

Statistics and Learning (slides)

Julien BECT, Laurent LE BRUSQUET & Arthur TENENHAUS

2023



Table of contents

1	Introduction and point estimation methods	5
2	Point estimation.....	35
3	Asymptotic distributions Confidence intervals.....	71
4	Bayesian estimation	105
5	Hypothesis testing	131
6	Introduction to supervised learning Linear models for regression	165
7	Classification: logistic regression Generalization error	195
8	Regularization and model selection	221
9	Some models for supervised learning	249
10	Unsupervised learning: two examples	275

Chapter 1

Introduction and point estimation methods



CentraleSupélec

Statistics and Learning

Arthur Tenenhaus[†], Julien Bect & Laurent Le Brusquet

(firstname.lastname@centralesupelec.fr)

Teaching: CentraleSupélec / Department of Mathematics

Research: Laboratory of signals and systems (L2S)

[†]: Course coordinator

1/42

Lecture 1/10

Introduction and point estimation methods

In this lecture you will learn how to...

- ▶ Introduce statistical inference and illustrate its usefulness
- ▶ Define the mathematical framework
- ▶ Present some commonly used estimation methods

2/42

Lecture outline

1 – Introduction

2 – The mathematical framework of statistical inference

3 – Some (classical) methods for point estimation

 3.1 – The substitution method

 3.2 – The method of moments

 3.3 – Maximum likelihood estimation

4 – Warming up exercise

3/42

Lecture outline

1 – Introduction

2 – The mathematical framework of statistical inference

3 – Some (classical) methods for point estimation

 3.1 – The substitution method

 3.2 – The method of moments

 3.3 – Maximum likelihood estimation

4 – Warming up exercise

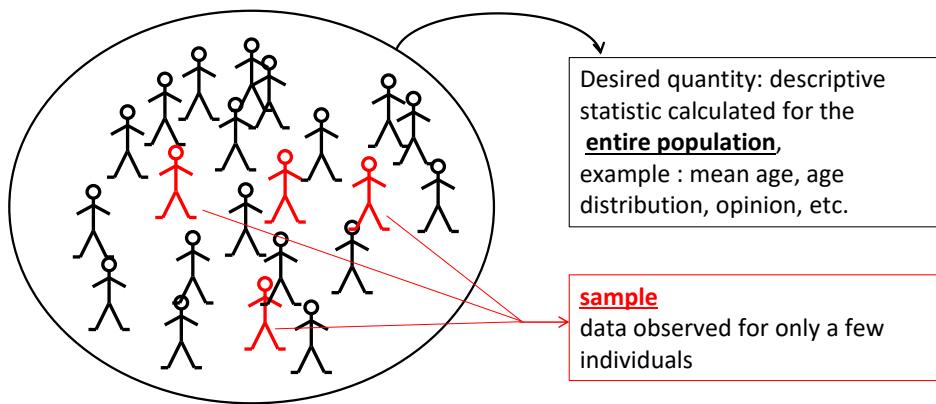
One word, several meanings...

- ▶ **One (or several) statistic(s)**: numerical indicators, often simple, computed from data.
Examples : average, standard deviation, median, etc....
- ▶ **statistics**: a mathematical discipline which has several branches, including
 - ⇒ descriptive statistics,
 - ⇒ **statistical inference** (part 1 of this course),
 - ⇒ design of experiments,
 - ⇒ **statistical learning** (part 2 of this course),
 - ⇒ ...

Remark: a mathematical definition of the word “statistic” (first meaning) will be given later.

4/42

Historical example: the opinion survey case



A descriptive statistic may be calculated on:

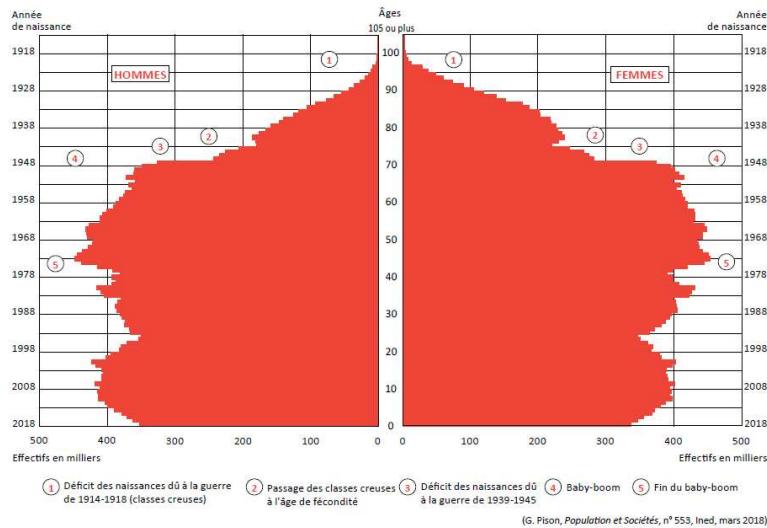
- ▶ the entire **population** → quantity of interest
- ▶ a **sample** → “approximate” value (sense to be defined)

To infer = to draw conclusions about a population
from data collected for a sample

5/42

Demographic statistics (census)

Population de la France - Évaluation provisoire au 1^{er} janvier 2018



Descriptive statistics are useful to “explore” data sets

Typical goals: obtain numerical summaries (of small dimension)
and/or easily interpretable visualizations.

6/42

Other example: estimation of a proportion

Context. Consider a box with W white balls and R red balls, where W and R are unknown.

Goal. Estimate the proportion $\theta = \frac{W}{W+R}$ of white balls.

Data (observations). We perform n draws with replacement
⇒ for the i -th draw, $x_i = 1$ if the ball is white, 0 otherwise.

Steps to estimate θ

① statistical modeling

x_i realization of a RV X_i , with $X_i \stackrel{\text{iid}}{\sim} \text{Ber}(\theta)$, $0 \leq \theta \leq 1$

② inference (here, estimation)

using the data $\underline{x} = (x_1, \dots, x_n)$ and the statistical model.

⇒ Consider $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n X_i$ (a possible descriptive statistic)

⇒ Is it reasonable to use it as a “substitute” for the unknown θ ?

7/42

Relation between statistical inference and probability theory

Probability theory provides the foundation for statistical inference:

- ▶ **probability theory**: a probability space is given;
- ▶ **statistical inference**: several probabilistic models are assumed possible; we want to extract (from data) information from data about the underlying probability measure.

Illustration on the “box” example:

	Probability (W and R known)	Inference (W and R unknown)
typical questions	<ul style="list-style-type: none"> • distribution of the number of white balls after n draws; • distribution of the number of draws to get the first white ball 	<ul style="list-style-type: none"> • estimate θ; • give an interval containing θ; • decide whether $\theta \leq 0.5$ or not.
type of conclusions	certain	for finite n , impossible to answer with certainty

8/42

Application fields & examples of statistical questions

Many fields of application:

- ▶ **Healthcare**: identify biomarkers responsible for a disease from data collected on cohorts.
- ▶ **Environment, safety**: estimate the probability of risk from measurement data.
- ▶ **Industry**: control the quality of a production line from data collected for only a few elements.
- ▶ **Opinion survey** : predict the winner of an election from a survey, quantify the uncertainty about the prediction.
- ▶ **Insurance** : evaluate the risk of ruin for an insurance company facing a disaster.

9/42

Lecture outline

1 – Introduction

2 – The mathematical framework of statistical inference

3 – Some (classical) methods for point estimation

- 3.1 – The substitution method
- 3.2 – The method of moments
- 3.3 – Maximum likelihood estimation

4 – Warming up exercise

From data to random variables

Data (observations)

Let $\underline{x} \in \mathcal{X}$ denote the data that must be analyzed. For instance:

- ① a scalar quantity, measured on n objects/individuals:
⇒ $\underline{x} = (x_1, \dots, x_n), \quad x_i \in \mathbb{R}, \quad \mathcal{X} = \mathbb{R}^n;$
- ② d scalar quantities, potentially of different natures, measured on n objects/individuals:
⇒ $\underline{x} = (x_1, \dots, x_n), \quad x_i \in \mathbb{R}^d, \quad \mathcal{X} = \mathbb{R}^{n \times d};$
- ③ any dataset of a more complex nature
(times series, symbolic data, graphs, etc.).

The data is modeled, **a priori**, by a **random variable** (RV) X

⇒ \underline{x} is considered as a realization of X .

Statistical model

The observation space ($\underline{\mathcal{X}}$, $\underline{\mathcal{A}}$)

It is the measurable space in which \underline{X} takes its values.

Most of the time, we will use:

- ▶ $\underline{\mathcal{X}} = \mathbb{R}^n$ with $\underline{\mathcal{A}} = \mathcal{B}(\mathbb{R}^n)$
- ▶ or, more generally, $\underline{\mathcal{X}} = \mathbb{R}^{n \times d}$ with $\underline{\mathcal{A}} = \mathcal{B}(\mathbb{R}^{n \times d})$.

Statistical modeling

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space carrying:

- ▶ the observed random variable \underline{X} ,
- ▶ any other (unobserved) RV that we might need.

