \mapsto P_\theta$ is injective.

In the remainder of this chapter, we will consider observable random variables with values in $\mathcal{X} \subset \mathbb{R}^d$. The parametric statistical model considered $\mathcal{M}_\Theta$ will be identifiable, dominated by ν , a σ -finite measure (Lebesgue measure or count measure) and represented by a family of densities on $(\mathcal{X}, \mathcal{A})$: $\mathcal{M}_\Theta = \{p_\theta, \theta \in \Theta\}$, with Θ a Borel set of $\mathbb{R}^p$.

2.1 Fisher Information

The statistical model is the prism through which the data is analyzed. Before detailing the estimation methods in the next section, we introduce the Fisher information concept, a very useful notion that defines the information that an observable random variable can provide about the unknown parameter, on which the random variable depends through the statistical model.

Definition 2.2 — Regularity Conditions. We define the regularity conditions of a $\mathcal{M}_\Theta$ statistical model, conditions that will be useful to us later on.

(C1) Θ is an open set and the distributions in $\mathcal{M}_\Theta$ satisfy $p_\theta(x) > 0 \Leftrightarrow p_{\theta'}(x) > 0, \forall x \in \mathcal{X}, \forall \theta, \theta' \in \Theta$. It follows that all laws in $\mathcal{M}_\Theta$ have the same support, denoted $\mathcal{S}$.

(C2) $\forall \theta \in \Theta, \forall k, 1 \leq k \leq p$, we can differentiate under the integral sign:

$$\frac{\partial}{\partial \theta_k} \int_{\mathcal{S}} p_{\theta}(x) \nu(dx) = \int_{\mathcal{S}} \frac{\partial p_{\theta}(x)}{\partial \theta_k} \nu(dx) .$$

(C3) For almost every $x \in \mathcal{S}$, the map $\theta \mapsto p_{\theta}(x)$ The function is twice continuously differentiable on Θ .

(C4) $\forall \theta \in \Theta, \forall i, j, 1 \leq i, j \leq p$, we can differentiate under the integral sign:

$$\frac{\partial}{\partial \theta_i} \int_{\mathcal{S}} \frac{\partial p_{\theta}(x)}{\partial \theta_j} \nu(dx) = \int_{\mathcal{S}} \frac{\partial^2 p_{\theta}(x)}{\partial \theta_i \partial \theta_j} \nu(dx) .$$

■ **Remark 2.1** The differentiation under the integral sign is ensured if the derivative is locally dominated (Billingsley, 1995). For example, for condition (C2), let $\theta \in \Theta$ and $k, 1 \leq k \leq p$, there exists a neighborhood of θ and a map g , integrable over $(\mathcal{S}, \mathcal{A}, \nu)$, such that for almost every $x \in \mathcal{S}$, $\left| \frac{\partial p_{\theta}(x)}{\partial \theta_k} \right| \leq g(x)$. Finer regularity conditions are also possible (Spokoiny and Dickhaus, 2015). ■

Definition 2.3 — Likelihood, Score. Let X be a random variable, $X \sim p_{\theta^*}$, with $p_{\theta^*} \in \mathcal{M}_{\Theta}$ and θ^* unknown. Let $x \in \mathcal{X}$ be an observation of X . The map defined from Θ to $[0; 1]$: $\theta \mapsto p_{\theta}(x)$ is called likelihood of the parameter θ . We will denote: $\mathcal{L}(\theta; x)$.

If the regularity condition (C1) is satisfied and if for almost every $x \in \mathcal{S}$ ($\mathcal{S}$ support of $\mathcal{M}_{\Theta}$), $\theta \mapsto \ln(\mathcal{L}(\theta, x))$ is differentiable, then the score is the random vector defined by:

$$S_{\theta}(X) = \nabla_{\theta} \ln(p_{\theta}(X)) = \left(\frac{\partial \ln(p_{\theta}(X))}{\partial \theta_1}, \dots, \frac{\partial \ln(p_{\theta}(X))}{\partial \theta_p} \right)^T ,$$

that is to say the gradient of the log-likelihood function with respect to the parameter vector.

Proposition 2.1 If the regularity conditions (C1,C2) are satisfied, the score is a centered random vector:

$$\mathbb{E}_{\theta}(S_{\theta}(X)) = 0 .$$

Preuve : We recall that $\mathbb{E}_{\theta}$ denotes the expectation of the random variable with respect to the probability law of density p_{θ} . For all its components $[S_{\theta}(X)]_k = \frac{\partial \ln(p_{\theta}(X))}{\partial \theta_k}$, we successively have:

$$\begin{aligned} \mathbb{E}_{\theta}([S_{\theta}(x)]_k) &= \int_{\mathcal{S}} \frac{\partial \ln(p_{\theta}(x))}{\partial \theta_k} p_{\theta}(x) \nu(dx) \\ &= \int_{\mathcal{S}} \frac{\partial p_{\theta}(x)}{\partial \theta_k} \nu(dx) \\ &= \frac{\partial}{\partial \theta_k} \left(\int_{\mathcal{S}} p_{\theta}(x) \nu(dx) \right) \end{aligned}$$

where the last line is obtained from condition (C2).

As $\int_S p_\theta(x) \nu(dx) = 1$, we obtain the desired result.

Note that we can write $\int_S \frac{\partial \ln(p_\theta(x))}{\partial \theta_k} p_\theta(x) \nu(dx)$ since $\frac{\partial \ln(p_\theta(x))}{\partial \theta_k}$ is defined almost everywhere (with respect to P_θ). $\square$

The Fisher Information is defined from the score vector.

Definition 2.4 — Fisher Information. We call Fisher Information in θ the variance-covariance matrix of the score vector. It is denoted by $I(\theta)$:

$$I(\theta) = \mathbb{V}_\theta(S_\theta(X))$$

If condition (C2) is satisfied, we have directly by Proposition 2.1:

$$I(\theta) = \mathbb{E}_\theta[S_\theta(X)S_\theta(X)^T]$$

Proposition 2.2 If the regularity conditions (C1,C2,C3,C4) are verified, the Fisher Information is expressed as follows:

$$I(\theta) = -\mathbb{E}_\theta[\nabla_\theta [S_\theta(X)]^T]$$

that is to say for all $1 \leq i, j \leq p$:

$$I_{ij}(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2 \ln p_\theta(X)}{\partial \theta_i \partial \theta_j} \right]$$

Preuve : For all $1 \leq i, j \leq p$, we have successively:

$$\begin{aligned} I_{ij}(\theta) &= \int_S \frac{\partial \ln(p_\theta(x))}{\partial \theta_i} \frac{\partial \ln(p_\theta(x))}{\partial \theta_j} p_\theta(x) \nu(dx) \\ &= \int_S \frac{\partial \ln(p_\theta(x))}{\partial \theta_i} \frac{\partial p_\theta(x)}{\partial \theta_j} \nu(dx) \\ &= \int_S \frac{\partial}{\partial \theta_j} \left[\frac{\partial \ln(p_\theta(x))}{\partial \theta_i} p_\theta(x) \right] \nu(dx) - \int_S \frac{\partial^2 \ln(p_\theta(x))}{\partial \theta_i \partial \theta_j} p_\theta(x) \nu(dx) \\ &= \int_S \frac{\partial}{\partial \theta_j} \left[\frac{\partial(p_\theta(x))}{\partial \theta_i} \right] \nu(dx) - \int_S \frac{\partial^2 \ln(p_\theta(x))}{\partial \theta_i \partial \theta_j} p_\theta(x) \nu(dx) \\ &= \frac{\partial^2}{\partial \theta_j \partial \theta_i} \int_S p_\theta(x) \nu(dx) - \int_S \frac{\partial^2 \ln(p_\theta(x))}{\partial \theta_i \partial \theta_j} p_\theta(x) \nu(dx) \end{aligned}$$

by using the fact that $\theta \mapsto \ln p_\theta(x)$ is twice-differentiable (condition (C3)) and where the last equality is obtained from conditions (C4) and (C2).

And since $\int_S p_\theta(x) \nu(dx) = 1$, and we finally get the desired result. $\square$

■ **Example 2.2 — Gaussian law.** Let $X \sim \mathcal{N}(\mu, \sigma^2)$. We can check that the Gaussian model verifies the regularity conditions (C1,C2,C3,C4). Note that we consider directly σ^2 as the parameter, not σ .

We have:

$$\ln p_{(\mu, \sigma^2)}(x) = \ln(p|\mu, \sigma^2) = -\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \sigma^2 - \frac{1}{2\sigma^2}(x - \mu)^2,$$

$$\frac{\partial^2 \ln p(x|\mu, \sigma^2)}{\partial \mu^2} = -\frac{1}{\sigma^2}$$

thus

$$-\mathbb{E}_\theta \left(\frac{\partial^2 \ln p(X|\mu, \sigma^2)}{\partial \mu^2} \right) = \frac{1}{\sigma^2}.$$

Similarly:

$$\frac{\partial^2 \ln p(x|\mu, \sigma^2)}{(\partial \sigma^2)^2} = \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}(x - \mu)^2$$

and

$$-\mathbb{E}_\theta \left(\frac{\partial^2 \ln p(x|\mu, \sigma^2)}{(\partial \sigma^2)^2} \right) = \frac{1}{2\sigma^4}$$

by using the fact that $\mathbb{E}_\theta((x - \mu)^2) = \sigma^2$.

Finally:

$$\frac{\partial^2 \ln p(x|\mu, \sigma^2)}{\partial \mu \partial \sigma^2} = \frac{\mu - x}{\sigma^4}$$

which leads to

$$\mathbb{E}_\theta \left(\frac{\partial^2 \ln p(X|\mu, \sigma^2)}{\partial \mu \partial \sigma^2} \right) = 0.$$

and we obtain:

$$I(\mu, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$

■

An important property for Fisher information is its additivity for independent random variables.

Proposition 2.3 Let's consider two statistical models $\mathcal{M}_\Theta = \{p_\theta, \theta \in \Theta\}$ and $\mathcal{M}'_\Theta = \{q_\theta, \theta \in \Theta\}$ with same parameterization defined by Θ . Let X and Y be two independent random variables of laws p_θ in $\mathcal{M}_\Theta$ and q_θ in $\mathcal{M}'_\Theta$ respectively, then:

$$I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta), \quad \forall \theta \in \Theta,$$

where $I_{(X,Y)}(\theta)$ is the Fisher information in θ for the random variable (X, Y) of density $(x, y) \mapsto p_\theta(x)q_\theta(y)$.

Preuve : If we get back to the score, we have:

$$S_\theta(X, Y) = \nabla_\theta \ln(p_\theta(X, Y)) = \nabla_\theta \ln(p_\theta(X)q_\theta(Y)) = \nabla_\theta \ln(p_\theta(X)) + \nabla_\theta \ln(q_\theta(Y))$$

that is to say:

$$S_\theta(X, Y) = S_\theta(X) + S_\theta(Y)$$

and the to vectors $S_\theta(X)$ and $S_\theta(Y)$ are independent (images of two independent random vectors by measurable functions).

The variance of the sum of two independent random variables is the sum of the variances of each, hence the result. $\square$

An important application of this result concerns i.i.d. samples.

Proposition 2.4 Let $(X_1, \dots, X_N)$ be an i.i.d. sample for P_θ such that the Fisher information associated with each variable individually is $I(\theta)$. Then the Fisher information of the sample $(X_1, \dots, X_N)$ denoted $I_N(\theta)$ is:

$$I_N(\theta) = NI(\theta)$$

In a simple way, we can see that N independent observations provide N times more information than a single observation.

Finally, we introduce the concept of a regular model that will allow to simplify the presentation of results in the remainder of the document. Note however that the definition varies in the literature.

Definition 2.5 — Regular Statistical Model. A regular model is a dominated statistical model $\mathcal{M}_\Theta$ that verifies the regularity conditions (C1-C1-C3-C4) and for which the Fisher information matrix $I(\theta)$ is positive definite^a $\forall \theta \in \Theta$.

^aWe recall that a symmetric matrix A is positive definite if for all $x \neq 0$, $x^t A x > 0$.

2.2 Point Estimation

Point estimation aims at determining θ^* , or more precisely to approximate it as accurately as possible. To do this, an estimator is used.

Definition 2.6 — Estimator and estimation. Let $(X_1, \dots, X_N)$ be an i.i.d. sample for P_{θ^*} . An estimator is a statistic $T_N(X_1, \dots, X_N)$ with values in Θ whose objective is to approximate the unknown parameter θ^* . A realization $T_N(x_1, \dots, x_N)$ of T_N will be called an estimate of θ^* .

2.2.1 Substitution method, method of moments

Let X be a random variable in $(\mathcal{X}, \mathcal{A})$ and let $\mathcal{P}$ be the set of probability measures on $(\mathcal{X}, \mathcal{A})$. We consider a statistical model $\mathcal{M}_\Theta \subset \mathcal{P}$, $\Theta \subset \mathbb{R}^p$, and the map $\mathcal{G} : \mathcal{P} \rightarrow \mathbb{R}^m$ (usually $m = p$, the dimension of the parameter vector). Given $P_\theta \in \mathcal{M}_\Theta$, $\mathcal{G}(P_\theta)$ is a function of θ , and we denote $g(\theta) = \mathcal{G}(P_\theta)$. We then evaluate $\mathcal{G}$ at the empirical measure associated to observations $\hat{P}(x_1, \dots, x_N)$ and determine $\hat{g} = \mathcal{G}(\hat{P}(x_1, \dots, x_N))$. The estimate $\hat{\theta}$ is thus obtained by solving $g(\theta) = \hat{g}$ or by minimizing $\|g(\theta) - \hat{g}\|$. We sometimes speak of **substitution method**. It is a generalization of the method of moments which consists in choosing $\mathcal{G}$ as the moments of the probability distribution.

Method of moments

If the real-valued random variable $X : (\Omega, \mathcal{F}, \mathbb{P}) \rightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is in $L^p(\Omega, \mathcal{F}, \mathbb{P})$, we recall that it is in $L^q(\Omega, \mathcal{F}, \mathbb{P})$ for all $q < p$ and the q -th moment, $1 \leq q \leq p$, is defined by:

$$\mathbb{E}(X^q) = \int_{\Omega} X(\omega)^q \mathbb{P}(d\omega) .$$

If the distribution of X is in the statistical model $\mathcal{M}_{\Theta} = \{P_{\theta}, \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^p$, $X \sim P_{\theta}$, we also have:

$$\mathbb{E}(X^q) = \int_{\mathbb{R}} x^q P_{\theta}(dx) ,$$

and if the statistical model is dominated, X admits a probability density function p_{θ} , and we have:

$$\mathbb{E}(X^q) = \int_{\mathbb{R}} x^q p_{\theta}(x) dx .$$

We will use the following notation $\mathbb{E}_{\theta}(X)$ for $\mathbb{E}_{P_{\theta}}(X)$, that is to say the expectation of X with respect to P_{θ} or $\mathbb{E}(X)$ when $X \sim P_{\theta}$.

We choose the map $\mathcal{G}$ with values in $\mathbb{R}^p$ such that

$$\mathcal{G}(P_{\theta}) = (\mathbb{E}_{\theta}(X), \dots, \mathbb{E}_{\theta}(X^p))$$

Recall then (see section 1.4.1) that the application of the $\mathcal{G}$ function to the empirical measure for a sample $(x_1, \dots, x_N)$ gives:

$$\mathcal{G}(\hat{P}(x_1, \dots, x_N)) = \left(\frac{1}{N} \sum_{i=1}^N x_i, \dots, \frac{1}{N} \sum_{i=1}^N x_i^p \right) .$$

Estimation with the method of moments amounts to solve a system of p equations with p unknowns (since $\Theta \subset \mathbb{R}^p$), we look for $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_p)$ such that:

$$\begin{cases} \mathbb{E}_{\hat{\theta}}(X) &= \frac{1}{N} \sum_{i=1}^N x_i \\ \vdots & \\ \mathbb{E}_{\hat{\theta}}(X^p) &= \frac{1}{N} \sum_{i=1}^N x_i^p \end{cases}$$

■ **Example 2.3** Let $(x_1, \dots, x_N)$, the realization of an i.i.d. random sample for a Gaussian distribution of unknown parameters. We consider: $\mathcal{M}_{\Theta} = \{p_{(\mu, \sigma^2)}, (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*\}$ with

$$p_{(\mu, \sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

and we take $\mathcal{G}$:

$$p_{(\mu, \sigma^2)} \rightarrow \left(\int x p_{(\mu, \sigma^2)}(x) dx, \int x^2 p_{(\mu, \sigma^2)}(x) dx \right) .$$

For Gaussian distributions, we have:

$$\mathcal{G}(p_{(\mu, \sigma^2)}) = (\mu, \sigma^2 + \mu^2) .$$

With the empirical moments, we then directly deduce the estimates by the method of moments:

$$\begin{cases} \hat{\mu} = \frac{1}{N} \sum_{i=1}^N x_i \\ \hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N x_i^2 - \left(\frac{1}{N} \sum_{i=1}^N x_i \right)^2 \end{cases}$$

■

2.2.2 Maximum Likelihood

We recall the definition of the likelihood function.

Definition 2.7 — Likelihood. Let $\mathcal{M}_\Theta$ a parametric statistical model on $(\mathcal{X}, \mathcal{A})$, $\Theta \subset \mathbb{R}^p$, characterized by a family of probability density functions:

$$\mathcal{M}_\Theta = \{p_\theta, \theta \in \Theta\}.$$

Let X be a random variable, $X \sim p_{\theta^*}$ with $p_{\theta^*} \in \mathcal{M}_\Theta$. For an observed value x of the random variable X , the function defined on the parameter set Θ : $\theta \mapsto p_\theta(x)$ is called the likelihood (or likelihood function) of the parameter θ , we will denote $\mathcal{L}(\theta)$ ou $\mathcal{L}(\theta; x)$ to mean that the likelihood depends on the observed value x for which it is calculated.

For an i.i.d. sample of observations $(x_1, x_2, \dots, x_N)$, the likelihood can be written:

$$\mathcal{L}(\theta; x_1, x_2, \dots, x_N) = \prod_{i=1}^N p_\theta(x_i),$$

or

$$\mathcal{L}(\theta; x_1, x_2, \dots, x_N) = \prod_{i=1}^N p(x_i | \theta).$$

The likelihood of the parameter θ thus corresponds to its plausibility for a certain set of observed data.

Definition 2.8 — Maximum Likelihood Estimation. For an observed value or a sample of observations x , if $\hat{\theta}(x)$ maximizes the likelihood $\mathcal{L}(\theta; x)$ on Θ $\hat{\theta}(x)$ will be called maximum likelihood estimate and the associated estimator, maximum likelihood estimator (MLE).

We denote:

$$\hat{\theta}(x) = \arg \max_{\theta \in \Theta} \mathcal{L}(\theta; x),$$

■ Remark 2.4

- (i) The existence and uniqueness of the maximum are not necessarily assured.
- (ii) It may be easier to look for the maximum of $\ell(\theta) = \ln(\mathcal{L}(\theta))$ or the minimum of $-\ell(\theta)$.

(iii) If $\hat{\theta}$ is an interior point in Θ , a necessary condition for $\hat{\theta} = \arg \max_{\theta \in \Theta} \mathcal{L}(\theta)$ is that:

$$\nabla \mathcal{L}(\hat{\theta}) = \left(\frac{\partial \mathcal{L}(\hat{\theta})}{\partial \theta_1}, \dots, \frac{\partial \mathcal{L}(\hat{\theta})}{\partial \theta_N} \right)^T = 0 .$$

f $\mathcal{L}(\theta)$ is concave and if Θ is convex, the condition is sufficient. It is generally necessary to implement numerical optimization procedures to solve the maximization problem (gradient method, quasi-Newton method..., see for example Nocedal and Wright (2006)).

(iv) Finally, note that for a continuous random variable, the probability that $\hat{\theta} = \theta^*$, i.e., that the estimate is equal to the true value of the parameter, is null...

■

■ **Example 2.5** Let $X_1, \dots, X_N$ be an i.i.d. sample for $\mathcal{N}(\theta, 1)$, $\theta \in \mathbb{R}$, and let be $x = (x_1, \dots, x_N)$ the observed values. In x , the likelihood is given by:

$$\mathcal{L}(\theta; x) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i - \theta)^2} = \frac{1}{(2\pi)^{N/2}} e^{-\frac{1}{2} \sum_{i=1}^N (x_i - \theta)^2}$$

Our objective is to maximize $\mathcal{L}(\theta; x)$ for $\theta \in \mathbb{R}$, which is equivalent to maximize

$$l(\theta; x) = \ln(\mathcal{L}(\theta; x)) = -\frac{N}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^N (x_i - \theta)^2$$

or equivalent to minimize: $h(\theta) = \frac{1}{2} \sum_{i=1}^N (x_i - \theta)^2$.

$h'(\theta) = \sum_{i=1}^N (x_i - \theta) = N\theta - \sum_{i=1}^N x_i$. Thus $h'(\hat{\theta}) = 0$ gives $\hat{\theta} = \frac{1}{N} \sum_{i=1}^N x_i = \bar{x}$. Since h is convex on $\mathbb{R}$, the minimality necessary condition is sufficient, and $\bar{x}$ is the maximum likelihood estimate of θ .

Without the convexity argument, we can also simply draw a variation table, checking that h is strictly decreasing on $] -\infty; \bar{x}[$ and strictly increasing on $]\bar{x}; +\infty[$, it thus reaches its minimum in $\bar{x}$.

Another way to check the result without using the derivative is to use a well-known equality: ¹

$$\sum_{i=1}^N (x_i - \theta)^2 = \sum_{i=1}^N (x_i - \bar{x})^2 + \sum_{i=1}^N (\bar{x} - \theta)^2 .$$

we see that $\forall \theta \in \mathbb{R}$,

$$\sum_{i=1}^N (x_i - \theta)^2 \geq \sum_{i=1}^N (x_i - \bar{x})^2$$

¹This equality is sometimes called the generalized Koenig-Huygens formula: it is demonstrated in a simple way by noticing that

$$\sum_{i=1}^N (x_i - \theta)^2 = \sum_{i=1}^N (x_i - \bar{x} + \bar{x} - \theta)^2 = \sum_{i=1}^N (x_i - \bar{x})^2 + \sum_{i=1}^N (\bar{x} - \theta)^2 + 2 \sum_{i=1}^N (x_i - \bar{x})(\bar{x} - \theta)$$

Since:

$$\sum_{i=1}^N (x_i - \bar{x})(\bar{x} - \theta) = (\bar{x} - \theta) \left(\sum_{i=1}^N x_i - N\bar{x} \right) = 0$$

and we obtain the equality if and only if $\theta = \bar{x}$, which gives $h(\bar{x}) \leq h(\theta), \forall \theta \in \mathbb{R}$ and equivalently $\mathcal{L}(\bar{x}; x) \geq \mathcal{L}(\theta; x), \forall \theta \in \mathbb{R}$, with the equality only true when $\theta = \bar{x}$. The maximum likelihood estimator is therefore given by the empirical mean statistic $\bar{X}$. ■

We are sometimes interested not in estimating the parameter θ directly, but in a function of it, $g(\theta)$. Then we have a very interesting result.

Theorem 2.5 If $\hat{\theta}$ is the maximum likelihood estimator for θ , then $g(\hat{\theta})$ is the maximum likelihood estimator for $g(\theta)$.

Preuve : Let $x = (x_1, \dots, x_N)$, be a sample of observations for p_θ .

Let $\eta = g(\theta)$, and let's denote $\mathcal{L}^*(\eta, x)$, the likelihood function in η .

If g is a bijection (one-to-one mapping) from Θ onto $g(\Theta)$, then the result is direct:

$$\mathcal{L}^*(\eta; x) = \prod_{i=1}^N p(x_i | g^{-1}(\eta)) = \mathcal{L}(g^{-1}(\eta); x),$$

and thus:

$$\sup_{\eta} \mathcal{L}^*(\eta; x) = \sup_{\theta} \mathcal{L}(\theta; x).$$

Finally, if $\hat{\theta}$ maximizes $\mathcal{L}(\theta; x)$, $g(\hat{\theta})$ maximizes $\mathcal{L}^*(\eta; x)$.

If g is not one-to-one, the situation is more complex, since for a given η , there can be several θ such that $g(\theta) = \eta$ and the likelihood function $\mathcal{L}^*(\eta; x) = \prod_{i=1}^N p(x_i | g^{-1}(\eta))$ is not uniquely defined.

In such case, we can introduce the induced likelihood:

$$\mathcal{L}^*(\eta; x) = \sup_{\theta: g(\theta)=\eta} \mathcal{L}(g^{-1}(\eta); x),$$

and the result follows similarly as for the bijective case. □

■ **Example 2.6** Let $(x_1, \dots, x_N)$ be a sample of observations for the Bernoulli distribution $\mathcal{B}(p)$. Suppose that we have determined the maximum likelihood estimate $\hat{p}$ for p . Then the MLE for the standard deviation of the distribution is directly $\sqrt{\hat{p}(1 - \hat{p})}$ (knowing that the standard deviation is obtained as a function of p , $g(p) = \sqrt{p(1 - p)}$). ■

2.3 Properties of point estimators

We have given some methods to estimate the unknown parameters of a probability law from the data. Nevertheless, we still need to study the properties of these estimators in order to characterize them. This study will be carried out both in the non-asymptotic framework, i.e. for a finite dimension of the vector of observations, but also in the asymptotic framework, considering that the dimension of the vector of observations tends towards infinity.

In the rest of the section, we consider X to be a random variable in $(\mathcal{X}, \mathcal{A})$, $\mathcal{X} \subset \mathbb{R}^d$, of unknown distribution P_{θ^*} such that $P_{\theta^*} \in \mathcal{M}_\Theta$, a parametric statistical model represented by a family of density functions on $(\mathcal{X}, \mathcal{A})$: $\mathcal{M}_\Theta = \{p_\theta(x), \theta \in \Theta\}$, with $\Theta \subset \mathbb{R}^p$.

From an i.i.d. sample $X_1, \dots, X_N$ de loi P_{θ^*} , we consider the estimator $T_N = T(X_1, \dots, X_N)$ of θ .

2.3.1 Quadratic Risk and Bias

The first criterion considered in evaluating an estimator T is its risk, in particular the quadratic risk.

Definition 2.9 — Quadratic Risk. For T , estimator of the parameter θ^* , we call quadratic risk:

$$R(T, \theta^*) = \mathbb{E}_{\theta^*} (\|T - \theta^*\|^2)$$

We also talk of d'mean square error.

The concept of risk will allow for the comparison of two estimators, T and T' .

Definition 2.10 If for all $\theta \in \Theta$, we have

$$R(T, \theta) \leq R(T', \theta) ,$$

and if there exists $\theta' \in \Theta$ such that

$$R(T, \theta') < R(T', \theta')$$

then T is said to be a better estimator than T' . T' is thus said inadmissible and T is said admissible if there is no better estimator.

The quadratic risk of T decomposes into two terms, the squared bias and the estimator variance.

Definition 2.11 — Bias. If the estimator T is integrable, $\mathbb{E}_{\theta^*}(\|T\|) < +\infty$, $b_{\theta^*}(T) = \mathbb{E}_{\theta^*}(T) - \theta^*$ is called the bias of T . An estimator T of θ is said unbiased if $\forall \theta \in \Theta$:

$$b_{\theta}(T) = 0, \text{ that is to say } \mathbb{E}_{\theta}(T) = \theta.$$

This means that on average, for independent replications of the same experiment, the estimator neither overestimates nor underestimates the true parameter, whatever it is.

Proposition 2.6 — Bias-variance Decomposition. When T is square-integrable, i.e. $\mathbb{E}_{\theta^*}(\|T\|^2) < +\infty$, we can decompose the quadratic risk as follows:

$$\mathbb{E}_{\theta^*} (\|T - \theta^*\|^2) = \text{Tr}(\mathbb{V}_{\theta^*}^*(T)) + \|b_{\theta^*}(T)\|^2 ,$$

where $\mathbb{V}_{\theta^*}^*(T)$ is the variance-covariance matrix of T . We talk about bias-variance decomposition.

Preuve : Indeed, we have:

$$\begin{aligned} \mathbb{E}_{\theta^*} (\|T - \theta^*\|^2) &= \mathbb{E}_{\theta^*} (\|T - \theta^* - b_{\theta^*}(T) + b_{\theta^*}(T)\|^2) \\ &= \mathbb{E}_{\theta^*} (\|T - \theta^* - b_{\theta^*}(T)\|^2) + \|b_{\theta^*}(T)\|^2 \end{aligned}$$

since $b_{\theta^*}(T) = \mathbb{E}_{\theta^*}(T - \theta^*)$, then:

$$\begin{aligned}\mathbb{E}_{\theta^*}(\|T - \theta^*\|^2) &= \mathbb{E}_{\theta^*}[(T - \theta^* - b_{\theta^*}(T))^T(T - \theta^* - b_{\theta^*}(T))] + \|b_{\theta^*}(T)\|^2 \\ &= \text{Tr}(\mathbb{V}_{\theta^*}(T - \theta^*)) + \|b_{\theta^*}(T)\|^2 = \text{Tr}(\mathbb{V}_{\theta^*}(T)) + \|b_{\theta^*}(T)\|^2\end{aligned}$$

□

■ **Remark 2.7** If $\Theta \subset \mathbb{R}$, we simply:

$$\mathbb{E}_{\theta^*}((T - \theta^*)^2) = \mathbb{V}_{\theta^*}(T) + b_{\theta^*}(T)^2.$$

■

■ **Example 2.8 — (Continuation of example 2.3).**

Consider the Gaussian model, $\mathcal{M} = \{\mathcal{N}(\mu, \sigma^2), (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*\}$. For an i.i.d. sample, we have seen that by the method of moments we can construct the following estimators:

$$\begin{cases} \hat{\mu} = \frac{1}{N} \sum_{i=1}^N X_i := \bar{X} \\ \hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^N X_i^2 - \left(\frac{1}{N} \sum_{i=1}^N X_i\right)^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 := S^2 \end{cases}$$

Let's calculate the expectations and variances of the two estimators (denoting $\theta = (\mu, \sigma^2)$). For the estimator of the mean, $\hat{\mu} = \bar{X}$:

$$\begin{cases} \mathbb{E}_{\theta}(\bar{X}) = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{\theta}(X_i) = \mu \\ \mathbb{V}_{\theta}(\bar{X}) = \frac{1}{N} \sum_{i=1}^N \mathbb{V}_{\theta}(X_i) = \frac{\sigma^2}{N} \end{cases}$$

The estimator $\bar{X}$ is thus an unbiased estimator of the mean.

For the estimator of the variance $\hat{\sigma}^2 = S^2$:

$$\mathbb{E}_{\theta}(S^2) = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{\theta}(X_i^2) - \mathbb{E}_{\theta}(\bar{X}^2) = (\mu^2 + \sigma^2) - \left(\mathbb{V}_{\theta}(\bar{X}) + \mathbb{E}(\bar{X})^2\right) = \frac{N-1}{N} \sigma^2$$

S^2 is thus a biased estimator of the variance. We then sometimes introduce the unbiased estimator: $S'^2 = \frac{N}{N-1} S^2$.

Regarding the variance of the S^2 estimator, a longer calculation gives:

$$\mathbb{V}_{\theta}(S^2) = \frac{2(N-1)}{N^2} \sigma^4 \quad (2.1)$$

Finally, an application of Cochran's theorem (Lehmann and Casella, 2006) would show that for the family of Gaussian laws, $\bar{X}$ and S^2 are independent random variables. The variance covariance matrix is then given by:

$$\mathbb{V}_{\theta}(\bar{X}, S^2) = \begin{pmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2(N-1)}{N^2} \sigma^4 \end{pmatrix}$$

Note that the results obtained for $\mathbb{E}_\theta(\bar{X})$, $\mathbb{V}_\theta(\bar{X})$ and $\mathbb{E}_\theta(S^2)$ are true whatever the distribution, provided that the second moment exists.

For $\mathbb{V}_\theta(S^2)$ however, the general result is given by the following formula, if the fourth moment exists:

$$\mathbb{V}_\theta(S^2) = \frac{(N-1)}{N^3} \left((N-1)\mu_4 - (N-3)\sigma^4 \right)$$

where $\mu_4 = \mathbb{E}_\theta((X - \mu)^4)$ is the centered fourth moment. For the Gaussian law, we know that $\mu_4 = 3\sigma^4$, which gives (2.1).

For the Gaussian family, we can thus quantify the quadratic risks:

$$R(\bar{X}, \mu) = \frac{\sigma^2}{N} \text{ et } R(S^2, \sigma^2) = \frac{2N-1}{N^2} \sigma^4 .$$

Interestingly, we have directly that $R(S'^2, \sigma^2) = \frac{2}{N-1} \sigma^4$, and since $\frac{1}{N-1} > \frac{N-1}{N^2}$, we see that even if it is biased S^2 is a better estimator of the variance than the unbiased estimator S'^2 , regarding the quadratic risk. For a particular problem, it is therefore important to clearly identify the criterion one wishes to favour. Finally, it should be noted that when the sample size increases, the difference between the two estimators becomes very small. ■

2.3.2 Efficiency of Unbiased Estimators

For unbiased estimators, the bias-variance decomposition allows to directly compare variances to compare estimators.

Definition 2.12 — Efficiency. Let T, T' be two unbiased estimators of θ , $\theta \in \Theta$. The estimator T is said to be more efficient than T' if:

$$\mathbb{V}_\theta(T) \leq \mathbb{V}_\theta(T'), \forall \theta \in \Theta$$

and

$$\exists \theta' \in \Theta, \mathbb{V}_{\theta'}(T) < \mathbb{V}_{\theta'}(T') .$$

The unbiased estimator T is said to have minimum variance (or **best unbiased estimator**) if $\mathbb{V}_\theta(T) \leq \mathbb{V}_\theta(T')$ for all unbiased estimators T' and for all $\theta \in \Theta$.

Remember that for two symmetric matrices A and B , we denote $A \leq B$ if $B - A$ is positive semi-definite² and $A < B$ if $B - A$ is positive definite³).

In regular statistical models, there is a lower bound on the variance of unbiased estimators.

Theorem 2.7 — Cramér-Rao Bound. Let $\mathcal{M}_\Theta$, $\Theta \subset \mathbb{R}^p$, be a regular statistical model, dominated by the σ -finite measure ν , with support $\mathcal{S}$.

Let T be an unbiased estimator of θ for $\theta \in \Theta$, with T square-integrable, $\mathbb{E}_\theta \|T\|^2 < \infty$, and

² $\forall x, (x^t(B - A)x \geq 0$

³ $\forall x \neq 0, (x^t(B - A)x > 0$

such that one can differentiate its expectation under the integral sign, $\forall 1 \leq i, j \leq p$:

$$\frac{\partial \mathbb{E}_\theta(T_i)}{\partial \theta_j} = \int_S T_i(x) \frac{\partial p_\theta(x)}{\partial \theta_j} \nu(dx) .$$

Then we have :

$$\mathbb{V}_\theta(T) \geq I^{-1}(\theta) .$$

The quantity $I^{-1}(\theta)$ is called **Cramér-Rao bound** (or Cramér-Rao lower bound)

Preuve : Let's denote $S = S(X, \theta) = \nabla_\theta \ln p_\theta(X)(\theta)$ the score vector. We know that $\mathbb{E}_\theta(S) = 0$ (from condition C2) and $\mathbb{V}_\theta(S) = I(\theta)$ for all $\theta \in \Theta$.

Moreover, T being an unbiased estimator of θ , we have $\mathbb{E}_\theta(T) = \theta$, and thus $\forall 1 \leq i, j \leq p$, $\frac{\partial \mathbb{E}_\theta(T_i)}{\partial \theta_j} = \delta_{ij}$, with δ_{ij} the Kronecker symbol⁴. But thanks to the condition of differentiation under the integral sign,

$$\begin{aligned} \frac{\partial \mathbb{E}_\theta(T_i)}{\partial \theta_j} &= \int_S T_i(x) \frac{\partial p_\theta(x)}{\partial \theta_j} \nu(dx) \\ &= \int_S T_i(x) \frac{\frac{\partial p_\theta(x)}{\partial \theta_j}}{p_\theta(x)} p_\theta(x) \nu(dx) \\ &= \int_S T_i(x) \frac{\partial \ln p_\theta(x)}{\partial \theta_j} p_\theta(x) \nu(dx) \\ &= \mathbb{E}_\theta(T_i S_j) \end{aligned}$$

Using matrix notation, we have:

$$\mathbb{E}_\theta(T S^T) = \mathbb{E}_\theta(S T^T) = Id \quad (2.2)$$

Thanks to a writing trick, by decomposing the matrix variance⁵, and by using the symmetry of $I^{-1}(\theta)$ and the fact that $\mathbb{E}(S) = 0$, we finally get (without denoting the dependency on θ of I to simplify the writing):

$$\begin{aligned} \mathbb{V}_\theta(I^{-1}S - T) &= \mathbb{E} \left((I^{-1}S - T)(I^{-1}S - T)^T \right) - \mathbb{E} \left(I^{-1}S - T \right) \mathbb{E} \left(I^{-1}S - T \right)^T \\ &= \mathbb{E} \left(I^{-1}S S^T I^{-1} \right) - \mathbb{E} \left(I^{-1}S T^T \right) - \mathbb{E} \left(T S^T I^{-1} \right) + \mathbb{E} \left(T T^T \right) - \mathbb{E}(T) \mathbb{E}(T)^T \end{aligned}$$

I^{-1} can be taken out of the expectation, and by using Equation (2.2) and the fact that $\mathbb{E}(S S^T) = \mathbb{V}(S) = I(\theta)$, we finally get:

$$\mathbb{V}_\theta(I^{-1}S - T) = I^{-1} \mathbb{E}(S S^T) I^{-1} - I^{-1} \mathbb{E}(S T^T) - \mathbb{E}(T S^T) I^{-1} + \mathbb{V}(T) = \mathbb{V}(T) - I^{-1}$$

$\mathbb{V}(T) - I^{-1}(\theta)$ is thus expressed as a matrix of variance-covariance, therefore it is positive definite, and we can note: $\mathbb{V}(T) \geq I^{-1}(\theta)$. $\square$

⁴ $\delta_{ij} = 1$ if $i = j$, 0 otherwise.

⁵ If X is in $\mathbb{R}^d$, $\mathbb{V}(X) = \mathbb{E}((X - \mathbb{E}(X))(X - \mathbb{E}(X))^T) = \mathbb{E}(X X^T) - \mathbb{E}(X) \mathbb{E}(X)^T$

■ **Remark 2.9** We hypothesized that $\mathcal{M}_\theta$ was regular for the sake of brevity, but we see in the proof that it is sufficient that $\mathcal{M}_\theta$ verifies the conditions of regularity (C1,C2), and that $I(\theta)$ is invertible. ■

Corollary 2.8 Under the same assumptions as in 2.7 Theorem, let's now assume that the T estimator is such that $\mathbb{E}_\theta(T) = g(\theta)$. So, by adapting the proof, we get that:

$$\mathbb{V}(T) \geq \left(\frac{\partial g(\theta)}{\partial \theta} \right) \left(\frac{\partial g(\theta)}{\partial \theta} \right)^T I^{-1}(\theta) ,$$

where $\left(\frac{\partial g(\theta)}{\partial \theta} \right)$ is the Jacobian matrix, $\left(\frac{\partial g(\theta)}{\partial \theta} \right)_{ij} = \frac{\partial g_i(\theta)}{\partial \theta_j}$.

This result concerns biased estimators, but also cases where one is more interested in estimating $g(\theta)$ rather than θ directly.

The important consequence in practice of the Cramér-Rao theorem is that the larger the information, the smaller the minimum variance and potentially a more accurate estimator can be found.

Definition 2.13 — Efficient estimator. An unbiased estimator T whose variance-covariance matrix satisfies:

$$\mathbb{V}_\theta(T) = I^{-1}(\theta)$$

is called an efficient estimator.

■ **Remark 2.10** There is no guarantee of the existenc of an estimator with variance equal to the Cramér-Rao bound. ■

Corollary 2.9 If the estimator is built from an i.i.d. sample, $T(X_1, \dots, X_N)$, we then immediately get:

$$\mathbb{V}(T(X_1, \dots, X_N)) \geq I_N^{-1}(\theta) = \frac{1}{N} I^{-1}(\theta) ,$$

where $I_N(\theta)$ is the Fisher information matrix for the i.i.d. sample and $I(\theta)$ that of the elementary random variable.

■ **Example 2.11 — (Continuation of examples 2.2 and 2.8).** From Example 2.2, we know that for a Gaussian model, the elementary Fisher information (i.e. for the realization of a random variable) is given by:

$$I(\mu, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$

Moreover, we have found in Example 2.8, that for an i.i.d. Gaussian sample,

$$\mathbb{V}_{\mu, \sigma^2}(\bar{X}, S^2) = \begin{pmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2(N-1)}{N^2} \sigma^4 \end{pmatrix}$$

Since $I(\theta)$ and $\mathbb{V}_{\theta}(\bar{X}, S^2)$ are diagonal, matrix inequality translates into inequality on the diagonal terms

For the estimator of the mean, we therefore have:

$$\mathbb{V}_{(\mu, \sigma^2)}(\bar{X}) = \frac{I_{(1,1)}^{-1}(\mu, \sigma^2)}{N} = \frac{1}{N} \sigma^2 ,$$

$\bar{X}$ is an efficient estimator of the mean.

For the variance estimator S^2 , we have

$$\mathbb{V}_{(\mu, \sigma^2)}(S^2) = \frac{2(N-1)}{N^2} \sigma^4 < \frac{I_{(2,2)}^{-1}(\mu, \sigma^2)}{N} .$$

But we recall that the estimator S^2 is biased for the variance, so we are not in the framework of Theorem 2.7 ! Note that in the same way, we have for the unbiased estimator $(\bar{X}, S'^2)$:

$$\mathbb{V}_{\mu, \sigma^2}(\bar{X}, S'^2) = \begin{pmatrix} \frac{\sigma^2}{N} & 0 \\ 0 & \frac{2}{N-1} \sigma^4 \end{pmatrix}$$

Considering the term corresponding to the variance:

$$\mathbb{V}_{(\mu, \sigma^2)}(S'^2) = \frac{2}{N-1} \sigma^4 > \frac{I_{(2,2)}^{-1}(\mu, \sigma^2)}{N} = \frac{2}{N} \sigma^4 .$$

S'^2 is not an efficient estimator of the variance. However, it could be shown to be the unbiased estimator of minimal variance, because there exists no unbiased efficient estimator for the variance. ■

In many cases, we will be interested in the properties of an estimator not for a fixed sample size of N , but in the limit when this size tends to infinity. The limit theorems in probability (the law of the large numbers, the central limit theorem) will often make it possible to infer properties of the estimator in this infinite limit. We will talk about asymptotic properties. Before looking in detail at the asymptotic properties of estimators, we recall several classical theorems on the convergences of random variable sequences that will prove useful in practice.