The probability \mathbb{P} is not perfectly known: we consider a

- ▶ set \mathcal{P} of probability distributions sur (Ω, \mathcal{F})

11/42

Statistical model (cont'd)

Distribution of the observations

Let $\mathbb{P}^{\underline{X}}$ denote the distribution of \underline{X} when $\mathbb{P} \in \mathcal{P}$ is the underlying probability measure.

- ⇒ We have a set $\mathcal{P}^{\underline{X}} = \{\mathbb{P}^{\underline{X}}, \mathbb{P} \in \mathcal{P}\}$ of possible distributions.

Definition: Statistical model

Formally, we call **statistical model** the triplet

$$\mathcal{M} = (\underline{\mathcal{X}}, \underline{\mathcal{A}}, \mathcal{P}^{\underline{X}}).$$

Remarks:

- ▶ We can construct several models $(\Omega, \mathcal{F}, \mathcal{P}, \underline{X})$ for a given \mathcal{M} .
- ▶ In particular, when we only care about the observed RV \underline{X} , we can work on the *canonical* model: $\Omega = \underline{\mathcal{X}}$, $\mathcal{F} = \underline{\mathcal{A}}$, $\mathcal{P} = \mathcal{P}^{\underline{X}}$, $\underline{X} = \text{Id}_{\underline{\mathcal{X}}}$.

12/42

Statistical inference

Reminder: the data $\underline{x} \in \mathcal{X}$ is seen as a realization of $\underline{X} \sim \mathbb{P}^{\underline{X}}$, for a certain (unknown) probability $\mathbb{P} \in \mathcal{P}$.

The goal of statistical inference

Goal: to construct procedures allowing to extract information about $\mathbb{P}^{\underline{X}}$ from

- ▶ one realization of \underline{X} ,
- ▶ the knowledge of the set $\mathcal{P}^{\underline{X}}$ of all possible distributions.

Important

Since the true probability \mathbb{P} is unknown, we must design statistical procedures that are “applicable” to **any** probability $\mathbb{P} \in \mathcal{P}$.

13/42

Family of distributions

The set \mathcal{P} is represented by a **parameterized family**:

$$\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}.$$

Parametric model

If Θ is finite-dimensional, the model is called **parametric**.

- ▶ the parameter vector θ is often of small size.
- ▶ we will denote by p the number of parameters ($\Theta \subset \mathbb{R}^p$).

Example. Family of **Gaussian distributions** on $\mathcal{X} = \mathbb{R}$

$$\mathcal{P}^{\underline{X}} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_*^+\}$$

(In this example we consider only one scalar observation.)

14/42

Assumptions on the family of distributions

Dominated model

The model

$$\mathcal{M} = (\underline{\mathcal{X}}, \underline{\mathcal{A}}, \{\mathbb{P}_\theta^X, \theta \in \Theta\})$$

is said to be **dominated** if there exists a (σ -finite) measure ν on $(\underline{\mathcal{X}}, \underline{\mathcal{A}})$ such that

$$\forall \theta \in \Theta, \quad \forall A \in \underline{\mathcal{A}}, \quad \mathbb{P}_\theta^X(A) = \int_A f_\theta(x) \nu(dx).$$

► f_θ is the **density** of \mathbb{P}_θ^X with respect to ν .

In this course, we will consider the following cases:

- ▶ “continuous” RV: reference measure $\nu = \text{Lebesgue's measure}$,
- ▶ discrete RV: reference measures $\nu = \text{counting measure}$.

15/42

Assumptions on the family of distributions (cont'd)

Identifiable model

The model

$$\mathcal{M} = (\underline{\mathcal{X}}, \underline{\mathcal{A}}, \{\mathbb{P}_\theta^X, \theta \in \Theta\})$$

is **identifiable** if the mapping $\theta \mapsto \mathbb{P}_\theta^X$ is **injective**.

In the rest of this course, all the models will be

- ▶ **dominated** by a reference measure ν ,
- ▶ **identifiable**.

16/42

Sampling models

n-sample

If $\underline{X} = (X_1, \dots, X_n)$ is such that:

- ▶ the X_i 's are (mutually) independent,
- ▶ all the X_i 's have the same distribution P ,

then the X_i 's are called **independent et identically distributed (iid)** and we say that \underline{X} is an (iid) ***n*-sample**.

Distribution of an *n*-sample.

Consider the model that describes each of the X_i 's individually:

- ▶ $(\mathcal{X}, \mathcal{A}, \{P_\theta, \theta \in \Theta\})$

Then we have:

- ▶ $(\underline{\mathcal{X}}, \underline{\mathcal{A}}) = (\mathcal{X}^n, \mathcal{A}^{\otimes n})$ (product space),
- ▶ $\forall \theta \in \Theta, \mathbb{P}_\theta^{\underline{X}} = P_\theta^{\otimes n}$ (product distribution).

17/42

Example: component reliability

This application will be used as an illustration in several lectures.

Context

- ▶ We are interested in the reliability of components from a production line.
- ▶ Reliability: measured by the **lifetime of the components**.
- ▶ Data (observations): a sample of $n = 10$ components, for which the lifetime has been recorded : $\underline{x} = (x_1, \dots, x_n)$.

Modeling

- ▶ Each x_i is modeled by a scalar RV X_i .
- ▶ The X_i 's are assumed **iid**, with values in $(\mathcal{X}, \mathcal{A}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

18/42

Example: component reliability

Modeling (cont'd): family of distributions

Typical* assumption for the lifetime of a component:

$$X_1 \sim \mathcal{E}(\theta), \quad \theta > 0.$$

Hence the statistical model for one observation:

$$(\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\mathcal{E}(\theta), \theta > 0\}).$$

Note: this assumption on X_1 holds for all the X_i 's, $i \geq 1$.

Density. The exponential distribution $\mathcal{E}(\theta)$ has the density:

$$f_\theta(x) = \theta \exp(-\theta x) \mathbb{1}_{[0, \infty]}(x).$$

* in the case of unpredictable failures, not related to the age of the component

Example: component reliability

A few problems of (statistical) interest

- ▶ estimate θ , or
- ▶ estimate $\eta = \frac{1}{\theta} = \mathbb{E}(X_1)$ (average lifetime)
 - ⇒ lectures #1 et #2
- ▶ provide confidence intervals for θ and η
 - ⇒ lecture #3
- ▶ estimate θ given prior information on its value
(e.g., provided by the manufacturer of the production line)
 - ⇒ lecture #4 on Bayesian estimation
- ▶ test the hypothesis $\eta \leq 10$, in order to assess the value of an optional warranty extension
 - ⇒ lecture #5 on hypothesis testing

Data.

0.5627	16.1121	5.4943	7.9374	1.2658
2.9885	8.6266	43.8877	2.1641	8.9138

Table – Measured values (arbitrary units) for a sample of size $n = 10$ **Estimating η** : a first **estimator**[†].

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow[n \rightarrow \infty]{\text{a.s.}} \mathbb{E}_{\theta}(X_1) = \eta \quad (\text{SLLN}).$$

⇒ $\hat{\eta}^{(1)} = \bar{X}$ seems to be a “reasonable” estimator of η .**Numerical application** $\hat{\eta}^{(1)} = 10.1960$ [†] see Lecture 2 for a definition**Notations / vocabulary****Notations.** We will often use notations such as

- ▶ $\mathbb{E}_{\theta}(\cdot)$ (expectation),
- ▶ $\text{var}_{\theta}(\cdot)$ (variance ou covariance matrix),
- ▶ $f_{\theta}(\cdot)$ (density), ...

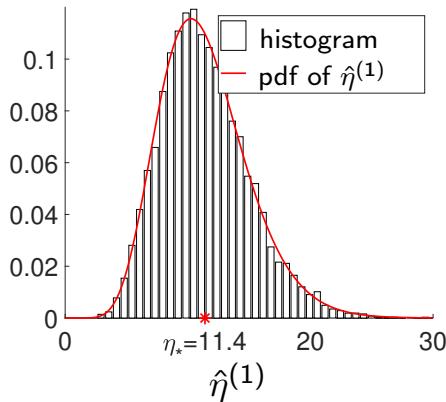
to indicate that these operators or functions depend on a probability \mathbb{P}_{θ} for a particular value of θ .**Definition: Statistic**A **statistic** is a random variable (often scalar- or vector-valued) that can be computed from \underline{X} alone*.Example: the estimator $\hat{\eta}^{(1)} = \bar{X}$ is a statistic.

* Technically: can be written as a measurable function of \underline{X} .
In particular, depends neither on other (unobserved) RVs nor on θ .

Numerical assessment of the performance of $\hat{\eta}^{(1)}$

With numerical simulations, (almost) everything is possible!

- ▶ we **choose** a particular value of η (here, $\eta_* = 11.4$), then
- ▶ we **simulate** on a computer a large number m of n -samples (here, $m = 10000$).



Remarks

- ▶ Our estimates are, in this case, **not very accurate**.
- ▶ Providing **confidence intervals** would be very relevant here.
- ▶ In this simple example we can compute the density of $\hat{\eta}^{(1)}$ analytically.

23/42

A few words on the Gamma distribution $\Gamma(p, \lambda)$

Let $X \sim \Gamma(p, \lambda)$, $p > 0$, $\lambda > 0$). Its pdf is

$$f(x) = \frac{\lambda^p}{\Gamma(p)} x^{p-1} \exp(-\lambda x) \mathbb{1}_{\mathbb{R}^+}(x).$$

Moments

- ▶ mean : $\mathbb{E}_\theta(X) = \frac{p}{\lambda}$
- ▶ variance : $\text{var}_\theta(X) = \frac{p}{\lambda^2}$

Particular cases

- ▶ $\mathcal{E}(\lambda) = \Gamma(p=1, \lambda)$
- ▶ $\Gamma(p=\frac{n}{2}, \lambda=\frac{n}{2}) = \chi^2(n)$

Properties

- ▶ Let $a > 0$. If $X \sim \Gamma(p, \lambda)$, then $aX \sim \Gamma(p, \frac{\lambda}{a})$.
- ▶ If $X \sim \Gamma(p, \lambda)$, $Y \sim \Gamma(q, \lambda)$, and X and Y are independent, then $X + Y \sim \Gamma(p+q, \lambda)$.

Application. $\hat{\eta}^{(1)} \sim \Gamma\left(n, \frac{n}{\eta}\right)$.

$\hat{\eta}^{(2)}$: another estimator

With a convergence argument similar to the one used earlier:

$$\frac{1}{n} \sum_{i=1}^n X_i^2 \xrightarrow[n \rightarrow \infty]{\text{a.s.}} \mathbb{E}_\theta(X_1^2) = \frac{2}{\theta^2} = 2\eta^2,$$

therefore using $\hat{\eta}^{(2)} = \sqrt{\frac{1}{2n} \sum_{i=1}^n X_i^2}$ seems “reasonable” as well.

Numerical application $\hat{\eta}^{(2)} = 11.2228$

Questions

- ▶ How can we compare two estimators ?
- ▶ If there an estimator that is “better” than the others ?
- ▶ How to construct “good” estimators ?

24/42

Lecture outline

1 – Introduction

2 – The mathematical framework of statistical inference

3 – Some (classical) methods for point estimation

- 3.1 – The substitution method
- 3.2 – The method of moments
- 3.3 – Maximum likelihood estimation

4 – Warming up exercise

Mathematical framework

In this section:

- ▶ we consider a statistical model

$$\mathcal{M} = \left(\underline{\mathcal{X}}, \underline{\mathcal{A}}, \left\{ \mathbb{P}_{\theta}^X, \theta \in \Theta \right\} \right),$$

most of the time assumed to be **parametric** ($\Theta \subset \mathbb{R}^p$);

- ▶ when X is an **IID n -sample**, we write

- ▶ $\underline{X} = (X_1, \dots, X_n)$
- ▶ $\underline{\mathcal{X}} = \mathcal{X}^n$, with $\mathcal{X} = \mathbb{R}$ or $\mathcal{X} = \mathbb{R}^d$,
- ▶ $\mathbb{P}_{\theta}^X = P_{\theta}^{\otimes n}$;

- ▶ we want to estimate a “**quantity of interest**”:

- ▶ either θ itself,
- ▶ or, more generally, $\eta = g(\theta)$.

25/42

Lecture outline

1 – Introduction

2 – The mathematical framework of statistical inference

3 – Some (classical) methods for point estimation

- 3.1 – The substitution method
- 3.2 – The method of moments
- 3.3 – Maximum likelihood estimation

4 – Warming up exercise

The substitution method

Assume that

- ▶ we already have an estimator $\hat{\eta}$ of $\eta = g(\theta)$
- ▶ and we want to estimate another quantity of interest η' that can be written as $\eta' = h(\eta)$, with h a continuous function.

The substitution method

The **substitution method** consists in using

$$\hat{\eta}' = h(\hat{\eta}) \text{ as an estimator of } \eta.$$

26/42

Example: component reliability

Reminder: $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{E}(\theta), \quad \theta > 0$.

We are interested in the probability that a failure occurs before t_0 :

$$\begin{aligned} \Rightarrow \eta' &= \mathbb{P}_\theta(X_1 \leq t_0) = \int_0^{t_0} \theta \exp(-\theta x) dx \\ &= 1 - \exp(-\theta t_0) = 1 - \exp\left(-\frac{t_0}{\eta}\right). \end{aligned}$$

Using $\hat{\eta}^{(1)} = \bar{X}$ as an estimator of η , we get

$$\hat{\eta}' = 1 - \exp\left(-\frac{t_0}{\bar{X}}\right).$$

27/42

Empirical measure

Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathbb{P}^{X_1}$.

Recall the **Dirac measure at $x \in \mathcal{X}$** :

$$\forall A \in \mathcal{A}, \quad \delta_x(A) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{otherwise.} \end{cases}$$

Definition: empirical measure

The **empirical measure** is the (random) measure defined by:

$$\hat{\mathbb{P}}^{X_1} = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}.$$

Usefulness: the empirical measure can be seen as an estimator of \mathbb{P}^{X_1} \Rightarrow allows us to **construct other estimators** using the **substitution method**.

28/42

Example : estimator of the k -th order moment

Assume $X_1 \in L^k$. Then

$$m_k = \mathbb{E}(X_1^k) = \mathcal{G}(\mathbb{P}^{X_1})$$

is well defined, with $\mathcal{G}(\mu) = \int_{\mathcal{X}} x^k \mu(dx)$. By substitution:

$$\hat{m}_k = \mathcal{G}(\hat{\mathbb{P}}^{X_1}) = \int_{\mathcal{X}} x^k \frac{1}{n} \sum_{i=1}^n \delta_{X_i}(dx) = \frac{1}{n} \sum_{i=1}^n X_i^k.$$

Similar example : the sample variance. If $X_1 \in L^2$ and $\eta' = \text{var}(X_1) = \mathcal{G}(\mathbb{P}^{X_1})$, where $\mathcal{G}(\mu) = \int_{\mathcal{X}} x^2 \mu(dx) - (\int_{\mathcal{X}} x \mu(dx))^2$, we get by substitution:

$$S^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (\text{sample variance}).$$

29/42

One last example : the empirical cdf

Let $x \in \mathbb{R}$. The cumulative distribution function (cdf) of X_1 at x is

$$F(x) = \mathbb{P}^{X_1}(X_1 \leq x) = \mathcal{G}_x(\mathbb{P}^{X_1}) \quad \text{with} \quad \mathcal{G}_x(\mu) = \int_{-\infty}^x \mu(dx).$$

Hence the **empirical cdf**:

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}}.$$

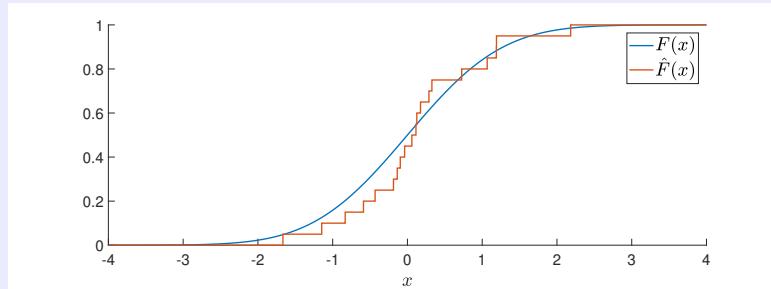


Figure – Empirical cdf for $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1)$ and $n = 20$.

Lecture outline

1 – Introduction

2 – The mathematical framework of statistical inference

3 – Some (classical) methods for point estimation

3.1 – The substitution method

3.2 – The method of moments

3.3 – Maximum likelihood estimation

4 – Warming up exercise

The method of moments

Assume that

- $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} P_\theta$, with $\theta \in \Theta$;
- the model is **parametric**: $\Theta \subset \mathbb{R}^p$,
- we want to estimate θ itself

Consider the function

$$\begin{aligned} h : \Theta \subset \mathbb{R}^p &\rightarrow h(\Theta) \subset \mathbb{R}^p, \\ \theta &\mapsto h(\theta) = \begin{pmatrix} \mathbb{E}_\theta(X_1) \\ \vdots \\ \mathbb{E}_\theta(X_1^p) \end{pmatrix}. \end{aligned}$$

Remark: sometimes other moments can be used (not necessarily the first p).

30/42

The method of moments (cont'd)

Assume $h : \Theta \rightarrow h(\Theta)$ injective, and thus **bijective**.

The method of moments

The method of moments consists in

- estimating the first p moments $\hat{m}_k = \frac{1}{n} \sum_{i=1}^n X_i^k$, $k \leq p$,
- then applying h^{-1} to construct an estimator of θ .

Hence **moment-of-moments estimator** : $\hat{\theta} = h^{-1}(\hat{m}_{1:p})$, where

$$\hat{m}_{1:p} = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n X_i \\ \vdots \\ \frac{1}{n} \sum_{i=1}^n X_i^p \end{pmatrix}.$$

Remark: well defined only if $\hat{m}_{1:p} \in h(\Theta)$ \mathbb{P}_θ -ps, pour tout θ .

Otherwise, minimization of some distance (generalized method of moments).