2.4 Limit theorems

We first recall the two fundamental theorems for the sequences of i.i.d. real-valued random variables seen in the probability course (Billingsley, 1995).

Theorem 2.10 — Strong Law of Large Number. Let $(X_i)_{i \in \mathbb{N}^*}$ be a sequence of i.i.d. real-valued random variables such that $\mathbb{E}(|X_1|) < +\infty$. We thus have:

$$\bar{X}_N \xrightarrow{p.s.} \mathbb{E}(X_1) ,$$

where we denoted $\bar{X}_N = \frac{1}{N} \sum_{i=1}^N X_i$.

Theorem 2.11 — Central Limit Theorem. Let $(X_i)_{i \in \mathbb{N}^*}$ be a sequence of i.i.d. real-valued random variables such that $0 < \mathbb{V}(X_1) < +\infty$. We thus have:

$$\sqrt{N} \left(\frac{\bar{X}_N - \mathbb{E}(X_1)}{\sqrt{\mathbb{V}(X_1)}} \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

or in an equivalent manner:

$$\sqrt{N} (\bar{X}_N - \mathbb{E}(X_1)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbb{V}(X_1)) . \quad (2.3)$$

■ **Remark 2.12** Note that in this second form (2.3), the Central Limit Theorem remains true for sequences of i.i.d. random vectors in $\mathbb{R}^d$, with finite expectation and invertible variance-covariance matrix. ■

We then recall some classical properties on the convergences of sequences of random variables. For the definition of the different types of convergence, we refer to (Billingsley, 1995)

Theorem 2.12 — Slutsky's theorem. Soient deux suites de variables aléatoires réelles Let Y_N and Z_N be two sequences of real-valued random variables, Y a random variable and $c \in \mathbb{R}$, such that:

$$Y_N \xrightarrow{\mathcal{L}} Y \text{ et } Z_N \xrightarrow{\mathcal{L}} c.$$

Then

$$Y_N + Z_N \xrightarrow{\mathcal{L}} Y + c \text{ et } Y_N Z_N \xrightarrow{\mathcal{L}} cY .$$

More generally, if f is continuous, $f(X_N, Y_N) \xrightarrow{\mathcal{L}} f(X, c)$.

Proposition 2.13 Let Y_N be a sequence of random variables with values in $\mathbb{R}^d$ and $c \in \mathbb{R}^d$ such that:

$$Y_N \xrightarrow{\mathcal{L}} c$$

then:

$$Y_N \xrightarrow{\mathbb{P}} c .$$

The reciprocal is of course always true, even if c is a non-degenerate random variable.

Proposition 2.14 Let Y_N be a sequence of random variables with values in $\mathbb{R}^d$. If there exists a random variable Z , a real-valued sequence $(a_N)_N$ such that $\lim_{N \rightarrow +\infty} a_N = +\infty$, and $c \in \mathbb{R}^d$ verifying:

$$a_N(Y_N - c) \xrightarrow{\mathcal{L}} Z$$

then:

$$Y_N \xrightarrow{\mathbb{P}} c.$$

The continuity theorem below makes it possible to transport limit laws by a function.

Theorem 2.15 — Continuity Theorem. Let Y_N be a sequence of random variables with values in $\mathbb{R}^d$ and h a continuous function on $\mathbb{R}^d$. Then:

- (i) $Y_N \xrightarrow{\mathcal{L}} Y \Rightarrow h(Y_N) \xrightarrow{\mathcal{L}} h(Y)$
- (ii) $Y_N \xrightarrow{\mathbb{P}} Y \Rightarrow h(Y_N) \xrightarrow{\mathbb{P}} h(Y)$
- (iii) $Y_N \xrightarrow{p.s.} Y \Rightarrow h(Y_N) \xrightarrow{p.s.} h(Y)$

Finally, the delta method allows in particular to extrapolate the results of the central limit theorem to transforms of random variables.

Theorem 2.16 — Delta Method. Let Y_N be a sequence of random variables with values in $\mathbb{R}$, $m \in \mathbb{R}$, and Z a random variable such that:

$$\sqrt{N}(Y_N - m) \xrightarrow{\mathcal{L}} Z$$

If $h : \mathbb{R} \rightarrow \mathbb{R}$ is of class $\mathcal{C}^1$ and $h'(m) \neq 0$, then:

$$\sqrt{N} \left(\frac{h(Y_N) - h(m)}{h'(m)} \right) \xrightarrow{\mathcal{L}} Z$$

Preuve : Let's consider the following function:

$$\psi(x) = \begin{cases} \frac{h(x) - h(m)}{x - m} & \text{si } x \neq m \\ h'(m) & \text{si } x = m. \end{cases}$$

It is continuous, since h is $\mathcal{C}^1$. Moreover, we know that $Y_N \xrightarrow{\mathbb{P}} m$ thanks to Proposition 2.14. Applying the continuity theorem, we have: $\psi(Y_N) \xrightarrow{\mathbb{P}} \psi(m) = h'(m)$. Then we write:

$$\sqrt{N} \left(\frac{h(Y_N) - h(m)}{h'(m)} \right) = \sqrt{N}(Y_N - m) \frac{\psi(Y_N)}{h'(m)}$$

Since $\sqrt{N}(Y_N - m) \xrightarrow{\mathcal{L}} Z$ and $\psi(Y_N) \xrightarrow{\mathbb{P}} h'(m)$, the application of Slutsky's theorem leads to the conclusion. $\square$

■ **Remark 2.13** To remember this result (this is not a rigorous demonstration!), we can apply

Taylor's formula to $h(Y_N)$ in m :

$$h(Y_N) \approx h(m) + h'(m) (Y_N - m)$$

and

$$\frac{h(Y_N) - h(m)}{h'(m)} \approx (Y_N - m) .$$

In the case where $h'(m) = 0$ but $h''(m) \neq 0$, we still have a limit law but we have to look for the term of order 2 and the limit law is proportional to that of Z^2 . ■

2.5 Asymptotic properties of estimators

First, we give the extension in the asymptotic framework of the notion of unbiased estimator.

We consider in this section $T_N = T(X_1, \dots, X_N)$, an estimator of θ for $X_1, \dots, X_N$ i.i.d. for an unknown law in $\mathcal{M}_\Theta$.

Definition 2.14 — Asymptotically unbiased. If for all $\theta \in \Theta$, $\lim_{N \rightarrow +\infty} b_\theta(T_N) = 0$, the estimator T is said asymptotically unbiased.

■ **Example 2.14** In the example ??, we saw that the variance estimator for an i.i.d. sample (Gaussian, but also more generally) of size N , S_N^2 was biased, with bias $b_{\sigma^2}(S_N^2) = \frac{\sigma^2}{N}$. We can see that the estimator is on the other hand asymptotically unbiased. ■

This is a probabilistic convergence in expectation, $\lim_{N \rightarrow +\infty} \mathbb{E}_\theta(T_N) = \theta$. It remains less strong than the convergence in probability and the almost-sure convergence, which define the consistency of the estimator.

Definition 2.15 — Consistency / Convergence.

T_N is said consistent (sometimes weakly consistent) or convergent if, for all $\theta \in \Theta$, it converges in probability to θ :

$$P_\theta(\|T_N - \theta\| > \epsilon) \xrightarrow{N \rightarrow +\infty} 0, \forall \epsilon > 0 .$$

We denote $T_N \xrightarrow{\mathbb{P}} \theta$.

T_N is said strongly consistent or strongly convergent if, for all $\theta \in \Theta$, T_N converges almost surely to θ :

$$P_\theta\left(\lim_{N \rightarrow +\infty} T_N = \theta\right) = 1 .$$

We denote $T_N \xrightarrow{p.s.} \theta$.

■ **Remark 2.15** A strongly consistent estimator is of course consistent, since the almost sure convergence implies the convergence in probability, cf. (Billingsley, 1995). ■

We have the following two results.

Proposition 2.17 If T_N is asymptotically unbiased and if $Tr(\mathbb{V}_\theta(T_N)) \xrightarrow{N \rightarrow +\infty} 0$, then:

- (i) $R(T_N, \theta) = \mathbb{E}(\|T_N - \theta\|^2) \xrightarrow{N \rightarrow +\infty} 0$. T_N converges in quadratic mean to θ ;
- (ii) T_N est consistant.

Preuve : The first point is simply the result of the bias-variance decomposition. The second point is a result of convergence theorems for the sequences of random variables: convergence in quadratic mean implies convergence in probability. $\square$

■ **Example 2.16** We saw that the estimator of the mean $\bar{X}$ was unbiased and the variance estimator S^2 was asymptotically unbiased. Furthermore, we also showed in the example ?? that their variances $\xrightarrow{N \rightarrow +\infty} 0$. Both estimators are therefore consistent. ■

Proposition 2.18 If T_N is consistent and if there exist $C_1, C_2 > 0$, such that $\forall N: \|b_\theta(T_N)\|^2 \leq C_1$ and $Tr(\mathbb{V}_\theta(T_N)) \leq C_2$, then T_N is asymptotically unbiased.

Preuve : Since the expectation of a random vector is the vector of the expectations of each component, we can get back to the case where $\theta \in \mathbb{R}$.

Let $\epsilon > 0$:

$$\begin{aligned} |\mathbb{E}_\theta(T_N - \theta)| &\leq \int_{x \in \mathcal{X}^N, |T_N(x) - \theta| \leq \epsilon} |T_N(x) - \theta| P_\theta(dx) + \int_{x \in \mathcal{X}^N, |T_N(x) - \theta| > \epsilon} |T_N(x) - \theta| P_\theta(dx) \\ &\leq \epsilon + \int_{x \in \mathcal{X}^N} \mathbb{I}_{\{|T_N(x) - \theta| > \epsilon\}}(x) |T_N(x) - \theta| P_\theta(dx) \\ &\leq \epsilon + \left(\int_{x \in \mathcal{X}^N} \mathbb{I}_{\{|T_N(x) - \theta| > \epsilon\}}^2(x) P_\theta(dx) \int_{x \in \mathcal{X}^N} |T_N(x) - \theta|^2 P_\theta(dx) \right)^{1/2} \\ &\leq \epsilon + (P_\theta(|T_N(x) - \theta| > \epsilon))^{1/2} \left(\|b_\theta(T_N)\|^2 + \mathbb{V}_\theta(T_N) \right)^{1/2} \\ &\leq \epsilon + \sqrt{C_1 + C_2} (P_\theta(|T_N(x) - \theta| > \epsilon))^{1/2} \end{aligned}$$

Therefore $\lim_{N \rightarrow +\infty} |\mathbb{E}_\theta(T_N - \theta)| \leq \epsilon$. That being true for all $\epsilon > 0$, we finally get $\lim_{N \rightarrow +\infty} |\mathbb{E}_\theta(T_N - \theta)| = 0$, the estimator is asymptotically unbiased. $\square$

We are now interested in the rate of convergence of consistent estimators.

Definition 2.16 — Rate of convergence of an estimator. . Let T_N be a consistent estimator of θ . Let's suppose that there exist a real-valued and positive sequence (a_N) such that:

$$\lim_{N \rightarrow \infty} a_N = +\infty \text{ et } a_N(T_N - \theta) \xrightarrow{\mathcal{L}} Z,$$

where Z is a non-degenerate random variable (that is to say that it is not almost surely equal to a constant). The sequence (a_N) is called the convergence rate of T_N .

In parametric models, convergence rate usually vary between $\sqrt{N}$ and N according to the regularity of the statistical model, the convergence in N being faster than the convergence in

$\sqrt{N}$.

■ **Example 2.17** We consider the Gaussian family of variance 1: $\mathcal{M}_\mu = \{\mathcal{N}(\mu, 1), \mu \in \mathbb{R}\}$. We saw in example 2.8 that $\bar{X}_N$ is an unbiased estimator of parameter μ , and we know that the sum of independent Gaussian variables is a Gaussian variable of mean the sum of means and variance the sum of variances⁶. Therefore $\bar{X}_N \sim \mathcal{N}(\mu, 1/N)$, and finally:

$$\sqrt{N}(\bar{X}_N - \mu) \sim \mathcal{N}(0, 1)$$

■

It remains for us to consider the asymptotic efficiency of an estimator. This notion is only valid for estimators that are asymptotically unbiased (i.e. for most estimators that are consistent thanks the proposition 2.18).

Thanks to the central limit theorem, it often boils down to estimators whose limit law is a Gaussian.

Definition 2.17 An estimator T_N of θ is asymptotically normal if there exist $\Sigma(\theta)$, a positive definite symmetric matrix of order p such that

$$\sqrt{N}(T_N - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma(\theta)) .$$

The variance-covariance matrix Σ does not depend on N . It will be called asymptotic variance associated with the normal estimator T_N (beware, it is not the variance limit of the T_N estimator). We have the following useful result.

Proposition 2.19 If an estimator is asymptotically normal, then it is consistent.

Preuve : If T_N is an asymptotically normal estimator of θ , then we can apply Proposition 2.14, and we have directly that $T_N \xrightarrow{\mathbb{P}} \theta$. □

For asymptotically normal estimators, it is natural to compare the associated asymptotic variances.

Definition 2.18 Let T_N and T'_N be estimators of θ which are asymptotically normal, of associated asymptotic variances Σ and Σ' . Then T_N is said asymptotically more efficient than T'_N if:

$$\Sigma(\theta) \leq \Sigma'(\theta) , \forall \theta \in \Theta \text{ et } \exists \theta' \in \Theta, \Sigma(\theta') < \Sigma'(\theta') .$$

For two symmetric matrices A, B we recall that we denote $A < B$ when $B - A$ is positive definite, $A \leq B$ when $B - A$ positive semi-definite.

Definition 2.19 An estimator is asymptotically efficient when it is asymptotically normal and

⁶This result can easily be shown by the characteristic functions, cf. Billingsley (1995)

when its associated asymptotic variance matrix

$$\Sigma(\theta) = I^{-1}(\theta) .$$

■ **Remark 2.18** This notion comes from the fact that if the estimator was efficient, we would have for all N :

$$\mathbb{V}_\theta(T_N) = \frac{1}{N} I(\theta)^{-1} ,$$

or équivalent:

$$\mathbb{V}_\theta \left(\sqrt{N} (T_N - \theta) \right) = I(\theta)^{-1} .$$

The extension of the concept to the asymptotic case is not obtained directly by the limit on variance. For an asymptotically efficient estimator, we consider that $\sqrt{N} (T_N - \theta)$ converges in distribution towards a normal distribution of variance $I(\theta)^{-1}$, which is différent, and does not necessarily imply the convergence of $\mathbb{V}_\theta \left(\sqrt{N} (T_N - \theta) \right)$ to $I(\theta)^{-1}$! ■

2.6 Asymptotic Properties of the Maximum Likelihood Estimator

We first define a very classical notion of divergence between laws of probability.

Definition 2.20 — Kullback-Leibler Divergence. We consider a statistical model $\mathcal{M}_\theta$, dominated by the σ -finite measure ν , and described by the family of densities:

$$\mathcal{M}_\theta = \{p_\theta, \theta \in \Theta\} .$$

$\forall \theta, \theta' \in \Theta$, is a measure of how one probability distribution $p_{\theta'}$ is different from a second, reference probability distribution p_θ and is given by:

$$\begin{aligned} K(\theta, \theta') &= \mathbb{E}_\theta \left[\ln \left(\frac{p_\theta(X)}{p_{\theta'}(X)} \right) \mathbb{I}_{\{p_{\theta'}(X) > 0\}} \right] \\ &= \int_{\{x, p_{\theta'}(x) > 0\}} \ln \left(\frac{p_\theta(x)}{p_{\theta'}(x)} \right) p_\theta(x) \nu(dx) . \end{aligned}$$

K is not a distance since it is not symmetric. On the other hand, one can easily verify (using Jensen's inequality) that $K(\theta, \theta') \geq 0$ and that $K(\theta, \theta) = 0$.

We then have the consistency result for the maximum likelihood estimator.

Theorem 2.20 — Consistency of the maximum likelihood estimator. Let $(X_i)_{i \in \mathbb{N}^*}$ be a sequence of i.i.d. random variables aléatoires for P_{θ^*} , with $P_{\theta^*} \in \mathcal{M}_\theta$, a dominated statistical model. We denote :

$$M_N(\theta) = \frac{1}{N} \sum_{i=1}^N \ln \left(\frac{p_\theta(X_i)}{p_{\theta^*}(X_i)} \right)$$

and $M(\theta) = -K(\theta^*, \theta)$. Let's suppose:

$$\sup_{\theta \in \Theta} |M_N(\theta) - M(\theta)| \xrightarrow{\mathbb{P}} 0$$

and that for all $\epsilon > 0$:

$$\sup_{\theta \in \Theta, |\theta - \theta^*| \geq \epsilon} M(\theta) < M(\theta^*)$$

Then the maximum likelihood estimator is weakly consistent:

$$\hat{\theta}_N \xrightarrow{\mathbb{P}} \theta^*$$

Preuve : By the definition of the Kullback-Leibler divergence, $M(\theta^*) = -K(\theta^*, \theta^*) = 0$ and $-M(\theta^*) = K(\theta^*, \theta) \geq 0, \forall \theta$. On the other hand, the definition of the maximum likelihood estimator gives us

$$M_N(\hat{\theta}_N) \geq M_N(\theta^*) .$$

We then have the following majoration:

$$\begin{aligned} 0 \leq M(\theta^*) - M(\hat{\theta}_N) &= M(\theta^*) - M_N(\theta^*) + M_N(\theta^*) - M_N(\hat{\theta}_N) + M_N(\hat{\theta}_N) - M(\hat{\theta}_N) \\ &\leq M(\theta^*) - M_N(\theta^*) + M_N(\hat{\theta}_N) - M(\hat{\theta}_N) \\ &\leq 2 \sup_{\theta \in \Theta} |M_N(\theta) - M(\theta)| \end{aligned}$$

Using the second assumption, for all $\epsilon > 0$, there's a $\delta > 0$ such that

$$|\theta - \theta^*| \geq \epsilon \Rightarrow M(\theta) < M(\theta^*) - \delta$$

The logical implication is the inclusion of the corresponding events:

$$\mathbb{P}_{\theta^*} \{ |\hat{\theta}_N - \theta^*| \geq \epsilon \} \leq \mathbb{P}_{\theta^*} \{ M(\theta^*) - M(\hat{\theta}_N) \geq \delta \} \leq \mathbb{P}_{\theta^*} \{ 2 \sup_{\theta \in \Theta} |M_N(\theta) - M(\theta)| \geq \delta \}$$

Now, by the first hypothesis of the theorem and by definition of convergence in probability, we have that:

$$\forall \delta > 0, \lim_{N \rightarrow \infty} \mathbb{P}_{\theta^*} \{ \sup_{\theta \in \Theta} |M_N(\theta) - M(\theta)| \geq \frac{\delta}{2} \} = 0.$$

Therefore, $\forall \epsilon > 0, \lim_{N \rightarrow \infty} \mathbb{P}_{\theta^*} \{ |\theta - \theta^*| \geq \epsilon \} = 0$, which concludes the proof. $\square$

Under regularity conditions, the asymptotic normality of the maximum likelihood estimator can be established (see e.g. Lehmann and Casella (2006) for demonstration).

Theorem 2.21 Let $\mathcal{M}_\Theta$ be a regular statistical model for random variables with values in $\mathbb{R}^d$. Let $(X_i)_{i \in \mathbb{N}^*}$ be a sequence of i.i.d. random variables aléatoires with density $p_{\theta^*} \in \mathcal{M}_\Theta$, and with Fisher information $I(\theta^*)$, and with $I(\theta)$ is well-defined and positive definite on a neighborhood of θ^* . Then, if the maximum likelihood estimator $\hat{\theta}_N^{MV}$ is a consistent estimator of θ^* , we have:

$$\sqrt{N}(\hat{\theta}_N^{MV} - \theta^*) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I(\theta^*)^{-1}) .$$

Under these conditions, the maximum likelihood estimator is asymptotically normal and asymptotically efficient.

2.7 Bayesian Estimation

2.7.1 Bayesian Model and Conditional Probability Distributions

Let X be a random variable in $(\mathcal{X}, \mathcal{A})$, of unknown law $P_\theta \in \mathcal{M}_\Theta$, $\Theta \subset \mathbb{R}^p$. Since θ is unknown, it can be considered as a random variable in $(\Theta, \mathcal{F}_\Theta)$, where $\mathcal{F}_\Theta$ is Θ (generally $\mathcal{F}_\Theta = \mathcal{B}(\Theta)$, Borel sets of Θ) if we give ourselves a probability law reflecting a priori knowledge on this parameter. This is the basis of the Bayesian approach. To clarify things, we introduce ϑ , a random variable in $\Theta, \mathcal{B}(\Theta)$ and we reserve the θ notation for the realizations of the random variable.

Definition 2.21 A Bayesian statistical model is a parametric statistical model $\mathcal{M}_\Theta$ and a law of probability Π_ϑ on $(\Theta, \mathcal{F}_\Theta)$, called prior distribution (or a priori distribution).

Proposition 2.22 If $(\mathcal{M}_\Theta, \pi_\vartheta)$ is a Bayesian statistical model and if $\forall A \in \mathcal{A}$, the mapping $\theta \mapsto P_\theta(A)$ is $\mathcal{F}_\Theta$ -measurable, then the mapping $\nu : \Theta \times \mathcal{A} \rightarrow [0, 1]$, $\nu(\theta, A) = P_\theta(A)$ is a transition kernel.

The transition kernel thus defined allows considering P_θ as a conditional distribution:

$$\forall \theta \in \Theta, \forall A \in \mathcal{A} : \nu(\theta, A) = P_\theta(A) = P_{X|\vartheta=\theta}(A)$$

by considering the joint distribution $\forall (E, A) \in (\mathcal{F}_\Theta, \mathcal{A})$:

$$P_{(\vartheta, X)}(E, A) = \int_E P_\theta(A) \Pi_\vartheta(d\theta) . \quad (2.4)$$

Preuve : The result is immediate by from the definition of a transition kernel (Definition 1.4):

(i) $\forall \theta \in \Theta, \nu(\theta, \cdot) = P_\theta$ is a probability law on $(\mathcal{X}, \mathcal{A})$.

(ii) $\forall A \in \mathcal{A}$, the mapping $\nu(\cdot, A)$ is $\mathcal{F}_\Theta$ -measurable.

Equation (2.4) of the joint distribution allows to define ν as a conditional distribution of X given ϑ (see section 1.2.3). $\square$

■ **Remark 2.19** If $\mathcal{M}_\Theta$ is dominated by a measure μ σ -finite and if Π_ϑ admits a density function π , we have owing to (2.4) that the joint density of (ϑ, X) admits a density function:

$$p_{(\vartheta, X)}(\theta, x) = p_\theta(x)\pi(\theta),$$

p_θ is the probability density function of the conditional distribution $P_{X|\vartheta=\theta}$ and we denote:

$$p_\theta(x) = p_{X|\vartheta=\theta}(x) = p(x|\theta) .$$

■ **Remark 2.20** The condition requiring that $\forall A \in \mathcal{A}$, the mapping $\theta \mapsto P_\theta(A)$ should be $\mathcal{F}_\Theta$ -measurable is easy to ensure: for example, in the case of a model dominated by μ , if for almost every x , $\theta \mapsto p_\theta(x)$ is continuous. The condition is therefore satisfied in particular for a regular model. ■

2.7.2 Bayesian Method

For the rest of this section, we therefore consider a Bayesian model, whose statistical model is dominated and the prior admits a density, and for which the application condition of the proposition 2.22 is also verified. These assumptions then allow us to manipulate conditional densities in a classical way. The model is denoted: $(\{p_\theta, \theta \in \Theta\}, \pi)$. Under these conditions, we recall the Bayes formula, which is central to Bayesian estimation.

Theorem 2.23 — Bayes Formula. Let $x \in \mathcal{X}$. If $p_X(x) \neq 0$, then we can define the conditional density of the distribution $P_{\vartheta|X=x}$, and we have:

$$p(\theta|x) = \frac{p_{(\vartheta, X)}(\theta, x)}{p_X(x)} = \frac{p(x|\theta)\pi(\theta)}{p_X(x)} .$$

For x given, we have by the formula of marginalization:

$$p_X(x) = \int_{\theta \in \Theta} p(x|\theta)\pi(\theta) d\theta ,$$

$p_X(x)$ can be seen as a normalization constant (since $p(\theta|x)$ is a density, its integral is equal to 1). We then often simply denote:

$$p(\theta|x) \propto p(x|\theta)\pi(\theta) \tag{2.5}$$

Equation (2.5) allows us to define the framework of Bayesian estimation.

Definition 2.22 — Bayesian Estimation. Let $(\{p_\theta, \theta \in \Theta\}, \pi)$ be a Bayesian model. Let x be an observation for a random variable for a distribution of the Bayesian model. Then Bayesian estimation consists in determining the posterior distribution $p(\theta|x)$, from the prior distribution $\pi(\theta)$ and the likelihood $\mathcal{L}(\theta; x) = p(x|\theta)$ en x .

■ **Example 2.21 — Bernoulli model with uniform prior.** Let $X_1, \dots, X_N$ be an i.i.d. sample of random variables following a Bernoulli distribution $B(\theta)$, $\theta \in [0, 1]$, and let $x = (x_1, \dots, x_N)$ the corresponding sample of observations. Considering a prior distribution on the parameter θ amounts to consider that θ is the realization of a random variable ϑ . We suppose here that $\vartheta \sim \mathcal{U}([0, 1])$.

We recall that if $X_i \sim B(\theta)$, $P_{X_i}(1) = \theta$ and $P_{X_i}(0) = 1 - \theta$. The density of the law (with respect to the count measure on $\{0; 1\}$), can thus be written concisely: $p(x_i|\theta) = \theta^{x_i}(1 - \theta)^{1-x_i}$.

Et donc pour l'échantillon x , on a:

$$p(x|\theta) = \prod_{i=1}^N \theta^{x_i} (1 - \theta)^{1-x_i} .$$

Let S denote the statistics $S(X_1, \dots, X_N) = \sum_{i=1}^N X_i$ and $S(x)$, its realization in x , we have:

$$p(x|\theta) = \theta^{S(x)} (1 - \theta)^{N-S(x)} .$$

By Bayes formula, we get:

$$p(\theta|x) \propto p(x|\theta)p(\theta) = \theta^{S(x)} (1 - \theta)^{N-S(x)}$$

We recall that the density of the bêta distribution $B(a, b)$ is given by⁷:

$$p_{(a,b)}(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1 - \theta)^{b-1} .$$

$p(\theta|x)$ is thus proportional to the density of the bêta distribution $B(S(x) + 1, N - S(x) + 1)$. But we know that **two probability density function that are proportional are necessary equal** (since their integrals are both equal to 1). We thus deduce the posterior distribution:

$$[\vartheta|X = x] \sim B(S(x) + 1, N - S(x) + 1) .$$

Note that the conditional distribution given the sample $(X_1, \dots, X_N)$ depends on the sample only through the S statistic. It is called a **sufficient statistic** with respect to the statistical model and its associated unknown parameter.

In Figure 2.1, we illustrate this example. By simulation of random variables, we construct a sample of size $N = 100$, $(x_1, \dots, x_{100})$ for the Bernoulli distribution $B(0.6)$, and we deduce the posterior distributions using the first data $p(\theta|x_1)$, the first ten $p(\theta|x_1, \dots, x_{10})$ and the whole sample $p(\theta|x_1, \dots, x_{100})$, with a uniform prior on $[0; 1]$. We can clearly see the evolution of the posterior distribution as with the increase of the sample size, it refocuses towards the true unknown value of 0.6. Note that we had $x_1 = 1$, which led to $p(\theta|x_1) = 2\theta\mathbb{I}_{[0;1]}(\theta)$. ■

2.7.3 Choice of the prior distribution

In the Bayesian method, the determination of the prior distribution is crucial. It generally has a strong influence on the posterior distribution. There are favourable situations in which one has a relatively precise idea of the prior. For example, if the parameters have a particular physical meaning, the distribution of these parameters can be evaluated in comparable and already known situations. This is for example the case when we want to estimate for a particular individual, a parameter for which we know the distribution in the population to which the individual belongs.

■ **Example 2.22** We find classic examples in pharmacokinetics (Lavielle, 2014). A simple model (PK) gives the concentration y of a drug in the blood as a function of time t and of the administered dose D as follows:

$$y(t, D) = \frac{Dk_a}{V k_a - C_l} \left[\exp\left(-t \frac{C_l}{V}\right) - \exp(-k_a t) \right] + \epsilon$$

with:

⁷The Γ function can be seen as a generalization to positive real numbers (and even to the complex half-plane) of the factorial function: $\forall x \in \mathbb{R}_+, \Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt$, with $\Gamma(n+1) = n!$ if $n \in \mathbb{N}$.

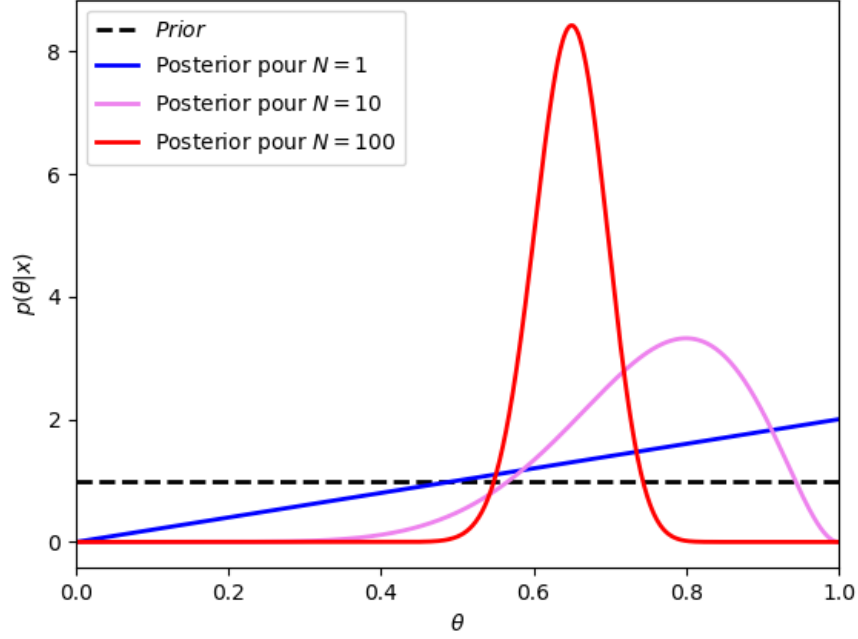


Figure 2.1: Posterior Distributions $p(\theta|x)$ for different sample sizes.

- $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$.
- k_a (absorption rate), C_l (elimination rate), V (distribution volume) are parameters specific to each individual.

An example of a curve obtained for this model is given in figure 2.2. Let us suppose that experiments on a certain number of individuals have allowed us to determine the distributions within the population⁸, for example:

$$(k_a, V, C_l)^T \sim \mathcal{N}\left((k_a^{pop}, V^{pop}, C_l^{pop})^T, \Sigma^{pop}\right). \quad (2.6)$$

Thus, if for a new patient we need to estimate his specific parameters, we can use the distribution in the population as a prior in a Bayesian approach. ■

In many cases, however, it is not possible to determine a precise prior distribution. In such cases, technical arguments can be used. For some types of likelihood functions, the prior and posterior distributions belong to the same family of probability laws, and the calculation of the posterior distribution amounts to a simple updating of the distribution parameters from the data. These are called conjugate distributions.

⁸We do not go into these details here, but the estimation of such population parameters corresponds to a very interesting field of Statistics, we speak of random effects models (or mixed effects models if certain parameters are stable in the population) or hierarchical models, see Lavielle (2014)

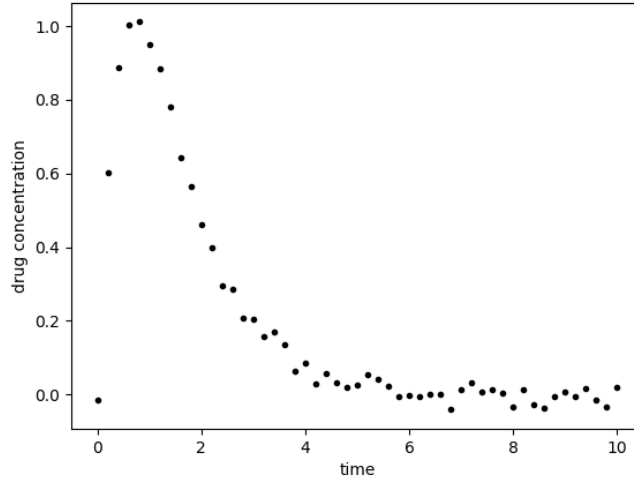


Figure 2.2: Example of a simulation for the PK model(2.6)

Definition 2.23 — Conjugate Distributions. A family of distributions $\mathcal{F}$ is said conjugate with respect to a likelihood function $p(x|\theta)$ if for all prior distribution $\pi \in \mathcal{F}$, the posterior distribution $p(\theta|x)$ is also in $\mathcal{F}$.

■ **Example 2.23** Let X be a random variable aléatoire in $\mathbb{N}$ following a Poisson distribution $\mathcal{P}(\lambda)$. Let x be an observation. We recall that for $x \in \mathbb{N}$, $p(x|\lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$. We want to determine the posterior distribution $p(\lambda|x)$ from a prior distribution on λ given by the Gamma distribution $\mathcal{G}(\alpha, \beta)$:

$$\pi(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

We thus have:

$$p(\lambda|x) \propto p(x|\lambda)\pi(\lambda) = \frac{\lambda^x}{x!} e^{-\lambda} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \propto \lambda^{x+\alpha-1} e^{-(\beta+1)\lambda}.$$

where we kept only the terms containing λ , the density being entirely determined by them, thanks to the normalization constraint of the density integral being 1. Once again, we see that the distribution is proportional to the density of a Gamma $\mathcal{G}(\alpha + x, \beta + 1)$ distribution, so the posterior distribution is $\mathcal{G}(\alpha + x, \beta + 1)$. The family of Gamma distributions is therefore conjugate for the likelihood of Poisson distributions. ■

There are many conjugate families. We refer, for example, to Robert (2007) for more examples. In practice, if conjugate families are chosen in order to simplify the calculations, these distributions should however be parameterized a priori as accurately as possible, for example by the method of moments. In our example 2.23, if we expect the parameter λ to equal a reference value λ_0 for example, we will choose the parameters (α, β) of the Gamma distribution such that its expectation is λ_0 , that is to say such that $\frac{\alpha}{\beta} = \lambda_0$ (since the expectation of $\mathcal{G}(\alpha, \beta)$ is given by $\frac{\alpha}{\beta}$, see the appendix).

Finally, one last way is the **non-informative prior distribution**. For example, we can choose a uniform distribution over all the possible parameters: none of the parameters is favored a priori. This is the approach used in Example 2.21.

■ **Remark 2.24** It is interesting to note that even this choice of a uniform prior is not neutral: if we get back to Example 2.21, for an observation $x_1 = 1$, we found $p(\theta|X_1 = 1) = 2\theta\mathbb{I}_{[0;1]}(\theta)$, for which the expectation is $2/3$, while if we had performed a maximum likelihood estimation, we would have found $\hat{\theta} = 1$. ■

2.7.4 Point Bayesian Estimation

One of the interests of Bayesian estimation lies in its contribution to decision theory (Robert, 2007). The estimation of a parameter in a statistical model is often only the prelude to a decision a user has to make: to give a forecast for a meteorologist, to take a position on the market for a trader, to prescribe a specific treatment to a patient for a doctor... Decision-making under uncertainty can be characterized by a so-called loss function, which measures the gap between the decision taken and the optimal decision. In a simplified way, for parameter estimation, one can assimilate the decision space to the Θ parameter space and the decision can be reduced to the choice of δ which minimizes the distance to the true parameter θ : $L(\theta, \delta)$. Since θ is unknown, we minimize $L(\theta, \delta)$ in expectation. In particular, in the Bayesian framework, and for an observation sample $X = x$, we will try to minimize:

$$J(\delta) = \mathbb{E}_{P_{\theta|X=x}}(L(\theta, \delta)) = \int_{\Theta} L(\theta, \delta) p(\theta|x) d\theta .$$

The loss function is generally considered to be the opposite of a utility function, the definition of which depends on the problem (e.g. the gain for taking a position in a market or the quality of the weather forecast). Two of the most common examples of loss functions are given below.

■ **Example 2.25 — Quadratic loss function.** For $\Theta \subset \mathbb{R}^p$, we can choose $L(\theta, \delta) = \|\delta - \theta\|^2$. This function can be used to penalize large deviations. We search $\hat{\theta}$ minimizing over Θ the loss:

$$J(\delta) = \int_{\Theta} \|\delta - \theta\|^2 p(\theta|x) d\theta .$$

But $\int_{\Theta} \theta p(\theta|x) d\theta = \mathbb{E}(\theta|x)$, therefore $\int_{\Theta} (\delta - \mathbb{E}(\theta|x))^T (\mathbb{E}(\theta|x) - \theta) p(\theta|x) d\theta = 0$ and finally:

$$J(\delta) = \int_{\Theta} \left(\|\delta - \mathbb{E}(\theta|x)\|^2 + \|\mathbb{E}(\theta|x) - \theta\|^2 \right) p(\theta|x) d\theta ,$$

that is to say

$$J(\delta) = \|\delta - \mathbb{E}(\theta|x)\|^2 + \int_{\Theta} \|\mathbb{E}(\theta|x) - \theta\|^2 p(\theta|x) d\theta .$$

The second term in the sum does not depend on δ , therefore $\hat{\theta}$ is obtained by minimizing $\|\delta - \mathbb{E}(\theta|x)\|^2$, and

$$\hat{\theta} = \mathbb{E}(\theta|x) = \int_{\Theta} \theta p(\theta|x) d\theta ,$$

$\hat{\theta}$ is the expectation of the prior distribution, also called posterior expectation. ■

■ **Example 2.26 — Mean absolute loss.** When $\Theta \subset \mathbb{R}$, we can also choose the mean absolute loss $L(\theta, \delta) = |\delta - \theta|$. We search $\hat{\theta}$ minimizing over Θ the loss:

$$J(\delta) = \int_{\Theta} |\delta - \theta| p(\theta|x) d\theta .$$

We have:

$$J(\delta) = \int_{-\infty}^{\delta} (\delta - \theta) p(\theta|x) d\theta + \int_{\delta}^{+\infty} (\theta - \delta) p(\theta|x) d\theta ,$$

and therefore $\forall h \in \mathbb{R}$:

$$\begin{aligned} \frac{J(\delta+h) - J(\delta)}{h} &= \frac{1}{h} \left[\int_{-\infty}^{\delta+h} (\delta+h - \theta) p(\theta|x) d\theta + \int_{\delta+h}^{+\infty} (\theta - \delta - h) p(\theta|x) d\theta \right. \\ &\quad \left. - \int_{-\infty}^{\delta} (\delta - \theta) p(\theta|x) d\theta - \int_{\delta}^{+\infty} (\theta - \delta) p(\theta|x) d\theta \right] \\ &= \frac{1}{h} \left[\int_{-\infty}^{\delta} h p(\theta|x) d\theta + \int_{\delta}^{+\infty} (-h) p(\theta|x) d\theta \right. \\ &\quad \left. + \int_{\delta}^{\delta+h} (\delta+h - \theta) p(\theta|x) d\theta - \int_{\delta}^{\delta+h} (\theta - \delta - h) p(\theta|x) d\theta \right] \\ &= \int_{-\infty}^{\delta} p(\theta|x) d\theta - \int_{\delta}^{+\infty} p(\theta|x) d\theta + \frac{2}{h} \int_{\delta}^{\delta+h} (\delta+h - \theta) p(\theta|x) d\theta . \end{aligned}$$

We have $\forall \theta \in [\delta; \delta+h]$, $0 \leq \delta+h - \theta \leq h$, thus

$$\lim_{h \rightarrow 0} \left(\frac{2}{h} \int_{\delta}^{\delta+h} (\delta+h - \theta) p(\theta|x) d\theta \right) = 0 ,$$

so finally:

$$J'(\delta) = \lim_{h \rightarrow 0} \frac{J(\delta+h) - J(\delta)}{h} = \int_{-\infty}^{\delta} p(\theta|x) d\theta - \int_{\delta}^{+\infty} p(\theta|x) d\theta .$$

A necessary condition for $\hat{\theta}$ to minimize J is that $J'(\hat{\theta}) = 0$, that is to say

$$\int_{-\infty}^{\hat{\theta}} p(\theta|x) d\theta = \int_{\hat{\theta}}^{+\infty} p(\theta|x) d\theta .$$

But

$$\int_{-\infty}^{\delta} p(\theta|x) d\theta + \int_{\delta}^{+\infty} p(\theta|x) d\theta = 1 ,$$

therefore $J'(\hat{\theta}) = 0$ when :

$$\int_{-\infty}^{\hat{\theta}} p(\theta|x) d\theta = \int_{\hat{\theta}}^{+\infty} p(\theta|x) d\theta = 1/2 ,$$

that is to say when $\hat{\theta}$ is the median of the posterior distribution. We have moreover that $J'(\delta) \leq 0$ when $\delta < \hat{\theta}$ and $J'(\delta) \geq 0$ when $\delta > \hat{\theta}$. The median of the posterior distribution is therefore the estimator minimizing the mean absolute loss. ■

2.8 Confidence Region

Let X be a random variable aléatoire or an i.i.d. sample for an unknown distribution inconnue P_{θ^*} in a statistical model $\mathcal{M}_\Theta$, with $\Theta \subset \mathbb{R}^p$. The point estimation of θ^* by $\hat{\theta}$ from data x , that is to say from an observation of X , remains unsatisfactory: the randomness of the observations from which the estimate is made necessarily results in an estimation uncertainty. In order to better characterize it, a confidence region can be constructed for the parameter.