31/42

Method of moments: examples

Example: component reliability

We have $\mathbb{E}_\theta(X_1) = \theta^{-1}$ (exponential distribution), therefore

$$\theta = (\mathbb{E}_\theta(X_1))^{-1} \quad \text{and} \quad \hat{\theta} = (\bar{X})^{-1}.$$

Example: $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, with $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$

$$\text{We have } h(\theta) = \begin{pmatrix} \mathbb{E}_\theta(X_1) \\ \mathbb{E}_\theta(X_1^2) \end{pmatrix} = \begin{pmatrix} \mu \\ \mu^2 + \sigma^2 \end{pmatrix},$$

$$\text{therefore } \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} = \begin{pmatrix} \mathbb{E}_\theta(X_1) \\ \mathbb{E}_\theta(X_1^2) - (\mathbb{E}_\theta(X_1))^2 \end{pmatrix},$$

$$\text{and finally } \begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n X_i \\ \frac{1}{n} \sum_{i=1}^n X_i^2 - \left(\frac{1}{n} \sum_{i=1}^n X_i \right)^2 \end{pmatrix}$$

Exercise. $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{U}_{[a,b]}$. Method-of-moments estimator of (a, b) ? 32/42

Lecture outline

1 – Introduction

2 – The mathematical framework of statistical inference

3 – Some (classical) methods for point estimation

3.1 – The substitution method

3.2 – The method of moments

3.3 – Maximum likelihood estimation

4 – Warming up exercise

Maximum likelihood estimation

Reminder: **dominated model** $\rightarrow \mathbb{P}_{\theta}^X$ admits a pdf f_{θ} .

Definition: likelihood

We call **likelihood** the function:

$$\begin{aligned}\mathcal{L} : \Theta \times \underline{\mathcal{X}} &\rightarrow \mathbb{R}_+ \\ (\theta; \underline{x}) &\mapsto f_{\theta}(\underline{x})\end{aligned}$$

Remark. Si $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} P_{\theta}$, then $\mathcal{L}(\theta; \underline{x}) = \prod_{i=1}^n f_{\theta}(x_i)$.

(usual abuse of notation: here $f_{\theta} = f_{\theta}^{X_1}$)

Definition: MLE

If $\hat{\theta}$ is a maximizer of $\theta \mapsto \mathcal{L}(\theta; \underline{X})$, then
 $\hat{\theta}$ is a **maximum likelihood estimator** (MLE) of θ .

33 / 42

MLE: practical details

- ▶ **Existence and uniqueness** of the MLE are not guaranteed in general.
- ▶ For an IID n -sample, we often use the **log-likelihood**:

$$\ln \mathcal{L}(\theta; \underline{x}) = \sum_{i=1}^n \ln f_{\theta}(x_i).$$

- ▶ If \mathcal{L} is C^2 wrt θ and $\Theta \subset \mathbb{R}^p$ is open, a **necessary condition** for $\hat{\theta}$ to be an MLE is:

$$\begin{cases} (\nabla_{\theta} (\ln \mathcal{L}))(\hat{\theta}; \underline{X}) = 0, \\ (\nabla_{\theta} \nabla_{\theta}^{\top} (\ln \mathcal{L}))(\hat{\theta}; \underline{X}) \text{ has negative eigenvalues.} \end{cases}$$

(locally concave function;
 $\nabla_{\theta} \nabla_{\theta}^{\top}$ is the Hessian operator)

34 / 42

MLE example: component reliability

For $x_1, \dots, x_n \geq 0$, we have $\mathcal{L}(\theta; \underline{x}) = \prod_{i=1}^n \theta \exp(-\theta x_i)$, and thus

$$\ln \mathcal{L}(\theta; \underline{x}) = n \ln(\theta) - \theta \sum_{i=1}^n x_i.$$

Stationarity condition ("likelihood equation")

$$\frac{\partial(\ln \mathcal{L})}{\partial \theta}(\theta; \underline{x}) = 0 \iff \frac{n}{\theta} - \sum_{i=1}^n x_i = 0.$$

⇒ If $\sum_{i=1}^n x_i > 0$, unique solution in $\Theta = \mathbb{R}_+^*$ at $\theta = n (\sum_{i=1}^n x_i)^{-1}$.

⇒ It is indeed a maximum of the likelihood function (cf. sign of the derivative).

⇒ Since $\sum_{i=1}^n X_i > 0$ a.s., a unique MLE exists: $\hat{\theta} = (\bar{X})^{-1}$.

Remark: the same estimator was obtained by the method of moments.

35/42

MLE example: Gaussian IID n -sample, $\theta = (\mu, \sigma^2)$

Same approach as in the previous example:

$$\ln \mathcal{L}(\theta; \underline{x}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2},$$

$$(\nabla_\theta \ln \mathcal{L})(\theta; \underline{x}) = \frac{n}{\sigma^2} \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n x_i - \mu \\ -\frac{1}{2} + \frac{1}{2\sigma^2} \cdot \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \end{pmatrix}.$$

Solving the likelihood equation yields:

$$\hat{\theta} = \begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n X_i \\ \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu})^2 \end{pmatrix}$$

and it can be proved that the maximum is attained at this point.

Remark: the same estimator was obtained by the method of moments.

36/42

Proof: the maximum is attained at $\hat{\theta}$

a) Let $\bar{x} = \sum_{i=1}^n x_i$. For any given σ^2 , we have:

$$\ln \mathcal{L}(\theta; \underline{x}) = -\frac{n}{2\sigma^2} (\mu - \bar{x})^2 + \text{const}(\underline{x}, n, \sigma^2).$$

$\Rightarrow \mu \mapsto \ln \mathcal{L}(\theta; \underline{x})$ is maximal at $\mu = \bar{x}$.

b) Consider then

$$\begin{aligned} g(\sigma^2) &= \max_{\mu} \ln \mathcal{L}((\mu, \sigma^2); \underline{x}) = \ln \mathcal{L}((\bar{x}, \sigma^2); \underline{x}) \\ &= -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{2\sigma^2}. \end{aligned}$$

The function g is differentiable, with derivative

$$g'(\sigma^2) = \frac{n}{2\sigma^4} \left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 - \sigma^2 \right).$$

We conclude from the sign of g' that g is maximal at $\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$.

$\Rightarrow \theta \mapsto \mathcal{L}(\theta; \underline{x})$ is maximal at $(\bar{x}, \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2)$. □

Lecture outline

1 – Introduction

2 – The mathematical framework of statistical inference

3 – Some (classical) methods for point estimation

- 3.1 – The substitution method
- 3.2 – The method of moments
- 3.3 – Maximum likelihood estimation

4 – Warming up exercise

Exercise 1 (Bernoulli model)

Let X_1, \dots, X_n be an n -sample of binary observations, independent and identically distributed according to the Bernoulli $\mathcal{B}(p)$ distribution, with $p \in [0, 1]$.

Questions

- ① Specify a formal statistical model $\mathcal{M} = (\underline{\mathcal{X}}, \underline{\mathcal{A}}, \mathcal{P}^{\underline{\mathcal{X}}})$ corresponding to this description.
- ② Construct an estimator of p using the method of moments.
- ③ Construct an estimator of p using the maximum likelihood method.
- ④ Compute the expectation and variance of $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$.

37/42

Solution of Exercise 1

① Statistical model $\mathcal{M} = (\underline{\mathcal{X}}, \underline{\mathcal{A}}, \mathcal{P}^{\underline{\mathcal{X}}})$

The “natural” (minimal) set to describe the values of a binary variable is $\mathcal{X} = \{0, 1\}$.

⇒ $\underline{\mathcal{X}} = \{0, 1\}^n$ for an n -sample

On a finite or countable set, we use in general the discrete σ -algebra, i.e., the set of all subsets of $\underline{\mathcal{X}}$.

⇒ $\underline{\mathcal{A}} = \mathcal{P}(\{0, 1\}^n) = \mathcal{P}(\{0, 1\})^{\otimes n}$

The distribution of an n -tuple (X_1, \dots, X_n) of independent RVs is the product measure $P^{X_1} \otimes \dots \otimes P^{X_n}$.

⇒ $\mathcal{P}^{\underline{\mathcal{X}}} = \{\mathcal{B}(p)^{\otimes n}, p \in [0, 1]\}$

Remark: another possible choice would have been $\underline{\mathcal{X}} = \mathbb{R}^n$, $\underline{\mathcal{A}} = \mathcal{B}(\mathbb{R}^n)$.

38/42

Solution of Exercise 1 (cont'd)

② Method of moments

If $X \sim \mathcal{B}(p)$, then $\mathbb{E}_p(X) = p$.

⇒ The method of moments, applied to the first-order moment, directly yields the estimator $\hat{p}_n = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n$.

③ Maximum likelihood

First write the likelihood:

$$\begin{aligned}\mathcal{L}(p; \underline{X}) &= \prod_{i=1}^n p^{X_i} (1-p)^{1-X_i} \\ &= p^N (1-p)^{n-N},\end{aligned}$$

where $N = \sum_{i=1}^n X_i$ and $0^0 = 1$,

39/42

Solution of Exercise 1 (cont'd)

then the log-likelihood for $p \in (0, 1)$:

$$\begin{aligned}\ell(p; \underline{X}) &= \ln(\mathcal{L}(p; \underline{X})) \\ &= N \ln(p) + (n - N) \ln(1 - p).\end{aligned}$$

The log-likelihood is differentiable on $(0, 1)$, with derivative

$$\begin{aligned}\frac{\partial \ell}{\partial p}(p; \underline{X}) &= \frac{N}{p} - \frac{n - N}{1 - p} \\ &= \frac{n}{p(1 - p)} \cdot (\bar{X}_n - p).\end{aligned}$$

We have $\frac{\partial \ell}{\partial p}(p; \underline{X}) > 0$ iff $p < N/n = \bar{X}_n$,
 $\frac{\partial \ell}{\partial p}(p; \underline{X}) < 0$ iff $p > N/n = \bar{X}_n$.