2.8.1 Exact Confidence Region

Definition 2.24 — Confidence Region. Let $X = (X_1, \dots, X_N)$ be a i.i.d. sample of P_{θ^*} . Let $\alpha \in]0, 1[$ a given risk level. A confidence region $I_\alpha(X)$ of confidence **level** $1 - \alpha$ is a statistic of the sample X with values in $\mathcal{P}(\Theta)$ such that:

$$\mathbb{P}_{\theta^*}(\theta^* \in I(X)) \geq 1 - \alpha.$$

If $\mathbb{P}_{\theta^*}(\theta^* \in I(X)) = 1 - \alpha$, the confidence region will be said of **size** $1 - \alpha$.

If $\Theta \subset \mathbb{R}$, we choose the confidence region in the set of intervals on $\mathbb{R}$ and we talk of **confidence interval**.

In the remainder of this chapter we'll restrict ourselves to the case where $\Theta \subset \mathbb{R}$, so we'll talk about confidence intervals. Concepts can be generalized to the multidimensional case.

■ **Remark 2.27** We point out that θ^* is fixed, the interval is a random variable. Typically, we can choose a level of risk $\alpha = 0.05$, and if we determine a statistic $I_\alpha(X)$ which is a confidence interval of size α , then for 95% of observation samples x , $I_\alpha(x)$ contains the true parameter θ^* .

■ **Remark 2.28** For a given law and a given confidence level, there is an infinity of confidence intervals satisfying the definition: if I is a confidence interval of level $1 - \alpha$, any I' interval containing I is a confidence interval too. The interval of minimum length is generally preferred, giving a better precision on the value of the true parameter. ■

The explicit determination of a confidence interval is based on the existence of a pivotal function.

Definition 2.25 — Pivotal function and distribution free of the parameter. For a model whose unknown parameter is θ , a pivotal function is a statistic $S(X; \theta)$ that depends on θ , but whose probability distribution does not depend on θ . We say that its distribution is free of θ . It is also called an **ancillary statistic**.

In practice, we choose the pivotal function whose expression according to the parameter is the simplest possible.

■ **Example 2.29 — Pivotal function for the Gaussian model.** Let $X = (X_1, \dots, X_N)$ be an i.i.d. sample for $\mathcal{N}(\mu, \sigma^2)$. We consider several situations:

(i) We want to estimate μ , with σ^2 known.

The sum of N independent Gaussian variables is a Gaussian of mean the sum of the means and variance the sum of the variances. However, the Gaussian family is stable with respect

to changes in location and scale.⁹, then

$$\frac{\sqrt{N}(\bar{X} - \mu)}{\sigma} \sim \mathcal{N}(0, 1) .$$

$\mathcal{N}(0, 1)$ is a distribution free of θ , and $\frac{\sqrt{N}(\bar{X} - \mu)}{\sigma}$ is a pivotal function for μ .

- (ii) We want to estimate σ^2 , with μ known.

We still have:

$$\frac{\sqrt{N}(\bar{X} - \mu)}{\sigma} \sim \mathcal{N}(0, 1) .$$

Therefore, $\frac{\sqrt{N}(\bar{X} - \mu)}{\sigma}$ is also a pivotal function for σ . But another pivotal function is usually used to facilitate the management of the confidence interval. If $Z_1, \dots, Z_N$ are i.i.d. variables of distribution $\mathcal{N}(0, 1)$, then $\sum_{i=1}^N Z_i^2 \sim \chi^2(N)$, a χ^2 distribution with N degrees of freedom. Therefore:

$$\frac{\sum_{i=1}^N (X_i - m)^2}{\sigma^2} \sim \chi^2(N) ,$$

and finally $\frac{\sum_{i=1}^N (X_i - m)^2}{\sigma^2}$ is a pivotal function for σ^2 .

- (iii) We want to estimate σ^2 , with μ unknown.

The pivotal function σ^2 can no longer involve μ . We can show using Cochran's theorem that for a Gaussian law:

$$\frac{NS^2}{\sigma^2} = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(N - 1) .$$

This result is detailed in Appendix A.2, Theorem A.2.

$\frac{NS^2}{\sigma^2}$ is thus a pivotal function for σ^2

- (iv) We want to estimate μ , with σ^2 unknown.

We can then combine the pivotal functions obtained in (i) and (iii). As a matter of fact, if $U \sim \mathcal{N}(0, 1)$, $Z \sim \chi^2(\nu)$ and Z is independent of U , then $\frac{U}{\sqrt{Z/\nu}} \sim T(\nu)$, Student's t-distribution with ν degrees of freedom. By Cochran's theorem, we also know that for a Gaussian distribution, $\bar{X}$ and S^2 are independent (cf. Appendix A.2). Thus

$$\frac{\frac{\sqrt{N}(\bar{X} - \mu)}{\sigma}}{\sqrt{\frac{NS^2}{(N-1)\sigma^2}}} = \frac{\sqrt{N-1}(\bar{X} - \mu)}{S} \sim T(N-1)$$

■

■ **Example 2.30** Let us consider the family of exponential distributions:

$$\mathcal{M} = \left\{ \mathcal{E}(\lambda), p_\lambda(x) = \lambda e^{-\lambda x} \mathbb{I}_{[0; +\infty[}(x), \lambda > 0 \right\} ,$$

⁹If $X \sim \mathcal{N}(\mu, \sigma^2)$, $\frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$.

and let $X_1, \dots, X_N$ be an i.i.d. sample for $\mathcal{E}(\lambda)$.

It is easy to check that if $X \sim \mathcal{E}(\lambda)$, and if $\beta > 0$, $\beta X \sim \mathcal{E}(\lambda/\beta)$, and thus in particular, $2\lambda X_i \sim \mathcal{E}(1/2)$. We also know that $\mathcal{E}(1/2) = \chi^2(2)$ (cf. Appendix A.1). On the other hand the sum of independent χ^2 variables is a χ^2 distribution of degree of freedom equal to the sum of the degrees of freedom. Therefore: $2\lambda \sum_{i=1}^N X_i \sim \chi^2(2N)$. $2\lambda \sum_{i=1}^N X_i$ is a pivotal function for λ of distribution $\chi^2(2N)$, free of λ . ■

The quantiles of the distributions of the pivotal functions will allow us to determine the confidence intervals.

Definition 2.26 — Quantile. Let F be the distribution function of a probability law P on $\mathbb{R}$. For $r \in]0; 1[$, the r -quantile of F is defined by:

$$q_r = \inf \{x \in \mathbb{R}, F(x) \geq r\}.$$

If F is continuous, we have $F(q_r) = r$. If moreover F is strictly increasing, q_r is the unique point such that $F(q_r) = r$, that is to say $q_r = F^{-1}(r)$.

■ **Remark 2.31** Historically, the quantiles of the usual laws were determined from statistical tables. Now, these quantiles are obtained directly from software. We give below a small code in Julia (<https://julialang.org/>) allowing to determine the 0.975-quantiles for a reduced centered Gaussian distribution and 0.95-quantile for a χ^2 distribution with 10 degrees of freedom.

```
using Distributions
q1=quantile(Normal(0,1),0.975)
q2=quantile(Chisq(10),0.95)
```

The following proposition gives us a way to construct confidence intervals from the pivotal functions and their distributions. ■

Proposition 2.24 Let α , $0 < \alpha < 1$, be given. Let $S(X; \theta)$ be a pivotal function for a parametric statistical model, of parameter $\theta \in \Theta$. If its cumulative distribution function is continuous and if q_r denotes the r -quantile, then $\forall \gamma$, $0 \leq \gamma \leq \alpha$:

$$I_\alpha^\gamma(X) = S^{-1}(X; [q_\gamma; q_{\gamma+1-\alpha}]) = \{a \in \Theta, q_\gamma \leq S(X; a) \leq q_{\gamma+1-\alpha}\}$$

is a confidence interval of size $(1 - \alpha)$ for θ .

Preuve : It is enough to check that:

$$P_\theta(\theta \in I_\alpha^\gamma(X)) = P_\theta(q_\gamma \leq S(X; \theta) \leq q_{\gamma+1-\alpha}) = F(q_{\gamma+1-\alpha}) - F(q_\gamma) = 1 - \alpha$$

where F denotes the continuous cumulative distribution function of $S(X; \theta)$. □

■ **Remark 2.32** One generally adapts the choice of γ so that, for a given level of risk α , I_α^γ is as small as possible, corresponding to better precision. ■

■ **Example 2.33** We go back to case (i) of Example 2.29. We consider a Gaussian i.i.d. sample $X_1, \dots, X_N \sim \mathcal{N}(\mu, 1)$, of parameter μ unknown. We saw that $S(X_1, \dots, X_N; \mu) = \sqrt{N}(\bar{X} - \mu)$ is a pivotal function for μ with distribution $\mathcal{N}(0, 1)$ free of μ , for which the cumulative distribution function is continuous. We then use Proposition 2.24. Let α be given, $\forall \gamma, 0 \leq \gamma \leq \alpha$, $I_\alpha^\gamma(X) = S^{-1}(X_1, \dots, X_N; [q_\gamma; q_{\gamma+1-\alpha}])$. The difficulty lies in calculating this reciprocal image, which is not always easy.

Here, we have:

$$\mathbb{P}\left(q_\gamma \leq \sqrt{N}(\bar{X} - \mu) \leq q_{\gamma+1-\alpha}\right) = 1 - \alpha$$

which is equivalent to write:

$$\mathbb{P}\left(\bar{X} - \frac{q_{\gamma+1-\alpha}}{\sqrt{N}} \leq \mu \leq \bar{X} - \frac{q_\gamma}{\sqrt{N}}\right) = 1 - \alpha$$

By symmetry of the density of $\mathcal{N}(0, 1)$, we have $-q_\gamma = q_{1-\gamma}$, therefore

$$I_\alpha^\gamma = \left[\bar{X} - \frac{q_{\gamma+1-\alpha}}{\sqrt{N}} \leq \mu \leq \bar{X} + \frac{q_{1-\gamma}}{\sqrt{N}} \right].$$

The length of the confidence interval is $(q_{\gamma+1-\alpha} + q_{1-\gamma})/\sqrt{N}$. In the case where the parameter-free distribution is a Gaussian density and if $\alpha \leq 0.5$ (which is usually the case, since we often take $\alpha = 0.01, 0.05$ ou 0.1 for example), we can show that the smallest (and therefore the most accurate) interval is obtained for $\gamma = \alpha/2$. Indeed, on $[1/2; 1]$, the quantile function of the Gaussian distribution (that is to say F^{-1} , the reciprocal of the cumulative distribution function) is convex¹⁰ (cf. Figure 2.3) and thus:

$$q_{\gamma+1-\alpha} + q_{1-\gamma} = F^{-1}(\gamma + 1 - \alpha) + F^{-1}(1 - \gamma) = 2 \left(\frac{1}{2} F^{-1}(\gamma + 1 - \alpha) + \frac{1}{2} F^{-1}(1 - \gamma) \right)$$

For $\alpha \leq 0.5$, since $0 \leq \gamma \leq \alpha$, we have $\gamma + 1 - \alpha$ and $1 - \gamma$ in $[1/2; 1]$, and therefore by convexity of F^{-1} over this interval:

$$\left(\frac{1}{2} F^{-1}(\gamma + 1 - \alpha) + \frac{1}{2} F^{-1}(1 - \gamma) \right) \geq F^{-1} \left(\frac{1}{2}(\gamma + 1 - \alpha) + \frac{1}{2}(1 - \gamma) \right) = F^{-1}(1 - \alpha/2).$$

So finally, $\forall 0 \leq \gamma \leq \alpha \leq 0.5$ and for a parameter-free centered Gaussian distribution, we have that the smallest confidence interval is obtained for $\gamma = \alpha/2$, that is to say:

$$I_\alpha^\gamma = \left[\bar{X} - \frac{q_{1-\alpha/2}}{\sqrt{N}} \leq \mu \leq \bar{X} + \frac{q_{1-\alpha/2}}{\sqrt{N}} \right].$$

Numerical Example: let's suppose that for example, $\alpha = 0.05$, $\bar{X} = 10$, and $N = 100$, we obtain that $q_{0.975} \approx 1.96$ and the optimal confidence interval is given by:

$$I_{0.05} \approx [9.804; 10.196].$$

Now assume that $N = 1000$, for the same observed empirical mean. The confidence interval becomes more precise:

$$I_{0.05} \approx [9.938; 10.062].$$

■

¹⁰Let C be a convex set in a vector space E , and $f : C \rightarrow \mathbb{R}$. We say that f is convex if $\forall x_1, x_2 \in C, \forall t \in [0; 1], f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$. In particular, f is convex if and only if its epigraph $\mathcal{E}_f = \{(x, \alpha), f(x) \leq \alpha\}$ is convex.

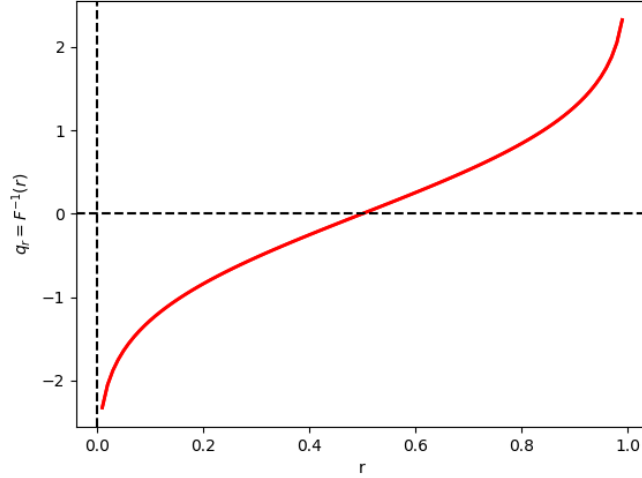


Figure 2.3: Quantile function of the standard Gaussian distribution.

■ **Example 2.34** We now return to the (ii) case in Example 2.29. We consider a Gaussian i.i.d. sample $X_1, \dots, X_N \sim \mathcal{N}(0; \sigma^2)$, of parameter σ^2 unknown. We saw that $S(X_1, \dots, X_N; \mu) = \frac{\sum_{i=1}^N X_i^2}{\sigma^2}$ is a pivotal function pivotalet for σ^2 with distribution $\chi^2(N)$, whose cumulative distribution function is continuous. Let α be gicen and let γ , $0 \leq \gamma \leq \alpha$. We have then:

$$\mathbb{P} \left(q_\gamma \leq \frac{\sum_{i=1}^N X_i^2}{\sigma^2} \leq q_{\gamma+(1-\alpha)} \right) = 1 - \alpha$$

where q_r denotes the r -quantile for $\chi^2(N)$. We deduce:

$$\mathbb{P} \left(\frac{\sum_{i=1}^N X_i^2}{q_{\gamma+(1-\alpha)}} \leq \sigma^2 \leq \frac{\sum_{i=1}^N X_i^2}{q_\gamma} \right) = 1 - \alpha$$

and the confidence interval:

$$I_\alpha^\gamma = \left[\frac{\sum_{i=1}^N X_i^2}{q_{\gamma+(1-\alpha)}} ; \frac{\sum_{i=1}^N X_i^2}{q_\gamma} \right].$$

Note that it's much harder to pick the best γ . If N is large, we can approach the $\chi^2(N)$ distribution by a normal law, and we also get that the best *gamma* is close to $\alpha/2$. ■

■ **Example 2.35 — (Continuation of Example 2.30).** We consider the exponential distributions model $\mathcal{E}(\lambda)$ seen in Example 2.30. We know that: $2\lambda \sum_{1 \leq i \leq N} X_i$ is a pivotal function for λ with $\chi^2(2N)$ distribution. Let α be given, and let γ , $0 \leq \gamma \leq \alpha$. We have:

$$\mathbb{P} \left(q_\gamma \leq 2\lambda \sum_{1 \leq i \leq N} X_i \leq q_{\gamma+(1-\alpha)} \right) = 1 - \alpha$$

where q_r denotes the r -quantile of the $\chi^2(2N)$ distribution, which gives us the confidence interval:

$$I_\alpha^\gamma = \left[\frac{q_\gamma}{2 \sum_{1 \leq i \leq N} X_i} ; \frac{q_{\gamma+(1-\alpha)}}{2 \sum_{1 \leq i \leq N} X_i} \right].$$

As in the previous example 2.34, it is not easy to find the optimal *gamma*, but for N big enough, we can take *gamma* = $\alpha/2$, and:

$$I_\alpha = \left[\frac{q_{\alpha/2}}{2 \sum_{1 \leq i \leq N} X_i} ; \frac{q_{(1-\alpha/2)}}{2 \sum_{1 \leq i \leq N} X_i} \right].$$

Numerical example: take for example $\alpha = 0.05$, $N = 100$, and suppose $\sum_{1 \leq i \leq N} X_i = 18.34$, we obtain $I_{0.05} = [4.44; 6.57]$. (Note that the values were obtained by simulation for $X_i \sim \mathcal{E}(5)$). ■

2.8.2 Asymptotic confidence region

It is often not possible to determine a pivotal function for the parameter to be estimated. Thanks to the convergence theorems on the sequences of random variables (in particular the central limit theorem), one can often consider the asymptotic framework.

Definition 2.27 — Asymptotic Confidence Region. Let $(X_i)_{i \in \mathbb{N}}$ be a sequence of i.i.d. random variables of distribution P_{θ^*} . Let $\alpha \in]0, 1[$ be a given risk level. A confidence region $I_\alpha(X)$ of **asymptotic level** $1 - \alpha$ is a statistic on sample X with values in $\mathcal{P}(\Theta)$ such that:

$$\lim_{N \rightarrow +\infty} \mathbb{P}_{\theta^*}(\theta^* \in I(X)) \geq 1 - \alpha.$$

In the same way as for the exact confidence regions, if $\lim_{N \rightarrow +\infty} \mathbb{P}_{\theta^*}(\theta^* \in I(X)) = 1 - \alpha$, the confidence region is said of **asymptotic size** $1 - \alpha$.

Confidence regions of asymptotic levels or asymptotic sizes are called asymptotic confidence regions.

If $\Theta \subset \mathbb{R}$, the confidence region will be chosen in the set of intervals on $\mathbb{R}$ and we talk of **asymptotic confidence intervals**.

However, the confidence region construction is carried out for a finite sample size of N . We then define the coverage probability, corresponding to the real size, which will allow us to measure the deviation from the asymptotic size.

Definition 2.28 — Coverage Probability. Let $X = (X_1, \dots, X_N)$ be an i.i.d. sample for P_{θ^*} . Let $I(X)$ be a confidence region for θ^* . The coverage probability τ_c is defined by:

$$\tau_c = P_{\theta^*}(\theta^* \in I(X)).$$

If $I_\alpha(X)$ is an exact confidence region of size $(1 - \alpha)$, then $\tau^c = 1 - \alpha$. Of course, this is no longer the case if $I(X)$ is a confidence region of asymptotic size $1 - \alpha$, we only have $\lim_{N \rightarrow +\infty} \tau^c = 1 - \alpha$. We therefore often use τ^c to characterize the construction of asymptotic confidence regions. It's illustrated in figure 2.4.

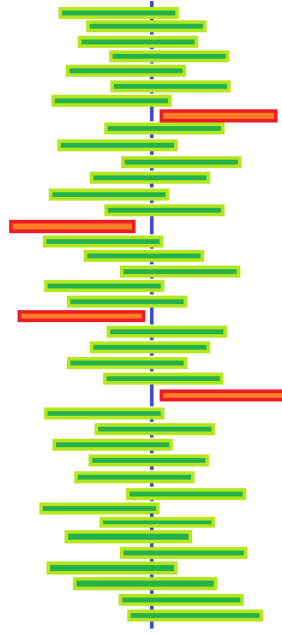


Figure 2.4: Confidence intervals and coverage probability: the true parameter θ^* is indicated by the blue line and several realizations of $I(X)$ are shown, in green if the interval contains θ^* , in red otherwise.

These asymptotic confidence regions are built from asymptotic pivotal functions.

Definition 2.29 — Asymptotic pivotal function. Let us consider a model of parameter θ , $\mathcal{M}_\Theta = \{P_\theta, \theta \in \Theta\}$, and $(X_i)_{i \in \mathbb{N}^*}$, a sequence of i.i.d. random variables for $P_\theta \in \mathcal{M}_\Theta$. An asymptotic pivotal function is a statistic $S(X_1, \dots, X_N; \theta)$ which may depend on θ , but whose limit distribution does not depend on θ , that is to say that there exists a random variable Y of known and θ -free distribution (that is to say independent of θ) such that:

$$S(X_1, \dots, X_N) \xrightarrow{\mathcal{L}} Y .$$

■ **Example 2.36** Let us consider the model of Bernoulli distributions, $\mathcal{M} = \{\mathcal{B}(\theta), \theta \in]0; 1[\}$. If $X_1, \dots, X_N$ is an i.i.d. sample of distribution $\mathcal{B}(\theta)$, it is not possible to find a pivotal function. We know for example that $\sum_{1 \leq i \leq N} X_i \sim \mathcal{B}(N, \theta)$ (binomial law), but it does not allow to determine a pivotal function. On the other hand, we can use the central limit theorem: we know that $\mathbb{E}(X_i) = \theta$ and $\mathbb{V}(X_i) = \theta(1 - \theta)$, therefore

$$\sqrt{N} \left(\frac{\bar{X} - \theta}{\sqrt{\theta(1 - \theta)}} \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) ,$$

$\sqrt{N} \left(\frac{\bar{X} - \theta}{\sqrt{\theta(1 - \theta)}} \right)$ is an asymptotic pivotal function: ■

■ **Example 2.37 — (Continuation of example 2.30).** If we take the model of exponential distributions of Example 2.30, let $X_1, \dots, X_N$ be an i.i.d sample for $\mathcal{E}(\lambda)$. We found an exact pivotal

function, but we also have an asymptotic pivotal function using the central limit theorem. Since $\mathbb{E}(X_i) = 1/\lambda$ and $\mathbb{V}(X_i) = 1/\lambda^2$, we have

$$\sqrt{N} \left(\frac{\bar{X} - 1/\lambda}{1/\lambda} \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0; 1) ,$$

$\sqrt{N} (\lambda \bar{X} - 1)$ is an asymptotic pivotal function. In this case, it is of course preferable to choose the exact pivotal function. We will see below the difference that it implies in the construction of confidence intervals, in terms of coverage probabilities. ■

The following theorem will help us to construct asymptotic pivotal functions in the case where $\theta \in \mathbb{R}$.

Theorem 2.25 Let $M_\Theta = \{P_\theta, \theta \in \Theta \subset \mathbb{R}\}$ be a statistical model. Let T_N be an estimator of θ , asymptotically normal, satisfying $\forall \theta \in \Theta$:

$$\sqrt{N} (T_N - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V(\theta))$$

with $V(\theta) > 0$ and continuous in θ , then:

$$\sqrt{N} \frac{(T_N - \theta)}{\sqrt{V(\theta)}} \text{ et } \sqrt{N} \frac{(T_N - \theta)}{\sqrt{V(T_N)}}$$

are asymptotic pivotal functions for θ with distribution $\mathcal{N}(0; 1)$.

Preuve : It is direct for $\frac{\sqrt{N}(T_N - \theta)}{\sqrt{V(\theta)}}$. Since T_N is asymptotically normal, it converges in probability to θ (cv. Theorem 2.19). By the continuity theorem 2.15, we then have that

$$\frac{\sqrt{V(\theta)}}{\sqrt{V(T_N)}} \xrightarrow{\mathbb{P}} 1$$

and since

$$\frac{\sqrt{N} (T_N - \theta)}{\sqrt{V(\theta)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0; 1)$$

using Slutsky's theorem 2.12

$$\sqrt{N} \frac{(T_N - \theta)}{\sqrt{V(T_N)}} = \sqrt{N} \frac{(T_N - \theta)}{\sqrt{V(\theta)}} \frac{\sqrt{V(\theta)}}{\sqrt{V(T_N)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0; 1) .$$

□

The second asymptotic pivotal function is obviously easier to inverse.

Corollary 2.26 Let α be given,

$$I_\alpha = \left[T_N - q_{1-\alpha/2} \sqrt{V(T_N)/N} ; T_N + q_{1-\alpha/2} \sqrt{V(T_N)/N} \right]$$

is a confidence interval for θ of asymptotic size $(1 - \alpha)$.

Preuve : Using the convergence in distribution of the pivotal function to the standard Gaussian law, and noting q_r the r -quantile of the latter, we have:

$$\lim_{N \rightarrow +\infty} \mathbb{P} \left(q_{\alpha/2} \leq \sqrt{N} \frac{(T_N - \theta)}{\sqrt{V(T_N)}} \leq q_{1-\alpha/2} \right) = 1 - \alpha$$

by using the symmetry of the standard Gaussian distribution, $q_{\alpha/2} = -q_{1-\alpha/2}$, and we directly get the result. $\square$

■ **Remark 2.38 — 'Plug-in' Method.** In a pivotal function that is difficult to invert, the use of the estimator in place of the parameter to be estimated is often made possible by Slutsky's theorem, as in the demonstration of theorem 2.25, which can simplify the expression and ease the determination of the asymptotic confidence interval. This is called a substitution or 'plug-in' method. ■

■ **Example 2.39 — (Continuation of Example 2.36).** For the model of Bernoulli distributions, we saw that

$$\sqrt{N} \left(\frac{\bar{X} - \theta}{\sqrt{\theta(1-\theta)}} \right)$$

is an asymptotic pivotal function of asymptotic distribution $\mathcal{N}(0, 1)$ or equivalently, that $\bar{X}$ is an asymptotically normal estimator of θ with:

$$\sqrt{N} (\bar{X} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \theta(1-\theta))$$

so using Theorem 2.25 and its corollary 2.26, we directly have the confidence interval of asymptotic size $(1 - \alpha)$:

$$\left[\bar{X} - q_{1-\alpha/2} \sqrt{\bar{X}(1-\bar{X})/\sqrt{N}}; \bar{X} + q_{1-\alpha/2} \sqrt{\bar{X}(1-\bar{X})/\sqrt{N}} \right] \quad (2.7)$$

where q_r denotes the r -quantile of the standard Gaussian distribution.

Numerical example: for $N = 100$ and $\alpha = 0.05$, suppose $\bar{X} = 0.53$ (for example 53% of positive answers in a sample of 100 people), we find the asymptotic confidence interval $[0.432; 0.628]$. We can see that the accuracy is very low. With $N = 1000$ and the same percentage $\bar{X} = 0.53$ observed, the confidence interval becomes $[0.499; 0.561]$.

By simulation, we can calculate an approximation of the coverage probability τ_c : for example, for a true parameter value of $p_0 = 0.53$, we simulate a very large number of times $N_s = 10^6$ a sample of $N = 100$ variables $\mathcal{B}(p_0)$, we calculate the confidence interval given by Equation (2.7), and we count the number of times where p_0 belongs to the confidence interval thus constructed. Here we find $\tau_c(100) = 0.944$. We do the same for $N = 1000$ and we find $\tau_c(1000) = 0.947$. The coverage probabilities here are very good (the asymptotic value being 0.95). ■

■ **Example 2.40 — (Continuation of Examples 2.35, 2.37).** For the model of exponential distributions, we showed that $\sqrt{N} (\lambda \bar{X} - 1)$ is an asymptotic pivotal function for λ . We directly have:

$$\lim_{N \rightarrow +\infty} \mathbb{P} \left(-q_{1-\alpha/2} \leq \sqrt{N} (\lambda \bar{X} - 1) \leq q_{1-\alpha/2} \right) = 1 - \alpha \quad (2.8)$$

where q_r denotes the r -quantile of the standard Gaussian distribution.
 And finally, the asymptotic confidence interval:

$$\left[\frac{1}{\bar{X}} \left(1 - \frac{q_{1-\alpha/2}}{\sqrt{N}} \right) ; \frac{1}{\bar{X}} \left(1 + \frac{q_{1-\alpha/2}}{\sqrt{N}} \right) \right]$$

Numerical example: if we take the numerical values from Example 2.35, $\alpha = 0.05$, $N = 100$, $\sum_{1 \leq i \leq N} X_i = 18.34$, and we obtain by Equation (2.8) the asymptotic confidence interval $[4.38; 6.52]$, to compare with the exact confidence interval, that is to say $I_{0.05} = [4.44; 6.57]$. (Recall that the values were obtained by simulation by taking $X_i \sim \mathcal{E}(5)$).

We calculate the coverage probability by simulation and find $\tau_c(100) = 0.951$. Even for $N = 20$, we find $\tau_c(20) = 0.953$: here again, the asymptotic approximation proves to be of good quality, even for a small sample size. ■

3. Statistical Hypothesis Testing

This chapter introduces the basic concepts of statistical hypothesis testing theory. This field covers a wide range of different configurations and many methods (Lehmann and Romano, 2006). We introduce the general principles here by illustrating them through a few simple examples.

3.1 Problem Formulation

Hypothesis testing is used to formalize decision making based on experimental data and the risk measures associated with this decision making.

3.1.1 Statistical Hypothesis and Test

Let us consider a statistical model $\mathcal{M}$, family of probability distribution on $(\mathcal{X}, \mathcal{A})$, and let $P^* \in \mathcal{M}$ be an unknown probability distribution describing the phenomenon of interest.

Definition 3.1 — Statistical Hypothesis. A statistical hypothesis about a distribution $P^* \in \mathcal{M}$ is a statement of the type $P^* \in \mathcal{M}_0$, where $\mathcal{M}_0 \subset \mathcal{M}$ is a subfamily of distributions in $\mathcal{M}$.

■ **Example 3.1** These hypotheses can be very diverse, for example $\mathcal{M}_0$ can be reduced to a singleton: P^* is a Gaussian distribution of zero mean and variance 10. $\mathcal{M}_0$ can also be a particular parametric model $\mathcal{M}_0 = \{\mathcal{E}(\lambda), \lambda > 0\}$ or relate only to features of the distribution: P^* is a symmetric distribution. ■

The purpose of a statistical hypothesis test (or more simply statistical test or test) is to indicate whether a statistical hypothesis can be considered true. Usually, the test will confront two disjoint statistical hypotheses, which is the framework we are considering here.

Definition 3.2 — Statistical Hypothesis Test. Let $\mathcal{M}_0 \subset \mathcal{M}$, $\mathcal{M}_1 \subset \mathcal{M}$, $\mathcal{M}_0 \cap \mathcal{M}_1 = \emptyset$. A statistical hypothesis test is a procedure that compares two statistical hypotheses: $H_0 : P^* \in \mathcal{M}_0$ and $H_1 : P^* \in \mathcal{M}_1$ and defines a decision rule allowing from the observation of an independent identically distributed sample of size N for P^* to indicate whether H_0 can be accepted (assumed to be true) or rejected (assumed to be false) in favor of H_1 .

We call H_0 the **null or conservative hypothesis** and H_1 the **alternative hypothesis**.

Formally, a test on $(\mathcal{X}^N, \mathcal{A}^{\otimes N})$, is thus defined by the triplet $(\mathcal{M}_0, \mathcal{M}_1, \delta)$ where $\mathcal{M}_0, \mathcal{M}_1$ are two disjoint family of distributions on $(\mathcal{X}, \mathcal{A})$ and δ a statistic $\delta : \mathcal{X}^N \rightarrow \{0, 1\}$ such that for all $(x_1, \dots, x_N) \in \mathcal{X}^N$, if $\delta(x_1, \dots, x_N) = 0$ H_0 is accepted, whereas if $\delta(x_1, \dots, x_N) = 1$, H_0 is rejected.

■ **Remark 3.2** An important aspect of this formulation is the asymmetry of the two assumptions. Indeed, depending on whether the null hypothesis is $P^* \in \mathcal{M}_0$ or $P^* \in \mathcal{M}_1$, we will have different tests and potentially different conclusions. The important thing to remember is that the null hypothesis must represent the conservative hypothesis, the one we would least like to falsely reject. A classic analogy is that of the presumption of innocence in a court of law: the basic assumption is that the person on trial is innocent, and convincing evidence must be provided to show that this is not true. This presumption is based on the principle that it is more serious to convict an innocent person than to release a guilty one. Another typical situation is, for example, the decision to put a new product into production for a company: since this potentially involves large additional costs, it is important to prove that the new product is better than an old product, for example. In this case, the conservative null hypothesis would be H_0 : the new product is not better than the old one. ■

Definition 3.3 — Critical region. Let $(\mathcal{M}_0, \mathcal{M}_1, \delta)$ be a test on $(\mathcal{X}^N, \mathcal{A}^{\otimes N})$. The set $\mathcal{R} \subset \mathcal{X}^N$, $\mathcal{R} = \delta^{-1}(\{1\})$, is called critical region (or region of rejection). Practically, since the statistic δ is completely defined by $\mathcal{R}$, ($\forall (x_1, \dots, x_N) \in \mathcal{X}^N$, $\delta(x_1, \dots, x_N) = 1$ if $(x_1, \dots, x_N) \in \mathcal{R}$, $\delta(x_1, \dots, x_N) = 0$ otherwise), a test will be equivalently considered a triplet $(\mathcal{M}_0, \mathcal{M}_1, \mathcal{R})$, specifying instead the critical region.

Generally, the critical region $\mathcal{R}$ leading to the rejection of the hypothesis H_0 is constructed from a **test statistic** $T : \mathcal{X}^N \rightarrow \mathbb{R}$, the distribution of which is known, and the rejection region will be in the form:

$$\mathcal{R} = \left\{ (x_1, \dots, x_N) \in \mathcal{X}^N : T(x_1, \dots, x_N) \in \mathcal{W} \right\}, \text{ avec } \mathcal{W} \subset \mathbb{R}.$$

We will later see several ways to determine the T test statistic.

3.1.2 Risk of a test

The problem of constructing a hypothesis test is closely related to the risk of false conclusion. Indeed, if $\mathcal{M}_0$ and $\mathcal{M}_1$ have the same support, any observation can be generated from any distribution of $\mathcal{M}_0$ and $\mathcal{M}_1$. Thus, in the decision to accept or reject the null hypothesis, there is always a probability of error, which should be controlled. In the test of H_0 versus H_1 , there are two types of error: type I error, which consists in rejecting H_0 when it is true (convicting an

innocent), and type II error, which consists in accepting H_0 when H_1 is true (releasing a guilty person).

Definition 3.4 Let $(\mathcal{M}_0, \mathcal{M}_1, \mathcal{R})$ be a hypothesis test regarding P^* on $(\mathcal{X}^N, \mathcal{A}^{\otimes N})$.

A type I risk is probability of rejection of a true null hypothesis, that is to say $\mathbb{P}((X_1, \dots, X_N) \in \mathcal{R})$ while $P^* \in \mathcal{M}_0$.

A type II risk is the probability of non-rejection of a false null hypothesis, that is to say $\mathbb{P}((X_1, \dots, X_N) \notin \mathcal{R})$ while $P^* \in \mathcal{M}_1$.

The power of the test is (1 - type II risk), that is to say the probability of rejecting a false H_0 .

The greater the power of the test, the more it is able to challenge the conservative assumption when it is false. We can see that there is a compromise to be found between type I risk and power.

In many cases, it is not possible to determine these risks because the true P^* law is unknown. Instead, we will calculate an upper bound of them. As mentioned above, the most serious error is the type I error. Also, in the construction of the test, once the assumptions H_0 and H_1 have been determined (i.e. $\mathcal{M}_0$ and $\mathcal{M}_1$ given) we will generally impose an upper bound α on the type I risk (typically $\alpha = 0.05$ or $\alpha = 0.01$) and the critical region will be constructed so that the type I risk remains below the threshold α . So we give α and look for a critical region $\mathcal{R}_\alpha$ such that:

$$\sup_{P^* \in \mathcal{M}_0} \mathbb{P}((X_1, \dots, X_N) \in \mathcal{R}_\alpha) \leq \alpha . \quad (3.1)$$

■ **Remark 3.3** If $X \sim P^*$ and $(X_1, \dots, X_N)$ is an i.i.d. sample, we denote:

$$\mathbb{P}((X_1, \dots, X_N) \in \mathcal{R}_\alpha) \text{ for } (P^*)^{\otimes N}(\mathcal{R}_\alpha) .$$

The notation $\mathbb{P}((X_1, \dots, X_N) \in \mathcal{R}_\alpha)$ is preferred here because it is more explicit. ■

As a test should be as powerful as possible (as discriminating as possible), it is appropriate to choose among the critical regions verifying the condition (3.1) of type I risk majoration, the one that ensures that $\mathbb{P}((X_1, \dots, X_N) \in \mathcal{R}_\alpha)$ is the greatest possible when $P^* \in \mathcal{M}_1$. In particular, if $\mathcal{R} \subset \mathcal{R}'$ and if for the two regions the majoration condition (3.1) is satisfied, then, as $\forall P^*, \mathbb{P}((X_1, \dots, X_N) \in \mathcal{R}) \leq \mathbb{P}((X_1, \dots, X_N) \in \mathcal{R}')$, the test associated with $\mathcal{R}'$ will be more powerful.

Definition 3.5 — Size and significance level of a test. For a test $(\mathcal{M}_0, \mathcal{M}_1, \mathcal{R})$ sur $(\mathcal{X}^N, \mathcal{A}^{\otimes N})$, we call size of the test

$$\alpha = \sup_{P^* \in \mathcal{M}_0} \mathbb{P}((X_1, \dots, X_N) \in \mathcal{R}) .$$

Any upper-bound of the test size will be called significance level of the test.

In practice, we will therefore often set a test significance level $\alpha \in]0; 1[$, and look for the most powerful test of that level (there is not necessarily a test of size α , especially if the distributions involved are discrete).

We can also not define a significance level α a priori, but rather consider a collection of tests of critical regions of different sizes, and see from the data, which tests lead to accept H_0 and which ones lead to reject H_0 . We use the notion of p -value (for probability-value).

Definition 3.6 — p -value. Let $\{(\mathcal{M}_0, \mathcal{M}_1, \mathcal{R}_t), t \in \mathcal{T}\}$ be a collection of tests on $(\mathcal{X}^N, \mathcal{A}^{\otimes N})$ indexed by $t \in \mathcal{T} \subset \mathbb{R}$, and α_t the size of the test of critical region $\mathcal{R}_t$. The p -value α^* is the statistic defined on $\mathcal{X}^N$ by:

$$\alpha^*(X_1, \dots, X_N) = \inf \{ \alpha_t, t \in \mathcal{T} : (X_1, \dots, X_N) \in \mathcal{R}_t \}$$

For a given sample of observations $x = (x_1, \dots, x_N)$, $\alpha^*(x)$ is therefore the lower bound of the test sizes for which we reject H_0 . If $\{ \alpha_t, t \in \mathcal{T} \}$ is an interval, it is also the upper bound of the test sizes for which we accept H_0 . (This is not necessarily the case if the law of test statistics is discrete, for example).

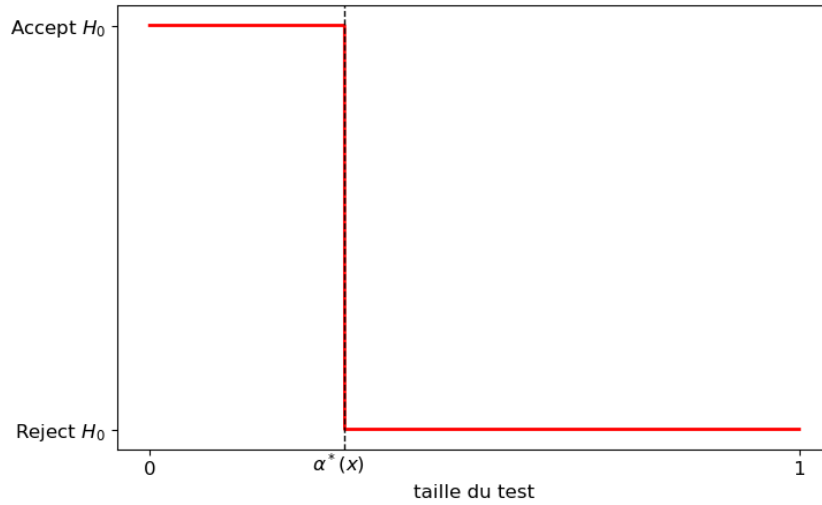


Figure 3.1: Acceptation / Rejection of H_0 as a function of the test size - $\alpha^*(x)$ denotes the p -value.

The p -value can then be interpreted as a measure of evidence against H_0 . The smaller α^* is, the stronger the evidence against H_0 (a low test level means that one must be very conservative not to reject H_0). In practice, we draw the conclusions described below based on the p -value:

p -valeur	évidence
$\alpha^*(x) < 0.01$	very strong evidence against H_0
$0.01 \leq \alpha^*(x) < 0.05$	strong evidence against H_0
$0.05 \leq \alpha^*(x) < 0.1$	weak evidence against H_0
$0.1 \leq \alpha^*(x)$	no real evidence against H_0

■ **Remark 3.4** In the classical test approach presented here, the p -value is not the probability that H_0 is true. H_0 is true with probability 1 or 0, because even if we don't know it, the unknown distribution P^* is either in $\mathcal{M}_0$ or in $\mathcal{M}_1$, it cannot be in both. Bayesian tests, on the contrary, provide a probabilistic framework giving the probability that H_0 is conditionally true given the x observation. We will not present these concepts here. ■

In the following, we will consider two particular and complementary types of tests: on the one hand parametric tests, for which we consider a given parametric model and the hypotheses of the tests relate to the values of parameters, and on the other hand goodness of fit tests, for which we check whether a distribution actually belongs to a parametric family.

3.2 Parametric Tests

We consider a parametric statistical model. $\mathcal{M}_\Theta = \{P_\theta, \theta \in \Theta\}$.

3.2.1 Parametric Tests and Power Function

Definition 3.7 — Parametric test. Let $\mathcal{M}_\Theta$ be a parametric statistical model. A parametric test is a statistical test $(\mathcal{M}_0, \mathcal{M}_1, \mathcal{R})$ such that there exist $\Theta_0 \subset \Theta$, $\Theta_1 \subset \Theta$, $\Theta_0 \cap \Theta_1 = \emptyset$ with $\mathcal{M}_0 = \{P_\theta, \theta \in \Theta_0\}$ and $\mathcal{M}_1 = \{P_\theta, \theta \in \Theta_1\}$. It will be noted then $(\Theta_0, \Theta_1, \mathcal{R})$.

In other words, for a parametric model $\mathcal{M}_\Theta$, and denoting P_θ the unknown distribution, a parametric test is a hypothesis test of the form:

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_1 \end{cases}$$

where Θ_0 and Θ_1 are two disjoint subsets of Θ .