40/42

Solution of Exercise 1 (cont'd)

If $\bar{X}_n = 0$, the log-likelihood is strictly decreasing

⇒ the likelihood is maximal at $p = 0$.

If $\bar{X}_n = 1$, the log-likelihood is strictly increasing

⇒ the likelihood is maximal at $p = 1$.

If $0 < \bar{X}_n < 1$, the log-likelihood is maximal at $p = \bar{X}_n$.

Summary: $\hat{p}_n = \bar{X}_n$ is the unique MLE.

Remark: the log-likelihood takes infinite values at $p = 0$ and/or $p = 1$, but the likelihood itself is well defined and continuous on $[0, 1]$.

41/42

Solution of Exercise 1 (cont'd)

④ Expectation and variance of \bar{X}

Reminders

- ▶ $\mathbb{E}_p(X_1) = p$ and $\text{var}_p(X_1) = p(1 - p)$.
- ▶ independence ⇒ decorrelation ⇒ $\text{var}(\sum_i X_i) = \sum_i \text{var}(X_i)$.

Using that the X_i 's are identically distributed:

$$\mathbb{E}_p(\bar{X}_n) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_p[X_1] = p.$$

Using that the X_i 's are IID:

$$\text{var}_p(\bar{X}_n) = \frac{1}{n^2} \text{var}_p\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{var}_p(X_i) = \frac{p(1 - p)}{n}.$$

42/42

Chapter 2

Point estimation



CentraleSupélec

Statistics and Learning

Arthur Tenenhaus[†], Julien Bect & Laurent Le Brusquet

(firstname.lastname@centralesupelec.fr)

Teaching: CentraleSupélec / Department of Mathematics

Research: Laboratory of signals and systems (L2S)

[†]: Course coordinator

1/55

Lecture 2/10

Point estimation

In this lecture you will learn how to...

- ▶ Learn how to quantify the performance of an estimator.
- ▶ Learn how to compare estimators.
- ▶ Introduce the asymptotic approach.

2/55

Lecture outline

- 1 – Point estimation: definition and notations
- 2 – Quadratic risk of an estimator
- 3 – A lower bound on the quadratic risk
- 4 – Asymptotic properties
- 5 – Warming up exercises

3/55

Lecture outline

- 1 – Point estimation: definition and notations
- 2 – Quadratic risk of an estimator
- 3 – A lower bound on the quadratic risk
- 4 – Asymptotic properties
- 5 – Warming up exercises

Recap: mathematical framework

Data

- ▶ Formally, an element \underline{x} in a set \mathcal{X} .
- ▶ ex: $\mathcal{X} = \mathbb{R}^n, \mathbb{R}^{n \times d}, \{\text{words}\}$, some functional space, etc.

From data to random variables

- ▶ **A priori** point of view: before the data is actually collected.
- ▶ Modeling: RV \underline{X} taking values in $(\mathcal{X}, \mathcal{A})$,
- ▶ **but the distribution of \underline{X} is unknown.**

Statistical modeling

- ▶ \underline{X} is assumed to be defined on $(\Omega, \mathcal{F}, \mathbb{P})$, with $\mathbb{P} \in \mathcal{P}$.
- ▶ \mathcal{P} : a set of possible probability measures on (Ω, \mathcal{F})
- ▶ Formally, $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}^{\underline{X}})$, with $\mathcal{P}^{\underline{X}} = \{\mathbb{P}^{\underline{X}}, \mathbb{P} \in \mathcal{P}\}$.

Canonical construction: $\Omega = \mathcal{X}, \mathcal{F} = \mathcal{A}, \underline{X} = \text{Id}_{\mathcal{X}}$ et $\mathcal{P} = \mathcal{P}^{\underline{X}}$.

4/55

Recap: mathematical framework (cont'd)

Important

Since $\mathbb{P} \in \mathcal{P}$ is unknown, we must design statistical procedure that “work well” (in a sense to be specified) for **any** distribution $\mathbb{P} \in \mathcal{P}$.

Parameterized family of probability distributions

- ▶ Usually, we write $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$.
- ▶ θ : **unknown parameter** (scalar, vector, function...)
- ▶ In the following, we assume a **parametric model**: $\Theta \subset \mathbb{R}^p$.

Important case: d -variate (iid) n -sample ($\rightarrow n \times d$ data table)

- ▶ $\mathcal{X} = \mathcal{X}^n$, with $\mathcal{X} \subset \mathbb{R}^d$, endowed with their Borel σ -algebras,
- ▶ $\underline{X} = (X_1, \dots, X_n)$ with $X_i \stackrel{\text{iid}}{\sim} P_\theta$, and thus $\mathbb{P}_\theta^{\underline{X}} = P_\theta^{\otimes n}$.

5/55

Point estimation

Parameter of interest

- ▶ We are interested in parameter $\eta = g(\theta)$, where $g : \Theta \mapsto \mathbb{R}$ ou \mathbb{R}^q .
- ▶ Its value is unknown, since θ is unknown.

Informal definition: estimation

Guess (infer) the value of η based on a realization \underline{x} of \underline{X} .

Definition: estimator

We call estimator any statistic $\hat{\eta} = \varphi(\underline{X})$ taking value in the set $N = g(\Theta)$ of possible values for η .

Remark: the word “estimator” can refer either to the RV $\hat{\eta}$ or to the function φ . In practice, we identify the two and write (abusively) $\hat{\eta} = \hat{\eta}(\underline{X})$.

6/55

Example 1 (reminder)

IID Gaussian n -sample: $\underline{X} = (X_1, \dots, X_n)$ with

- ▶ $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$,
- ▶ $\theta = (\mu, \sigma^2)$,
- ▶ $\Theta = \mathbb{R} \times]0; +\infty[$.

In this example, we assume that we want to estimate the mean μ :

- ▶ here $\eta = \mu$ and $g : \theta = (\mu, \sigma^2) \mapsto \mu$,
- ▶ σ^2 is unknown too (nuisance parameter).

7/55

Example 1 (cont'd)

Some possible estimators . . .

- ▶ $\hat{\mu}_1 = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ (method of moments / MLE),
- ▶ $\hat{\mu}_2 = \mu_0$ for a given $\mu_0 \in \mathbb{R}$,
- ▶ $\hat{\mu}_3 = \frac{1}{2}\mu_0 + \frac{1}{2}\bar{X}_n$,
- ▶ $\hat{\mu}_4 = \bar{X}_n + c$ for a given $c \neq 0$,
- ▶ $\hat{\mu}_5 = \text{med}(X_1, \dots, X_n)$,
- ▶ ...

Questions

- ▶ Is one these estimators “better” than the others?
- ▶ Can we find an “optimal” estimator ?
- ▶ In what sense ?

8/55

Other examples

Example 1'

- ▶ Same statistical model as in Example 1, but
- ▶ $g(\theta) = \sigma^2$.
- ▶ In this case, μ is seen as a nuisance parameter.

Example 1''

- ▶ Again the same statistical model, but
- ▶ $g(\theta) = \theta = (\mu, \sigma^2)$.
- ▶ Here, the parameter to be estimated is a **vector**.

9/55

Other examples (cont'd)

Example 2

- ▶ $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{E}(\theta)$, i.e., $f_\theta(x) = \theta e^{-\theta x} \mathbb{1}_{x \geq 0}$,
- ▶ $\Theta = (0, +\infty)$,
- ▶ $g(\theta) = \mathbb{E}_\theta(X_1) = 1/\theta$.

Example 2'

- ▶ Same statistical model, but
- ▶ $g(\theta) = \mathbb{P}_\theta(X_1 > x_0) = e^{-\theta x_0}$ for a given $x_0 > 0$.

10/55

Other examples (cont'd)

Example 3

- ▶ $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} P$,
- ▶ $\theta = P$, unknown distribution,
- ▶ $\Theta = \{\text{distributions on } (\mathbb{R}, \mathcal{B}(\mathbb{R}))\}$,
- ▶ $g(\theta) = F$: cumulative distribution functions of the X_i 's.

Example 4

- ▶ $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} P_\theta$,
- ▶ P_θ : probability density functions $\theta(x)$
- ▶ $\Theta = \{\text{pdf on } \mathbb{R}, \text{ of class } \mathcal{C}^2, \text{ with } \int \theta''(x)^2 dx < +\infty\}$
- ▶ $g(\theta) = \theta$.

Examples 3 et 4: non-parametric statistics (not treated in this course).

11/55

Lecture outline

1 – Point estimation: definition and notations

2 – Quadratic risk of an estimator

3 – A lower bound on the quadratic risk

4 – Asymptotic properties

5 – Warming up exercises

General concept of risk

Goal

Quantify the performance of an estimator

Consider a **loss function** $L : N \times N \rightarrow \mathbb{R}$.