■ Example 3.5

- A parametric test is said with **simple hypotheses** when Θ_0 and Θ_1 are reduced to singletons, that is to say:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{cases}$$

- Any other parametric test is said **composite** (or with composite hypotheses). A general example is

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_1 = \Theta_0^c, \end{cases}$$

where Θ_0^c is the complementary set of Θ_0 in Θ , and a frequent case is such that: $\Theta_0 = \{\theta_0\}$:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

If $\Theta \subset \mathbb{R}$, this last configuration corresponds to a **two-tailed** test, as opposed to **one-tailed** tests:

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}$$

where

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta < \theta_0. \end{cases}$$

■

Definition 3.8 — Power function. If $(\Theta_0, \Theta_1, \mathcal{R})$ is a parametric test on $(\mathcal{X}^N, \mathcal{A}^{\otimes N})$, then its power function π is defined $\forall \theta \in \Theta_0 \cup \Theta_1$ by:

$$\pi(\theta) = \mathbb{P}_\theta(X \in \mathcal{R}),$$

■ **Remark 3.6** The test size α is thus given by:

$$\alpha = \sup_{\theta \in \Theta_0} \pi(\theta)$$

In the special case of simple hypothesis testing, $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$, we get:

- (i) $\alpha = \pi(\theta_0)$ is the test size, that is to say the type I risk,
- (ii) $\beta = 1 - \pi(\theta_1)$ is the type II risk,
- (iii) $\pi(\theta_1)$ is the power of the test.

■

In the case of tests with simple hypotheses, it is then easy to compare the power of the tests.

Definition 3.9 Let $\mathcal{T} = (\Theta_0, \Theta_1, \mathcal{R})$ and $\mathcal{T}' = (\Theta_0, \Theta_1, \mathcal{R}')$ be two parametric tests both confronting the same hypotheses and their respective power functions π and π' .

If $\Theta_1 = \{\theta_1\}$: $\mathcal{T}$ is more powerful than $\mathcal{T}'$ if $\pi(\theta_1) > \pi'(\theta_1)$

In the general case (Θ_1 not necessarily a singleton): $\mathcal{T}$ is said uniformly more powerful than $\mathcal{T}'$ if $\pi(\theta) > \pi'(\theta)$, $\forall \theta \in \Theta_1$.

3.2.2 Method to Build Critical Regions

Let $\mathcal{M}_\Theta$ be a parametric statistical model, and suppose defined the unknown distribution P_θ , $P_\theta \in \mathcal{M}_\Theta$ and the two alternative hypotheses:

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_1 \end{cases}$$

with $\Theta_0 \subset \Theta$, $\Theta_1 \subset \Theta$ et $\Theta_0 \cap \Theta_1 = \emptyset$.

Exact Method

Consider an i.i.d. sample of size N for P_θ . Our objective is to determine the critical region $\mathcal{R} \subset \mathcal{X}^N$ of the test. Generally, the test size α is given, and we look for a test statistic $T : \mathcal{X}^N \rightarrow \mathbb{R}$ whose distribution is known la loi, and the critical region will be obtained as:

$$\mathcal{R} = \{(x_1, \dots, x_N) \in \mathcal{X}^N : T(x_1, \dots, x_N) \in \mathcal{W}\}, \text{ avec } \mathcal{W} \subset \mathbb{R}. \quad (3.2)$$

such that

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta((X_1, \dots, X_N) \in \mathcal{R}) \leq \alpha.$$

Among all the critical regions verifying constraint (3.2), we will look for the one that maximizes the power.

■ **Example 3.7** We consider a Gaussian statistical model $\mathcal{M} = \{\mathcal{N}(\theta, 1) : \theta \in \mathbb{R}\}$ for which the variance is given, equal to 1. Let $P_\theta \in \mathcal{M}$ and $X = (X_1, \dots, X_N)$, an i.i.d. sample for P_θ . We test the following hypotheses:

$$\begin{cases} H_0 : \theta = -1 \\ H_1 : \theta = 1 \end{cases}$$

We are in a simple hypothesis test. Since we know that the maximum likelihood estimator of θ is the empirical mean $\bar{X}$, we propose a test based on it and it seems natural to consider a rejection region of the type:

$$\mathcal{R} = \left\{ (x_1, \dots, x_N) \in \mathbb{R}^N : \frac{1}{N} \sum_{i=1}^N x_i > x_0 \right\}$$

with $x_0 \in \mathbb{R}$. Le test consiste donc à faire la moyenne des observations et à rejeter l'hypothèse $\theta = -1$ au profit de $\theta = 1$ si cette moyenne dépasse un certain seuil x_0 .

A reasonable value for x_0 is the midpoint between θ_0 and θ_1 , that is to say 0. Let's calculate the type I risk α and the power π associated with this test:

$$\alpha = \mathbb{P}_{\theta=-1}((X_1, \dots, X_N) \in \mathcal{R}) = \mathbb{P}_{\theta=-1}(\bar{X} > 0)$$

and

$$\pi = \mathbb{P}_{\theta=1}((X_1, \dots, X_N) \in \mathcal{R}) = \mathbb{P}_{\theta=1}(\bar{X} > 0)$$

We must evaluate the probability of the same event but under different distributions. In the first case, under hypothesis H_0 , we have $\bar{X} \sim \mathcal{N}(-1, 1/N)$, that is to say $\sqrt{N}(\bar{X} + 1) \sim \mathcal{N}(0, 1)$ and in the second case, under hypothesis H_1 , we have $\bar{X} \sim \mathcal{N}(1, 1/N)$, and $\sqrt{N}(\bar{X} - 1) \sim \mathcal{N}(0, 1)$. The two distributions are represented (for $N = 1$) in Figure 3.2. We denote Φ the cumulative distribution function of the standard Gaussian distribution $\mathcal{N}(0, 1)$ and we have:

$$\alpha = \mathbb{P}_{\theta=-1}\{\sqrt{N}(\bar{X} + 1) > \sqrt{N}\} = 1 - \Phi(\sqrt{N})$$

$$\pi = \mathbb{P}_{\theta=1}\{\sqrt{N}(\bar{X} - 1) > -\sqrt{N}\} = 1 - \Phi(-\sqrt{N}) = \Phi(\sqrt{N})$$

We notice that if the sample size N tends towards infinity, the type I risk tends towards 0 and the power tends towards 1. But for $N = 1$ for example, we find $\alpha = 0.159$, which is generally too great a risk. By taking a threshold $x_0 = 0.5$, which gives the risks represented by the Figure 3.2, we then have: $\alpha = 0.068$, which is already more acceptable.

If we set the test size at $\alpha = \alpha_0$, a critical region $\mathcal{R}_{\alpha_0}$ is then defined by choosing a rejection threshold c_α such that $\alpha_0 = \mathbb{P}_{\theta=-1}(\bar{X} > c_\alpha) = \mathbb{P}_{\theta=-1}(\sqrt{N}(\bar{X} + 1) > \sqrt{N}(c_\alpha + 1)) = 1 - \Phi(\sqrt{N}(c_\alpha + 1))$. We can deduce that

$$c_\alpha = -1 + \frac{1}{\sqrt{N}} q_{1-\alpha_0}$$

where q_u here designates the u -quantile of the standard normal distribution, $q_u = \Phi^{-1}(u)$. For $N = 1$, we find $c_\alpha = 0.644$. ■

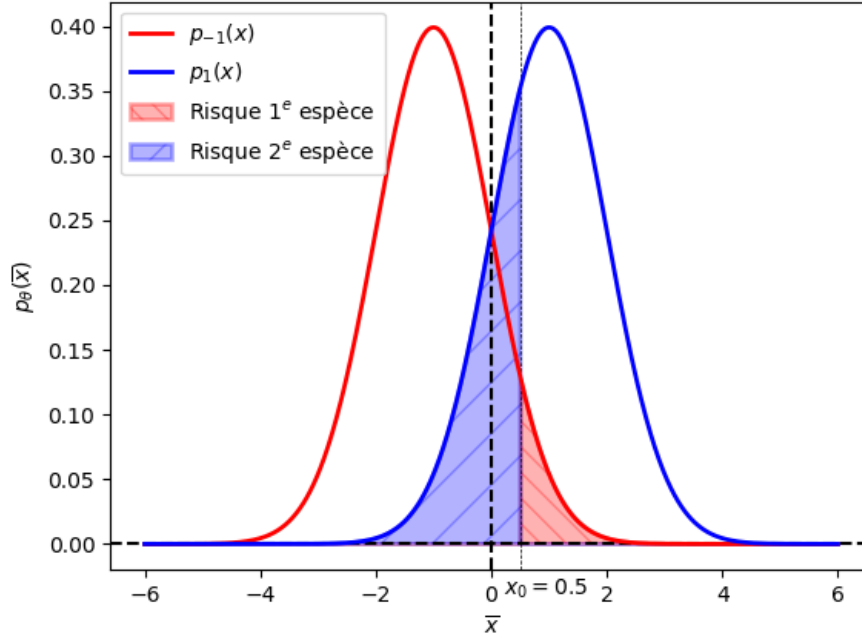


Figure 3.2: Test of simple hypotheses, Example 3.7. We show the density function of the test statistic under hypotheses H_0 and H_1 . The rejection rule ($\bar{X} > 0.5$) leads to two types of errors whose probabilities under H_0 and under H_1 are respectively represented here by the areas hatched in red (type I risk) and blue (type II risk). It can be seen that these two quantities vary in opposite directions when the threshold defining the rejection region shifts, they cannot be minimized simultaneously.

Asymptotic Method

In the previous example, the distribution quantiles of the statistic $T(X)$ were seen to be involved in the calibration of thresholds specifying the region of rejection. If the exact law of $T(X)$ is not known, its limit distribution can be used. The use of the quantiles of the limit distribution instead of the exact one defines an approximation of the critical region of the test. This is referred to as an asymptotic test.

Likelihood ratio test and Neyman-Pearson method

We consider the case of a simple hypothesis test on the N -sample $(X_1, \dots, X_N)$ i.i.d. for the unknown law P_θ in a dominated parametric model (P_θ admits a density p_θ).

$$\begin{cases} H_0 & : \theta = \theta_0 \\ H_1 & : \theta = \theta_1 \end{cases}$$

Definition 3.10 A likelihood ratio test of significance level α is a test that uses the test statistic:

$$T^*(X_1, \dots, X_N) = \frac{\mathcal{L}(\theta_1; X_1, \dots, X_N)}{\mathcal{L}(\theta_0; X_1, \dots, X_N)}$$

and defines the critical region by

$$\mathcal{R}_\alpha^* = \left\{ (x_1, \dots, x_N) \in \mathcal{X}^N, T^*(x_1, \dots, x_N) > c_\alpha \right\},$$

where c_α is chosen such that $\mathbb{P}_{\theta_0}((X_1, \dots, X_N) \in \mathcal{R}_\alpha^*) \leq \alpha$.

Theorem 3.1 — Neyman-Pearson Lemma. The likelihood ratio test of size α is the most powerful test of all tests of significance level α .

Preuve : Consider the sample $X = (X_1, \dots, X_N)$ and its realization $x = (x_1, \dots, x_N)$. We know that $\mathcal{L}(\theta; x) = p_\theta(x) = \prod_{i=1}^N p_\theta(x_i)$.

Let $\mathcal{R}^*$ be the region of rejection associated with the likelihood ratio test...la région de rejet associée au test du rapport de vraisemblance, $\mathcal{R}^* = \{x \in \mathbb{R}^N : T^*(x) > c_\alpha\}$ de taille α . On veut montrer que, pour toute région de rejet $\mathcal{R}$ associée à un test de niveau α , on a:

$$\mathbb{P}_{\theta_1}(X \in \mathcal{R}) \leq \mathbb{P}_{\theta_1}(X \in \mathcal{R}^*).$$

Or

$$\mathbb{P}_\theta(X \in \mathcal{R}) = \int_{\mathcal{R}} p_\theta(x) dx.$$

On the other hand, by definition of the likelihood ratio test, if $x \in \mathcal{R}^*$ then $p_{\theta_1}(x) > c_\alpha p_{\theta_0}(x)$. The difference can then be computed:

$$\begin{aligned} \mathbb{P}_{\theta_1}(X \in \mathcal{R}^*) - \mathbb{P}_{\theta_1}(X \in \mathcal{R}) &= \int_{\mathcal{R}^*} p_{\theta_1}(x) dx - \int_{\mathcal{R}} p_{\theta_1}(x) dx \\ &= \int_{\mathcal{R}^* \setminus \mathcal{R}} p_{\theta_1}(x) dx - \int_{\mathcal{R} \setminus \mathcal{R}^*} p_{\theta_1}(x) dx \end{aligned}$$

$\mathcal{R}^* \setminus \mathcal{R} \subset \mathcal{R}^*$, so $p_{\theta_1}(x) > c_\alpha p_{\theta_0}(x)$ on $\mathcal{R}^* \setminus \mathcal{R}$, while $\mathcal{R} \setminus \mathcal{R}^* \cap \mathcal{R}^* = \emptyset$, and thus $p_{\theta_1}(x) \leq c_\alpha p_{\theta_0}(x)$ on $\mathcal{R} \setminus \mathcal{R}^*$. Therefore:

$$\begin{aligned} \mathbb{P}_{\theta_1}(X \in \mathcal{R}^*) - \mathbb{P}_{\theta_1}(X \in \mathcal{R}) &\geq c_\alpha \left(\int_{\mathcal{R}^* \setminus \mathcal{R}} p_{\theta_0}(x) dx - \int_{\mathcal{R} \setminus \mathcal{R}^*} p_{\theta_0}(x) dx \right) \\ &= c_\alpha \left(\int_{\mathcal{R}^*} p_{\theta_0}(x) dx - \int_{\mathcal{R}} p_{\theta_0}(x) dx \right) \\ &= c_\alpha (\mathbb{P}_{\theta_0}(X \in \mathcal{R}^*) - \mathbb{P}_{\theta_0}(X \in \mathcal{R})) \geq 0 \end{aligned}$$

where we used the fact that $\mathbb{P}_{\theta_0}(X \in \mathcal{R}^*) = \alpha$ (likelihood ratio test of size α) and $\mathbb{P}_{\theta_0}(X \in \mathcal{R}) \leq \alpha$ (test of significance level α). $\square$

■ **Example 3.8** To illustrate this result, we use the Gaussian statistical model:

$$\mathcal{M} = \{\mathcal{N}(\theta, 1) : \theta \in \mathbb{R}\}$$

studied in Example 3.7 and the test of simple hypotheses: $H_0 : \theta = \theta_0$, versus $H_1 : \theta = \theta_1$. Neyman-Pearson lemma applies. The likelihood ratio is then

$$T^*(X_1, \dots, X_N) = \frac{\prod_{i=1}^N p_{\theta_1}(X_i)}{\prod_{i=1}^N p_{\theta_0}(X_i)} = \exp \left(-\frac{1}{2} \sum_{i=1}^N (X_i - \theta_1)^2 + \frac{1}{2} \sum_{i=1}^N (X_i - \theta_0)^2 \right)$$

that is to say after simplifying

$$T^*(X_1, \dots, X_N) = \exp \left(N(\theta_1 - \theta_0) \bar{X} \right) \exp \left(\frac{N}{2} (\theta_0^2 - \theta_1^2) \right)$$

The exponential is a strictly increasing function. Two cases must be considered: if $\theta_1 > \theta_0$, we have $T^*(X_1, \dots, X_N) > c$ if and only if $\bar{X} > c'$, for a given c' , and if $\theta_1 < \theta_0$, we have $T^*(X_1, \dots, X_N) > c$ if and only if $\bar{X} < c''$, for a given c'' .

We therefore find a test based on the empirical mean, whose threshold only needs to be calibrated to be α . We now know that this natural test is the most powerful test for these hypotheses. ■

3.2.3 p -value for Parametric Tests

In the previous section, we placed ourselves in the situation where we gave ourselves the level of the test, and constructed a rejection region. Now suppose that we have defined a test statistic, but that we do not want to set the test level a priori, but determine the p -value from the data.

Theorem 3.2 Consider a collection of tests $\{(\mathcal{M}_0, \mathcal{M}_1, \mathcal{R}_t), t \in I\}$ where I is an interval in $\mathbb{R}$ such that $\forall t \in I$ the critical region $\mathcal{R}_t$ is of the following form:

$$\mathcal{R}_t = \left\{ (x_1, \dots, x_N) \in \mathcal{X}^N : T(x_1, \dots, x_N) > t \right\},$$

for a test statistic T on $\mathcal{X}^N$.

Let y be an observation sample of size N such that $T(y) \in I$. Then, the p -value $\alpha^*(y)$ is given by:

$$\alpha^*(y) = \sup_{\theta \in \Theta_0} \{P_\theta(T(X_1, \dots, X_N) \geq T(y))\}.$$

If $\Theta_0 = \{\theta_0\}$,

$$\alpha^*(y) = P_{\theta_0}(T(X_1, \dots, X_N) \geq T(y)).$$

Preuve : We consider directly the simplified case where $\Theta_0 = \{\theta_0\}$.

For $t \in I$, let's denote α_t the size of the test associated with $\mathcal{R}_t$.

If $t \geq T(y)$: then $y \notin \mathcal{R}_t$ and

$$\alpha_t = P_{\theta_0}(X \in \mathcal{R}_t) = P_{\theta_0}(T(X) > t) \leq P_{\theta_0}(T(X) \geq T(y)).$$

If $t < T(y)$: then $y \in \mathcal{R}_t$ and

$$\alpha_t = P_{\theta_0}(X \in \mathcal{R}_t) = P_{\theta_0}(T(X) > t) \geq P_{\theta_0}(T(X) \geq T(y)).$$

Moreover,

$$\lim_{t \rightarrow T(y)^-} P_{\theta_0}(T(X) > t) = 1 - \lim_{t \rightarrow T(y)^-} P_{\theta_0}(T(X) \leq t) = 1 - P_{\theta_0}(T(X) < T(y)) = P_{\theta_0}(T(X) \geq T(y)) .$$

We thus have that

$$\inf \{ \alpha_t, t \in I : y \notin \mathcal{R}_t \} = \inf \{ P_{\theta_0}(T(X) > t), t < T(y) \} = P_{\theta_0}(T(X_1, \dots, X_N) \geq T(y)) .$$

□

■ **Example 3.9** We get back to Example 3.7: in the Gaussian statistical model $\mathcal{M} = \{\mathcal{N}(\theta, 1) : \theta \in \mathbb{R}\}$ where the variance is given, equal to 1. We want to test the hypotheses:

$$\begin{cases} H_0 : \theta = -1 \\ H_1 : \theta = 1 \end{cases}$$

We defined $\mathcal{R}_\alpha = \{x \in \mathbb{R}^N : \bar{x} > c_\alpha\}$ with

$$c_\alpha = -1 + \frac{1}{\sqrt{N}} q_{1-\alpha}$$

and reciprocally:

$$\alpha = P_{\theta=-1}(\bar{X} > c_\alpha) = 1 - \Phi(\sqrt{N}(c_\alpha + 1))$$

We thus have a collection of tests as in the hypothesis of Theorem 3.2. For a given sample of observations y , we thus have that the p -value $\alpha^*(y)$ is given by:

$$\alpha^*(y) = P_{\theta=-1}(\bar{X} \geq \bar{y}) = 1 - \Phi(\sqrt{N}(\bar{y} + 1))$$

Thus, we move from acceptance to rejection as soon as the size of the test exceeds the value of $\alpha = \alpha^*(y)$ and we reject H_0 for all $\alpha > \alpha^*(y)$. ■

3.2.4 Pearson's χ^2 Test for the Multinomial Model

The multinomial model is considered here. The multinomial distribution of parameters (N, p) corresponds to the following experiment: we consider an urn with balls of k different colors in proportions $p = (p_1, \dots, p_k)$. We make N draws with replacement and note the number of X_i of ball draws of each color in the vector $X = (X_1, \dots, X_k)$, we have $\sum_{i=1}^k X_i = N$.

Definition 3.11 — Multinomial distribution. Let $N \in \mathbb{N}^*$, $p \in]0; 1[^k$ such that $\sum_{i=1}^k p_i = 1$. We call multinomial distribution of parameters (N, p) , the probability law on $\{0; 1; \dots; N\}^k$ defined by the mass function:

$$P(x_1, \dots, x_k) = \frac{N!}{\prod_{i=1}^k x_i!} \prod_{i=1}^k p_i^{x_i},$$

for all $(x_1, \dots, x_k) \in \{0; 1; \dots; N\}^k$ such that $\sum_{i=1}^k x_i = N$. We denote: $X \sim M(N, p)$.

■ **Remark 3.10** The i^{th} marginal distribution of a multinomial variable $M(N, p)$ is a binomial distribution $B(N, p_i)$. ■

For the multinomial model, we introduce the Pearson statistic.

Definition 3.12 — Pearson Statistic. Let $X \sim M(N, p)$, $p \in]0; 1[^k$, $\sum_{i=1}^k p_i = 1$. The Pearson statistic T is defined by:

$$T(X) = \sum_{i=1}^k \frac{(X_i - Np_i)^2}{Np_i}.$$

T is a relative measure of the difference between actual and theoretical numbers of elements in each category.

Proposition 3.3 Let $X \sim M(N, p)$, $p \in]0; 1[^k$, $\sum_{i=1}^k p_i = 1$ and $T(X)$, the Pearson statistic. Then, the limit distribution of $T(X)$ is a chi-squared distribution with $k - 1$ degrees of freedom:

$$T(X) = \sum_{i=1}^k \frac{(X_i - Np_i)^2}{Np_i} \xrightarrow{\mathcal{L}} \chi^2(k - 1)$$

Preuve : We give a sketch of the proof below and refer to (Lehmann and Romano, 2006) for details. First, it requires the use of the Multidimensional Central Limit Theorem, which gives us:

$$\frac{1}{\sqrt{N}} \begin{pmatrix} \frac{X_1 - Np_1}{\sqrt{p_1}} \\ \vdots \\ \frac{X_k - Np_k}{\sqrt{p_k}} \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}(0, I - \sqrt{p}\sqrt{p}^T)$$

The result is established by noting that the vector X is the sum of i.i.d. vectors whose mean and covariance matrix can be calculated. The proposition is then a consequence of the multidimensional TCL and Cochran's theorem. The limit Gaussian covariance matrix is the matrix of the orthogonal projector on the orthogonal of the vector $\sqrt{p}$ thus on a subspace of dimension $k - 1$. The norm of the limit Gaussian vector is thus distributed as a $\chi^2(k - 1)$ distribution. We conclude by using the continuous mapping theorem with the Euclidean norm. $\square$

An asymptotic test procedure can then be implemented for the multidimensional model.

Corollary 3.4 — Pearson's chi-squared test for the multidimensional model. Let $X \sim M(N, p)$, N and p given. Let $p_0 \in]0; 1[^k$ such that $\sum_{i=1}^k p_{0i} = 1$. We consider the test:

$$\begin{cases} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{cases}$$

Then, under H_0 ,

$$T(X) = \sum_{i=1}^k \frac{(X_i - Np_{0i})^2}{Np_{0i}} \xrightarrow{\mathcal{L}} \chi^2(k - 1).$$

For a given test size α , we deduce the asymptotic region of rejection:

$$\mathcal{R}_\alpha = \left\{ x : T(x) > q_{1-\alpha}^{\chi^2(k-1)} \right\}$$

with $q_{1-\alpha}^{\chi^2(k-1)}$ the $(1 - \alpha)$ -quantile for the $\chi^2(k - 1)$ distribution. We have:

$$P_{p_0}(X \in \mathcal{R}_\alpha) \xrightarrow{N \rightarrow +\infty} \alpha.$$

Note that we choose the one-tailed region of rejection for large $T(X)$, because T measures the difference between realized and theoretical enrolments under H_0 , and the smaller this measure is, the more H_0 should be preferred.

3.3 Goodness of Fit Tests

Goodness of fit tests (Test for fit of a distribution) belong to the family of non-parametric tests. Let $(X_1, \dots, X_N)$, be an i.i.d sample for an unknown distribution P^* . In this part, we want to check if P^* belongs to a specific parametric family $\mathcal{M}_\Theta = \{P_\theta, \theta \in \Theta\}$:

$$\begin{cases} H_0 : P^* \in \mathcal{M}_\Theta \\ H_1 : P^* \notin \mathcal{M}_\Theta \end{cases}$$

In practice, this type of test generally considers first an estimation phase: for a given sample $(x_1, \dots, x_N)$, among all the P_θ , one chooses the one that maximizes $P_\theta(x_1, \dots, x_N)$. Then, by taking $P_{\hat{\theta}(x_1, \dots, x_N)}$, with $\hat{\theta}(x_1, \dots, x_N)$ the maximum likelihood estimate of θ , we are back to a test of the type:

$$\begin{cases} H_0 : P^* = P_{\theta_0} \\ H_1 : P^* \neq P_{\theta_0} \end{cases}$$

Note that despite this particular form, it is not a parametric test strictly speaking, because the alternative hypothesis does not assume that $P^* \in \mathcal{M}_\Theta$ as would be the case for a parametric test.

3.3.1 χ^2 -Test

The χ^2 goodness of fit test adapts the test defined for the multinomial distribution presented in Section 3.2.4.

Let $\mathcal{M}_\Theta$ be a parametric model, and $\theta_0 \in \Theta$, the χ^2 -test allows to test:

$$\begin{cases} H_0 : P^* = P_{\theta_0} \\ H_1 : P^* \neq P_{\theta_0} \end{cases}$$

Definition 3.13 — Generalized Pearson statistic. Let $(X_1, \dots, X_N)$ be an i.i.d. sample of distribution P^* .

Let $(I_1, \dots, I_k)$ be a partition of the support of P^* such that $\forall j, 1 \leq j \leq k, P^*(I_j) > 0$. Let's denote $p_j = P^*(I_j)$ and $p = (p_1, \dots, p_k)$.

We consider $Y = (Y_1, \dots, Y_k)$, the statistic taking values in $\{0; 1; \dots; N\}^k$ and counting the

number of X_i in every I_j , $1 \leq j \leq k$:

$$Y_j(X_1, \dots, X_N) = \sum_{i=1}^N \mathbb{I}_{I_j}(X_i) .$$

Then $Y \sim M(N, p)$ and

$$\tilde{T}(X_1, \dots, X_N) := T(Y(X_1, \dots, X_N)) = \sum_{i=1}^k \frac{(Y_i - Np_i)^2}{Np_i}$$

is the generalized Pearson statistic.

Under hypothesis H_0 , $Y \sim M(N, p_0)$ with $p_{0j} = P_{\theta_0}(X_1 \in I_j)$, $\forall 1 \leq j \leq k$. And thus:

$$T(Y(X_1, \dots, X_N)) = \sum_{i=1}^k \frac{(Y_i - Np_{0i})^2}{Np_{0i}} \xrightarrow{\mathcal{L}} \chi^2(k-1) .$$

For a given test level α , we form the asymptotic region of rejection:

$$\mathcal{R}_\alpha = \left\{ (x_1, \dots, x_N) : T(Y(x_1, \dots, x_N)) > q_{1-\alpha}^{\chi^2(k-1)} \right\} ,$$

with $q_{1-\alpha}^{\chi^2(k-1)}$ the $(1 - \alpha)$ -quantile of $\chi^2(k-1)$.

We have:

$$\lim_{N \rightarrow \infty} P_{\theta_0}((X_1, \dots, X_N) \in \mathcal{R}_\alpha) = \alpha$$

Note again that since the Pearson statistic measures the difference between observed class sizes and theoretical class sizes under H_0 , the rejection region is one-tailed: H_0 is rejected when the differences are too large.

■ **Remark 3.11** As underlined, the χ^2 -test is often preceded by an estimation phase of θ_0 . If Θ is of dimension q , q degrees of freedom are thus lost, and we have:

$$T(Y(X_1, \dots, X_N)) = \sum_{i=1}^k \frac{(Y_i - Np_{0i})^2}{Np_{0i}} \xrightarrow{\mathcal{L}} \chi^2(k-1-q) .$$

■

■ **Remark 3.12** We consider that the asymptotic approximation is valid as soon as we have at least $Np_{0i} \geq 5$ for all classes. If this is not the case, some classes can be grouped together so that the asymptotic approximation is more realistic. ■

3.3.2 Goodness of Fit Test of Kolmogorov-Smirnov

For an i.i.d. sample $X_1, \dots, X_N$ of distribution P^* , we consider a test of the form:

$$\begin{cases} H_0 : P^* = P_0 \\ H_1 : P^* \neq P_0 \end{cases}$$

where P_0 is a distribution whose cumulative function F_0 is known.

Let $\hat{F}_N$ be the empirical cumulative distribution function computed from the sample. We recall that:

$$\hat{F}(x) = \frac{1}{N} \sum_{1 \leq i \leq N} \mathbb{I}_{]-\infty; X_i]}(x) .$$

The Kolmogorov-Smirnov test consists in rejecting the null hypothesis H_0 as soon as the statistic:

$$D_N = \sup_{x \in \mathbb{R}} |\hat{F}_N(x) - F_0(x)|$$

exceeds a certain threshold.

This threshold can be calculated numerically since the limit law of $\sqrt{N}D_N$ under the null hypothesis is known. Indeed, we have:

$$\lim_{N \rightarrow \infty} \mathbb{P} \left(\sqrt{N} \sup_{x \in \mathbb{R}} |\hat{F}_N(x) - F_0(x)| \leq y \right) = \left(\sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2 y^2} \right) \mathbb{I}_{[0; +\infty[}(y) . \quad (3.3)$$

that is to say

$$\sqrt{N}D_N \xrightarrow{\mathcal{L}} \mathcal{K}$$

with $\mathcal{K}$, the Kolmogorov-Smirnov cumulative distribution function given by the second member of the equation (3.3). This distribution is tabulated and accessible in the different statistical software.

■ **Remark 3.13** Effectively, for the calculation of D_N , as $\hat{F}_N$ is piecewise constant and F_0 is increasing and continuous, it is enough to look for the maximum among a finite number of $2N$ values:

$$D_N = \max \left\{ \left| \frac{i}{N} - F_0(X_{(i)}) \right|, \left| \frac{i-1}{N} - F_0(X_{(i)}) \right| : i \in \{1, \dots, N\} \right\}$$

where $X_{(i)}$ is the statistic which to the sample $(X_1, \dots, X_N)$ maps the i^{th} element for the order relationship, $X_{(1)} \leq \dots \leq X_{(i)} \leq X_{(i+1)} \leq \dots \leq X_{(N)}$. ■

■ **Example 3.14** Consider the following sample $x = (x_1, \dots, x_{10})$:

$$6.42 ; 5.12 ; 2.99 ; 2.00 ; 9.79 ; 5.78 ; 5.54 ; 6.44 ; 8.92 ; 6.19 .$$

We wish to test the fit to the $\mathcal{N}(5.2^2)$ Gaussian distribution. We compute the value of the D_{10} statistic in the sample and we find $D_{10} \approx 0.325$ (cf. Figure 3.14), that is to say $\sqrt{10}D_{10} \approx 1.027$. By using the asymptotic approximation (even if $N = 10$ remains pretty low), we determine the p -value $\alpha^*(x)$, obtained by $\alpha^*(x) = \mathbb{P}(\mathcal{K} > 1.027) = 0.243$. There is no evidence against H_0 . Note that if we perform the test with the Gamma distribution $\gamma(5.1)$, we find $D_{10} = 0.381$ and a p -value of 0.110. For the classical test levels (0.01, 0.05, 0.1), we cannot reject H_0 , whereas in practice the sample has actually been simulated with a $\mathcal{N}(5.2^2)$ distribution. We need a N large enough for the goodness of fit test with a $\gamma(5.1)$ to be rejected. This is due to the proximity of the cumulative distribution functions of $\mathcal{N}(5.2^2)$ and $\gamma(5.1)$, as shown in Figure 3.14. ■

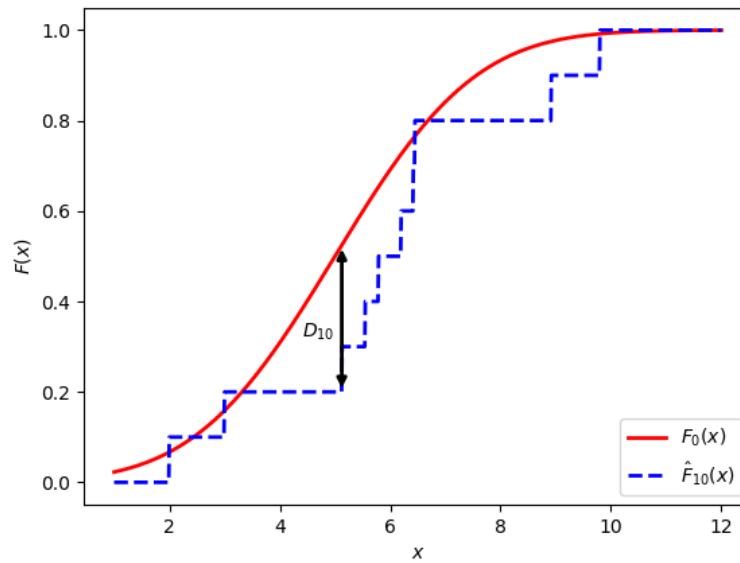


Figure 3.3: Goodness of fit test of the observation sample $\mathcal{N}(5, 2^2)$: the Kolmogorov-Smirnov distance is represented by the arrow.

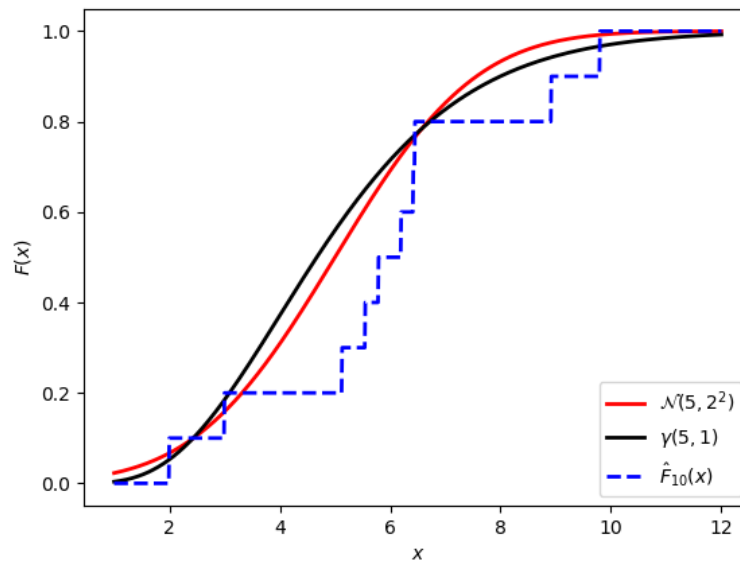


Figure 3.4: Comparison of cumulative distribution functions for $\mathcal{N}(5, 2^2)$ and $\gamma(5, 1)$.

Part II

Statistical Learning

4. Supervised learning

Let $(\mathcal{X}, \mathcal{A})$, $(\mathcal{Y}, \mathcal{B})$ be two measurable spaces.

In this chapter, we look at the problem of predicting a variable $y \in \mathcal{Y}$ from a variable $x \in \mathcal{X}$, that is, determining a **decision function** $h : \mathcal{X} \rightarrow \mathcal{Y}$, such that $y = h(x)$.

x is generally multidimensional, we will speak of explanatory variables (or covariates, or predictors...) and y is(are) the explained variable(s). If $\mathcal{Y}$ is a continuous set, we talk about a regression problem, if $\mathcal{Y}$ is a finite set (for example $\{0; 1\}$), we talk about a classification problem. Explanatory variables can also be continuous or discrete, or a combination of both.

■ Example 4.1

- We can try to automatically predict from the gray levels of an image pixels representing a handwritten digit which digit it is (LeCun et al., 1990).
- We can try to predict the survival time of a cancer patient based on clinical variables (such as age, smoking, treatment, etc.), genomic variables (such as the level of expression of genes), imaging variables (corresponding to radiology images of the affected organ)... (Zhu et al., 2017).

■

To carry out this prediction task, we have N observations of pairs of explanatory variables - explained variables $\{(x_i, y_i)\}_{1 \leq i \leq N}$: these are the data. The learning task will be to determine the decision function h from these data.

We will consider this prediction problem in a probabilistic framework.

4.1 Probabilistic Framework of Statistical Learning

We consider two random variables X and Y with values respectively in $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$, such that (X, Y) admits a density with respect to a product measure on $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \otimes \mathcal{B})$.

4.1.1 Regression Model, Classification Model

For all $x \in \mathcal{X}$ such that the marginal density $p_X(x) \neq 0$, we recall that the conditional density can be defined: $p_{Y|X=x}(y) = p(x, y)/p_X(x)$ (see Section 1.2.3).

Definition 4.1 — Statistical regression model. Under the above assumptions, we call a statistical regression model a family of conditional probability distributions given X . In particular, if the family is parameterized and the conditional distributions admit densities, we will have:

$$\mathcal{M}_\Theta = \{p_\theta(y|x), \theta \in \Theta, x \in \mathcal{X}\}$$

where we use the notation $p_{Y|X=x}(y) = p(y|x)$.

■ **Remark 4.2** This definition is general, regardless of the type of explained variable. Nevertheless, as mentioned previously, the usage differentiates **regression model** when Y is a continuous random variable from **classification model** when Y is a discrete random variable, with values in a finite set. The values taken by Y are then called **labels**. If the mathematical framework is identical, we will see that the methods differ. ■

The determination of the decision function h is then based on the determination of a family of conditional laws, and learning from the data set $\{(x_i, y_i)\}_{1 \leq i \leq N}$ is then a problem of statistical inference. If the model is parametric, we're faced with an estimation problem. We look for $\hat{\theta}$ such that the conditional distribution $p_{Y|X}$ (family of conditional probability distributions $\{p_{Y|X=x}(y), x \in \mathcal{X}\}$) is given by the family $\{p_{\hat{\theta}}(y|x), x \in \mathcal{X}\}$.

4.1.2 Decision Theory

As with the Bayesian estimate, the conditional distribution may not be sufficient, and in order to make a decision, a point value may be needed. For $x \in \mathcal{X}$, how to choose $h(x)$ according to $p(y|x)$?

We consider a loss function $L(y, h(x))$ which measures the difference between the true value and the predicted value and depends on the intended application. For example, for regression problems, we often choose $L(y, h(x)) = \|y - h(x)\|^2$, which will strongly penalize large errors. $L(y, h(x)) = \|y - h(x)\|$ will rather consider all errors in the same way.

We then look for h^* which minimizes the expectation of the loss function, that is to say:

$$h^* = \arg \min_h \int_{\mathcal{X} \otimes \mathcal{Y}} L(y, h(x)) p(x, y) dx dy.$$

Theorem 4.1 Let $x \in \mathcal{X}$. For the loss function $L(y, h(x)) = \|y - h(x)\|^2$, the prediction function h^* which minimizes the expectation of the loss function is given by:

$$h^*(x) = \mathbb{E}(Y|X = x) .$$

Preuve : Let us denote $\alpha(x) = \mathbb{E}(Y|X = x)$:

$$\begin{aligned} \int_{\mathcal{X} \otimes \mathcal{Y}} L(y, h(x)) p(x, y) dx dy &= \int_{\mathcal{X}} \int_{\mathcal{Y}} \|y - h(x)\|^2 p(y|x) dy p(x) dx \\ &= \int_{\mathcal{X}} \int_{\mathcal{Y}} [\|y - \alpha(x)\|^2 + \|\alpha(x) - h(x)\|^2 + 2(y - \alpha(x), \alpha(x) - h(x))] p(y|x) dy p(x) dx . \end{aligned}$$

Since:

- (i) $\int_{\mathcal{Y}} \|y - \alpha(x)\|^2 p(y|x) dy$ does not depend on h ,
- (ii) $\int_{\mathcal{Y}} \|\alpha(x) - h(x)\|^2 p(y|x) dy = \|\alpha(x) - h(x)\|^2$,
- (iii) $\int_{\mathcal{Y}} 2(y - \alpha(x), \alpha(x) - h(x)) p(y|x) dy = 0$ because $\int_{\mathcal{Y}} yp(y|x) dy = \mathbb{E}(Y|X = x) = \alpha(x)$.

Minimizing $\int_{\mathcal{X} \otimes \mathcal{Y}} L(y, h(x)) p(x, y) dx dy$ is equivalent to minimize $\int_{\mathcal{X}} \|\alpha(x) - h(x)\|^2 p(x) dx$, and the minimum is attained for $h^*(x) = \alpha(x) = \mathbb{E}(Y|X = x)$. $\square$

■ **Remark 4.3** Thus, if the objective is essentially that of prediction, one can directly seek to approximate the conditional expectation simply from the data, without going through the construction of a probabilistic model and its inference. ■

■ **Example 4.4 — k -nearest neighbours.** An elementary way is to search for the nearest x_i in the training data for a given x , and define $h(x) = y_i$. In order to smooth things out, one can also instead choose the nearest k neighbors with $k \geq 1$. If we call $V_k(x)$ the set of nearest k neighbors in the training set, we thus take:

$$h(x) = \frac{1}{k} \sum_{x_i \in V_k(x)} y_i$$

(A priority rule can be specified in the case where two points are equidistant from x , e.g. choose the lowest index, so as to uniquely determine h).

One can also weight the influence of the points by taking into account the distance at x for example by the Gaussian kernel method:

$$h(x) = \frac{\sum_{i=1}^N \omega_i y_i}{\sum_{i=1}^N \omega_i},$$

where the weights w_i are given by:

$$w_i = \exp\left(-\frac{\|x_i - x\|^2}{2\sigma^2}\right).$$

Note that in these last two methods, the number k which defines the number of points in the neighborhood and the σ , standard deviation of the Gaussian distribution, are two parameters to be determined: we speak of hyperparameters. We will see in Section 4.3.1 how to determine them. ■

Other methods of direct approximation of conditional expectation exist (e.g. neural networks described in Section 4.6). However, we will first consider the use of parametric models of conditional densities to build decision functions, starting with the most classical of them, the linear model.

4.2 Models of Linear Regression

In this section, we consider real-valued explained and explanatory variable, and we consider the Borel σ -algebra.

4.2.1 Definition

Definition 4.2 — Linear regression model. Let $X = (X_1, X_2, \dots, X_p)$, random variable with values in $\mathbb{R}^p$ and Y , random variable in $\mathbb{R}$. We call linear regression model of Y with respect to X the family of conditional distributions

$$\mathcal{M} = \left\{ p_{(\beta_0, \beta, \sigma^2)}(y|x), (\beta_0, \beta, \sigma^2) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}_+^*, x \in \mathbb{R}^p \right\}$$

in which densities are given by

$$p_{(\beta_0, \beta, \sigma^2)}(y|x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{y - \beta_0 - \beta^t x}{\sigma} \right)^2}.$$

The model defines an affine relationship between the conditional expectation of Y given X and X :

$$\mathbb{E}(Y|X) = \beta_0 + \beta^t X,$$

and there exists a centered Gaussian random variable $\xi \sim \mathcal{N}(0, \sigma^2)$, ξ independent of X , such that:

$$Y = \beta_0 + \beta^t X + \xi.$$

The random variable ξ is called **residual**.

If the explained variable Y is in $\mathbb{R}^d$, $d > 1$, we speak of multivariate regression. $\beta_0 \in \mathbb{R}^d$, $\beta \in \mathcal{M}_{p,d}(\mathbb{R})$ (matrix with p rows, d columns), and the variable ξ is a multivariate Gaussian variable of variance Σ : $\xi \sim \mathcal{N}(0, \Sigma)$.

Definition 4.3 — Multivariate Linear Regression Model. Let $X = (X_1, X_2, \dots, X_p)$ be a random variable with values in $\mathbb{R}^p$ and Y random variable in $\mathbb{R}^d$. We call multivariate linear regression model of Y with respect to X the family of conditional distributions

$$\mathcal{M} = \left\{ p_{(\beta_0, \beta, \Sigma)}(y|x), (\beta_0, \beta, \Sigma) \in \mathbb{R}^d \times \mathcal{M}_{p,d}(\mathbb{R}) \times \mathcal{S}_d^+(\mathbb{R}), x \in \mathbb{R}^p \right\}$$

whose densities are given by

$$p_{(\beta_0, \beta, \Sigma)}(y|x) = \frac{1}{|\Sigma|^{1/2} (2\pi)^{d/2}} e^{-\frac{1}{2} (y - \beta_0 - \beta^t x)^t \Sigma^{-1} (y - \beta_0 - \beta^t x)}.$$

■ **Remark 4.5** If Σ is diagonal, then the multivariate linear regression model is equivalent to d independent univariate linear regression models, corresponding to the d components of Y . ■

■ **Remark 4.6** It is also possible to have the relationship

$$Y = \beta_0 + \beta^t X + \xi ,$$

with $\mathbb{E}(\xi) = 0$, ξ independent of X , but ξ not a Gaussian random variable, we then speak of **non-Gaussian linear regression**. ■

We can extend the definition of the linear model to the case where it is a function of the expectation of the response Y which depends linearly on the explanatory variables (Nelder and Wedderburn, 1972).

Definition 4.4 If there exists g such that

$$g(\mathbb{E}(Y|X)) = \beta_0 + \beta^t X ,$$

we speak of **generalized linear regression model**. The function g is called **link function**.

4.2.2 Estimation

Univariate Case

We first consider the univariate case: X with values in $\mathbb{R}^p$ and Y in $\mathbb{R}$.

From the observation couples $(x_i, y_i)_{1 \leq i \leq N}$, our objective is to estimate $\theta = (\beta_0, \beta, \sigma^2)$, that will completely characterize the linear model.

Let's denote $\hat{\theta} = (\hat{\beta}_0, \hat{\beta}, \hat{\sigma}^2)$, the maximum likelihood estimate. It is obtained by minimizing the opposite of the log-likelihood:

$$J(\theta) = -\ln(\mathcal{L}(\theta; (x_i, y_i)_{1 \leq i \leq N})) = \frac{N}{2} \ln(2\pi) + \frac{N}{2} \ln(\sigma^2) + \frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - \beta_0 - \beta^t x_i)^2 . \quad (4.1)$$

Lemme 4.2 If $(\hat{\beta}_0, \hat{\beta})$ minimizes the sum of squared residuals $R(\beta_0, \beta) = \sum_{i=1}^N (y_i - \beta_0 - \beta^t x_i)^2$, then for all $(\beta_0, \beta, \sigma^2) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}_+^*$,

$$J(\hat{\beta}_0, \hat{\beta}, \sigma^2) \leq J(\beta_0, \beta, \sigma^2) .$$

Preuve : This simply stems from the fact that J can be written:

$$J(\theta) = J_1(\sigma^2) + \frac{1}{2\sigma^2} R(\beta_0, \beta) .$$

Therefore, if $R(\hat{\beta}_0, \hat{\beta}) \leq R(\beta_0, \beta)$ for all (β_0, β) , then $\frac{1}{2\sigma^2} R(\hat{\beta}_0, \hat{\beta}) \leq \frac{1}{2\sigma^2} R(\beta_0, \beta)$ for all $(\beta_0, \beta, \sigma^2)$, and similarly $J_1(\sigma^2) + \frac{1}{2\sigma^2} R(\hat{\beta}_0, \hat{\beta}) \leq J_1(\sigma^2) + \frac{1}{2\sigma^2} R(\beta_0, \beta)$ for all $(\beta_0, \beta, \sigma^2)$. □

This small result is very interesting in practice, as it allows to decouple the estimation of the regression parameters (β_0, β) and the residual parameter σ^2 . We obtain:

$$(\hat{\beta}_0, \hat{\beta}) = \arg \min_{(\beta_0, \beta)} R(\beta_0, \beta) , \text{ puis } \hat{\sigma}^2 = \arg \min_{\sigma^2} J(\hat{\beta}_0, \hat{\beta}, \sigma^2) .$$

Let x_{ij} be the j -th component of the explanatory variables of x_i , the i -th element of the observation sample. Let us denote:

$$A = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{N1} & \dots & x_{Np} \end{pmatrix}, \quad B = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, \quad \gamma = (\beta_0, \beta) .$$

Alors, $R(\beta_0, \beta) = R(\gamma)$ se écrit:

$$R(\gamma) = \|A\gamma - B\|^2 .$$

■ **Remark 4.7** Matrix A , with or without the column of 1, is called **design matrix**. Without the column of 1, it is often denoted by X in the literature, with the slight ambiguity that it does not correspond to the matrix of vectors x_i in columns, but in rows:

$$X = (x_1, \dots, x_N)^T .$$

■

Theorem 4.3 In the univariate case, if A is injective, then the maximum likelihood estimator of the linear regression model is given by:

$$\hat{\gamma} = (\hat{\beta}_0, \hat{\beta}) = (A^t A)^{-1} A^t B \quad (4.2)$$

and

$$\hat{\sigma}^2 = \frac{1}{N} R(\hat{\gamma}) = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{\beta}_0 - \hat{\beta}^t x_i)^2 . \quad (4.3)$$

Preuve : We first look for $\hat{\gamma}$ minimizing $R(\gamma) = \|A\gamma - B\|^2$.

i) Let us first show that there is a solution to the problem of minimization. We denote $\|A\gamma\|^2 = \gamma^t A^t A \gamma$, and since A is injective, $A^t A$ is symmetrical and invertible. (If $A^t A z = 0$, then $z^t A^t A z = \|A z\|^2 = 0$, and $z = 0$ because of the injectivity of A . $A^t A$ is injective, thus bijective as an endomorphism, and finally invertible). $A^t A$ is thus positive definite, and according to the spectral representation theorem: $\forall \gamma, \gamma^t A^t A \gamma \geq \lambda_{\min} \|\gamma\|^2$, where $\lambda_{\min} > 0$ is the smallest eigenvalue of $A^t A$.

Therefore, $\lim_{\|\gamma\| \rightarrow +\infty} \|A\gamma\| = +\infty$, and since $\|A\gamma - B\| \geq \|A\gamma\| - \|B\|$, we finally have that

$\lim_{\|\gamma\| \rightarrow +\infty} R(\gamma) = +\infty$. This property is called **coercivity**.

Let $\gamma_0 \in \mathbb{R}^{p+1}$. By the coercivity property, there exists M_0 such that $\forall \gamma : \|\gamma\| > M_0, R(\gamma) > R(\gamma_0)$. So the lower bound of the function is to be searched on $\{\gamma : \|\gamma\| \leq M_0\}$, i.e. a compact set, and thus R being continuous, it reaches its minimum on $\{\gamma : \|\gamma\| \leq M_0\}$.

ii) Having proved the existence of the minimum, we can show that a necessary condition for R to reach its minimum in γ is that $\nabla R(\gamma) = 0$ (that is to say the gradient ¹ of R cancels out in

¹Let $f : H \rightarrow \mathbb{R}$, with H a Hilbert space. We say that f admits a gradient in x , if there exists an element of H , denoted $\nabla f(x)$, such that $\forall h \in H, f(x+h) = f(x) + (\nabla f(x), h) + o(\|h\|)$. (Frechet differentiability. Slightly less strict conditions can be given with Gâteaux differentiability). In Euclidean spaces, the gradient is classically given by the vector of partial derivatives.

γ). Indeed, if R admits a gradient, then $R(\gamma + th) = R(\gamma) + t(\nabla R(\gamma), h) + t^2\alpha(t)$, with $\alpha(t)$ bounded (Taylor expansion). Therefore:

$$R(\gamma + th) - R(\gamma) = t(\nabla R(\gamma), h) + t\alpha(t) ,$$

and if $\nabla R(\gamma) \neq 0$, $\exists h$ such that $(\nabla R(\gamma), h) \neq 0$ and if there exists t small enough, positive or negative, such that the member on the right side takes any sign, and therefore the minimum (nor the maximum for that matter) cannot be reached in γ . A necessary condition for the minimum to be reached in γ is that $\nabla R(\gamma) = 0$.

iii) We now compute ∇R , using scalar product notation:

$$\begin{aligned} R(\gamma + h) &= (A(\gamma + h) - b, A(\gamma + h) - b) = (A\gamma - b + Ah, A\gamma - b + Ah) \\ &= R(\gamma) + 2(A\gamma - b, Ah) + (Ah, Ah) \\ &= R(\gamma) + (2A^t(A\gamma - b), h) + o(\|h\|) \end{aligned}$$

soit $\nabla R(\gamma) = 2A^t(A\gamma - b)$.

iv) Finally,,

$$\nabla R(\hat{\gamma}) = 0 \iff \hat{\gamma} = (A^t A)^{-1} A^t B ,$$

thanks to the invertibility of $A^t A$.

$\hat{\gamma}$ is the only point verifying the necessary conditions of minimality, but the existence of the minimum has been shown, so it is the only point where R reaches its minimum, and it is therefore the maximum likelihood estimate for the γ parameter.

Using Lemma 4.2, we get $\hat{\sigma}^2$ by minimizing $J(\hat{\gamma}, \alpha)$ with respect to α . From Equation 4.1, we have:

$$\frac{\partial J(\hat{\gamma}, \alpha)}{\partial \alpha} = \frac{N}{2\alpha} - \frac{1}{2\alpha^2} R(\hat{\gamma}) ,$$

and

$$\frac{\partial J(\hat{\gamma}, \hat{\alpha})}{\partial \alpha} = 0 \iff \hat{\alpha} = \hat{\sigma}^2 = \frac{1}{N} R(\hat{\gamma}) .$$

We see that $\alpha \mapsto J(\hat{\gamma}, \alpha)$ is strictly decreasing for $\alpha < \hat{\alpha}$ and strictly increasing for $\alpha > \hat{\alpha}$, J reaches its minimum in $(\hat{\gamma}, \hat{\alpha})$, which concludes the proof. $\square$

■ **Remark 4.8** An important assumption of Theorem 4.3 is the injectivity of A . In practice, it corresponds to the fact that we have enough data to inverse the model. But there are many problems in which the number of covariates is much larger than the number of observations ($p > N$), so A is not invertible: this is called overfitting. We will see in Part 4.4 that regularization techniques allow to get around this problem. ■

■ **Remark 4.9 — Estimation of the maximum likelihood or minimization of a loss function.** Note that to characterize the decision function for the quadratic loss, we only need to estimate $\gamma = (\beta_0, \beta)$, since

$$h_\gamma(X) = \mathbb{E}(Y|X) = \beta_0 + \beta^t X .$$

It is a pretty common situation in some statistical learning: many methods will not propose a parametric family of probability densities, but directly a family of parametric functions. $\{h_\theta : \mathcal{X} \rightarrow \mathcal{Y}, \theta \in \Theta\}$ corresponding to the decision function. And, a linear regression model has also come to mean a model of the form $Y = \beta_0 + \beta^t X$ (en identifying Y with $\mathbb{E}(Y|X)$), and

we solve Equation 4.2. The same applies, for example, to neural networks, Section 4.6. Learning will no longer be strictly an estimation of parameters by maximum likelihood, but the calibration of a model by minimizing a loss function, generally quadratic: we look for $h^* = h_{\theta^*}$ such that

$$h_{\theta^*} = \arg \min_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^N |h_{\theta}(x_i) - y_i|^2 .$$

where $\frac{1}{N} \sum_{i=1}^N |h_{\theta}(x_i) - y_i|^2$ is the empirical approximation of $\mathbb{E}(L(h_{\theta}(X), Y))$.

We can always go back to the probabilistic framework by assuming the existence of a random variable $\xi \sim \mathcal{N}(0; \sigma^2)$ such that

$$Y = h_{\theta}(X) + \xi .$$

As mentioned above for the linear case, it is easy to see that for this model, the estimate of θ by the maximum likelihood corresponds to the minimization of the quadratic loss function.

It is important to keep in mind the underlying probabilistic formulation, especially when characterizing the properties of the estimator or when performing an uncertainty analysis of our predictions. ■

For the linear regression, we can deduce several important properties of the estimator.

Theorem 4.4 If $(\hat{\gamma}, \hat{\sigma}^2)$ is the maximum likelihood estimator for the linear regression as defined in Theorem 4.3, then $\hat{\gamma}$ and $\hat{\sigma}^2$ are two independent random variables:

$$\hat{\gamma} \sim \mathcal{N}(\gamma, \sigma^2 (A^t A)^{-1}) \quad (4.4)$$

$$\frac{N \hat{\sigma}^2}{\sigma^2} \sim \chi^2(N - p - 1) \quad (4.5)$$

Preuve : If η is a Gaussian vector of dimension N , all components of which are independent and of variance σ^2 , $\eta \sim \mathcal{N}(0, \sigma^2 I_N)$ (I_N , identity matrix of order N).

We then have:

$$\hat{\gamma} = (A^t A)^{-1} A^t B = (A^t A)^{-1} A^t (A \gamma + \eta) = \gamma + (A^t A)^{-1} A^t \eta .$$

We know that if Z is a real-valued random vector of dimension k and of covariance Σ , and if M is a real matrix $l \times k$, then MZ is a random vector of covariance $M \Sigma M^t$ (see for example Anderson (1962)). We thus have:

$$\mathbb{V}((A^t A)^{-1} A^t \eta) = (A^t A)^{-1} A^t \sigma^2 I_N (A^t A)^{-1} A^t = \sigma^2 (A^t A)^{-1} ,$$

and $\hat{\gamma} \sim \mathcal{N}(\gamma, \sigma^2 (A^t A)^{-1})$. The independence of random variables and the fact that $\frac{N \hat{\sigma}^2}{\sigma^2} \sim \chi^2(N - p - 1)$ results from Cochran's Theorem (cf. Appendix A.2). We have $N - p - 1$ degrees of freedom since $p + 1$ parameters were estimated for γ . □

Corollary 4.5 $\forall i = 0, \dots, p$, let us denote:

$$\hat{\sigma}_i^2 = \frac{N \hat{\sigma}^2}{N - p - 1} [(A^t A)^{-1}]_{ii} , \quad (4.6)$$

where for a matrix M , M_{ii} represents the i -th term of the diagonal.

Therefore:

$$\forall i = 0, \dots, p : \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_i} \sim T(N - p - 1) ,$$

that is to say a Student distribution with $N - p - 1$ degrees of freedom.

Preuve : From 4.4, $\frac{\hat{\beta}_i - \beta_i}{\sigma \sqrt{[(A^t A)^{-1}]_{ii}}} \sim \mathcal{N}(0, 1)$, thus by using 4.5 and the independence of the two variables:

$$\frac{\frac{\hat{\beta}_i - \beta_i}{\sigma \sqrt{[(A^t A)^{-1}]_{ii}}}{\left(\frac{N \hat{\sigma}^2}{(N - p - 1) \sigma^2} \right)^{1/2}} \sim T(N - p - 1)$$

by definition of the Student distribution, which gives us the result. $\square$

■ **Remark 4.10 — Test on the nullity of a regression coefficient.** An important application of this result to the distribution of $\frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_i}$ is that it allows us to implement tests on the nullity of the regression coefficients, if we want to test, for example, whether a certain covariate has an influence on our variable to be explained. Let i be the index of the covariate of interest, then we construct the hypothesis test:

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0 \end{cases}$$

Under H_0 , $\frac{\hat{\beta}_i}{\hat{\sigma}_i} \sim T(N - p - 1)$ and we build the region of rejection at significance level α :

$$\mathcal{R}_\alpha = \left\{ (x_1, y_1), \dots, (x_N, y_N) \in (\mathcal{X} \times \mathcal{Y})^N : \left| \frac{\hat{\beta}_i}{\hat{\sigma}_i} \right| > q_{1-\alpha/2}^{T(N-p-1)} \right\} ,$$

where $q_u^{T(k)}$ represents the u -quantile for a Student distribution with k degrees of freedom. If the test leads to reject hypothesis H_0 , then the regression coefficient for the covariate i cannot be considered zero. If, on the contrary, we cannot reject H_0 , we can consider a simplified regression model, which does not take into account the i -th covariate.

For given observations, we obtain the p -value α associated with our hypothesis test by:

$$\alpha = 2 \left(1 - \Phi^{T(N-p-1)} \left(\left| \frac{\hat{\beta}_i}{\hat{\sigma}_i} \right| \right) \right)$$

where $\Phi^{T(N-p-1)}$ represents the cumulative distribution function of the Student distribution with $(N - p - 1)$ degrees of freedom.

Note that the tests are valid only for each coefficient individually, as the coefficients obtained are correlated with each other. If, for example, a first test leads us to accept the hypothesis that the coefficient linked to the i -th covariate is zero, then we must perform a new regression without taking this covariate into account before implementing the test for a new covariate. ■

■ **Example 4.11 — Linear regression and k -nearest neighbours.** We consider two random variables X and Y with values in $\mathbb{R}$. Let $N = 50$ and consider N couples of observations $(x_i, y_i)_{1 \leq i \leq N}$ represented in Figure 4.1.

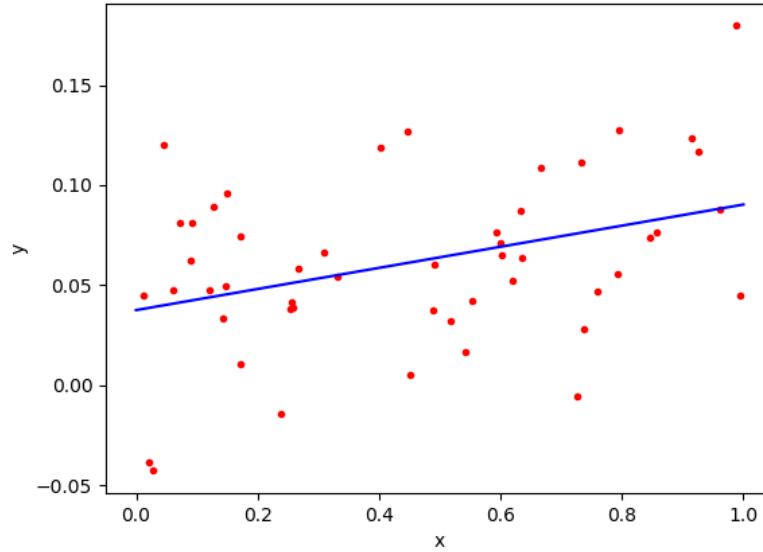


Figure 4.1: Scatterplot of 50 experimental points (x_i, y_i) . In blue, regression line $y = \hat{\beta}_0 + \hat{\beta}_1 x$.

Linear Regression:

First of all, we consider an affine regression model, i.e:

$$Y = \beta_0 + \beta_1 X + \xi, \text{ avec } \xi \sim \mathcal{N}(0, \sigma^2).$$

We form the matrix A by adding a column of 1 to the columns of explanatory variables and the vector B given by the y_i , and we get:

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (A^t A)^{-1} A^t B = \begin{pmatrix} 0.03756 \\ 0.05275 \end{pmatrix}, \hat{\sigma} = 0.03969,$$

and using Equation 4.6, standard errors can be calculated on the regression coefficients:

$$\begin{pmatrix} \hat{\sigma}_0 \\ \hat{\sigma}_1 \end{pmatrix} = \begin{pmatrix} 0.008232 \\ 0.01791 \end{pmatrix}.$$

Testing for the nullity of a regression coefficient:

When looking at the scatterplot, one may wonder whether Y varies with X or not. This corresponds to testing the nullity of β_1 . The test statistic takes the value $t_1 = \frac{\hat{\beta}_1}{\hat{\sigma}_1} = 2.9444$, and the test statistic follows a Student distribution with $N - 2$ degrees of freedom, and we deduce the p -value α :

$$2(1 - \Phi^{T(48)}(|t_1|)) = 0.00497 .$$

For classical levels of tests, even very conservative ones (for example at a level of 0.01), we reject the hypothesis H_0 , and the coefficient cannot be considered as null.

Prevision:

We now consider a prediction test. Suppose we have $N_t = 20$ new points $(x_i, y_i)_{N+1, N+N_t}$, in black in Figure 4.2. We wish to evaluate the prediction capacity of the linear model. The estimation of the coefficients is carried out on the first 50 points, then the model thus obtained is tested by estimating $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ for $i = N + 1, \dots, N + N_t$. The points on which we evaluate the prediction capacity, not used to calibrate the model, are often referred to as "test points". Predicted values are indicated with blue crosses in Figure 4.2.

Different criteria can be used to assess the quality of the prediction. We consider the *mean squared error of prediction* (MSEP).

$$MSEP = \frac{1}{N_t} \sum_{i=N+1}^{N+N_t} \|y_i - \hat{y}_i\|^2 .$$

In this example, we find: $MSEP = 0.0181$.

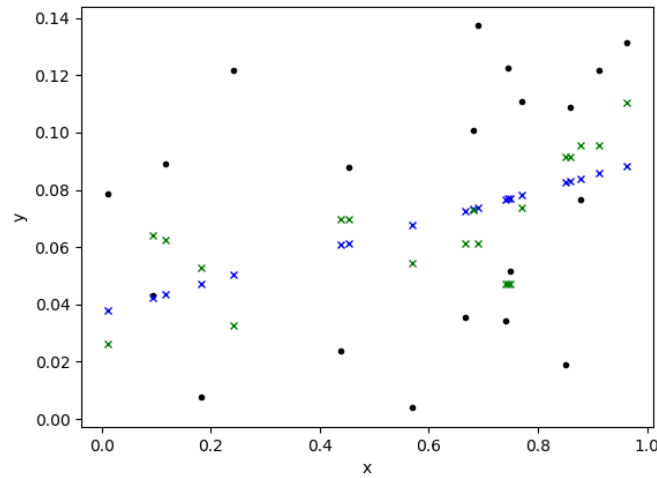


Figure 4.2: The 20 points in black are the test points. The blue crosses are the linear regression predictions, the green crosses are the 5-closest neighbor predictions.

Method of the k nearest neighbors and selection of the hyperparameter k :

It may be interesting to compare linear regression with another method, for example the elementary method of k -nearest neighbors (cf. example 4.4). For $k = 5$, we find $MSEP = 0.00202$, the

predictions are indicated by green crosses on the figure. The result is therefore slightly worse. However, the choice $k = 5$ may not be optimal.

In figure 4.3, we show how the mean squared error of prediction evolves as a function of the hyperparameter k for the k -nearest neighbor method.

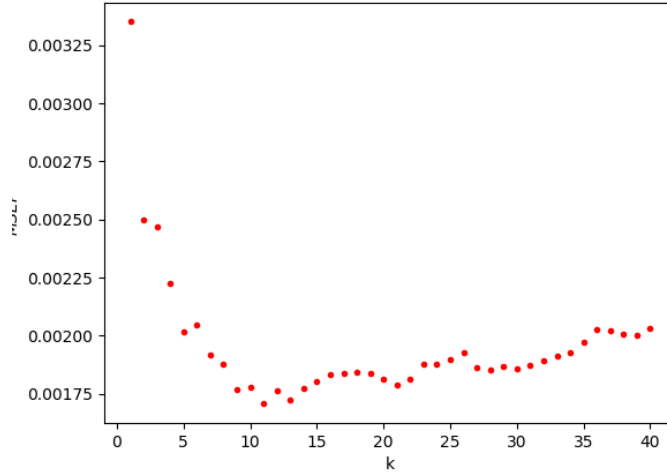


Figure 4.3: Mean squared error of prediction ($MSEP$) as a function of k , the number of nearest neighbors.

We see that the best result is obtained for $k = 11$, with $MSEP = 0.00171$. We are therefore tempted to conclude that the nearest neighbor method with $k = 11$ is a better prediction model than linear regression. This would correspond to a very classical reasoning error in statistical learning. Indeed, we used the test data to select the hyperparameter. In other words, we are not really in a pure prediction approach, because of course the nearest neighbor method is not generally the best for $k = 11$, it depends on both the training data set and the test data set. When the prediction model depends on hyperparameters, then the correct approach is to use the training dataset to select the hyperparameters, and then set these hyperparameters to perform the predictions on the test dataset.

Thus, among the $N = 50$ observations of the training set, we put aside for example 20% as validation data, in order to choose the hyperparameters. Here, we take $N_v = 10$ observations for validation. Therefore, for each value of hyperparameter (or a search grid), we use $(x_i, y_i)_{1 \leq i \leq N - N_v}$ to evaluate the model, and the $MSEP$ is calculated for $(x_i, y_i)_{N - N_v + 1 \leq i \leq N}$. The best k is obtained for $k = 4$, but the $MSEP$ on the test data is then 0.00221...

Rather than take the last N_v of data from the learning set, a more robust method is called **cross-validation**: the learning set is divided into m parts of equal size (or roughly when N is not divisible by m), and successively, each part is used for validation, while the other $m - 1$ parts are used for learning. The validation $MSEP$ is then obtained by averaging the $MSEP$ obtained for all m configurations (cf. Figure 4.5). This generally increases robustness, especially when the training data set is small.

We apply this strategy with $m = 5$ on our training data. The mean squared error of prediction

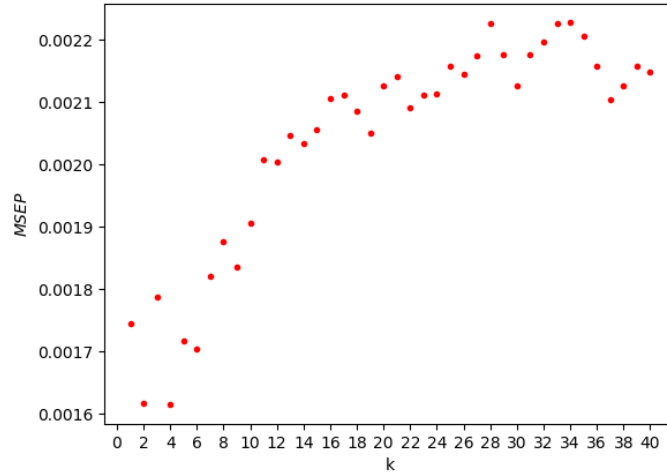


Figure 4.4: Mean squared error of prediction ($MSEP$) as a function of k , the number of nearest neighbors.

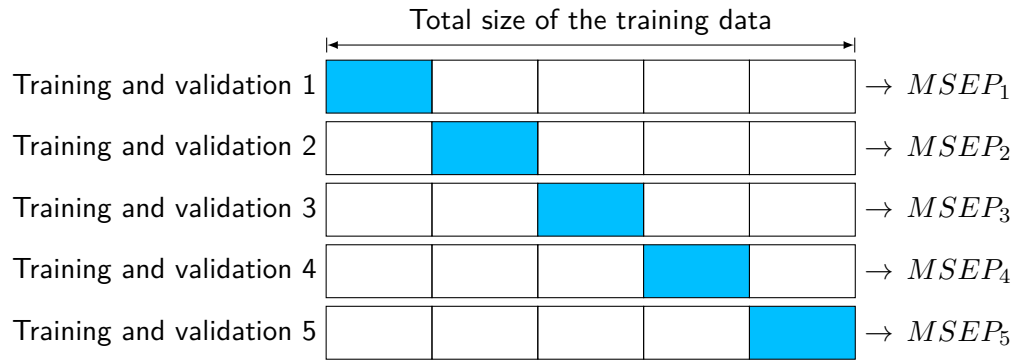


Figure 4.5: 5-fold cross validation: for each row j , $1 \leq j \leq 5$, data in white are used for training and data in grey are used for validation and computation of $MSEP_j$. We then compute the mean $MSEP$: $MSEP^{cv} = \frac{1}{5} \sum_{j=1}^5 MSEP_j$

obtained by cross-validation averaging, denoted $MSEP^{cv}$, is given as a function of the number of neighbors in Figure 4.6. The best result is then obtained for $k = 24$! We can see the very high variability in the selected hyperparameter. It is due to the very noisy data. We then find in prediction (on the test data this time) $MSEP = 0.188$. A special case of cross validation is when $m = N$, we talk about **leave-one-out cross validation**, it is normally the most robust because we use almost the whole training set for each validation, but it is of course much more expensive. With the leave-one-out method on our example, we find that the best hyperparameter is $k = 29$! The $MSEP$ on the test set is then $MSEP = 0.187$. In this case, there is little difference with the 5-fold cross-validation.

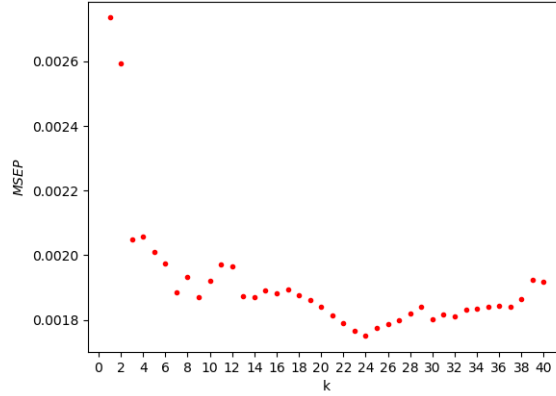


Figure 4.6: Mean squared error of prediction ($MSEP$) as a function of k , the number of nearest neighbors.

■

Multivariate Regression

In this course, we will not go into the details of multivariate linear regression. However, it is interesting to keep in mind that when the residuals are correlated, the regression of an explained variable Y of dimension d is more interesting than d one-dimensional regressions performed independently: the regression for each of the component of Y is enriched by the information on the other components. We give in the theorem below the coefficients of multivariate regression, whose expression is very similar to that of one-dimensional regression.

Theorem 4.6 Let X, Y be random variables with values respectively in $\mathbb{R}^p$ and $\mathbb{R}^d$, such that there exist $\beta_0 \in \mathbb{R}^d$, $\beta \in \mathcal{M}_{p,d}(\mathbb{R})$, $\xi \sim \mathcal{N}(0, \Sigma)$ such that

$$Y = \beta_0 + \beta^t X + \xi .$$

Let $(x_i, y_i)_{1 \leq i \leq N}$, be N independent couples of observations of (X, Y) . $x_i \in \mathbb{R}^p$ and we denote x_{ij} , $1 \leq j \leq p$, the j -th component of x_i . Similarly, $y_i \in \mathbb{R}^d$ and we denote y_{ij} , $1 \leq j \leq d$, the j -th component of y_i .

Once more, let:

$$A = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{N1} & \dots & x_{Np} \end{pmatrix} , \quad B = \begin{pmatrix} y_{11} & \dots & y_{1d} \\ \vdots & & \vdots \\ y_{N1} & \dots & y_{Nd} \end{pmatrix} , \quad \gamma = \begin{pmatrix} \beta_{01} & \dots & \beta_{0d} \\ \beta_{11} & \dots & \beta_{1d} \\ \vdots & & \vdots \\ \beta_{p1} & \dots & \beta_{pd} \end{pmatrix}$$

Then the maximum likelihood estimate of γ is given by:

$$\hat{\gamma} = (A^t A)^{-1} A^t B \quad (4.7)$$

et

$$\hat{\Sigma} = \frac{1}{N} (A\hat{\gamma} - B)^t (A\hat{\gamma} - B) \quad (4.8)$$

The proof can be found for example in Anderson (1962), chapter 8.

4.3 Linear Models for Classification

We first consider a two-class classification problem. Let X be a random variable with values in $\mathcal{X}$, $\mathcal{X} \subset \mathbb{R}^p$, and Y a random variable with values in $\{0; 1\}$ (one class will correspond to the 0 label and the other to the 1 label). The classification problem corresponds to the determination of the conditional distribution $\mathbb{P}(Y|X)$, i.e. the determination of $\{\mathbb{P}(Y = 0|X), \mathbb{P}(Y = 1|X)\}$. Since Y only takes two values, we have $\forall x, \mathbb{P}(Y = 0|X = x) = 1 - \mathbb{P}(Y = 1|X = x)$ and thus the two-class classification simply boils down to the determination of $\mathbb{P}(Y = 1|X) = \{\mathbb{P}(Y = 1|X = x), x \in \mathcal{X}\}$.

Definition 4.5 — Linear binary classification model. We will call any parametric family of conditional probability distributions a linear binary classification model:

$$\mathcal{M}_\psi = \left\{ \mathbb{P}(Y|X) : \mathbb{P}(Y = 1|X) = \psi(\beta_0 + \beta^t X), (\beta_0, \beta) \in \mathbb{R}^{p+1} \right\}$$

where $\psi : \mathbb{R} \rightarrow [0; 1]$ is a strictly increasing function.

Definition 4.6 — Decision function. As for linear regression, we define for a parametric model $\mathbb{P}_{(\beta_0, \beta)}(Y|X)$, the **decision function** (also called classification function) $\delta : \mathcal{X} \rightarrow \{0; 1\}$ which to $x \in \mathcal{X}$ maps one labels as follows:

$$\delta : x \mapsto \begin{cases} \delta(x) = 1 & \text{if } \mathbb{P}_{(\beta_0, \beta)}(Y = 1|X = x) > \delta_0 \\ \delta(x) = 0 \text{ or } 1 & \text{if } \mathbb{P}_{(\beta_0, \beta)}(Y = 1|X = x) = \delta_0 \\ \delta(x) = 0 & \text{if } \mathbb{P}_{(\beta_0, \beta)}(Y = 1|X = x) < \delta_0 \end{cases}, \quad (4.9)$$

where $\delta_0 \in [0; 1]$ is a decision threshold.

For the sake of simplicity and without loss of generality, we choose $\delta(x) = 1$ if $\mathbb{P}_{(\beta_0, \beta)}(Y = 1|X = x) = \delta_0$.

■ **Remark 4.12** Classically, we take $\delta_0 = 1/2$, which is equivalent to write:

$$\delta : x \mapsto \begin{cases} \delta(x) = 1 & \text{si } \mathbb{P}_{(\beta_0, \beta)}(Y = 1|X = x) \geq \mathbb{P}_{(\beta_0, \beta)}(Y = 0|X = x) \\ \delta(x) = 0 & \text{si } \mathbb{P}_{(\beta_0, \beta)}(Y = 0|X = x) > \mathbb{P}_{(\beta_0, \beta)}(Y = 1|X = x) \end{cases}.$$

But in some configurations (unbalanced data between labels, asymmetric risks associated to labels 0 and 1 ...), it may be more relevant to have an asymmetric decision function, considering a decision threshold different from 1/2. ■

Corollary 4.7 For the decision function of a linear classification model, if we consider the partitioning of the covariate space $\mathcal{X}$ in $\mathcal{X}_0 = \{x \in \mathcal{X} : \delta(x) = 0\}$ and $\mathcal{X}_1 = \{x \in \mathcal{X} : \delta(x) = 1\}$, then the boundary between $\mathcal{X}_0$ and $\mathcal{X}_1$ is an affine hyperplane.

Preuve : ψ is increasing: $x \in \mathcal{X}_1$ is equivalent to $\psi(\beta_0 + \beta^t x) \geq \delta_0$, that is to say $\beta_0 + \beta^t x \geq \psi^-(\delta_0)$ (where $\psi^-(\delta_0)$ is the generalized inverse, $\psi^-(\delta_0) = \inf_z \{z \in \mathbb{R} : \psi(z) \geq \delta_0\}$).

Therefore, the affine hyperplane $(\beta_0 - \psi^-(\delta_0)) + \beta^t x = 0$ represents the boundary between $\mathcal{X}_0$ and $\mathcal{X}_1$. $\square$

4.3.1 Logistic Model for Classification

The most classical model is the logistic one where $\psi : z \mapsto \frac{e^z}{1+e^z}$. We thus have:

$$\begin{cases} \mathbb{P}_{(\beta_0, \beta)}(Y = 1|X = x) = \frac{\exp(\beta_0 + \beta^t x)}{1 + \exp(\beta_0 + \beta^t x)} \\ \mathbb{P}_{(\beta_0, \beta)}(Y = 0|X = x) = \frac{1}{1 + \exp(\beta_0 + \beta^t x)} \end{cases}$$

We notice that:

$$\ln \left(\frac{\mathbb{P}_{(\beta_0, \beta)}(Y = 1|X = x)}{1 - \mathbb{P}_{(\beta_0, \beta)}(Y = 1|X = x)} \right) = \beta_0 + \beta^t x. \quad (4.10)$$

The function $p \mapsto \ln \left(\frac{p}{1-p} \right)$ is called the **logit** function. It is interesting in that it transforms a probability (between 0 and 1) into a value between $-\infty$ and $+\infty$, which makes possible the representation by the real number line.

We see that we have:

$$\text{logit}(\mathbb{E}(Y|X)) = \beta_0 + \beta^t X,$$

the logistic classification model is a generalized linear model (cf. Definition 4.4), the logit function is a link function.

■ **Remark 4.13** For $\delta_0 = 1/2$, the boundary between $\mathcal{X}_0$ and $\mathcal{X}_1$ for the decision function is thus directly $\beta_0 + \beta^t x = 0$. ■

■ **Remark 4.14 — Extension to the multi-label case.** Note that the equation 4.10 allows to easily extend the logistic model to multiclass (multilabel) classification, i.e. for any number of classes. Let us now consider that Y takes its values (the labels) dans $\mathcal{Y} = \{0, 1, 2, \dots, K\}$. A class (for example the class corresponding to label 0) is used as a reference, and the model determined by:

$$\begin{cases} \ln \left(\frac{\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = 0|X = x)} \right) = \beta_{1,0} + \beta_1^t x \\ \ln \left(\frac{\mathbb{P}(Y = 2|X = x)}{\mathbb{P}(Y = 0|X = x)} \right) = \beta_{2,0} + \beta_2^t x \\ \dots \\ \ln \left(\frac{\mathbb{P}(Y = K|X = x)}{\mathbb{P}(Y = 0|X = x)} \right) = \beta_{K,0} + \beta_K^t x \end{cases}$$

The model corresponding to the conditional distribution $\mathbb{P}(Y|X)$ is totally determined because $\sum_{k=0}^K \mathbb{P}(Y = k|X = x) = 1$. ■

Estimation

We consider the 2-class classification problem, X takes values in $\mathbb{R}^p$ and Y in $\{0; 1\}$. Based on the pairs of observations $(x_i, y_i)_{1 \leq i \leq N}$, we try to estimate $\gamma = (\beta_0, \beta)$, the parameters of the logistic models such that:

$$\begin{cases} \mathbb{P}_\gamma(Y = 1|X = x) = \frac{\exp(\beta_0 + \beta^t x)}{1 + \exp(\beta_0 + \beta^t x)} \\ \mathbb{P}_\gamma(Y = 0|X = x) = \frac{1}{1 + \exp(\beta_0 + \beta^t x)} \end{cases} \quad (4.11)$$

We denote $p(\gamma; x) = \mathbb{P}_\gamma(Y = 1|X = x)$ and form the likelihood using the classic rewriting trick for the Bernoulli distribution:

$$\begin{aligned} \mathcal{L}(\gamma; (x_i, y_i)_{1 \leq i \leq N}) &= \prod_{i=1}^N \mathbb{P}(Y = y_i|X = x_i) \\ &= \prod_{i=1}^N p(\gamma; x_i)^{y_i} (1 - p(\gamma; x_i))^{1-y_i} . \end{aligned}$$

The log-likelihood is then written as follows:

$$\ell(\gamma; (x_i, y_i)_{1 \leq i \leq N}) = \sum_{i=1}^N y_i \ln[p(\gamma, x_i)] + (1 - y_i) \ln[1 - p(\gamma, x_i)] . \quad (4.12)$$

Denoting $z_i = (1, x_i)$ and using the probability expressions given in 4.11, we obtain after simplification:

$$\ell(\gamma; (x_i, y_i)_{1 \leq i \leq N}) = \sum_{i=1}^N y_i (\gamma^t z_i) - \ln(1 + \exp(\gamma^t z_i)) . \quad (4.13)$$

In what follows, we omit the data in the ℓ argument for the sake of simplicity.

Since the function is concave, we can show that a necessary and sufficient condition for ℓ to reach its maximum in $\text{hat}\gamma$ is that $\nabla \ell(\hat{\gamma}) = 0$ (where $\nabla \ell(\gamma)$ represents the gradient of ℓ in γ). This equation can be solved by Newton-Raphson's method, by constructing a sequence $(\gamma_n)_{n \in \mathbb{N}}$ such that γ_0 is given (for example $\gamma_0 = 0$) and that

$$\gamma_{n+1} = \gamma_n + [\nabla^2 \ell(\gamma_n)]^{-1} \nabla \ell(\gamma_n) . \quad (4.14)$$

with $\nabla^2 \ell(\gamma_n)$, the Hessian² of ℓ in γ_n .

We have:

$$\nabla \ell(\gamma) = \sum_{i=1}^N y_i z_i - \frac{\exp(\gamma^t z_i)}{1 + \exp(\gamma^t z_i)} z_i$$

and

$$\nabla^2 \ell(\gamma) = - \sum_{i=1}^N \frac{\exp(\gamma^t z_i)}{(1 + \exp(\gamma^t z_i))^2} z_i z_i^t .$$

²Let $f : \mathcal{H} \rightarrow \mathbb{R}$, $\mathcal{H}$ an Hilbert space, f twice differentiable in $x \in \mathcal{H}$, there exists a unique linear application $\nabla^2 f : \mathcal{H} \rightarrow L(\mathcal{H})$, ($L(\mathcal{H})$: set of endomorphisms on $\mathcal{H}$), called Hessian such that $\forall h, k \in \mathcal{H}$, $d^2 f(x)(h, k) = (\nabla^2 f(x)h, k)$. If $\mathcal{H} = \mathbb{R}^n$ and if the scalar product is the Euclidean product, the Hessian simply corresponds to the Hessian matrix.