- ▶ Reminder: $N = g(\Theta)$ is the set of all possible values for η .
- ▶ Interpretation: we lose $L(\eta, \eta')$ if we choose η' as our estimate while η is the true value.

Risk

For a given loss function L , we define the risk $R_\theta(\hat{\eta})$ of the estimator $\hat{\eta}$, for the value $\theta \in \Theta$ of the unknown parameter, by

$$R_\theta(\hat{\eta}) = \mathbb{E}_\theta(L(g(\theta), \hat{\eta})).$$

Quadratic risk

Quadratic risk

We call **quadratic risk** the risk associated with the loss function

$$L(\eta, \eta') = \|\eta - \eta'\|^2,$$

that is,

$$R_\theta(\hat{\eta}) = \mathbb{E}_\theta (\|g(\theta) - \hat{\eta}\|^2).$$

Remarks

- ▶ Also called “mean square error” (MSE).
- ▶ **Most commonly used** notion of risk
(for the sake of simplicity, as we will see);
- ▶ in the rest of the lecture, **we will consider this risk exclusively.**

13/55

Example 1 (reminder)

IID Gaussian n -sample: $\underline{X} = (X_1, \dots, X_n)$ with

- ▶ $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2),$
- ▶ $\theta = (\mu, \sigma^2),$
- ▶ $\Theta = \mathbb{R} \times]0; +\infty[.$

In this example, we assume that we want to **estimate the mean μ** ;

- ▶ here $\eta = \mu$ and $g : \theta = (\mu, \sigma^2) \mapsto \mu,$
- ▶ σ^2 is unknown too (*nuisance parameter*).

14/55

Example 1: risk of the estimator $\hat{\mu}_1$

Consider the estimator

$$\hat{\mu}_1 = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

For all $\theta = (\mu, \sigma^2) \in \Theta$, we have the following result:

Quadratic risk of the sample mean

$$R_\theta(\hat{\mu}_1) = \mathbb{E}_\theta((\hat{\mu}_1 - \mu)^2) = \frac{\sigma^2}{n}.$$

Remark: the result holds as soon as the X_i 's have finite second order moments
 (Gaussianity is not actually used)

15/55

Example 1: risk of the estimator $\hat{\mu}_1$ (computation)

Notice that

$$\mathbb{E}_\theta(\hat{\mu}_1) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\theta(X_i) = \mu.$$

Therefore

$$\begin{aligned} R_\theta(\hat{\mu}_1) &= \text{var}_\theta(\hat{\mu}_1) = \frac{1}{n^2} \text{var}_\theta\left(\sum_{i=1}^n X_i\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{var}_\theta(X_i) = \frac{\sigma^2}{n} \end{aligned}$$

□

16/55

Bias of an estimator

Let $\hat{\eta}$ be an estimator of $\eta = g(\theta)$ st $\mathbb{E}_\theta (\|\hat{\eta}\|) < +\infty$, $\forall \theta \in \Theta$.

Definition: bias / unbiased estimator

The **bias** of an estimator $\hat{\eta}$ at $\theta \in \Theta$ is defined as

$$b_\theta(\hat{\eta}) = \mathbb{E}_\theta(\hat{\eta}) - g(\theta).$$

We will say that $\hat{\eta}_n$ is an **unbiased estimator** (UE) if

$$b_\theta(\hat{\eta}) = 0, \quad \forall \theta \in \Theta.$$

Example 1

- ▶ We have already seen that $\hat{\mu}_1 = \bar{X}_n$ is an UE of μ .
- ▶ More generally (exercise): $\hat{\mu} = \alpha + \beta \bar{X}_n$ is an UE of μ if, and only if, $\alpha = 0$ et $\beta = 1$.

17/55

Bias-variance decomposition

Reminder: we still consider the **quadratic risk**.

Let $\hat{\eta}$ be an estimator of $\eta = g(\theta)$ st $\mathbb{E}_\theta (\|\hat{\eta}\|^2) < +\infty$, $\forall \theta \in \Theta$.

Proposition: Bias-variance decomposition (scalar case)

If the quantity of interest is scalar ($\eta \in \mathbb{R}$), we have:

$$R_\theta(\hat{\eta}) = \mathbb{E}_\theta((\hat{\eta} - g(\theta))^2) = \text{var}_\theta(\hat{\eta}) + b_\theta(\hat{\eta})^2.$$

Remark: we can generalize to the vector case by summing over the components:

$$R_\theta(\hat{\eta}) = \mathbb{E}_\theta(\|\hat{\eta} - g(\theta)\|^2) = \text{tr}(\text{var}_\theta(\hat{\eta})) + \|b_\theta(\hat{\eta})\|^2,$$

where $\text{var}_\theta(\hat{\eta})$ is the covariance matrix of $\hat{\eta}$.

18/55

Example 1: risk of some estimators

$$\hat{\mu}_1 = \bar{X}_n$$

$$R_\theta(\hat{\mu}_1) = \frac{\sigma^2}{n} + 0^2$$

$$\hat{\mu}_2 = \mu_0$$

$$R_\theta(\hat{\mu}_2) = 0^2 + (\mu - \mu_0)^2$$

$$\hat{\mu}_3 = \frac{1}{2}\mu_0 + \frac{1}{2}\bar{X}_n$$

$$R_\theta(\hat{\mu}_3) = \frac{1}{4}\frac{\sigma^2}{n} + \frac{1}{4}(\mu - \mu_0)^2$$

$$\hat{\mu}_4 = \bar{X}_n + c$$

$$R_\theta(\hat{\mu}_4) = \frac{\sigma^2}{n} + c^2$$

$$\hat{\mu}_5 = \text{med}(X_1, \dots, X_n) \quad R_\theta(\hat{\mu}_5) \approx 1.57 \frac{\sigma^2}{n} + 0^2 \quad (n \rightarrow +\infty)$$

Exercise: Compute $R_\theta(\hat{\mu}_j)$, $2 \leq j \leq 4$

Remark: only the result for $\hat{\mu}_5$ actually uses the Gaussianity assumption.

19/55

Admissible estimators

Definition: order relation on the set of estimators

We will say that $\hat{\eta}'$ is (weakly) **preferable** to $\hat{\eta}$ if

- $\forall \theta \in \Theta, R_\theta(\hat{\eta}') \leq R_\theta(\hat{\eta})$,

We will say that it is **strictly preferable** to $\hat{\eta}$ if, in addition,

- $\exists \theta \in \Theta, R_\theta(\hat{\eta}') < R_\theta(\hat{\eta})$,

Remarks

- The relation "is preferable to" is a partial order on risk functions.
- In general there is no optimal estimator, i.e., no estimator that is preferable to all the others (unless we restrict the class of estimators that is considered)

Admissibility

We will say that $\hat{\eta}$ is **admissible** if there is no estimator $\hat{\eta}'$ that is strictly preferable to it.

20/55

Example 1 (cont'd)

$$\hat{\mu}_1 = \bar{X}_n \quad R_\theta(\hat{\mu}_1) = \frac{\sigma^2}{n} + 0^2$$

$$\hat{\mu}_2 = \mu_0 \quad R_\theta(\hat{\mu}_2) = 0^2 + (\mu - \mu_0)^2$$

$$\hat{\mu}_3 = \frac{1}{2}\mu_0 + \frac{1}{2}\bar{X}_n \quad R_\theta(\hat{\mu}_3) = \frac{1}{4}\frac{\sigma^2}{n} + \frac{1}{4}(\mu - \mu_0)^2$$

$$\hat{\mu}_4 = \bar{X}_n + c \quad R_\theta(\hat{\mu}_4) = \frac{\sigma^2}{n} + c^2$$

- ▶ $\hat{\mu}_1$ is strictly preferable to $\hat{\mu}_4$, therefore $\hat{\mu}_4$ is not admissible.
- ▶ $\hat{\mu}_1, \hat{\mu}_2$, et $\hat{\mu}_3$ are pairwise incomparable.
- ▶ It can be proved that all three are admissible.
Exercise: Prove that $\hat{\mu}_2$ is admissible.

21/55

Lecture outline

1 – Point estimation: definition and notations

2 – Quadratic risk of an estimator

3 – A lower bound on the quadratic risk

4 – Asymptotic properties

5 – Warming up exercises

Motivation

We will present in this section a lower bound of the form

$$\text{var}_\theta(\hat{\eta}) \geq v_{\min}(\theta), \quad \forall \theta \in \Theta,$$

that holds for (nearly) **all unbiased estimators** of $g(\theta)$.

Remark: for an UE, $R_\theta(\hat{\eta}) = \text{var}_\theta(\hat{\eta})$.

Usefulness of such a bound?

- ① Prove that a certain level of accuracy cannot be met by an unbiased estimator.
- ② Prove that a given UE is **optimal** (rare situation).
- ③ Prove that a given UE is **nearly optimal**.