The algorithm doesn't necessarily converge to a point $\hat{\gamma}$ such that $\nabla \ell(\hat{\gamma}) = 0$. Nevertheless, we will see in the optimization course (or for example in Nocedal and Wright (2006)) that we can always choose a $\rho_n > 0$ small enough in the recurrence equation (4.15):

$$\gamma_{n+1} = \gamma_n + \rho_n \left[\nabla^2 \ell(\gamma_n) \right]^{-1} \nabla \ell(\gamma_n) . \quad (4.15)$$

such that at each iteration $\ell(\gamma_{n+1}) > \ell(\gamma_n)$ and that the method converges towards $\hat{\gamma}$ such that $\nabla \ell(\hat{\gamma}) = 0$ and therefore attains the maximum likelihood.

■ **Remark 4.15 — Cross-entropy.** Equation 4.12 illustrates an important concept, that of cross-entropy. In the particular case of distributions from the same parametric model, we have given the definition of the Kullback-Leibler divergence (cf. Definition 2.20). In the more general case, the Kullback-Leibler divergence $D_{\text{KL}}(Q \parallel P)$ measures how much a distribution P deviates from a reference distribution Q .

If P and Q are probability distributions on $\mathcal{X}$, absolutely continuous with respect to a reference measure μ , and denoting p, q their densities with respect to μ , we have:

$$\begin{aligned} D_{\text{KL}}(Q \parallel P) &= \int_{\mathcal{X}} q \ln \left(\frac{q}{p} \right) d\mu \\ &= - \int_{\mathcal{X}} q \ln(p) d\mu + \int_{\mathcal{X}} q \ln(q) d\mu \end{aligned}$$

The first term $H(Q, P)$ is called **cross entropy**:

$$H(Q, P) = - \int_{\mathcal{X}} q \ln(p) d\mu .$$

The second term, denoted $H(Q)$ is called **entropy**:

$$H(Q) = \int_{\mathcal{X}} q \ln(q) d\mu .$$

In the case of binary classification, if we consider a sample $(x_i, y_i) \in \mathcal{X} \times \{0, 1\}$, we take as reference distribution Q the reality (known through the empirical distribution), and as approximation distribution P , the model. We then have:

$$H(Q, P) = - \sum_{c=0,1} Q(y_i = c|x_i) \ln(P(y_i = c|x_i)) ,$$

and since $Q(Y_i = 1|x_i) = y_i$ (it is 1 if $y_i = 1$ and 0 if $y_i = 0$) and $Q(Y_i = 0|x_i) = 1 - y_i$, we get:

$$H(Q, P) = -y_i \ln(p(\gamma; x_i)) - (1 - y_i) \ln(1 - p(\gamma; x_i)) .$$

We have $H(Q) = 0$ since

$$H(Q) = - \sum_{c=0,1} Q(y_i = c|x_i) \ln(Q(y_i = c|x_i)) ,$$

and $Q(y_i = c|x_i) = 0$ or $Q(y_i = c|x_i) = 1$.

Finally, we see that for the maximum likelihood estimation of the parameters of the classification model, we see that the log-likelihood (cf. equation 4.12) is the opposite of the sum of the cross entropy on all samples, maximizing the likelihood will therefore be equivalent to minimizing the cross entropy, which can thus be considered as a loss function.

In the same way as the quadratic loss function is the loss function conventionally chosen for regression, the cross-entropy is the loss function conventionally used for classification. It extends to the multi-class case.

Definition 4.7 — Cross entropy loss function. Let X a random variable with values in $\mathcal{X}$, Y with values in $\{0; 1; \dots; K\}$. Let $\mathbb{P}(Y|X)$ be a statistical model for the multi-class classification problem.

Then, the cross entropy loss function for this model is defined for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ by:

$$L((x, y), \mathbb{P}(Y|X)) = - \sum_{k=0}^K \delta_y(k) \ln[\mathbb{P}(Y = k|X = x)] .$$

The best model $\mathbb{P}^*(Y|X)$ with respect to this loss function is classically obtained by minimizing the expectation of the loss function:

$$\mathbb{P}^*(Y|X) = \arg \min_{\mathbb{P}(Y|X)} \mathbb{E}(L((X, Y), \mathbb{P}(Y|X))) .$$

Let $(x_i, y_i)_{1 \leq i \leq N}$ be an observation sample, by taking the empirical mean:

$$\mathbb{P}^*(Y|X) = \arg \min_{\mathbb{P}(Y|X)} - \frac{1}{N} \sum_{i=1}^N \sum_{k=0}^K \delta_{y_i}(k) \ln[\mathbb{P}(Y = k|X = x_i)] .$$

If $\mathbb{P}(Y|X)$ is parametric, $\mathbb{P}(Y|X) = \mathbb{P}_\gamma(Y|X)$ and if $\mathbb{P}_\gamma(Y|X)$ is obtained by minimizing the empirical mean of the cross-entropies on all samples, $\hat{\gamma}$ is the maximum likelihood estimator for the model. ■

■ **Example 4.16 — Logistic model for classification.**

Generation of training data:

Let $X_0 \sim \mathcal{N}(m_0; \Sigma_0)$ with $m_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\Sigma_0 = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, and $X_1 \sim \mathcal{N}(m_1; \Sigma_1)$, with $m_1 = \begin{pmatrix} 3 \\ 3 \end{pmatrix}$ et $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Denote respectively p_0 and p_1 the associated densities.

Let $(x_i)_{1 \leq i \leq N_0}$ an i.i.d. sample for X_0 giving the elements in class 0, and $(x_i)_{N_0+1 \leq i \leq N_0+N_1}$ an i.i.d. sample for X_1 giving the elements in class 1. Figure 4.7 illustrates the samples for $N_0 = N_1 = 200$.

Estimation of model parameters:

The parameters of a logistic classification model are estimated from the training data $\{(x_i, 0)_{1 \leq i \leq N_0}, (x_i, 1)_{N_0+1 \leq i \leq N_0+N_1}\}$ using Newton-Raphson method.

We find: $\hat{\gamma} = (-6.09, 1.43, 1.56)$, and the associated line: $\hat{\beta}_0 + \hat{\beta}^t x = 0$ visible in Figure 4.7 represents the boundary of the decision function obtained for $\delta_0 = 1/2$ (cf. Remark 4.13). That is to say that for new entries $(\tilde{x}_i)_{1 \leq i \leq N_{test}}$, if $\hat{\beta}_0 + \hat{\beta}^t \tilde{x}_i > 0$, we take $\delta(\tilde{x}_i) = 1$ and if $\hat{\beta}_0 + \hat{\beta}^t \tilde{x}_i < 0$, $\delta(\tilde{x}_i) = 0$. If the point is on the boundary (but it is a zero probability event...), we define the decision rule we want. As the supports of the distributions of X_0 and X_1 overlap, we see that it will not be possible to obtain a perfect classifier.

Predictino Test:

Let's perform a prediction test by randomly generating $N_{test} = 1000$ test points $(\tilde{x}_i)_{1 \leq i \leq N_{test}}$, choosing also randomly according to X_0 or X_1 . We know the distribution from which each of the

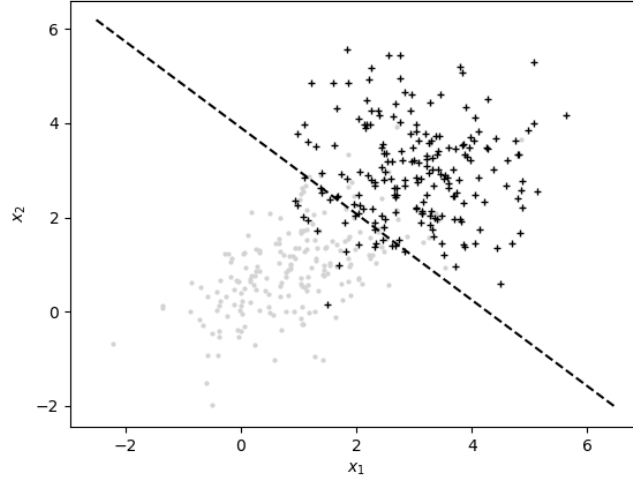


Figure 4.7: Light grey dots: i.i.d. sample for X_0 (class 0) and black crosses: i.i.d. sample for X_1 (class 1). The dotted line indicates the boundary of the decision function. $\hat{\beta}_0 + \hat{\beta}^t x = 0$.

dots is drawn: that gives us the truth $\tilde{y}_i$ for each of the $\tilde{x}_i$'s. For each of them, the prediction is calculated from the sign of $\hat{\beta}_0 + \hat{\beta}^t \tilde{x}_i$, and we calculate our rate of correct classifications, called **Accuracy**:

$$\text{Accuracy} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \mathbb{I}_{\{j: \delta(\tilde{x}_j) = \tilde{y}_j\}}(i) .$$

On our test case, we find: Accuracy = 0.892. The results are shown in Figure 4.8. As you would expect, the misclassifications are in the overlapping area of the two distributions, around the border. Note that in this example, we could calculate the theoretical decision function δ^{th} at any point x by calculating the density ratio:

$$\delta^{th}(x) = \begin{cases} 1 & \text{si } p_1(x) \geq p_0(x) \\ 0 & \text{sinon.} \end{cases}$$

But $p_1(x) \geq p_0(x)$ is equivalent to $\ln(|\Sigma_0|) - \ln(|\Sigma_1|) + (x - m_0)^t \Sigma_0^{-1} (x - m_0) - (x - m_1)^t \Sigma_1^{-1} (x - m_1) \geq 0$. The theoretical decision boundary is not linear. It is shown in Figure 4.9. We find Accuracy = 0.901 on our 1000 test samples. Classically, we have $\mathbb{E}(\text{Accuracy}(\delta)) \leq \mathbb{E}(\text{Accuracy}(\delta^{th}))$ for any classifier δ . We can estimate those expectations at the limit when N_{test} tends towards $+\infty$. Here, for the given training set, $\mathbb{E}(\text{Accuracy}(\delta)) = 0.899$ and $\mathbb{E}(\text{Accuracy}(\delta^{th})) = 0.903$.

Comparison with the k -nearest neighbors classifier:

We can as well compare with the k -nearest neighbors classifier. Let $V_k(x)$ be the set of the k nearest neighbors of x in the training set. We can similarly choose a decision threshold δ_0^{NN} to

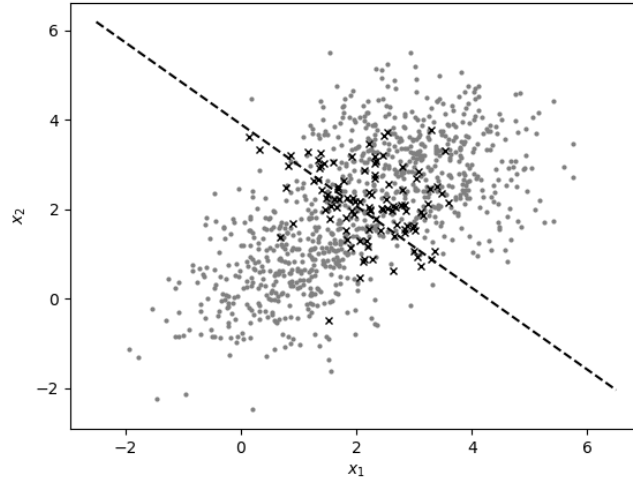


Figure 4.8: Grey dots are test samples that are well classified, black crosses are those that are misclassified. Dotted lines indicate the boundary of the decision function. $\hat{\beta}_0 + \hat{\beta}^t x = 0$.

determine the decision function δ^{NN} :

$$\delta^{NN}(x) = \begin{cases} 1 & \text{si } \frac{1}{k} \sum_{x_i \in V_k(x)} y_i \geq \delta_0^{NN} \\ 0 & \text{sinon.} \end{cases}$$

The classical choice is again $\delta_0 = 1/2$, which then corresponds to a majority vote approach to determine the decision function. Note that the nearest neighbors approach by majority vote extends directly to multi-class classification.

On the test data set, we find for $k = 5$, an Accuracy = 0.880. Note that as for the regression, we could try to choose a better hyperparameter k beforehand by cross-validation for example. In Figure 4.10, we see the areas associated with the decision function obtained by the 5-nearest neighbors method. It shows local irregularities (e.g. non-connex areas). If k increases, the decision regions become more regular, as shown in Figure 4.11 for $k = 100$, and in this case, becomes close to those of the logistic classification. ■

4.3.2 Performance criteria and classification decision

In this section, we look in more detail at the evaluation of the performance of classification methods (whether linear or not) and the criteria for constructing the decision rule.

If we consider a classifier that seeks to automatically recognize handwritten numbers, the problem is (normally) symmetrical, each error has the same severity, each success is of equal interest. This is clearly no longer the case if we now consider the detection of a tumour in a radiology scanner, the detection of anomalies in an airbag production line or a problem of insurance fraud detection. It should be noted that in these different cases, positive detection generally leads to additional examinations which make a false alarm less serious than missing a case (especially in the case of detection of a disease). On the other hand, the methods of detecting spam in emails leads to the deletion of the email or the filing of the email in a folder that is rarely

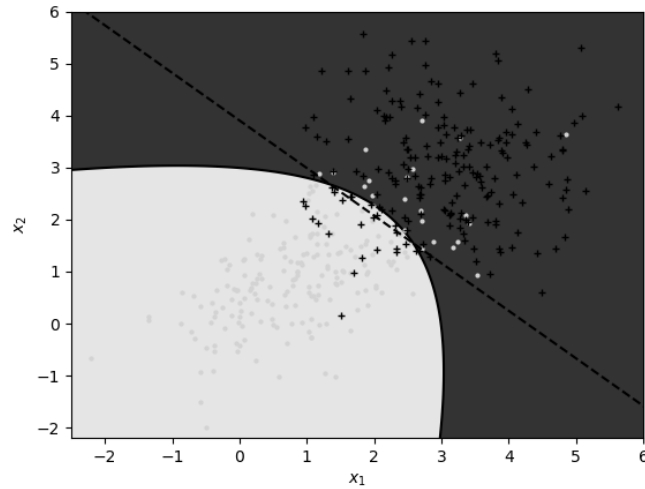


Figure 4.9: **Theoretical Classifier.** The light grey dots are the training samples for the 0 class and the black crosses are the learning samples for the 1 class. The light gray area corresponds to the set of x such that $\delta^{th}(x) = 0$ and the dark gray area corresponds to the set of x such that $\delta^{th}(x) = 1$. In solid line is given the boundary of the theoretical classifier and in dotted line is given the boundary of the decision function for the logistic model.

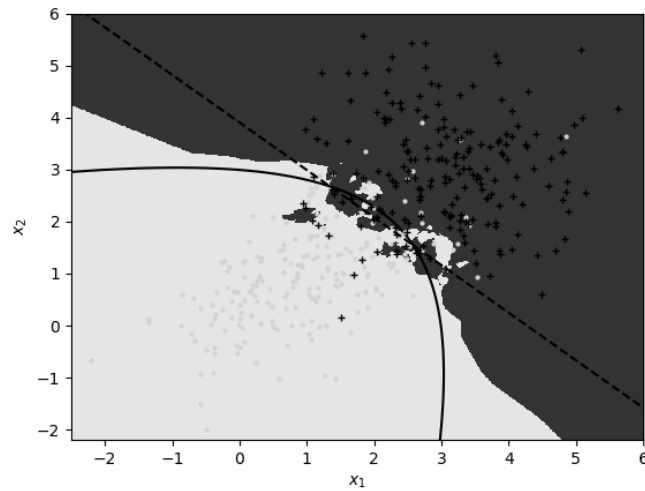


Figure 4.10: **5-nearest neighbors classifier.** The area in dark grey corresponds to the set of x such that $\delta^{NN}(x) = 1$ and the area in light gray corresponds to the set of x such that $\delta^{NN}(x) = 0$. The dashed line shows the boundary of the decision function for the logistic model.

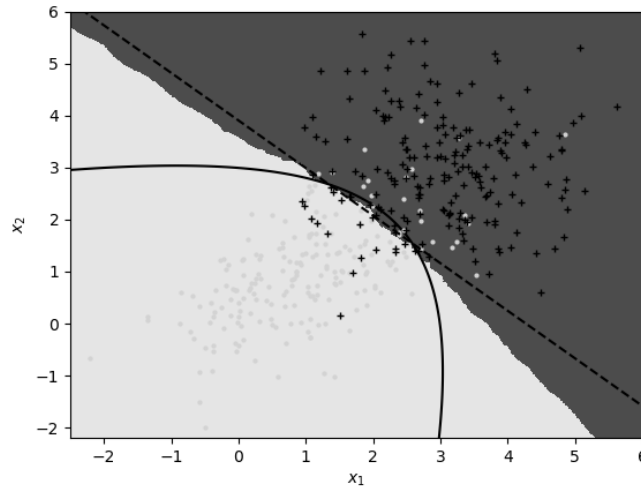


Figure 4.11: **100-nearest neighbors classifier.** The dark gray area corresponds to the set of x such that $\delta^{NN}(x) = 1$ and the light gray area corresponds to the set of x such that $\delta^{NN}(x) = 0$. In dotted line, is given the boundary of the decision function for the logistic model.

consulted: in this case, it is preferable that the algorithm misses the detection of a spam rather than treating an important email as spam! We can see that the situations are different, and that the classification decision must be adapted.

Definition 4.8 — Confusion matrix. Consider a classification problem with $K + 1$ classes, X taking values in $\mathcal{X} \subset \mathbb{R}^p$ and Y in $\{0; 1; \dots; K\}$. Consider the test samples $(\tilde{x}_i, \tilde{y}_i)_{1 \leq i \leq N_{test}}$. The confusion matrix M associated to a classifier is a matrix of order $K + 1$, $M \in \mathcal{M}_{K+1, K+1}(\mathbb{N})$, such that $\forall k, l, 0 \leq k, l \leq K$:

$$M_{k+1, l+1} = \text{Card} \{ (\tilde{x}_i, \tilde{y}_i), 1 \leq i \leq N_{test} : \delta(\tilde{x}_i) = k \text{ et } \tilde{y}_i = l \} ,$$

i.e. the number of items classified in k when their real class is l .

The ideal confusion matrix, corresponding to a 100% Accuracy (success rate) is therefore diagonal.

In the case of the binary classification, Y with values in $\{0; 1\}$, and more particularly in a detection case, the class corresponding to the 1 label is said to be the positive class (case corresponding to anomaly, disease, fraud, spam ...), and the class corresponding to the 0 label is said to be the negative class. The confusion matrix then takes the following form (with notations TP and TN for True Positive and True Negative):

		Real Cases	
		negative (N)	positive (P)
Predictions	predicted negative (PN)	true negative (TN)	false negative (FN)
	predicted positive (PP)	false positive (FP)	true positive (TP)

Definition 4.9 If we make the analogy with the test theory, the case 0 corresponds to the conservative (normal) hypothesis, i.e. to the hypothesis H_0 and the case 1 to the alternative hypothesis H_1 . A false positive is therefore a type I error, i.e. the rejection of H_0 whereas H_0 is true. It is also called **false alarm**. A false negative corresponds to the non-rejection of H_0 whereas it is false, it is type II error.

From this matrix, several rates are calculated (the rate always refers to real cases, e.g. a false positive is a negative real case, so the false positive rate will be FP/N).

Definition 4.10 On appelle:

- True Positive Rate, also called Sensitivity or Recall:

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} .$$

- True Negative Rate, also called Specificity:

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} .$$

- False Positive Rate or false alarm rate or antispecificity:

$$FPR = \frac{FP}{N} = \frac{FP}{TN + FP} = 1 - TNR .$$

- False Negative Rate:

$$FNR = \frac{FN}{P} = \frac{FN}{TP + FN} = 1 - TPR .$$

- Precision:

$$\text{Precision} = \frac{TP}{PP} = \frac{TP}{TP + FP} .$$

Finally, we recall the definition of Accuracy which is the success rate (or 1 - error rate) and is given by: $\text{Accuracy} = (TN + TP)/(N + P)$.

■ **Remark 4.17** The confusion matrix and the different criteria defined above are statistics defined on $(\mathcal{X} \times \mathcal{Y})^{N_{test}}$. ■

We often look for a compromise between Sensitivity and False Alarm Rate or between Recall and Precision, i.e. we want to detect the maximum number of positive cases (from cases to problems...), while avoiding crying wolf for no reason...

One way to represent this compromise is the ROC curve for Receiving Operator Characteristic, which plots the TPR (true positive rate) against the FPR (false alarm rate) when the detection threshold varies. For example in the binary classification, we have defined the decision function (see 4.6):

$$\delta : x \mapsto \begin{cases} \delta(x) = 1 & \text{si } \mathbb{P}(Y = 1|X = x) \geq \delta_0 \\ \delta(x) = 0 & \text{si } \mathbb{P}(Y = 1|X = x) < \delta_0 \end{cases}$$

For a specific test configuration, we evaluate TPR and FPR , when δ_0 varies in $[0; 1]$.

■ **Example 4.18 — ROC curve, continuation of Example 4.16.** We make δ_0 vary from 0 to 1 and perform the classification for our test data, using the logistic model and the 5-nearest neighbors method. Figure 4.12 shows the result. One way to measure the overall performance

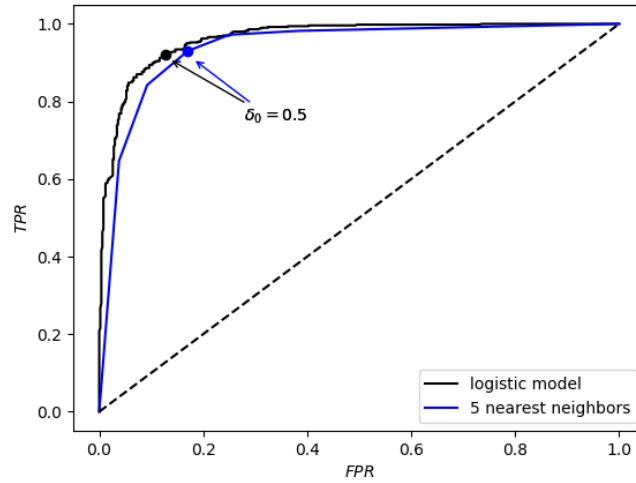


Figure 4.12: **ROC curves.** In black, the ROC curve of the logistic model, in blue that of the 5 nearest neighbours method. The dots represent the pairs (TPR, FPR) for $\delta_0 = 0.5$. The first bisector corresponds to a random classification strategy.

of the classification model on the test dataset is to give the *AUC*, the **Area Under Curve**. For example, $AUC = 0.963$ for the logistic model and $AUC = 0.927$ for the 5-nearest neighbors model for the ROC curves on the figure 4.12. ■

To illustrate the compromise between detection and false alarms, it is sometimes preferable to plot the precision curve as a function of the recall. Similarly the F_1 -score is the harmonic mean between precision and recall, and is thus an indicator of the trade-off between the two:

$$F_1 = 2 \frac{\text{Precision} \times \text{Rappel}}{\text{Precision} + \text{Rappel}}.$$

What decision should be made in terms of classification, in the case of asymmetric binary classification? The following theorem makes the link with the decision theory presented at the beginning of Section 4.1.2.

Theorem 4.8 Let's consider a case of binary classification, such that we are able to evaluate costs for each of the configurations: cost of true positive α_{TP} , cost of false positive α_{FP} , cost of true negative α_{TN} and cost of false negative α_{FN} , then we can calculate the δ^* that

minimizes the expectation of the loss function by:

$$\delta^*(x) = \begin{cases} 1 & \text{si } \mathbb{P}(Y = 1|x) \geq \delta_0 = \frac{(\alpha_{FP} - \alpha_{TN})}{(\alpha_{FP} - \alpha_{TP}) + (\alpha_{FN} - \alpha_{TN})} \\ 0 & \text{sinon,} \end{cases}$$

Preuve : As discussed in Section 4.1.2, we're trying to solve the minimization problem:

$$\begin{aligned} \delta^* &= \arg \min_{\delta} \int_{\mathcal{X} \otimes \mathcal{Y}} L(y, h(x)) p(x, y) dx dy \\ &= \arg \min_{\delta} \int_{\mathcal{X}} \mathbb{P}(Y = 1|x) [\delta(x)\alpha_{TP} + (1 - \delta(x))\alpha_{FN}] + \mathbb{P}(Y = 0|x) [\delta(x)\alpha_{FP} + (1 - \delta(x))\alpha_{TN}] p(x) dx, \end{aligned}$$

where we use $\mathbb{P}(Y|x)p(x)$ (from the model) as an approximation of $p(y|x)p(x) = p(x, y)$ (the real distribution).

We see that the minimum is attained in δ^* if for almost all $x \in \mathcal{X}$,

$$\delta(x) [\mathbb{P}(Y = 1|x) (\alpha_{TP} - \alpha_{FN}) + \mathbb{P}(Y = 0|x) (\alpha_{FP} - \alpha_{TN})] + \mathbb{P}(Y = 1|x)\alpha_{FN} + \mathbb{P}(Y = 0|x)\alpha_{TN}$$

is minimum. It is an affine function in $\delta(x)$. Its minimum therefore depends on the sign of the slope, and we take:

$$\delta^*(x) = \begin{cases} 1 & \text{si } \mathbb{P}(Y = 1|x) (\alpha_{TP} - \alpha_{FN}) + \mathbb{P}(Y = 0|x) (\alpha_{FP} - \alpha_{TN}) \leq 0 \\ 0 & \text{sinon.} \end{cases}$$

which we can rewrite:

$$\delta^*(x) = \begin{cases} 1 & \text{si } \mathbb{P}(Y = 1|x) (\alpha_{TP} - \alpha_{FN}) + (1 - \mathbb{P}(Y = 1|x)) (\alpha_{FP} - \alpha_{TN}) \leq 0 \\ 0 & \text{sinon.} \end{cases}$$

and

$$\delta^*(x) = \begin{cases} 1 & \text{si } \mathbb{P}(Y = 1|x) \geq \frac{(\alpha_{FP} - \alpha_{TN})}{(\alpha_{FP} - \alpha_{TN}) + (\alpha_{FN} - \alpha_{TP})} \\ 0 & \text{sinon,} \end{cases}$$

where we used the fact that $\alpha_{FP} > \alpha_{TP}$ and $\alpha_{FN} > \alpha_{TN}$: the cost of making a mistake is greater than the cost of not making a mistake. The cost of not making a mistake may possibly be non-zero, if for example it induces additional tests that need to be considered in the decision function. $\square$

■ **Remark 4.19** En posant

$$\delta_0 = \frac{(\alpha_{FP} - \alpha_{TN})}{(\alpha_{FP} - \alpha_{TP}) + (\alpha_{FN} - \alpha_{TN})},$$

we find the general expression of the decision function given in 4.6. The classical case is the symmetrical one, that is to say $\alpha_{FN} = \alpha_{FP}$ et $\alpha_{TN} = \alpha_{TP}$: we then get $\delta_0 = 1/2$. ■

4.4 Regularization, Variable Selections, Model Selection

Let's get back to the regression case: let X, Y be a random variables taking values in $\mathcal{X} \subset \mathbb{R}^p$, and in $\mathcal{Y} \subset \mathbb{R}$.

Let $(x_i, y_i)_{1 \leq i \leq N}$, be an N -sample of observations for (X, Y) which we use as a training data set for a linear parameter regression model $\gamma = (\beta_0, \beta) \in \mathbb{R}^{p+1}$.

x_{ij} is the j -th component of explanatory variables x_i . Let's recall the notations:

$$A = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{N1} & \dots & x_{Np} \end{pmatrix}, \quad B = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

In the 4.3 theorem, we needed the assumption that A was injective in order to invert $(A^t A)$ in the regression formula. In practice, the non-injective case is very frequent, for example as soon as the number of samples $N \leq p$ or when the x_i 's are highly correlated with each other. This is called overfitting. One way to remedy this is to add constraints on the parameters, for example that they are not too large, which can result in a constraint on the norm of these parameters. The ℓ_2 or ℓ_1 norms are generally used. The linear regression problem is then written, in the case of ℓ_2 :

$$\hat{\beta} = \arg \min_{\gamma} \|A\gamma - B\|^2$$

sous la contrainte: $\|\gamma\|_2^2 \leq t$

and in the case of ℓ_1 :

$$\hat{\beta} = \arg \min_{\gamma} \|A\gamma - B\|^2$$

sous la contrainte: $\|\gamma\|_1 \leq t$.

Karush-Kuhn-Tucker Theorem in Optimization (Nocedal and Wright, 2006) in the convex case allows rewriting the equivalent formulations, respectively for the ℓ_2 and ℓ_1 cases:

$$\exists \lambda > 0 \text{ tel que } \hat{\beta} = \arg \min_{\gamma} \|A\gamma - B\|^2 + \lambda \|\gamma\|_2^2 \quad (4.16)$$

or

$$\exists \lambda > 0 \text{ tel que } \hat{\beta} = \arg \min_{\gamma} \|A\gamma - B\|^2 + \lambda \|\gamma\|_1. \quad (4.17)$$

The terms $\lambda \|\gamma\|_2^2$ or $\lambda \|\gamma\|_1$ are said penalty or regularization terms. We can also show that $\forall \lambda$, each of these two penalized problems admits one and only one solution.

Definition 4.11 — Ridge Estimator. Let $\lambda > 0$. The estimator determining $\hat{\gamma}^{ridge}$ by:

$$\hat{\gamma}^{ridge} = \arg \min_{\gamma} \|A\gamma - B\|^2 + \lambda \|\gamma\|_2^2 \quad (4.18)$$

is called **ridge** estimator.

Theorem 4.9 The ridge estimator exists and is unique for any A and $\lambda > 0$. It is given by:

$$\hat{\gamma} = (\hat{\beta}_0, \hat{\beta}) = (A^t A + \lambda I_{p+1})^{-1} A^t B. \quad (4.19)$$

Preuve : We simply adapt the proof of Theorem 4.3. We have coercivity directly through the term $\lambda \|\gamma\|^2$, from which we deduce the solution existence. The function is moreover convex, so the necessary and sufficient condition verified by the solution is::

$$\nabla R(\hat{\gamma}^{ridge}) + \lambda(\gamma^{ridge}, \gamma^{ridge}) = 0$$

that is to say

$$2A^t(A\hat{\gamma}^{ridge} - B) + 2\lambda\hat{\gamma}^{ridge} = 0$$

$(A^t A + \lambda I_{p+1})$ is invertible, which ends the proof. $\square$

Since the aim is to keep the parameters as small as possible, those corresponding to the least influential covariates will be smaller than those corresponding to the most influential covariates.

Definition 4.12 — Lasso estimator. Let $\lambda > 0$. The estimator determining $\hat{\gamma}^{lasso}$ by

$$\hat{\gamma}^{lasso} = \arg \min_{\gamma} \|A\gamma - B\|^2 + \lambda \|\gamma\|_1 \quad (4.20)$$

is said **lasso** estimator (for *Least Absolute Shrinkage and Selection Operator*.)

Unlike the ridge estimator, there is no explicit expression of the solution. Numerical estimation methods must therefore be used. One difficulty is related to the non-differentiability of the ℓ_1 norm.

We can show that if X is a continuous random variable, then the lasso solution is almost certainly unique (Tibshirani et al., 2013). But the solution is not necessarily unique if the inputs are discrete.

■ **Remark 4.20 — Variable Selection by Lasso Regression.** A very interesting result of the lasso regression is that it allows to select the most important covariates. Indeed, when λ increases, the β_i coefficients of the regression are set to 0 one after the other. The demonstration that explains it uses optimization results that are beyond the scope of this course, but it can for example be found in Tibshirani et al. (2013). On the other hand, it is possible to understand what happens thanks to the figure 4.13: the minimum of $R(\gamma)$ under the constraint $\|\gamma\|_1 \leq t$ is obtained when the curve of the isovalues closest to the minimum (represented by the point) crosses the area of the admissible parameters. We can see that this occurs at a corner of the square, therefore for a γ with a coefficient of zero. The same phenomenon occurs in any dimension. We see on the other hand that for the ridge regression, the admissible area of the parameters checking the constraint $\|\gamma\|_2^2 \leq t$ would be a disk, so the coefficients are generally not zero.

This property of the lasso estimator makes it a very interesting tool for the selection of the most relevant explanatory variables in a model. In figure 4.14, we have built a model with 10 explanatory covariates. For a data set $(x_i, y_i)_{1 \leq i \leq N}$, we perform the lasso estimation for different values of the regularization coefficient λ . We can see what happens when we increase

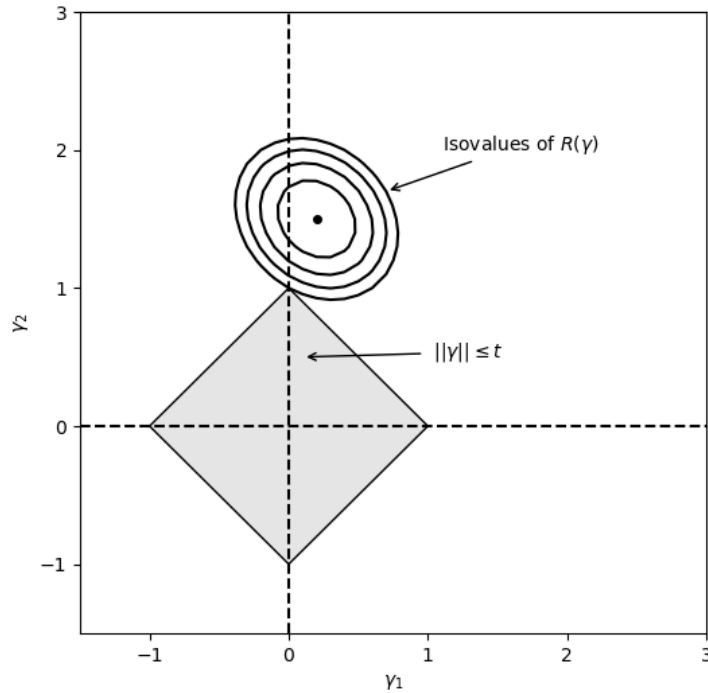


Figure 4.13: Lasso Regression: We represent the isovalues of the criterion to be minimized $R(\gamma)$ and the set of γ that verifies the constraint $\|\gamma\|_1 \leq t$. The point at which the minimum of the function is attained not being in the admissible area, we look for the first curve of isovalues meeting the admissible area. We see that the intersection takes place at an angle, i.e. in our configuration at a point where the coefficient $\gamma_1 = 0$.

λ : on the left, the coefficients correspond to those of a classical linear regression (because the penalty is negligible); when *lambda* increases, we see that little by little the coefficients are set to 0. The order in which this happens makes it possible to build nested models, keeping only the most relevant explanatory variables. Lasso regression is therefore a very important tool in the construction of parsimonious (or sparse) models³. ■

■ **Remark 4.21** In practice, for ridge and lasso regressions, the λ parameter is considered a **hyperparameter**. That is to say that it is generally not fixed a priori, but selected in order to give the best performance in prediction, for example by cross-validation (see Figure 4.5). For the lasso, once the λ has been determined, and the selection of the explanatory variables carried out, it may be interesting to determine $\hat{\gamma}^{sel}$ which minimizes the $R(\gamma^{sel})$ criterion not penalized this time (where γ^{sel} corresponds to the subset of components of γ selected by the lasso, the others being set to 0): This eliminates the bias introduced by the penalty in $\hat{\gamma}^{lasso}$. ■

³The principle of parsimony or sparsity consists in using only the minimum of elementary causes to explain a phenomenon. This principle was notably stated by Guillaume d'Ockham, and is also called Ockham's razor, because it aims at 'shaving', i.e. eliminating unnecessary explanations.

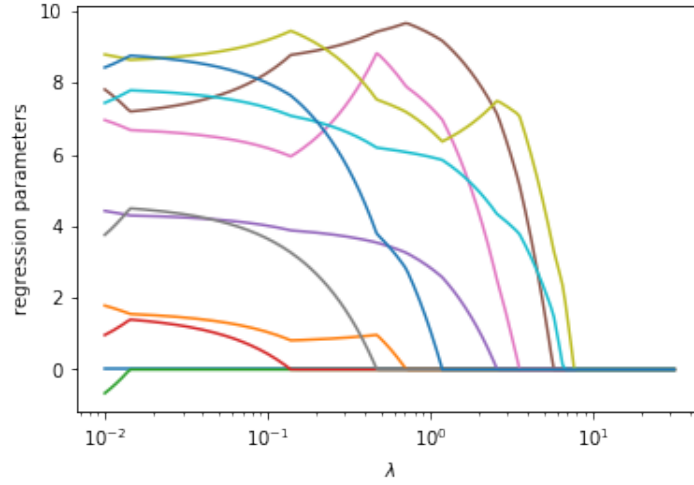


Figure 4.14: Evolution of regression coefficients $\hat{\beta}_0^{lasso}, \hat{\beta}_1^{lasso}, \dots, \hat{\beta}_{10}^{lasso}$ according to λ (in logarithmic scale).

■ **Remark 4.22 — Data normalization / standardization.** As the ridge and lasso penalties consider the norm of the vector of parameters, if the different explanatory variables correspond to totally different quantities (for example in the case of clinical variables: a person's age, body mass index, blood pressure, stage of a disease ...), the parameters β_i will also have very different orders of magnitude, which will therefore play on the norm. Therefore, it is important to standardize the input data to bring them back to comparable orders of magnitude. Several normalizations are possible.

The most classical one is called **standardization**, it consists in taking as covariates $w_{.j}$, for $1 \leq j \leq p$, with for all i , $1 \leq i \leq p$:

$$w_{ij} = \frac{x_{ij} - \bar{x}_{.j}}{s_{.j}},$$

where $\bar{x}_{.j}$ and $s_{.j}$ are respectively the empirical mean and standard deviation of the j -th covariate on the training data.

$$\bar{x}_{.j} = \frac{1}{N} \sum_{i=1}^N x_{ij}, s_{.j}^2 = \frac{1}{N} \sum_{i=1}^N (x_{ij} - \bar{x}_{.j})^2.$$

It is important to note that when switching to prediction on a new dataset, the same standardization must be applied to the input data in order to use the model, with the empirical means and standard deviations calculated on the training set.

Note that we can equally standardize the outputs: if we take

$$z_i = \frac{y_i - \bar{y}}{s_Y}$$

with $\bar{y}$ and s_Y the empirical mean and standard deviation of the output variable on the training data. Note that in this case β_0 is zero! ■

■ **Remark 4.23** It is **important** to note that for the sake of simplicity, the penalty methods have been presented in a general framework, but in practice the preferable framework for performing lasso and ridge regressions is to standardize the input variables X and to center the output variable Y . We can indeed show that some interesting properties of the estimator are preserved in this situation Hastie et al. (2009). In particular, we can see that the penalty of β_0 , the intercept in the regression, has no legitimate justification (perturbation of the geometry). ■

■ **Remark 4.24** Ridge and lasso penalties also apply to non-linear models and to classification models. This is done by simply adding the penalty term to the criterion to be optimized. For example, for the binary classification logistic model (cf. section 4.3.1), the criterion to be minimized is given by Equation 4.13:

$$\ell(\gamma; (x_i, y_i)_{1 \leq i \leq N}) = \sum_{i=1}^N y_i (\gamma^t z_i) - \ln(1 + \exp(\gamma^t z_i)) .$$

and the penalized criteria are given for the ridge regression by:

$$\ell(\gamma; (x_i, y_i)_{1 \leq i \leq N}) + \lambda \|\gamma\|_2^2$$

and for the lasso regression by:

$$\ell(\gamma; (x_i, y_i)_{1 \leq i \leq N}) + \lambda \|\gamma\|_1 .$$

■

Model Selection

We have seen that the ℓ_1 penalty in the lasso regression allows to cancel some coefficients, and thus to select explanatory variables, which in the end allows to obtain a simpler model. This is a major concern in modelling: finding the right compromise between the complexity of the model and its robustness in prediction. In the training phasis, a higher complexity of the model will generally allow a better minimization of the loss function and thus to capture the most specific behaviors. The counterpart is that some of these behaviors may simply be due to noise or non generalizable phenomena: this phenomenon is called "overfitting". In a prediction test on new inputs, this model will generally result in a lack of robustness. On the contrary, a model that is too simple will not be able to capture non-trivial variations, and will therefore be of limited interest.

Model selection theory is a very rich field of mathematics, (cf. for example Burnham and Anderson (2002)). We will restrict ourselves here with a few elementary concepts. One of the first objectives of a model is to predict. Generalization error is a measure of the ability of a trained (calibrated) model to predict in situations different from those represented in the training data.

Definition 4.13 — Generalization error. Let (X, Y) take values in $\mathcal{X} \times \mathcal{Y}$, and consider $\mathcal{T} = (x_i, y_i)_{1 \leq i \leq N}$, an i.i.d. sample providing the training data.

We consider a model whose prediction error is h . The generalization error associated with the model, conditionally to $\mathcal{T}$ is:

$$EG_{\mathcal{T}} = \mathbb{E}_{P_{(X,Y)}} [L(Y, h(X|\mathcal{T}))]$$

where L is the loss function (for example the quadratic loss for a regression problem or the cross entropy for a classification problem).

The notation $h(X|\mathcal{T})$ is used to emphasize that the decision function depends on the set of training data used for its calibration.

Note that generally, the generalization error is greater than the training error:

$$EA_{\mathcal{T}} = \frac{1}{N} \sum_{i=1}^N L(y_i, h(x_i|\mathcal{T})) .$$

Indeed, h was determined to minimize this training error. The difference between generalization and training errors generally increases with the complexity of the model.

■ **Remark 4.25 — Estimation of the generalization error by cross-validation.** As illustrated in Example 4.11 and Figure 4.5, a way to approximate $EG_{\mathcal{T}}$ is by cross-validation. For K -fold cross-validation, we build from $\mathcal{T}$, K pairs: $(\mathcal{T}_k, \mathcal{T}_{-k})_{1 \leq k \leq K}$, where $\mathcal{T}_{-k} = \{(x_j, y_j) \in \mathcal{T} : (x_j, y_j) \notin \mathcal{T}_k\}$. We obtain:

$$\hat{EG}_{\mathcal{T}}^{cv} = \frac{1}{K} \sum_{1 \leq k \leq K} \frac{1}{|\mathcal{T}_{-k}|} \sum_{j \in \mathcal{T}_{-k}} L(y_j, h(x_j|\mathcal{T}_k)) .$$

■

For parametric models $p_{\theta}(y|x)$, and under certain conditions, the generalization error can also be approximated by a formula combining the training error and the number of parameters. From this approximation we deduce a very useful index for model comparison, the AIC (Akaike Information Criteria).

Definition 4.14 — AIC (Akaike Information Criteria). Let Z be a random variable with values in $\mathcal{Z}$, and a parametric model on $\mathcal{Z}$, $\mathcal{M}_{\Theta} = \{p_{\theta}, \theta \in \Theta \subset \mathbb{R}^p\}$. Let $(z_1, \dots, z_N)$, an i.i.d. sample of Z , and $\hat{\theta}$ the maximum likelihood estimator of the model parameters. The *AIC* is defined by:

$$AIC = -2 \ln \mathcal{L}(\hat{\theta}; (z_1, \dots, z_N)) + 2p .$$

The smaller the *AIC*, the smaller the generalization error (and then better the model, or more exactly, bigger the probability for the model to have better prediction).

Suppose that we have $\mathcal{M}_1$ and $\mathcal{M}_2$, with computed criteria AIC_1 and AIC_2 and $AIC_1 < AIC_2$. Then:

$\mathcal{M}_1$ minimizes the information loss with probability: $\frac{1}{1 + \exp((AIC_1 - AIC_2)/2)}$

$\mathcal{M}_2$ minimizes the information loss with probability: $\frac{1 + \exp((AIC_1 - AIC_2)/2)}{1 + \exp((AIC_1 - AIC_2)/2)}$.

For example, for $AIC_2 - AIC_1 = 2$, $\mathcal{M}_1$ will be better with a probability of ≈ 0.73 and $\mathcal{M}_2$ will be better with a probability of ≈ 0.27 .

This is an asymptotic result, so in the case of a reduced number of data, the corrected *AIC* will be preferred, denoted *AIC_c*:

$$AIC_c = AIC + \frac{2p^2 + 2p}{N - p - 1} .$$

4.5 Basis Expansion for Non-linear Regression

Let X be a random variable with values in $\mathbb{R}^p$ and Y with values in $\mathbb{R}$. Linear models have the advantage of simplicity, readability and explicit learning. Unfortunately, they are sometimes too limited. One way to adapt linear models to give them more flexibility is to transform p explanatory variables into M derived explanatory variables, using for $1 \leq m \leq M$, well-chosen transformations $\psi_m : \mathbb{R}^p \rightarrow \mathbb{R}$. The decision function is then given by a linear model in these new derived variables. For the regression:

$$h(X) = \sum_{m=1}^M \beta_m \psi_m(X) .$$

et pour la classification logistique:

$$\text{logit} [\mathbb{P}(Y = 1|X)] = \sum_{m=1}^M \beta_m \psi_m(X) .$$

It is called basis expansion because the functional basis on which $\mathbb{E}(Y|X)$ is decomposed is enriched with new elements.

Definition 4.15 A generalized linear model with basis expansion is a model such that there exists a link function $g : \mathcal{Y} \rightarrow \mathbb{R}$, coefficients $\beta = (\beta_1, \dots, \beta_M) \in \mathbb{R}^M$, and functions $\psi_m : \mathbb{R}^p \rightarrow \mathbb{R}$, verifying:

$$g(\mathbb{E}(Y|X)) = \sum_{m=1}^M \beta_m \psi_m(X) .$$

■ **Example 4.26** We give below some classic examples of functions ψ_m :

- $\psi_j(X) = X^j$, where X^j represents the j -th component of X . We then get the classical linear model (without bias term, but we can of course add a constant function $\psi_0(X) = 1$).
- Quadratic interaction terms are often added to the classical linear model by taking $\psi_m(X) = (X^j)^2$ or $\psi_m(X) = X^i X^j$.
- Logarithmic transformations can be used $\psi_m(X) = \ln(X^j)$ or square root transformations $\psi_m(X) = \sqrt{X^j} \dots$
- To obtain a constant piecewise decision function, a space partition is used. Let $\{Z_m \subset \mathbb{R}^p, 1 \leq m \leq M\}$ be this partition, we then take $\psi_m(X) = \mathbb{I}_{Z_m}(X), 1 \leq m \leq M$.
- The **generalized additive models** are a special type of models with basis expansion:

$$g(\mathbb{E}(Y|X)) = \beta_0 + \sum_{i=1}^p \beta_i \psi_i(X^i) ,$$

i.e. such that the contribution of each variable can be additively decomposed. An important example is when each $\psi_i(X^i)$ is a cubic spline (i.e. a polynomial interpolation of degree 3 on intervals and $\mathcal{C}^2$, i.e. continuous, with continuous derivatives and second derivatives on $\mathbb{R}$).

■

■ **Remark 4.27** Let $(x_i, y_i)_{1 \leq i \leq N}$ be a training dataset.

If the link function is the identity function and if errors have a Gaussian distribution, i.e. if our model is of the classical form:

$$Y = \sum_{m=1}^M \beta_m \psi_m(X) + \xi, \text{ avec } \xi \sim \mathcal{N}(0, \sigma^2), \quad (4.21)$$

and if ψ_m does not depend on unknown parameters, then we can deduce the explicit regression formulas from the classical linear regression of Theorem 4.3.

Let's denote from now on:

$$A = \begin{pmatrix} \psi_1(x_1) & \psi_2(x_1) & \dots & \psi_M(x_1) \\ \vdots & \vdots & & \vdots \\ \psi_1(x_N) & \psi_2(x_N) & \dots & \psi_M(x_N) \end{pmatrix}, \quad B = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, \quad R(\beta) = \|A\beta - B\|^2.$$

Then, if A is injective, the maximum likelihood estimator of the linear regression model 4.21 is given by:

$$\hat{\beta} = (A^t A)^{-1} A^t B \quad (4.22)$$

and

$$\hat{\sigma}^2 = \frac{1}{N} R(\hat{\beta}) = \frac{1}{N} \sum_{i=1}^N \left(y_i - \sum_{m=1}^M \hat{\beta}_m \psi_m(x_i) \right)^2. \quad (4.23)$$

■

■ **Remark 4.28** The functional basis given by $\{\psi_1, \dots, \psi_M\}$ is sometimes called **dictionary**. It can be chosen very broadly, combining several types of functional decomposition: linear, quadratic, localized regression (such as splines), different transformations... During the estimation, a penalty term is then used, notably the lasso, which allows to select in the dictionary the most important elements to explain the response. ■

4.5.1 Decision Trees

We return to one of the particular types of basic linear expansion models, the piecewise constant models, especially those with recursively defined regions.

Definition 4.16 — Decision Tree. Let X, Y be random variables with values in $\mathcal{X} \subset \mathbb{R}^p$ and $\mathcal{Y} \subset \mathbb{R}$. A decision tree is a model such that

$$\mathbb{E}(Y|X) = \sum_{m=1}^M \beta_m \mathbb{I}_{Z_m}(X)$$

where $\{Z_m \subset \mathbb{R}^p, 1 \leq m \leq M\}$ is a partition of $\mathcal{X}$ which can be defined from a binary tree, in which each node partitions space based on a decision rule on one of the explanatory variables.

The terminal nodes are the leaves of the tree.

Let us call $\mathcal{E}^j$ the domain in which the explanatory variable X^j takes its values (for example $\mathbb{R}$ for a quantitative variable, but it can also be $\{0; 1\}$ for categorical variables or $\{1; 2; 3; 4\}$ for ordinal variables or something else).

Formally, we build a sequence of partitions $\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_M$, where $\mathcal{P}_1 = \{\mathcal{X}\}$, $\mathcal{P}_M = \{Z_1, \dots, Z_M\}$ and $\forall k, 1 \leq k \leq M-1$, if $\mathcal{P}_k = \{Z_1^k, \dots, Z_k^k\}$, then $\exists j, 1 \leq j \leq k$, $\exists D^j \subset \mathcal{E}^j$, and $\exists l, 1 \leq l \leq k$ such that

$$\mathcal{P}_{k+1} = \left\{ Z_i^k, 1 \leq i \leq k, i \neq l \right\} \cup \left\{ \left\{ x \in Z_l^k \mid x^j \in D^j \right\} \right\} \cup \left\{ \left\{ x \in Z_l^k \mid x^j \in \overline{D^j} \right\} \right\},$$

where $\overline{D^j}$ is the complementary set of D^j in $\mathcal{E}^j$. (The double braces comes from the fact that the set of sets is a singleton).

If X^j is a quantitative variable, D^j is generally taken in the form $D^j = \{x \in \mathcal{X} : x^j \leq \delta^j\}$.

Decision trees work for both regression and classification. In figure 4.15, we give a simple example of a 4-leaf (4 regions) decision tree for classification, for X taking values in $[0; 1]^2$ and Y in $\{0; 1\}$.

A difficulty in learning such models is that ψ_m 's now depend on unknown parameters, especially those defining the D^j regions. If we fix the number of regions M and consider quantitative variables, we then look for the parameters $(l_k, j_k, \delta_k)_{1 \leq k \leq M-1}$ defining the tree and the associated partition $\mathcal{P}_M = \{Z_1, \dots, Z_M\}$, as well as $\beta = (\beta_1, \dots, \beta_M)$ to minimize:

$$J((l_k, j_k, \delta_k)_{1 \leq k \leq M-1}, \beta) = \sum_{i=1}^N L \left(y_i, \sum_{m=1}^M \beta_m \mathbb{I}_{Z_m}(x_i) \right),$$

with L the loss function.

This problem is too complex to be solved directly, and an iterative approach is adopted (without, however, any guarantee of convergence towards the local minimum).

Estimation Procedure.

Note first that J rewrites:

$$J((l_k, j_k, \delta_k)_{1 \leq k \leq M-1}, \beta) = \sum_{m=1}^M \sum_{i: x_i \in Z_m} L(y_i, \beta_m).$$

If we consider a regression problem, and the quadratic loss function, it is easy to see that for a partition $\mathcal{P}_M$,

$$\sum_{m=1}^M \sum_{i: x_i \in Z_m} (y_i - \bar{y}^m)^2 \leq \sum_{m=1}^M \sum_{i: x_i \in Z_m} (y_i - \beta_m)^2, \forall \beta_1, \dots, \beta_M$$

where $\bar{y}^m$ is the mean of the y_i 's, for all $x_i \in Z_m$:

$$\bar{y}^m = \frac{\sum_{i: x_i \in Z_m} y_i}{|\{i : x_i \in Z_m\}|},$$

and where for a set E , $|E|$ denotes its cardinal.

We then build the nodes of the decision tree iteratively: we first have $l_1 = 1$ since there is only one region $Z_1^1 = \mathcal{X}$, and we look for (j_1, δ_1) such that

$$\sum_{i: x_i \in Z_1^2(j_1, \delta_1)} (y_i - \bar{y}^1)^2 + \sum_{i: x_i \in Z_2^2(j_1, \delta_1)} (y_i - \bar{y}^2)^2$$

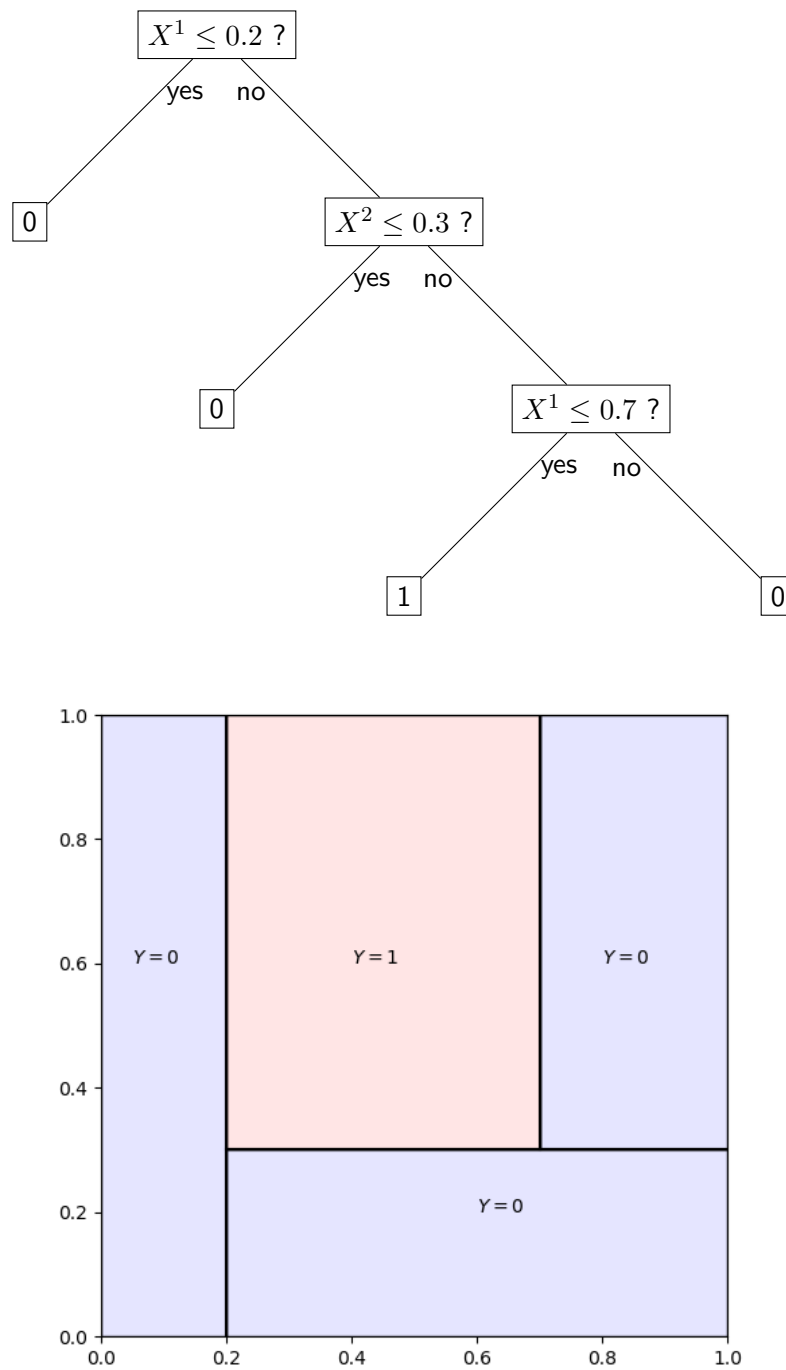


Figure 4.15: Above: decision tree for classification. Bottom: corresponding representation of the space partition: $\{Z_1, Z_2, Z_3, Z_4\}$.

is minimal, with $Z_1^2(j_1, \delta_1) = \{x \in \mathcal{X} : x^{j_1} \leq \delta_1\}$ and $Z_2^2(j_1, \delta_1) = \{x \in \mathcal{X} : x^{j_1} > \delta_1\}$. In practice, we solve p unidimensional minimization problems (by testing each of the explanatory variables), and take the most favorable case.

We repeat the operation iteratively: the difficulty is that at step k , we must also look for l^k , i.e. in which zone $Z_{l_k}^{k-1}$ we can find the best partition. The complexity thus increases with the number of nodes in the tree.

The procedure is the same in the case of classification, with an adapted loss function, e.g. cross-entropy:

$$L(y_i, \beta_m) = -y_i \ln(\beta_m) .$$

Similarly, we can see that for a given partition, the best β_m is also given by $\beta_m = \frac{\sum_{i: x_i \in Z_m} y_i}{|\{i : x_i \in Z_m\}|}$. We note that the procedure then seeks at each iteration to maximize the purity of each region (i.e. to have the greatest number of elements with the same label in the same region).

Choice of the number of terminal nodes

The number M of sets in the partition of the space $\mathcal{X}$ (which is also the number of terminal nodes) determines the complexity of the tree: in the learning procedure, by increasing M , one can always reduce the training error. At the limit, by taking $M = N$, we can find a partition such that the training error is null... Once again, the principle of parsimony applies, and one must find the number M ensuring the best generalization error. We can treat it as a hyperparameter and select it by cross validation for example.

■ **Remark 4.29** Decision trees are rarely used as such. However, they are often used as basic elements in so-called "ensemble methods", which are highly efficient statistical learning methods. The idea is then to use several trees for decision making, exploiting in some way the "wisdom of crowds". In a simplified way, there are two ways to build these sets, in parallel, we talk about **bagging**, and sequentially, trying to correct previous errors at each iteration, we call it boosting. (cf. Schapire et al. (1998)). The first type of method gave notably the (**random forest** method, Breiman (2001)) which is a method of very great robustness and often very good efficiency. The second type gave notably the AdaBoost method. ■

4.6 Artificial Neural Networks

In this part, we introduce some basic concepts about artificial neural networks which are the most used nonlinear models today, in very diverse contexts, and with often remarkable performances. We present in particular the most classical architecture, that of the **multilayer perceptrons**. For more sophisticated architectures, such as convolutional neural networks or recurrent neural networks, see the reference book (Goodfellow et al., 2016).

4.6.1 Multilayer Perceptron for Regression and Classification

Definition 4.17 — Multilayer Perceptron . A perceptron with $M + 1$ layers is a function $\psi : \mathcal{X} \subset \mathbb{R}^p \rightarrow \mathcal{Y} \subset \mathbb{R}^K$, such that there are:

- non-zero integers $m_0, \dots, m_M$, with $m_0 = p$ and $m_M = K$, corresponding to the number of neurons per layer;

- affine transformations $g_1, \dots, g_M$, with $g_k : \mathbb{R}^{m_{k-1}} \rightarrow \mathbb{R}^{m_k}$ and $g_k(z) = W_k z + B_k$ where $W_k \in \mathcal{M}_{m_k, m_{k-1}}(\mathbb{R})$ and $B_k \in \mathbb{R}^{m_k}$;
- non-linear transformations called activation functions $\sigma_1, \dots, \sigma_M$, with $\sigma_k : \mathbb{R} \rightarrow \mathbb{R}$ and $h_k : \mathbb{R}^{m_k} \rightarrow \mathbb{R}^{m_k}$,

$$h_k(z_1, \dots, z_{m_k}) = \begin{pmatrix} \sigma_k(z_1) \\ \vdots \\ \sigma_k(z_{m_k}) \end{pmatrix}$$

and such that $\forall x \in \mathcal{X}$:

$$\psi(x) = h_M \circ g_M \circ \dots \circ h_1 \circ g_1(x). \quad (4.24)$$

The first layer is the input layer, the last one is the output layer. The intermediate ones are called hidden layers. Taking the notations introduced in the definition 4.17, we have thus successively for the neural network illustrated in Figure 4.16:

$$\begin{aligned} z_l^{(1)} &= \sigma_1 \left(\sum_{i=1}^3 [W_1]_{li} x_i + [B_1]_l \right), \text{ for } l = 1, \dots, 4 \\ z_l^{(2)} &= \sigma_2 \left(\sum_{i=1}^4 [W_2]_{li} z_i^{(1)} + [B_2]_l \right), \text{ for } l = 1, \dots, 4 \\ y_l &= \sigma_3 \left(\sum_{i=1}^4 [W_3]_{li} z_i^{(2)} + [B_3]_l \right), \text{ for } l = 1, 2 \end{aligned}$$

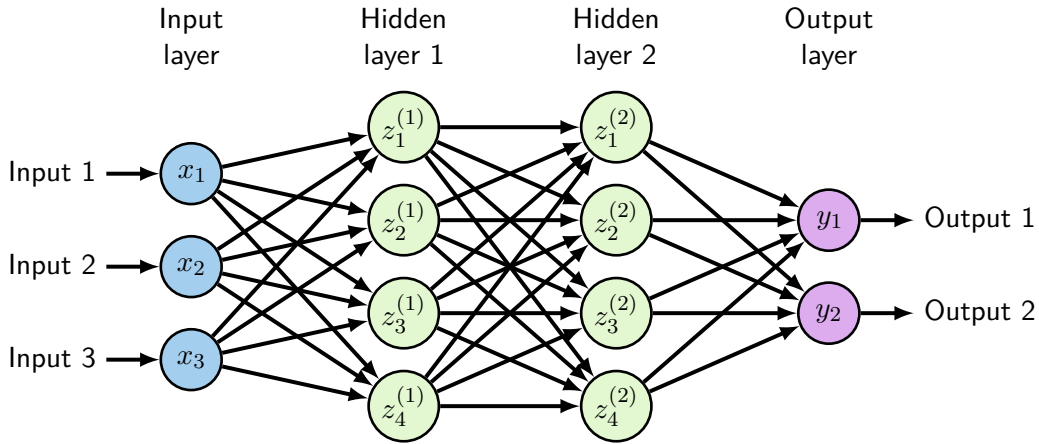


Figure 4.16: Example of a neural network with 3 inputs and 2 outputs, with two hidden layers. The arrows correspond to the links between variables.

In the case of regression, the variables output from the network correspond directly to those given by the decision function.

In the case of classification, we add a transformation such that the outputs of the network y^i 's correspond to the probabilities that $\mathbb{P}(Y = i | X = x)$. The most commonly used function is the softmax function:

Definition 4.18 — softmax. The softmax function, defined by $\mathbb{R}^K \rightarrow \mathbb{R}^K$ is such that:

$$\forall z \in \mathbb{R}^K, \text{softmax}(z) = \left(\frac{\exp(z_1)}{\sum_{k=1}^K \exp(z_k)}, \dots, \frac{\exp(z_K)}{\sum_{k=1}^K \exp(z_k)} \right),$$

where $(z_1, \dots, z_K)$ representing the K components of z .

We see that the components of $\text{softmax}(z)$ sum to 1.

Thus, keeping the previous notations for affine transformations and activation functions, a multi-layer perceptron for classification will generally be in the following form, for a problem with K classes:

$$\psi(x) = \text{softmax} \circ h_M \circ g_M \circ \dots \circ h_1 \circ g_1(x),$$

and $\forall k = 1, \dots, K, [\psi(x)]_k = \mathbb{P}(Y = k | X = x)$.

The associated decision function will generally be chosen as:

$$\delta(x) = \arg \max_{1 \leq k \leq K} [\psi(x)]_k.$$

In the binary case, as discussed in the section on the logistic model, if the two classes are dissymmetric, (cf. Theorem 4.8), the decision threshold for either class may not be 1/2.

4.6.2 Activation Function

Schematically, each neuron in the network receives information from the previous layer, this information is weighted, and according to these different signals, the neuron transmits a response. By analogy with the natural neurons, we speak about activation or inhibition: the σ_k transformations called activation functions play this role by predicting the response. We give a few examples of activation functions below.

- The Heaviside function is the ideal function to reproduce this inhibition-activation effect:

$$\forall x \in \mathbb{R}, f(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 & \text{si } x \geq 0. \end{cases}$$

Nevertheless, its discontinuity and the fact that it has a zero gradient does not make it numerically practical. The sigmoid function is preferred.

- The sigmoid function can be seen as a smooth version of the Heaviside function. (cf. Figure 4.17):

$$\forall x \in \mathbb{R}, f(x) = \frac{1}{1 + e^{-x}}.$$

It can be parameterized by $\alpha \in \mathbb{R}_{*+}$:

$$\forall x \in \mathbb{R}, f_\alpha(x) = \frac{1}{1 + e^{-\alpha x}}.$$

If we want the response values to be within $[-1; 1]$, we may prefer the hyperbolic tangent function $\tanh$.

One of the difficulties with the sigmoid and $\tanh$ activation functions is that for a wide range of values, the gradient is very small. It makes it more difficult to train deep networks.

- The *ReLU* function, for *Rectified Linear Unit* has emerged as one of the most effective activation functions for learning neural networks:

$$\forall x \in \mathbb{R}, f(x) = \max(0, x) .$$

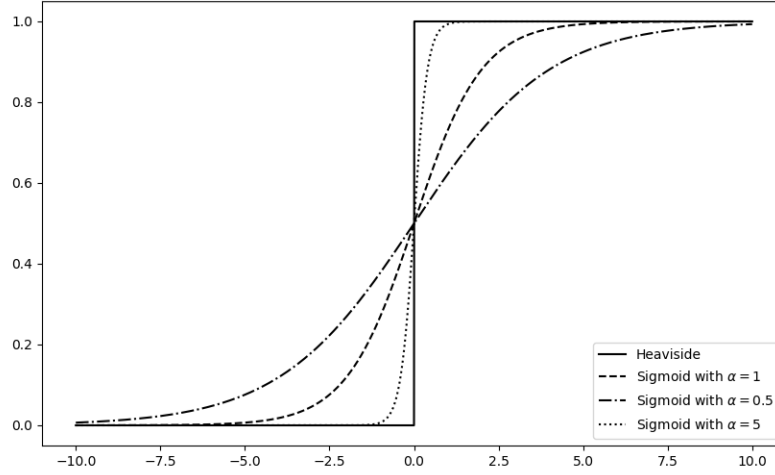


Figure 4.17: Heaviside and sigmoid ctivation functions.

4.6.3 Training Neural Networks

Let's consider a training set $(x_i, y_i)_{1 \leq i \leq N}$.

We assume that the architecture of the neural network is fixed (number of layers, number of neurons per layer), as well as the type of activation functions. These characteristics can be considered as hyperparameters and could be chosen as such, for example by cross-validation.

The parameters to estimate are thus those of the affine transformations, i.e. the weights of the links between neurons (the matrices W_k), and the biases (the vectors B_k). Let us note θ , the set of parameters. We can see that their number can increase very quickly: let us imagine a model of prediction of survival in oncology and as model inputs gene expression levels, which can be of the order of 20000. Let's take a first hidden layer hidden with 140 neurons (of the order of $\sqrt{20000}$), simply for this first layer, we then have: $20000 * 150 + 150 = 3000150$ parameters!

First, a loss function L is chosen: usually the quadratic loss function in the case of regression, and cross-entropy in the case of classification. We then look for:

$$\theta^* = \arg \min_{\theta} \frac{1}{N} \sum_{i=1}^N L(y_i, \psi_{\theta}(x_i))$$

with:

$$L(y_i, \psi_{\theta}(x_i)) = \|y_i - \psi_{\theta}(x_i)\|^2 \text{ pour la r\'egression,}$$

and

$$L(y_i, \psi_\theta(x_i)) = - \sum_{k=1}^K y_{ik} \ln([\psi_\theta(x_i)]_k) \text{ pour la classification,}$$

where y_{ik} and $[\psi_\theta(x_i)]_k$ are the k -th components of vectors y_i and $\psi_\theta(x_i)$.

As has been pointed out, the number of parameters is often very large and we are therefore faced with the risk of overfitting. Thus, we usually add a term of regularization $\lambda\Omega(\theta)$, with a ridge penalty $\Omega(\theta) = \|\theta\|_2^2$ or lasso penalty $\Omega(\theta) = \|\theta\|_1$. (Note that there exist other regularization methods, like the *early-stopping* or the *drop-out*, cf. Goodfellow et al. (2016)). Finally, we look for:

$$\theta^* = \arg \min_{\theta} \frac{1}{N} \sum_{i=1}^N L(y_i, \psi_\theta(x_i)) + \lambda\Omega(\theta) .$$

λ is also a hyper-parameter that can be selected by cross-validation.

Optimization methods to solve this problem are beyond the scope of this course. They are based on two extremely powerful concepts that have played a large role in the recent successes of neural networks (Goodfellow et al., 2016):

- stochastic gradient method: the numerical resolution of minimization problems usually involves an iterative algorithm based on the calculation of the criterion gradient and the determination of a descent direction from it. We recall that $\frac{1}{N} \sum_{i=1}^N L(y_i, \psi_\theta(x_i))$ is the empirical approximation of $\mathbb{E}(L(Y, \psi_\theta(X)))$. Since the algorithm to find the minimum is iterative, the determination of $\nabla \mathbb{E}(L(Y, \psi_\theta(X)))$ does not need to be done extremely precisely at each iteration, and a small subsample of points among the N is enough to approximate it (Robbins-Monroe's stochastic approximation theory). The efficiency gain can be very important when N is large.
- the gradient backpropagation algorithm: the calculation of $\nabla L(y_i, \psi_\theta(x_i))$ can be greatly facilitated by exploiting the composition formula 4.24, which greatly accelerates learning.

5. Unsupervised Learning

In this part, we consider a random variable X with values in $(\mathcal{X}, \mathcal{A})$, but without any variable Y to explain, contrary to the case of supervised learning. We are thus interested in the study of the law of probability of the variable X , as in chapters 2 and 3, but with two particular questions, representation learning and clustering (partitioning).

On the one hand, we're interested in representation learning (Bengio et al., 2013). It consists in carrying out a transformation of the variable X in a space more adapted to the resolution of a certain problem. This concept of representation is one of the bases of the success of deep learning methods: each layer of the neural network can be seen as a way to apprehend data in order to extract a certain type of information relevant to a prediction or partitioning task for example. One of the first uses of representation learning is that of dimension reduction. It has already been said that the dimension of $\mathcal{X}$ space can be very large (especially in image processing or genomics for example), but the relevant information can eventually be hidden in a variety of much smaller dimensions. In this framework, representation learning will be seen as dimension reduction.

On the other hand, it may be interesting to identify consistent subgroups within a large data set. One may want to identify subgroups of patients with similar molecular profiles in order to specify treatments by group, or on an online sales site, one can analyze customers' shopping baskets to define some typical profiles and adopt more targeted marketing strategies. This partitioning task will be called clustering.

5.1 Representation Learning

5.1.1 Principal Component Analysis

Principal Component Analysis (PCA, or *Analyse en Composantes Principales* en français) is still probably the most widely used method of data analysis. It was introduced by Pearson in 1901. It consists of orthogonally projecting multivariate data onto a new basis of variables that are decorrelated from each other, these new variables are called principal directions or principal axes, and the vector of coordinates of all observations on each of these principal axes will be called a

principal component. By ordering the principal components according to their variances, and keeping only a certain number of them, we obtain a representation of our data in a smaller space.

Let's consider a random variable X with values in $\mathbb{R}^p$ and $(x_1, \dots, x_N)$ a sample of observations. We look for a linear approximation of rank q , $q \leq p$, of these data, i.e. the projection of the data on an affine subspace of dimension q .

Let $\mathcal{A}_q$ be an affine space of dimension q in $\mathbb{R}^p$. It is characterized by a localization factor μ and a matrix $A_q \in \mathcal{M}_{p,q}$ of orthogonal vectors ($A_q^T A_q = I_q$, identity matrix of order q). The affine subspace can be written: $\mathcal{A}_q = \{y \in \mathbb{R}^p : y = \mu + A_q \lambda, \lambda \in \mathbb{R}^q\}$. Given a sample $(x_1, \dots, x_N)$ we look for:

- (i) the parameters μ and A_q characterizing the affine subspace;
- (ii) for a given point x_i , what $y_i \in \mathcal{A}_q$ approximates x_i the best, and what is λ_i such that $y_i = \mu + A_q \lambda_i$.

The quadratic error of approximation is given by:

$$\mathcal{E}(\mu, A_q, \{\lambda_i\}_{1 \leq i \leq N}) = \sum_{i=1}^N \|x_i - \mu - A_q \lambda_i\|^2. \quad (5.1)$$

We will see below how to minimize this error. For this purpose, we will need the following result.

Theorem 5.1 — Singular Value Decomposition. Let $X \in \mathcal{M}_{N,p}(\mathbb{R})$. Then, there exist: $U \in \mathcal{M}_N(\mathbb{R})$ orthogonal ($U^T U = I_p$), $D \in \mathcal{M}_{N,p}(\mathbb{R})$, diagonal matrix of rank r , $D = \text{diag}(d_1, \dots, d_r, 0, \dots, 0)$, avec $d_1 \geq d_2 \geq \dots \geq d_r > 0$, and $V \in \mathcal{M}_p$ orthogonal, such that:

$$X = U D V^T.$$

Columns of U are called left singular vectors and columns of V are called right singular vectors. The non-zero diagonal elements of D are the singular values of X . These singular values are unique.

We speak of *Singular Values Decomposition* (SVD).

Preuve : It results from the spectral decomposition. Indeed, we know that $X^T X$ is symmetrical and positive semi-definite of rank p , and therefore, there is $\tilde{V}$, orthogonal, and $\tilde{\Sigma} = \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix}$, with Σ diagonal, positive definite of rank r such that:

$$X^T X = \tilde{V} \tilde{\Sigma} \tilde{V}^T.$$

After some algebraic manipulations we can construct the appropriate U and V . Note that the non-zero singular values are then the roots of the non-zero eigenvalues of $X^T X$. $\square$

We then deduce the theorem on the approximation on an affine subspace by principal component analysis.

Theorem 5.2 Let $\tilde{X} \in \mathcal{M}_{N,p}$ be the matrix of row vectors $x_i - \bar{x}$, for $1 \leq i \leq N$. x_{ij} denotes the j -th component of x_i and $\bar{x}_j$ the j -th component of $\bar{x}$, for $1 \leq j \leq p$:

$$\tilde{X} = \begin{pmatrix} x_{11} - \bar{x}_1 & \dots & x_{1p} - \bar{x}_p \\ \vdots & & \vdots \\ x_{N1} - \bar{x}_1 & \dots & x_{Np} - \bar{x}_p \end{pmatrix}$$

Let's assume that $\tilde{X}$ is of rank r and consider a singular value decomposition of $\tilde{X}$:

$$\tilde{X} = U D V^T .$$

Let $q \leq r$ and V_q , the matrix corresponding to the first q right singular vectors (q first columns of V). Then, a solution to the minimization of quadratic approximation error (5.1) is given by:

$$\mu = \bar{x}, A_q = V_q, \lambda_i = V_q^T (x_i - \bar{x}), \forall 1 \leq i \leq N .$$

That is to say,

$$x_i \approx \bar{x} + V_q V_q^T (x_i - \bar{x}).$$

The column vectors of V_q are called **principal directions** or **principal axes**.

The vector $V_q^T (x_i - \bar{x})$ corresponds to the q coordinates of $x_i - \bar{x}$ with respect to the principal axes. We also have $V_q^T (x_i - \bar{x}) = (U_{i1}d_1, U_{i2}d_2, \dots, U_{iq}d_q)^T$. These coordinates are often called *loadings*.

Preuve :

$$\mathcal{E}(\mu, A_q, \lambda_{1 \leq i \leq N}) = \sum_{i=1}^N \|x_i - \mu - A_q \lambda_i\|^2$$

is convex in μ and λ_i . For A_q given, a necessary and sufficient condition for μ and $\{\lambda_i\}_{1 \leq i \leq N}$ to minimize $\mathcal{E}$ is therefore that:

$$\frac{\partial \mathcal{E}}{\partial \mu} = 2 \sum_{i=1}^N (x_i - \mu - A_q \lambda_i) = 0$$

and

$$\frac{\partial \mathcal{E}}{\partial \lambda_i} = -2A_q^T (x_i - \mu - A_q \lambda_i) = 0, \forall 1 \leq i \leq N.$$

We can see that this condition is realized for $\mu = \bar{x}$ and $\lambda_i = A_q^T (x_i - \bar{x})$. We are then faced

with the problem of minimization on the set of orthogonal matrices in $\mathcal{M}_{p,q}$ of $\sum_{i=1}^N \|x_i - \bar{x} -$

$A_q A_q^T (x_i - \bar{x})\|^2 = \|\tilde{X}^T - A_q A_q^T \tilde{X}^T\|_F^2$ en using Frobenius norm. So, using a singular value decomposition, then using the invariance of the Frobenius norm by orthogonal transformation:

$$\|\tilde{X}^T - A_q A_q^T \tilde{X}^T\|_F^2 = \|V D^T U^T - A_q A_q^T V D^T U^T\|_F^2 = \|D - V^T A_q A_q^T V D\|_F^2 .$$

The best approximation of rank q in the sense of the Frobenius norm for the matrix D^T is obtained for $D_q = \text{diag}(d_1, \dots, d_q, 0, \dots, 0)$. Then it is enough to check that by taking $A_q = V_q$,

we have:

$$V^T V_q V_q^T V D^T = D_q,$$

and the result is proved. $\square$

Definition 5.1 — Principal Component Analysis. Let $(x_1, \dots, x_N)$ be a sample of observations in $\mathbb{R}^p$ and $X \in \mathcal{M}_{N,p}$ the corresponding design matrix. Formally, we call Principal Component Analysis the research of the principal axes V , the sorted singular values $d_1 \geq d_2 \geq \dots \geq d_r$ and the matrix $UD = (X - \bar{X})V$, the columns of which being the principal components and the rows of which being the loadings. One can choose the number of principal components q to keep (corresponding to keep the first q columns of V and the first q columns of UD) before carrying out the analysis, or a posteriori, after the analysis of the results.

Figure 5.1 illustrates the construction of the data approximation from 3 principal axes.

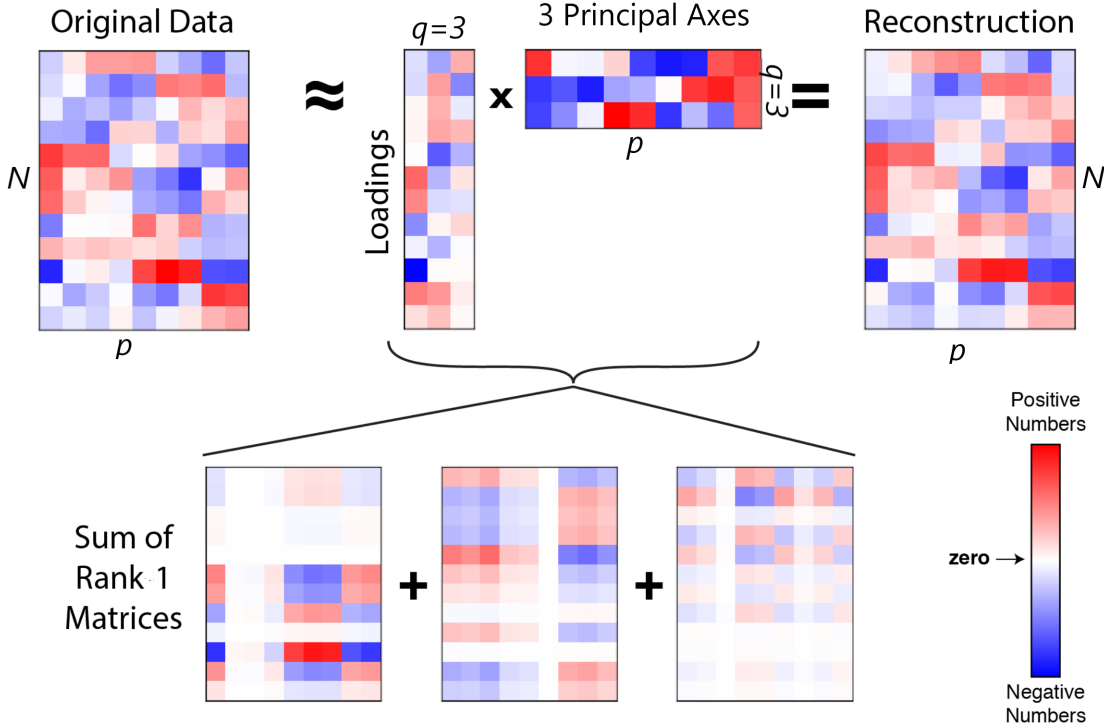


Figure 5.1: Example of data approximation from the 3 principal axes. The centered data matrix X on the left is approximated by $U_3 D_3$ called *Loadings* and V_3^T , the principal axes. This approximation is of rank at most 3. Image from the Alex Williams's blog: [itsneuronalblog](#).

■ **Remark 5.1** Based on the singular value decomposition, we have: $\tilde{X}V = UD$.

The name principal components is ambiguous. Traditionally in France, the vectors of the coordinates of each point $x_i - \bar{x}$ in the new basis of principal directions are called principal

components. According to this appellation, the vector $C = \begin{pmatrix} d_1 U_{11} \\ \vdots \\ d_1 U_{N1} \end{pmatrix}$ thus corresponds to the first component, and the columns of $X V_q = U_q D_q$ give the first q main components. This is for example the convention adopted by Saporta (2006) or Hastie et al. (2009). It is also the convention that we adopt. But in some documents, including a large number of online data science blogs, the principal components refer to the principal axes (the V column vectors) and in this case the UD column vectors are often called principal component scores. However, the ambiguity can be resolved, keeping in mind that V column vectors are of dimension p and $XV = UD$ column vectors are of dimension N . ■

■ **Remark 5.2** We can see that to simplify the manipulations, it is simpler to center the values before carrying out a PCA. In the same way, especially when the variables are of different orders of magnitude, it is also wise to standardize them, i.e. to perform the PCA on the $(\tilde{x}_1, \dots, \tilde{x}_N)$, where $\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$ and s_j is the empirical standard deviation of the j -th component. That's what we will do in the remainder of this section.. ■

■ **Remark 5.3** We obtained the principal axes by the singular value decomposition. But we also have: $X^T X = V D^T U^T U D V = V \Sigma V^T$, with $\Sigma = D^T D$, diagonal matrix of order N whose non-zero diagonal elements are d_i^2 . So we can also get V by diagonalizing $X^T X$ corresponding to the spectral decomposition. It turns out that numerically, this second solution is less stable, and we will thus privilege the singular value decomposition. ■

We have so far seen the principal component analysis as a process of approximating our sample in a space of smaller dimension. It can also be seen in terms of maximizing the variance of the sample in a space of lower dimension

We are now in the case where the variables have been centered. We denote by $X^1, \dots, X^p$ the random variables corresponding to each of the p components of our data. The empirical mean of each of these is zero on the sample. We consider the orthogonal projection of this random vector on a normalized vector w $w \in \mathbb{R}^p$, $\|w\| = 1$: $Z = \sum_{j=1}^p w_j X^j$. Then, we have the following result:

Proposition 5.3 Let $(x_1, \dots, x_N)$ be a centered sample of observations, $x_i \in \mathbb{R}^p$, et $\bar{x} = 0$. Let $X \in \mathcal{M}_{Np}$ be the corresponding design matrix:

$$X = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & & \vdots \\ x_{N1} & \dots & x_{Np} \end{pmatrix}$$

and its singular value decomposition:

$$X = U D V^T,$$

with $d_1 \geq d_2 \geq \dots \geq d_r > 0$, singular values of D .

The orthogonal projection of $(X^1, \dots, X^p)$ on the right singular vector v_1 is that of maximum empirical variance among the orthogonal projections on a subspace of dimension 1. It is equal to $\frac{d_1^2}{N}$.

Preuve : Let $w \in \mathbb{R}^p$ be a normed vector. The orthogonal projection of the random vector $(X^1, \dots, X^p)$ on the subspace of dimension 1 generated by w is given by the random vector $Z = \sum_{j=1}^p w_j X^j$.