22/55

Regularity condition C₁

Dominated model: there exists a (σ -finite) measure ν on $(\underline{\mathcal{X}}, \underline{\mathcal{A}})$ st

$$\forall A \in \underline{\mathcal{A}}, \quad \mathbb{P}_\theta(\underline{X} \in A) = \int_A f_\theta(\underline{x}) \nu(d\underline{x}).$$

Regularity condition C₁

The densities f_θ share a **common support**: $\exists \mathcal{S} \in \underline{\mathcal{A}}$,

$$\forall \theta \in \Theta, \quad f_\theta(\underline{x}) > 0 \Leftrightarrow \underline{x} \in \mathcal{S}.$$

Remarks:

- \mathcal{S} is only defined up to a ν -négligible set (as pdf's are).
- Strictly speaking, the “support” of the measure is the closure of \mathcal{S} .

23/55

Regularity condition C₁: examples / counter-example

Consider an IID univariate n -sample:

$$\underline{X} \sim f_{\theta}(\underline{x}) = \prod_{i=1}^n f_{\theta}(x_i)$$

(with a usual abuse of notation for the pdf's).

Remark: if C₁ holds for $n = 1$ with $\mathcal{S} = \mathcal{S}_1$,
then it also holds for all $n \geq 2$ with $\mathcal{S} = \mathcal{S}_1^n$.

A few examples...

- ① $\mathcal{N}(\mu, \sigma^2)$ with $\sigma^2 > 0$: C₁ holds with $\mathcal{S}_1 = \mathbb{R}$,
- ② $\mathcal{E}(\theta)$: C₁ holds with $\mathcal{S}_1 = [0, +\infty)$.
- ③ $\mathcal{U}_{[0, \theta]}$: C₁ does not hold!

24/55

Another regularity condition

We assume that C₁ holds.

Regularity condition C₂

- ❶ Θ is an open subset of \mathbb{R}^p ,
- ❷ $\theta \mapsto f_{\theta}(\underline{x})$ is differentiable for ν -almost all \underline{x} ,
- ❸ and, at any $\theta \in \Theta$, we have

$$\int_{\mathcal{S}} \nabla_{\theta} f_{\theta}(\underline{x}) \nu(d\underline{x}) = \nabla_{\theta} \int_{\mathcal{S}} f_{\theta}(\underline{x}) \nu(d\underline{x}) = 0.$$

In other words: $\forall \theta \in \Theta, \forall k \leq p$,

$$\int_{\mathcal{S}} \frac{\partial f_{\theta}(\underline{x})}{\partial \theta_k} \nu(d\underline{x}) = \frac{\partial}{\partial \theta_k} \int_{\mathcal{S}} f_{\theta}(\underline{x}) \nu(d\underline{x}) = 0.$$

25/55

Score

Definition / property: score

Assume that C₁, C₂-i and C₂-ii hold and define, for all $\underline{x} \in \mathcal{S}$

$$S_\theta(\underline{x}) = \nabla_\theta (\ln f_\theta(\underline{x})) = \begin{pmatrix} \frac{\partial \ln f_\theta(\underline{x})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ln f_\theta(\underline{x})}{\partial \theta_p} \end{pmatrix}.$$

Then

- ❶ We call **score** the random vector $S_\theta = S_\theta(\underline{X})$.
- ❷ C₂-iii $\Leftrightarrow \forall \theta \in \Theta$, the score S_θ is **centered** under \mathbb{P}_θ .

Remarks:

- ▶ Well defined, since $\underline{X} \in \mathcal{S}$ \mathbb{P}_θ -ps, $\forall \theta \in \Theta$.
- ▶ The score vanishes at the MLE (recall that $\Theta \subset \mathbb{R}^p$ is assumed open).

26/55

The score is centered (proof)

Notice that

$$\nabla_\theta (\ln f_\theta) = \frac{1}{f_\theta} \nabla_\theta f_\theta,$$

and thus, for all $\theta \in \Theta$,

$$\begin{aligned} \mathbb{E}_\theta (S_\theta) &= \int_{\mathcal{S}} S_\theta(\underline{x}) f_\theta(\underline{x}) \nu(d\underline{x}) \\ &= \int_{\mathcal{S}} \frac{1}{f_\theta(\underline{x})} \nabla_\theta f_\theta(\underline{x}) \mathbf{f}_\theta(\underline{x}) \nu(d\underline{x}) \\ &= \int_{\mathcal{S}} \nabla_\theta f_\theta(\underline{x}) \nu(d\underline{x}). \end{aligned}$$

Finally,

$$\mathbb{E}_\theta (S_\theta) = 0 \Leftrightarrow \int_{\mathcal{S}} \nabla_\theta f_\theta(\underline{x}) \nu(d\underline{x}) = 0 \quad (\text{C}_2\text{-iii}). \quad \square$$

27/55

Example 2

Recall that $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{E}(\theta)$ with $\theta \in \Theta =]0, +\infty[$.

We compute the **likelihood**, for any $x_1, \dots, x_n \geq 0$:

$$\mathcal{L}(\theta; \underline{x}) = f_\theta(\underline{x}) = \prod_{i=1}^n f_\theta(x_i) = \theta^n e^{-\theta \sum x_i},$$

then the **log-likelihood**:

$$\ln \mathcal{L}(\theta; \underline{x}) = \ln f_\theta(\underline{x}) = n \ln \theta - \theta \sum x_i,$$

and, finally, the **score**:

$$S_\theta(\underline{X}) = \sum_{i=1}^n S_\theta(X_i) = n \left(\frac{1}{\theta} - \bar{X}_n \right).$$

28/55

Remark on condition C₂-iii

Recall C₂-iii: $\forall \theta \in \Theta$,

$$\int_{\mathcal{S}} \nabla_\theta f_\theta(\underline{x}) \nu(d\underline{x}) = \nabla_\theta \int_{\mathcal{S}} f_\theta(\underline{x}) \nu(d\underline{x}) = 0,$$

or, equivalently: $\mathbb{E}_\theta(S_\theta) = 0$.

Two approaches are available to check this condition:

- ① Compute **explicitely** $\mathbb{E}_\theta(S_\theta) = \int_{\mathcal{S}} \nabla_\theta f_\theta(\underline{x}) \nu(d\underline{x})$.
- ② Use a **domination** condition: show that $\forall \theta_0 \in \Theta, \exists \mathcal{V} \subset \Theta$, neighbourhood of θ_0 , and a ν -integrable function $g : \underline{\mathcal{X}} \rightarrow \mathbb{R}$ st

$$\forall \theta \in \mathcal{V}, \forall \underline{x} \in \mathcal{S}, \forall k \leq p, \quad \left| \frac{\partial f_\theta(\underline{x})}{\partial \theta_k} \right| \leq g(\underline{x}).$$

29/55

Cramér-Rao inequality (scalar case)

Consider a statistical model where C_1 and C_2 hold, and $\forall \theta \in \Theta, \text{var}_\theta(S_\theta) > 0$.

Let $\hat{\eta}$ be an estimator of $\eta = g(\theta) \in \mathbb{R}$ st $\mathbb{E}_\theta(\hat{\eta}^2) < +\infty, \forall \theta \in \Theta$.

Definition: regular estimator

$\hat{\eta}$ is said to be **regular** if $\theta \mapsto \mathbb{E}_\theta(\hat{\eta})$ is differentiable, with

$$\nabla_\theta \mathbb{E}_\theta(\hat{\eta}) = \int_{\mathcal{S}} \hat{\eta}(\underline{x}) \nabla_\theta f_\theta(\underline{x}) \nu(d\underline{x}), \quad \forall \theta \in \Theta.$$

Theorem / definition: Cramér-Rao inequality

If $\hat{\eta}$ is **regular unbiased** estimator, then $\forall \theta \in \Theta$

$$R_\theta(\hat{\eta}) = \text{var}_\theta(\hat{\eta}) \geq \nabla g(\theta)^\top \text{var}_\theta(S_\theta)^{-1} \nabla g(\theta).$$

An unbiased estimator is called **efficient** if the bound is met for all θ .

30/55

Proof

Preliminary remark: since $\hat{\eta}$ is a regular UE of $g(\theta)$, g is differentiable.

Let $\theta \in \Theta$, and set $c = \text{cov}_\theta(S_\theta, \hat{\eta}) \in \mathbb{R}^P$. Then, $\forall a \in \mathbb{R}^P$,

$$\text{var}_\theta(\hat{\eta} - a^\top S_\theta) = \text{var}_\theta(\hat{\eta}) - 2a^\top c + a^\top \text{var}_\theta(S_\theta) a \geq 0.$$

In particular, for $a = \text{var}_\theta(S_\theta)^{-1} c \in \mathbb{R}^P$, we get:

$$\text{var}_\theta(\hat{\eta}) - c^\top \text{var}_\theta(S_\theta)^{-1} c \geq 0.$$

Finally, since S_θ is centered and $\hat{\eta}$ is a regular UE,

$$\begin{aligned} c &= \mathbb{E}_\theta(\hat{\eta} S_\theta) = \int_{\mathcal{S}} \hat{\eta}(\underline{x}) \cdot \frac{1}{f_\theta(\underline{x})} \nabla_\theta f_\theta(\underline{x}) \cdot f_\theta(\underline{x}) \nu(d\underline{x}) \\ &= \int_{\mathcal{S}} \hat{\eta}(\underline{x}) \nabla_\theta f_\theta(\underline{x}) \nu(d\underline{x}) = \nabla_\theta \mathbb{E}_\theta(\hat{\eta}) = \nabla g(\theta). \end{aligned} \quad \square$$

Fisher information (scalar case)

We still assume that C₁ and C₂ hold.