The realizations of Z for the observation samples are then $z_i = w^T x_i = \sum_{j=1}^p w_j x_{ij}$. Z is also of zero empirical mean, and its empirical variance is given by:

$$\frac{1}{N} \sum_{i=1}^N z_i^2 = \frac{1}{N} \sum_{i=1}^N w^T x_i x_i^T w.$$

But we know that $X^T X = \sum_{i=1}^N x_i x_i^T$, therefore:

$$\frac{1}{N} \sum_{i=1}^N z_i^2 = \frac{1}{N} w^T X^T X w = \frac{1}{N} w^T V D^T D V^T w.$$

Let's denote $\alpha = V^T w$ and $\|\alpha\|^2 = w^T V V^T w = 1$, since w is normed and V orthogonal. Therefore, $\forall 1 \leq i \leq p$, $0 \leq \alpha_i \leq 1$. The empirical variance is therefore given by:

$$\frac{1}{N} \sum_{i=1}^N z_i^2 = \frac{1}{N} \sum_{i=1}^p d_i^2 \alpha_i^2.$$

It is maximum when $\alpha_1 = 1$ and $\alpha_i = 0, \forall i \geq 2$. The vector $w = V\alpha$ is thus given by v_1 , the first column of V , which ends the proof. $\square$

If we seek a second vector w_2 , orthogonal to v_1 such that the empirical variance of the orthogonal projection of $(X^1, \dots, X^p)$ on w_2 is maximum, we will find by similar reasoning, $w_2 = v_2$, second right singular vector, for an empirical variance of d_2^2 , and so on.

We will formalize this result. To do so, we introduce the concept of total variance of a sample.

Definition 5.2 Let $(X_1, \dots, X_N)$ be a sample, X_i taking values in $\mathbb{R}^p$. The total variance of the sample is the statistic:

$$VT(X_1, \dots, X_N) = \frac{1}{N} \sum_i^N \|X_i - \bar{X}\|^2$$

Let $(x_1, \dots, x_N)$ a centered sample of observations, $\bar{x} = 0$, and let X be the corresponding design matrix.

We thus have:

$$VT(x_1, \dots, x_N) = \frac{1}{N} \sum_i^N \|x_i\|^2 = \frac{1}{N} \text{tr}(X^T X) = \frac{1}{N} \sum_i^p d_i^2,$$

where $\text{tr}(M)$ is the trace of M .

We have indeed:

$$\text{tr}(X^T X) = \text{tr}(V D^T U^T U D V^T) = \text{tr}(V D^T D V^T) = \text{tr}(D^T D) = \sum_i^p d_i^2.$$

Proposition 5.4 Let $(x_1, \dots, x_N)$ a centered sample of observations, $\bar{x} = 0$, and let X be the corresponding design matrix, $X \in \mathcal{M}_{N,p}$ of rank r .

Let $q \leq r$ and $X_q \in \mathcal{M}_{N,q}$ the matrix obtained by orthogonal projection of X onto the q first right singular vectors of the sample: we have $X_q = X V_q$, where V_q corresponds to the q first column vectors of V .

Then X_q is the projected sample of maximum total variance among all projection of X on subspaces of dimension q .

The total variance of the projected sample X_q is $VT_q = \frac{1}{N} \sum_i^q d_i^2$.

We call **fraction of variance explained** the ratio

$$\frac{VT_q}{VT(x_1, \dots, x_N)} = \frac{\sum_i^q d_i^2}{\sum_i^r d_i^2}.$$

Preuve : We have $VT(x_1, \dots, x_N) = \frac{1}{N} \sum_i^p d_i^2$.

Moreover $VT_q = \frac{1}{N} \text{tr}(V_q^T X^T X V_q) = \frac{1}{N} \text{tr}(V_q^T V D^T D V^T V_q)$.

$$\text{Or: } V_q^T V = (I_q, 0_{q(p-q)})$$

where I_q is the identity matrix of order q and $0_{q(p-q)}$ the zero matrix in $\mathcal{M}_{q,(p-q)}$.

	1	2	3	4	5	6	7	8
1	-0.941	1.336	0.219	0.387	-0.786	0.866	1.499	0.728
2	0.780	-0.125	1.089	1.003	0.946	-1.031	0.998	-0.818
3	1.205	-0.926	0.284	-2.310	0.403	-1.808	-0.128	-1.405
4	0.038	-1.409	0.630	0.327	0.823	-0.214	0.437	0.651
5	-1.695	-0.211	0.362	-0.447	0.502	0.472	-0.951	1.174
6	1.257	1.117	-1.398	-0.417	-1.654	-0.756	-1.075	-1.105
7	0.450	0.479	-0.727	-0.258	-1.223	0.886	-1.437	0.123
8	-0.261	1.019	-0.726	-0.544	-0.358	-0.390	0.340	0.745
9	-1.413	0.354	-1.464	1.366	-0.359	1.760	-0.905	1.224
10	0.578	-1.634	1.733	0.893	1.706	0.213	1.221	-1.317

Table 5.1: Example of a dataset: 10 standardized observations in $\mathbb{R}^8$.

	principal component 1	principal component 2
1	-1.258	-1.096
2	2.061	-0.488
3	2.444	2.332
4	1.098	-1.179
5	-0.989	-1.125
6	-0.865	2.953
7	-1.623	1.139
8	-0.945	0.501
9	-2.909	-1.367
10	2.986	-1.671

Table 5.2: First two principal components (columnss); each row gives the two coordinates in the basis of the first two principal axes for the approximation of each of the 10 points.

So finally:

$$V_q^T V D^T D V^T V_q = \begin{pmatrix} d_1^2 & 0 & \dots & 0 \\ 0 & d_2^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & d_q^2 \end{pmatrix},$$

which shows the result. □

■ **Example 5.4** We consider a sample of 10 observations in $\mathbb{R}^8$, given in 5.1. The data were standardized (zero mean and variance equal to 1 for each column). We perform a PCA. Figure 5.2 shows the fraction of variance explained as a function of the number of principal components retained. We see that only 2 components are needed to explain 75% of the variance. The main components are given by:

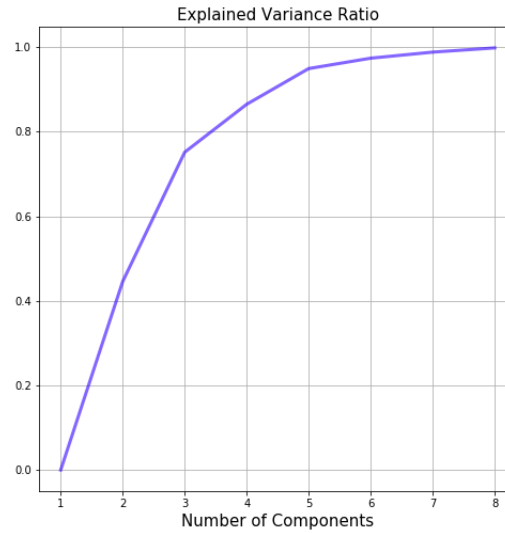


Figure 5.2: Fraction of variance explained as a function of the number of principal components retained in the reconstruction: for 8 components, we of course get a fraction of 1.

An interesting use of PCA is that the projection on the 2 principal axes allows a $2D$ representation of multivariate data. This is illustrated on our data in Figure 5.3. It is difficult to represent data in dimension 8 (let's not talk about dimension $1E5$ or $1E6$ as for genomics applications...). On the other hand, by using the first two main components, we can have a useful representation. Of course, it depends on the percentage of variance explained by these first two components! This data visualization is particularly used in the case of clustering, which we will see in the following part 5.2.

■

5.2 Clustering

Let a scatter plot $(x_1, \dots, x_N)$ corresponding to independent realizations of a random variable in $\mathcal{X}$. The objective of clustering is to group the most similar points in groups called clusters. Of course, this notion of similarity will strongly depend on the metric used to measure the distance between points. A good clustering will present clusters that will be homogeneous and as separated as possible. The basic idea will be to define a function on the set of point partitions in the data sample that measures the consistency of the partition and to find the partition that maximizes this consistency.

Definition 5.3 — Cluster, centroid, clustering. Let $E = \{x_1, \dots, x_N\}$, with $x_i \in \mathbb{R}^p$, $\forall 1 \leq i \leq N$. A cluster $E^{(k)}$ is a subset of points of E . We will identify a cluster with the set of

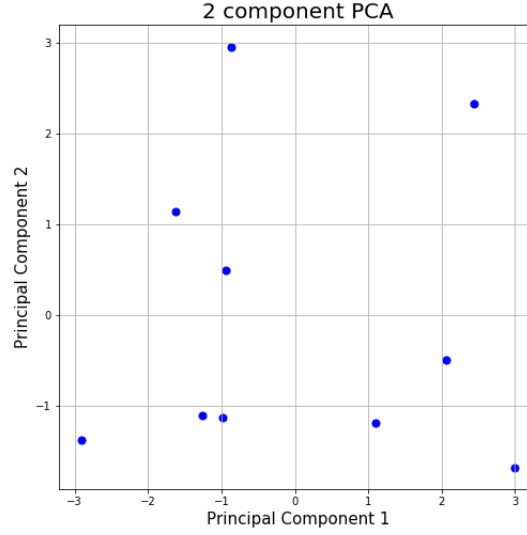


Figure 5.3: Scatterplots projected on the two principal axes.

indices of the points it contains: $\pi^{(k)}$:

$$\pi^{(k)} = \{j : x_j \in E^{(k)}\} .$$

The centroid of $E^{(k)}$ is the barycentre of the cluster:

$$\bar{x}^{(k)} = \frac{1}{|\pi^{(k)}|} \sum_{j \in \pi^{(k)}} x_j ,$$

where $|\pi^{(k)}|$ is the cardinal of $\pi^{(k)}$.

A clustering of size K is a partition of E in K non-empty clusters: $\Pi = \{\pi^{(1)}, \dots, \pi^{(K)}\}$

The K centroids are also a way to summarize a dataset of N points with only K points, K much smaller than N .

For a given clustering size K , the objective of clustering will be to maximize the overall consistency of the Π partition across all partitions.

The coherence of the partition can be qualified by measures of inertia, intra-cluster inertia (which characterizes the homogeneity of clusters) and inter-cluster inertia (which depends on their separation).

Definition 5.4 — Inertia. We call inertia of a sample $(X_1, \dots, X_N)$ the statistic

$$T(X_1, \dots, X_N) = \sum_{i=1}^N \|X_i - \bar{X}\|^2$$

Let $(x_1, \dots, x_N)$ be an observation sample. We consider a clustering for this set of points: $\Pi = \{\pi^{(1)}, \dots, \pi^{(K)}\}$. For the sake of simplicity, we denote $T := T(x_1, \dots, x_N)$ and we have:

$$\begin{aligned} T &= \sum_{k=1}^K \sum_{i \in \pi^{(k)}} \|x_i - \bar{x}\|^2 \\ &= \sum_{k=1}^K \sum_{i \in \pi^{(k)}} \left(\|x_i - \bar{x}^{(k)}\|^2 + \|\bar{x}^{(k)} - \bar{x}\|^2 \right) \\ &= \sum_{k=1}^K \sum_{i \in \pi^{(k)}} \|x_i - \bar{x}^{(k)}\|^2 + \sum_{k=1}^K |\pi^{(k)}| \|\bar{x}^{(k)} - \bar{x}\|^2 \end{aligned}$$

using König-Huygens decomposition.

We call *within-cluster* inertia of clustering Π , denoted by $W(\Pi)$, the term:

$$W(\Pi) = \sum_{k=1}^K \sum_{i \in \pi^{(k)}} \|x_i - \bar{x}^{(k)}\|^2$$

and we call *between-cluster* inertia of clustering Π , denoted by $B(\Pi)$, the term:

$$B(\Pi) = \sum_{k=1}^K |\pi^{(k)}| \|\bar{x}^{(k)} - \bar{x}\|^2.$$

We then have: $T = W(\Pi) + B(\Pi)$.

Let us recall that $T := T(x_1, \dots, x_N)$ is determined by the sample only and does not depend on the clustering.

■ **Remark 5.5** We see that if we add a $1/N$ factor to the inertia, we obtain the total variance of the sample given in Definition 5.2:

$$VT(X_1, \dots, X_N) = \frac{1}{N} \sum_{i=1}^N \|X_i - \bar{X}\|^2.$$

The term inertia is sometimes also used to refer to the total variance of the sample. In what follows, one might just as well reason about the variance, since the constant $1/N$ factor does not impact the reasoning and gives the same results. ■

For a given sample $(x_1, \dots, x_N)$, $T(x_1, \dots, x_N)$ is constant, and we have: $B(\Pi) = T(x_1, \dots, x_N) - W(\Pi)$.

To maximize clustering consistency, one possibility¹ is therefore to maximize between-cluster inertia or in an equivalent way to minimize within-cluster inertia, i.e. to solve the optimization problem:

$$\Pi^* = \arg \min_{\Pi \in \mathcal{P}_K(E)} W(\Pi)$$

where $\mathcal{P}_K(E)$ is the set of all partitions of E of size K .

¹There are other possibilities, based on other ways of measuring clustering consistency.

An enumeration calculation shows that the number of elements in this set (modulo a permutation of the indices, so as not to count several times the same partition obtained for different cluster numbers) is $\frac{1}{K!} \sum_{k=1}^K (-1)^{K-k} C_K^k k^N$. For a relatively modest problem size, for example 100 points to be divided into 4 clusters, there are already almost $5E17$ possibilities! It is therefore not possible to conduct an exhaustive search among all the possibilities. Empirical search algorithms have therefore been developed to solve this problem. We will only present here the most popular one, the K -means algorithm, but there are many other approaches such as hierarchical clustering or spectral clustering, for example (see Hastie et al. (2009)).

5.2.1 Algorithme des K moyennes ou K -means

The K -means algorithm has been developed for situations where all variables are quantitative. So here we consider $\mathcal{X} \subset \mathbb{R}^p$ and the distance is the Euclidean distance.

We describe the algorithm below.

Algorithm 5.1 — K -means.

The clustering size K is given.

Initialization: let $\Pi_0 = \{\pi_0^{(1)}, \dots, \pi_0^{(K)}\}$ be a random initial clustering, and $\bar{x}_0^{(1)}, \dots, \bar{x}_0^{(K)}$ the corresponding centroids.

Do

$t \leftarrow t+1$

 For $i = 1 : N$

 find the centroid $\bar{x}_{t-1}^{(p)}$ the closest to x_i and classify x_i in cluster p .

 end

$\Pi_t =$ the new clustering.

 Compute $\bar{x}_t^{(1)}, \dots, \bar{x}_t^{(K)}$, the new centroids.

While $|W(\Pi_t) - W(\Pi_{t-1})| > 0$

Proposition 5.5 Let $(\Pi_t)_{t \geq t_0}$ the sequence of clustering built during the K -means algorithm. Then, $\exists T \geq 0$ such that $\forall t < T$: $W(\Pi_t) < W(\Pi_{t-1})$ and $W(\Pi_T) = W(\Pi_{T-1})$. In other words, during iterations of the K -means algorithm, $W(\Pi_t)$ is strictly decreasing until it stabilizes.

Preuve : We have:

$$W(\Pi_{t-1}) = \sum_{k=1}^K \sum_{i \in \pi_{t-1}^{(k)}} \|x_i - \bar{x}_{t-1}^{(k)}\|^2$$

We can see that the two steps of the K -means algorithm decrease W .

If point j goes from cluster q to cluster p , $\|x_j - \bar{x}_{t-1}^{(p)}\|^2 < \|x_j - \bar{x}_{t-1}^{(q)}\|^2$. This is the case for all points that change of clusters. After putting the points in their new clusters (which translates

by the sum for $i \in \pi_t^{(k)}$ and no longer for $i \in \pi_{t-1}^{(k)}$ in the equation below), we have:

$$W(\Pi_{t-1}) > \sum_{k=1}^K \sum_{i \in \pi_t^{(k)}} \|x_i - \bar{x}_{t-1}^{(k)}\|^2.$$

The inequality is strict as soon as at least one point changes cluster.

Now we can easily see (using again the König-Huygens decomposition) that for any cluster k and for all y :

$$\sum_{i \in \pi_t^{(k)}} \|x_i - y\|^2 = \sum_{i \in \pi_t^{(k)}} \left(\|x_i - \bar{x}_t^{(k)}\|^2 + \|\bar{x}_t^{(k)} - y\|^2 \right) \geq \sum_{i \in \pi_t^{(k)}} \|x_i - \bar{x}_t^{(k)}\|^2$$

Taking $y = \bar{x}_{t-1}^{(k)}$, we have:

$$W(\Pi_{t-1}) > \sum_{k=1}^K \sum_{i \in \pi_t^{(k)}} \|x_i - \bar{x}_{t-1}^{(k)}\|^2 \geq \sum_{k=1}^K \sum_{i \in \pi_t^{(k)}} \|x_i - \bar{x}_t^{(k)}\|^2 = W(\Pi_t).$$

The within-cluster inertia is thus strictly decreasing as long as points change clusters. Since there is a finite number of possible partitions, $W(\Pi_t)$ can only take a finite number of values, and therefore there exists T such that $W(\Pi_T) = W(\Pi_{T-1})$, which corresponds to the step for which no point is found that is closer to the centroid of another cluster than to its own. $\square$

One of the difficulties is that the algorithm is not guaranteed to converge to the global minimum. It is only guaranteed to converge to local minima. However, in general, in the partition space, there are a very large number of local minima. A common precaution to take is therefore to launch the algorithm several times with different random initializations and to check the stability of the solutions found, or of the optimized criterion. The best configuration among the different tests will be chosen.

5.2.2 Choice of the Optimal Number of Clusters

We still consider in this section a given dataset $(x_1, \dots, x_N)$ of points in $\mathbb{R}^p$ and the objective of clustering these points.

Choosing the optimal number of clusters is tricky. Philosophically, it can be seen as a model selection problem (cf. 4.4): a high number of clusters refines the description, but corresponds to a larger number of parameters. The characterization of what a good clustering is is not always obvious.

Different criteria have been constructed to achieve this selection, with different formulations and objectives. We present two of the most classic ones, the Davies-Boulding index and the silhouette value.

The Davies-Boulding Index defines measures of cluster homogeneity and separability and proposes a criterion that ensures a compromise between the two.

Definition 5.5 — Homogénéité, Séparabilité. Let $\Pi = (\pi^{(1)}, \dots, \pi^{(K)})$ be a clustering of sample $(x_1, \dots, x_N)$.

The homogeneity of a cluster $\pi^{(k)}$ is defined by:

$$H^{(k)} = \frac{1}{|\pi^{(k)}|} \sum_{i \in \pi^{(k)}} \|x_i - \bar{x}^{(k)}\|.$$

The homogeneity of a clustering is thus:

$$H(\Pi) = \frac{1}{K} \sum_{k=1}^K H^{(k)}.$$

The separability between clusters $\pi^{(k)}$ and $\pi^{(l)}$ is simply given by the distance between their two centroids $\|\bar{x}^{(k)} - \bar{x}^{(l)}\|$, and the separability of a clustering is the average separability of all pairs of clusters. It is given by :

$$S(\Pi) = \frac{2}{K(K-1)} \sum_{k=1}^K \sum_{l=k+1}^K \|\bar{x}^{(k)} - \bar{x}^{(l)}\|.$$

Definition 5.6 — Davies-Bouldin Index. Let $\Pi = (\pi^{(1)}, \dots, \pi^{(K)})$ be a clustering of sample $(x_1, \dots, x_N)$.

The Davies-Bouldin index of cluster $\pi^{(k)}$ is given by:

$$DB^{(k)} = \max_{l: l \neq k} \frac{H^{(k)} + H^{(l)}}{\|\bar{x}^{(k)} - \bar{x}^{(l)}\|}.$$

And the Davies-Bouldin index of clustering Π is the average index of all clusters:

$$DB(\Pi) = \frac{1}{K} \sum_{k=1}^K DB^{(k)}.$$

The more homogeneous and well separated the clusters are, the smaller $DB(\Pi)$ will be. So the smaller the Davies-Bouldin Index, the better the clustering will be considered.

Le coefficient de silhouette adopte un point de vue un peu différent qui définit une sorte de score d'appartenance au bon cluster pour chaque point.

Definition 5.7 — Silhouette. Let $\Pi = (\pi^{(1)}, \dots, \pi^{(K)})$ be a clustering. Let $i \in \pi^{(k)}$. For x_i , we consider the average of the distances to other points in the same cluster

$$a(x_i) = \frac{1}{|\pi^{(k)}|} \sum_{j \in \pi^{(k)}, j \neq i} \|x_j - x_i\|$$

and the minimum value that this value could take if x_i belonged to another cluster:

$$b(x_i) = \min_{l: l \neq k} \frac{1}{|\pi^{(l)}|} \sum_{j \in \pi^{(l)}} \|x_j - x_i\|$$

The silhouette value of the point x_i is then given by: $s(x_i) = \frac{b(x_i) - a(x_i)}{\max(a(x_i), b(x_i))}$, and the silhouette value of the clustering by:

$$s(\Pi) = \frac{1}{N} \sum_{i=1}^N s(x_i) .$$

$s(\Pi) \leq 1$ and the closer to 1 $s(\Pi)$, the better the clustering.

■ **Remark 5.6** Be careful, depending on the index considered, there is not necessarily the same selection rule: the Davies-Bouldin index must be as small as possible, while the silhouette coefficient must be as large as possible.

Two other classical indices that we do not detail here are the Calinski-Harabasz and Dunn indices. These two indices increase with the quality of clustering.

■ **Example 5.7** We consider a set of points in $\mathbb{R}^2$ shown in Figure 5.4.

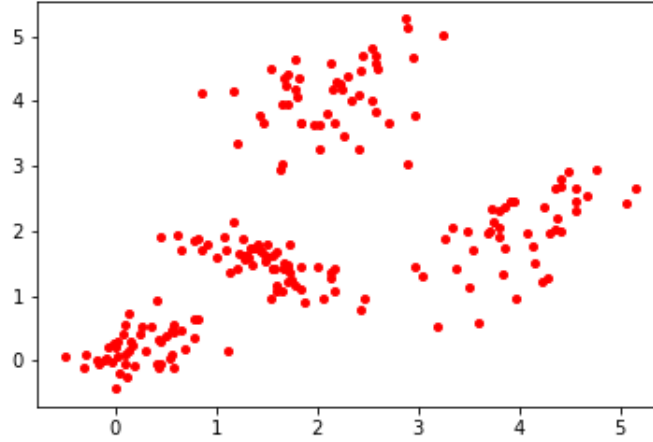


Figure 5.4: Dataset of 200 points in $\mathbb{R}^2$

In figure 5.5, we present the results of the K -means algorithm, for different values of K . We are in a simple case where the clusters are well separated, and at a glance, the 4-cluster configuration seems the best. We then plot the Davies-Bouldin index evolution curves and silhouette values for $2 \leq K \leq 10$. The results are given in Figure 5.6. We can clearly see that the highest silhouette values and the lowest Davies-Bouldin index are obtained for $K = 4$, which confirms in this simple case the visual impression.

■ **Remark 5.8 — Rand Index.** Finally, a last point that may be interesting is the comparison of two clusterings $\Pi = (\pi^{(1)}, \dots, \pi^{(K)})$ and $\Gamma = (\gamma^{(1)}, \dots, \gamma^{(L)})$. We measure the coherence between the two clusterings by comparing if the points that are in the same cluster for Π are

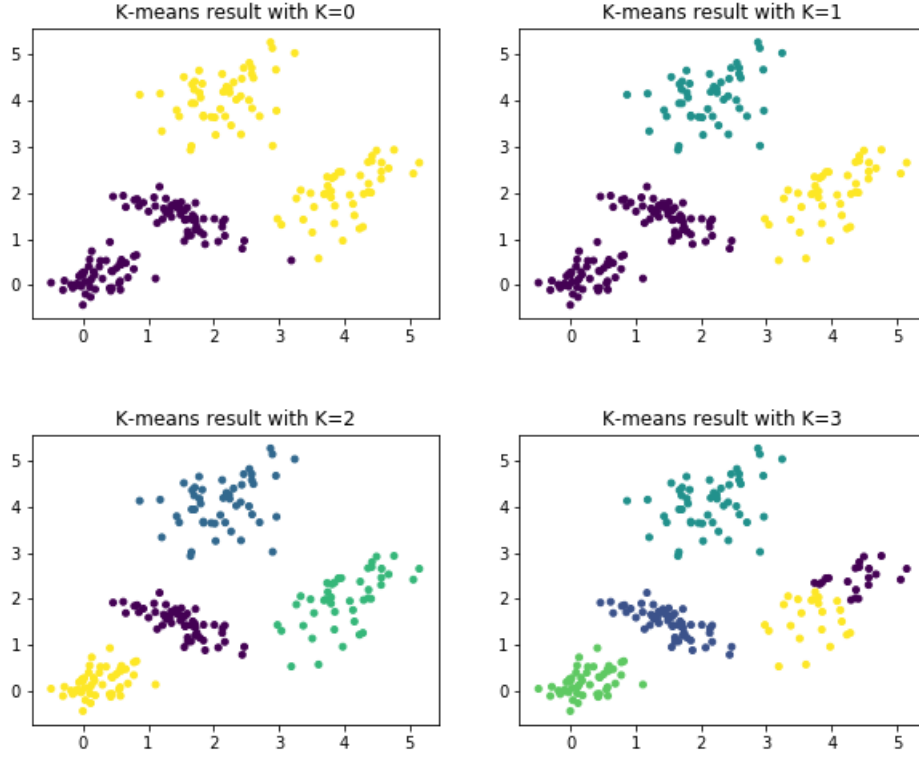


Figure 5.5: Clustering of the sample in 2,3,4,5 clusters by K -means

also in the same cluster for Γ and reciprocally, if when they are not in the same cluster for Π , they are not in the same cluster for Γ either. Let C_Π denote the function which to a point in the sample associates its cluster number in clustering Π : $C_\Pi : x_i \mapsto k$, if $i \in \pi^{(k)}$, and similarly C_Γ for Γ . The alignment between two points x_i, x_j is given by:

$$\alpha_{\Pi, \Gamma}(x_i, x_j) = \delta(C_\Pi(i) = C_\Pi(j)) \delta(C_\Gamma(i) = C_\Gamma(j)) + \delta(C_\Pi(i) \neq C_\Pi(j)) \delta(C_\Gamma(i) \neq C_\Gamma(j))$$

and the Rand index between the two clusterings $R(\Pi, \Gamma)$ is:

$$R(\Pi, \Gamma) = \frac{2}{N(N-1)} \sum_{i=1}^N \sum_{j=i+1}^N \alpha_{\Pi, \Gamma}(x_i, x_j).$$

It is 1, when the two clusterings are in perfect agreement. ■

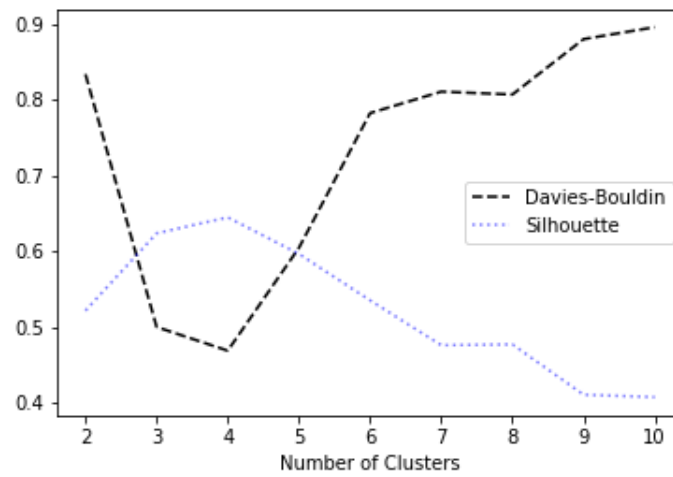


Figure 5.6: Evolution of the Davies-Bouldin index and the silhouette value for $2 \leq K \leq 10$

A. Annexes

A.1 Lois usuelles

A.1.1 Lois discrètes

–Loi de Bernoulli $\mathcal{B}(p)$

Support : $x \in \{0, 1\}$

Paramètre : $p \in [0, 1]$

Fonction de masse: $f(x) = p^x(1 - p)^{1-x}$

Espérance: p , Variance: $p(1 - p)$

–Loi binomiale $\mathcal{B}(n, p)$

Support : $x \in \{0, 1, \dots, n\}$

Paramètres : (n, p) , $n \in \mathbb{N}^*$, $p \in [0, 1]$

Fonction de masse: $f(x) = C_n^x p^x (1 - p)^{n-x}$

Espérance: np , Variance: $np(1 - p)$

–Loi de Poisson $\mathcal{P}(\lambda)$

Support : $x \in \mathbb{N}$

Paramètre : $\lambda \in \mathbb{R}_+^*$

Fonction de masse: $f(x) = e^{-\lambda} \lambda^x / x!$

Espérance: λ , Variance: λ

A.1.2 Lois continues

–Loi uniforme $\mathcal{U}([a, b])$

Support : $x \in [a, b]$

Paramètres : $(a, b) \in \mathbb{R}^2$, $a < b$

Densité $f(x) = (b - a)^{-1} \mathbb{I}_{[a, b]}(x)$

Espérance: $(a + b)/2$, Variance: $(b - a)^2/12$.

–Loi Gamma $\gamma(\alpha, \beta)$ Support : $x \in \mathbb{R}^+$ Paramètres : $(\alpha, \beta) \in (\mathbb{R}_+^*)^2$ Densité $f(x) = \frac{\beta^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta x} \mathbb{I}_{[0;+\infty[}(x)$ avec $\Gamma(x) = \int_0^{+\infty} u^{x-1} e^{-u} du$.Espérance: α/β , Variance: α/β^2 Propriété: $\gamma(\alpha_1 + \alpha_2, \beta) = \gamma(\alpha_1, \beta) + \gamma(\alpha_2, \beta)$ si ces deux dernières sont indépendantes.**–Loi exponentielle** $\mathcal{E}(\lambda)$ Support : $x \in \mathbb{R}^+$ Paramètre : $\lambda \in \mathbb{R}_+^*$ Densité: $f(x) = \lambda e^{-\lambda x}$ Espérance: $1/\lambda$, Variance: $1/\lambda^2$ Propriété: $\mathcal{E}(\lambda) = \gamma(1, \lambda)$ **–Loi gaussienne ou normale** $\mathcal{N}(m, \sigma^2)$ Support : $x \in \mathbb{R}$ Paramètres : $(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$ Densité: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$ Espérance: μ , Variance: σ^2 Propriété: Si $X \sim \mathcal{N}(\mu, \sigma^2)$, alors $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$.La loi $\mathcal{N}(0, 1)$ est dite gaussienne standard ou centrée réduite.**–Loi du chi-deux** $\chi^2(p)$ La variable aléatoire $Y = X_1^2 + \dots + X_p^2$, avec X_i i.i.d gaussiennes standards, est une variable dite du chi-deux de Pearson à p degrés de liberté, on note cette loi $\chi^2(p)$ Support : $x \in \mathbb{R}^+$ Paramètre : $p \in \mathbb{N}^*$ Densité $f(y) = C(p) y^{p/2-1} e^{-y/2} \mathbb{I}_{[0;+\infty[}(y)$ où $C(p) = (2^{p/2} \Gamma(p/2))^{-1}$, Espérance: p , Variance: $2p$ Propriété : la loi $\chi^2(p)$ est une loi $\gamma(p/2, 1/2)$.**–Loi de Student** $T(p)$ La variable aléatoire $Y = \frac{U}{\sqrt{V/p}}$ avec $U \sim \mathcal{N}(0, 1)$ et $V \sim \chi^2(p)$ deux variables aléatoires indépendantes est une variable de Student à p degrés de liberté, notée $T(p)$.Densité : $f(x) = C(p) \left(1 + \frac{x^2}{p}\right)^{-(p+1)/2}$ où $C(p) = (\sqrt{p} B(1/2, p/2))^{-1}$ et $B(p, q) = \Gamma(p)\Gamma(q)/\Gamma(p+q)$.**–Loi de Fisher-Snedecor** $F(p, q)$ La variable aléatoire $Y = \frac{U/p}{V/q}$ avec $U \sim \chi^2(p)$ et $V \sim \chi^2(q)$ deux v.a. indépendantes est une variable de Fisher-Snedecor à degrés de liberté p et q , notée $F(p, q)$.Support : $x \in \mathbb{R}^+$

Paramètre : $p, q \in \mathbb{N}^* \times \mathbb{N}^*$

Densité : $f(x) = C(p, q) \frac{x^{p/2-1}}{(q+px)^{(p+q)/2}} \mathbb{I}_{[0;+\infty[}(x)$ où $C(p, q) = \frac{p^{p/2} q^{q/2}}{B(p/2, q/2)}$.

—Loi bêta $\beta(a, b)$

Support : $x \in [0; 1]$

Paramètres : $(a, b) \in]0; +\infty[\times]0; +\infty[$

Densité: $f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1}$

Espérance: $\frac{a}{a+b}$, Variance: $\frac{ab}{(a+b)^2(a+b+1)}$.

A.2 Propriétés des échantillons gaussiens

Nous commençons par donner le théorème de Cochran.

Theorem A.1 — Théorème de Cochran. Soit X un vecteur aléatoire de loi gaussienne multivariée sur $\mathbb{R}^n$ donnée par $\mathcal{N}(m, \sigma^2 I_n)$ où $m \in \mathbb{R}^n, \sigma^2 > 0$ et I_n est la matrice identité sur $\mathbb{R}^n$. On note $E_1, \dots, E_p$ des sous-espaces orthogonaux en somme directe, tels que $\dim E_k = d_k$ et $\sum_{k=1}^p d_k = n$. On note π_k la matrice de projection sur E_k et $Y_k = \pi_k X$ représente la projection de X sur E_k . On a alors:

- (1) les vecteurs aléatoires $Y_1, \dots, Y_p$ sur $\mathbb{R}^n$ sont des vecteurs indépendants;
- (2) $\forall k \in \{1, \dots, p\}, Y_k \sim \mathcal{N}(\pi_k m, \sigma^2 \pi_k)$
- (3) les variables aléatoires $\|Y_1 - \pi_1 m\|^2, \dots, \|Y_p - \pi_p m\|^2$ sont indépendantes;
- (4) $\forall k \in \{1, \dots, p\}, \frac{1}{\sigma^2} \|Y_k - \pi_k m\|^2 \sim \chi^2(d_k)$.

Preuve : On peut supposer dans un premier temps que $m = 0$. On introduit le vecteur $Y = (Y_1, \dots, Y_p)^T \in \mathbb{R}^{np}$. On peut noter que $Y = AX$ où $A = (\pi_1, \dots, \pi_p)^T$ est une matrice $np \times n$ ce qui montre que Y est un vecteur gaussien sur $\mathbb{R}^{np}$ de loi $\mathcal{N}(0, \sigma^2 AA^T)$. En utilisant les hypothèses (les π_k sont des matrices de projection sur des espaces orthogonaux en somme directe), on obtient que la matrice $AA^T = \text{diag}(\pi_1, \dots, \pi_p)$ est une matrice constituée de matrices blocs positionnés sur la diagonale. Par suite, les vecteurs Y_k sont indépendants, de loi $\mathcal{N}(0, \sigma^2 \pi_k)$. L'indépendance des Y_k entraîne celle des variables $\|Y_k\|^2$. Il reste à calculer la loi de ces dernières. On note que la loi de X est invariante par changement de base orthonormée, donc quelle que soit la base de décomposition les coordonnées du vecteur suivent des lois $\mathcal{N}(0, \sigma^2)$. Soit $(u_1, \dots, u_{d_k})$ une base orthonormée de $E_k = \text{Im}(\pi_k)$ et $(u_{d_k+1}, \dots, u_n)$ une base orthonormée de $E_k^\perp = \text{Ker}(\pi_k)$. Nous avons alors $\|Y_k\|^2 = \sum_{i=1}^{d_k} |(X, u_i)|^2$.

$U = (u_1, \dots, u_n)$ est la matrice orthogonale du changement de base associée et : $U^T X \sim \mathcal{N}(0, U^T \sigma^2 I_n U) = \mathcal{N}(0, \sigma^2 I_n)$, c'est à dire que $U^T X$ est un vecteur de n composantes gaussiennes (X, u_i) indépendantes de loi $\mathcal{N}(0, \sigma^2)$, et donc Y_k est la somme de d_k composantes gaussiennes indépendantes de loi $\mathcal{N}(0, \sigma^2)$. En particulier, $\frac{\|Y_k\|^2}{\sigma^2} \sim \chi^2(d_k)$, par définition de la loi du χ^2 . Le cas m quelconque se traite de la même façon en raisonnant sur les variables (centrées) $X - m$ et $\tilde{Y}_k = \pi_k(X - m)$. $\square$

Le théorème suivant est fondamental pour l'estimation dans le modèle gaussien. Il fournit la loi de la moyenne empirique $\bar{X}$ et de la variance empirique S^2 .

Theorem A.2 Soit $X_1, \dots, X_N$ des variables aléatoires i.i.d. de loi $\mathcal{N}(m, \sigma^2)$. On pose:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i, \quad S^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2$$

On a alors:

- (1) $\bar{X}$ et S^2 sont indépendantes.
- (2) $\bar{X} \sim \mathcal{N}(m, \sigma^2/N)$.
- (3) $\frac{NS^2}{\sigma^2} \sim \chi^2(N-1)$.

Preuve : On introduit le sous-espace vectoriel $E_1 = \{\lambda \mathbf{1}_N, \lambda \in \mathbb{R}\}$ où $\mathbf{1}_N = (1, \dots, 1)^T \in \mathbb{R}^N$, et E_2 est le sous-espace de dimension $N-1$ orthogonal à E_1 . Les matrices de projection sur E_1 et E_2 sont les suivantes :

$$\pi_1 = \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^T$$

$$\pi_2 = I_N - \pi_1,$$

où I_N est l'identité dans $\mathbb{R}^N$. On en déduit:

$$Y_1 = \pi_1 X = \bar{X} \mathbf{1}_N.$$

$$Y_2 = \pi_2 X = (X_1 - \bar{X}, \dots, X_N - \bar{X})^T$$

On applique alors le théorème de Cochran pour conclure: Y_1 et Y_2 sont indépendantes, on sait alors que pour toutes fonctions mesurables f et g , $f(Y_1)$ et $g(Y_2)$ sont indépendantes et donc $\bar{X}$ et S^2 sont indépendantes. D'autre part:

$$Y_1 \sim \mathcal{N}(\pi_1 m, \sigma^2 \pi_1) = \mathcal{N}(m, \frac{\sigma^2}{N} \mathbf{1}_N \mathbf{1}_N^T)$$

soit le résultat voulu en prenant une des marginales.

Enfin, en utilisant le point (4) du théorème de Cochran:

$$\frac{NS^2}{\sigma^2} = \frac{\|Y_2\|^2}{\sigma^2} \sim \chi^2(N-1).$$

□

Corollary A.3 Soit $X_1, \dots, X_N$ des variables aléatoires i.i.d. de loi $\mathcal{N}(m, \sigma^2)$. Alors,

$$\frac{\sqrt{N-1}(\bar{X} - m)}{S} \sim T(N-1).$$

Preuve : $\frac{\sqrt{N}(\bar{X} - m)}{\sigma} \sim \mathcal{N}(0, 1)$, $\frac{NS^2}{\sigma^2} \sim \chi^2(N-1)$ et les variables aléatoires sont indépendantes d'après ce qui précède, donc

$$\frac{\sqrt{N}(\bar{X} - m)}{\sigma} / \frac{\sqrt{N}S}{\sqrt{N-1}\sigma} \sim T(N-1)$$

d'après la définition d'une loi de Student (cf. Annexe A.1), ce qui donne le résultat souhaité en simplifiant.

□

Bibliography

- T. W. Anderson. **An introduction to multivariate statistical analysis**. Wiley, New York, 1962.
- Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. **IEEE transactions on pattern analysis and machine intelligence**, 35(8):1798–1828, 2013.
- P. Billingsley. **Probability and measure**. John Wiley & Sons, 1995.
- L. Breiman. Random forests. **Machine learning**, 45(1):5–32, 2001.
- K. Burnham and D. Anderson. **Model selection and multi-model inference**. Springer-Verlag, 2002.
- I. Goodfellow, Y. Bengio, and A. Courville. **Deep learning**. MIT press, 2016.
- T. Hastie, R. Tibshirani, and J. Friedman. **The elements of statistical learning: data mining, inference, and prediction**. Springer New York, 2009.
- M. Lavielle. **Mixed effects models for the population approach: models, tasks, methods and tools**. CRC press, 2014.
- Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In **Advances in neural information processing systems (NIPS1990)**, pages 396–404, 1990.
- E. L. Lehmann and G. Casella. **Theory of point estimation**. Springer Science & Business Media, 2006.
- E. L. Lehmann and J. P. Romano. **Testing statistical hypotheses**. Springer Science & Business Media, 2006.

- J. A. Nelder and R. W. Wedderburn. Generalized linear models. **Journal of the Royal Statistical Society: Series A**, 135(3):370–384, 1972.
- J. Nocedal and S. Wright. **Numerical optimization**. Springer Science & Business Media, 2006.
- C. Robert. **The Bayesian choice: from decision-theoretic foundations to computational implementation**. Springer Science & Business Media, 2007.
- C. Robert and G. Casella. **Monte-Carlo statistical methods**. Springer New York, 1999.
- G. Saporta. **Probabilités, analyse des données et statistique**. Editions Technip, 2006.
- R. E. Schapire, Y. Freund, P. Bartlett, W. S. Lee, et al. Boosting the margin: A new explanation for the effectiveness of voting methods. **The annals of statistics**, 26(5):1651–1686, 1998.
- B. W. Silverman. **Density estimation for statistics and data analysis**, volume 26. CRC press, 1986.
- V. Spokoiny and T. Dickhaus. **Basics of modern mathematical statistics**. Springer, 2015.
- R. J. Tibshirani et al. The lasso problem and uniqueness. **Electronic Journal of Statistics**, 7: 1456–1490, 2013.
- B. Zhu, N. Song, R. Shen, A. Arora, M. J. Machiela, L. Song, M. T. Landi, D. Ghosh, N. Chatterjee, V. Baladandayuthapani, et al. Integrating clinical and multiple omics data for prognostic assessment across human cancers. **Scientific reports**, 7(1):16954, 2017.