Definition: Fisher information

We call **Fisher information** of \underline{X} the $p \times p$ matrix

$$I_{\underline{X}}(\theta) = \text{var}_{\theta}(S_{\theta}(\underline{X})) = \mathbb{E}_{\theta} \left(S_{\theta}(\underline{X}) S_{\theta}(\underline{X})^{\top} \right)$$

which appears in the Cramér-Rao lower bound.

Proposition

Let $I_n(\theta)$ denote the Fisher information in an IID n -sample. Then

$$I_n(\theta) = n I_1(\theta).$$

The CR inequality becomes: $\text{var}_{\theta}(\hat{\eta}) \geq \frac{1}{n} \nabla g(\theta)^{\top} I_1(\theta)^{-1} \nabla g(\theta)$.

31/55

Proof

Notice that the score is additive in an IID sample:

$$S_{\theta}(\underline{X}) = \sum_{i=1}^n S_{\theta}(X_i)$$

and thus

$$\text{var}_{\theta}(S_{\theta}(\underline{X})) = \sum_{i=1}^n \text{var}_{\theta}(S_{\theta}(X_i)) = n \text{var}_{\theta}(S_{\theta}(X_1))$$

since $S_{\theta}(X_1), \dots, S_{\theta}(X_n)$ are IID. □

32/55

Example 1: estimation of μ

Reminder: $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ and $\theta = (\mu, \sigma^2)$

- ▶ $\hat{\mu}_n = \bar{X}_n$ is the MLE of μ ,
- ▶ $\hat{\mu}_n$ is unbiased and $R_\theta(\hat{\mu}_n) = \text{var}_\theta(\hat{\mu}_n) = \frac{\sigma^2}{n}$.

Exercise: the **Fisher information matrix** in this model is

$$I_n(\theta) = n \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}.$$

Cramér-Rao inequality with $g(\theta) = \mu$: $\forall \hat{\mu}'_n$ UE of μ ,

$$R_\theta(\hat{\mu}'_n) = \text{var}_\theta(\hat{\mu}'_n) \geq \frac{\sigma^2}{n},$$

therefore $\hat{\mu}_n = \bar{X}_n$ is efficient.

33/55

Example 1': estimation of σ^2

Same statistical model, but we want to estimate $g(\theta) = \sigma^2$.

Exercise: show that

- ▶ the MLE $S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ is biased;
- ▶ $\hat{\sigma}_n^2 = (S'_n)^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ is an UE of σ^2 .

It is then possible to show (see TD 6) that

$$\text{var}_\theta(\hat{\sigma}_n^2) = \frac{2\sigma^4}{n-1},$$

therefore $\hat{\sigma}_n^2$ is not an efficient estimator, since

$$\text{var}_\theta(\hat{\sigma}_n^2) > \frac{2\sigma^4}{n}.$$

(Beware the misleading terminology: it can be proved, using Lehmann-Scheffé's theorem, that $\hat{\sigma}_n^2$ is a minimal variance UE for this problem, and therefore is optimal for the quadratic risk among all UE's.)

34/55

Exercise solution

Let us show that the sample variance S_n^2 is biased:

$$\begin{aligned}\mathbb{E}_\theta(S_n^2) &= \mathbb{E}_\theta \left(\frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2 \right) = \mathbb{E}_\theta(X_1^2) - \mathbb{E}_\theta(\bar{X}_n^2) \\ &= (\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n} + \mu^2 \right) = \frac{n-1}{n} \sigma^2 \neq \sigma^2.\end{aligned}$$

We conclude that the “corrected” sample variance is unbiased:

$$\mathbb{E}_\theta((S'_n)^2) = \frac{n}{n-1} \mathbb{E}_\theta(S_n^2) = \frac{n}{n-1} \cdot \frac{n-1}{n} \sigma^2 = \sigma^2. \quad \square$$

Lecture outline

1 – Point estimation: definition and notations

2 – Quadratic risk of an estimator

3 – A lower bound on the quadratic risk

4 – Asymptotic properties

5 – Warming up exercises

Motivation / notations

Problem

It is sometimes (often !) difficult to obtain the exact properties of statistical procedures.

(point estimators, but also CIs, tests, etc. (cf. next lectures))

Asymptotic approach(es) → approximate properties

- ▶ $X_1, X_2, \dots \stackrel{\text{iid}}{\sim} P_\theta$, defined on a common $(\Omega, \mathcal{F}, \mathbb{P}_\theta)$
- ▶ Sequences of estimators: $\hat{\eta}_n = \hat{\eta}_n(X_1, \dots, X_n)$
- ▶ Properties of the estimators when $n \rightarrow \infty$?

Remark: we have now not one but a **sequence** $(\mathcal{M}_n)_{n \geq 1}$ of statistical models

$$\mathcal{M}_n = (\mathcal{X}^n, \mathcal{A}^{\otimes n}, \{P_\theta^{\otimes n}, \theta \in \Theta\}),$$

that we instantiate on a common underlying probability space (Ω, \mathcal{F}) .

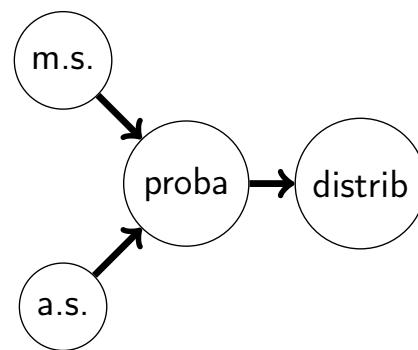
35/55

Probability refresher: convergence modes

Main convergence modes that are useful in Statistics:

- ▶ **almost sure** convergence ,
- ▶ convergence **in L^2** (in mean square),
- ▶ convergence **in probability**,
- ▶ convergence **in distribution**.

Implications between convergence modes:



36/55

Probability refresher: convergence modes

☞ **almost sure** convergence :

$$T_n \xrightarrow{\text{ps}} T \quad \text{if} \quad \mathbb{P}(T_n \rightarrow T) = 1$$

☞ **convergence in L^2** (in mean square):

$$\begin{aligned} T_n &\xrightarrow{L^2} T \quad \text{if} \quad \mathbb{E}(\|T_n - T\|^2) \rightarrow 0 \\ &\text{iff} \quad \forall j \leq p, \quad T_n^{(j)} \xrightarrow{L^2} T^{(j)} \end{aligned}$$

☞ **convergence in probability**:

$$T_n \xrightarrow{P} T \quad \text{if} \quad \forall \varepsilon > 0, \quad \mathbb{P}(\|T_n - T\| \geq \varepsilon) \rightarrow 0$$

☞ **convergence in distribution**:

$$T_n \xrightarrow{\text{loi}} T \quad \text{if} \quad \forall \varphi, \quad \mathbb{E}(\varphi(T_n)) \rightarrow \mathbb{E}(\varphi(T)),$$

with $\varphi : \mathbb{R}^p \rightarrow \mathbb{R}$ continuous and bounded.

Consistency

Let $(\hat{\eta}_n)$ denote a sequence of estimators of $\eta = g(\theta)$.

(weak) Consistency

We will say that $\hat{\eta}_n$ is a **consistent** estimator of $\eta = g(\theta)$ if, $\forall \theta \in \Theta$,

$$\hat{\eta}_n \xrightarrow[n \rightarrow \infty]{\mathbb{P}_\theta} g(\theta). \quad (\text{with an obvious abuse of terminology})$$

Strong and mean-square consistency

We will say that $\hat{\eta}_n$ is **strongly consistent**
(resp. **consistent in the mean-square sense**) if, $\forall \theta \in \Theta$,

$$\hat{\eta}_n \xrightarrow[n \rightarrow \infty]{\mathbb{P}_\theta-\text{a.s.}} g(\theta) \quad \left(\text{resp., } \hat{\eta}_n \xrightarrow[n \rightarrow \infty]{L^2(\mathbb{P}_\theta)} g(\theta) \right).$$

Remark: the word “convergent” is sometimes used instead of “consistent”.

Probability refresher: law of large numbers

Let $(X_k)_{k \geq 1}$ be a sequence of real- or vector-valued RV.

Strong law of large numbers

If the X_k 's are IID and $\mathbb{E}(\|X_1\|) < +\infty$, then

$$\bar{X}_n \xrightarrow[n \rightarrow \infty]{\text{a.s.}} \mathbb{E}(X_1).$$

Law of large numbers in L^2

If the X_k 's are IID and $\mathbb{E}(\|X_1\|^2) < +\infty$, then

$$\bar{X}_n \xrightarrow[n \rightarrow \infty]{L^2} \mathbb{E}(X_1).$$

Proof (scalar case): $\mathbb{E}((\bar{X}_n - \mathbb{E}(X_1))^2) = \text{var}_{\theta}(\bar{X}_n) = \frac{1}{n} \text{var}_{\theta}(X_1) \rightarrow 0$.

□

38/55

Consistency: examples

A) IID n -sample with finite first order moment

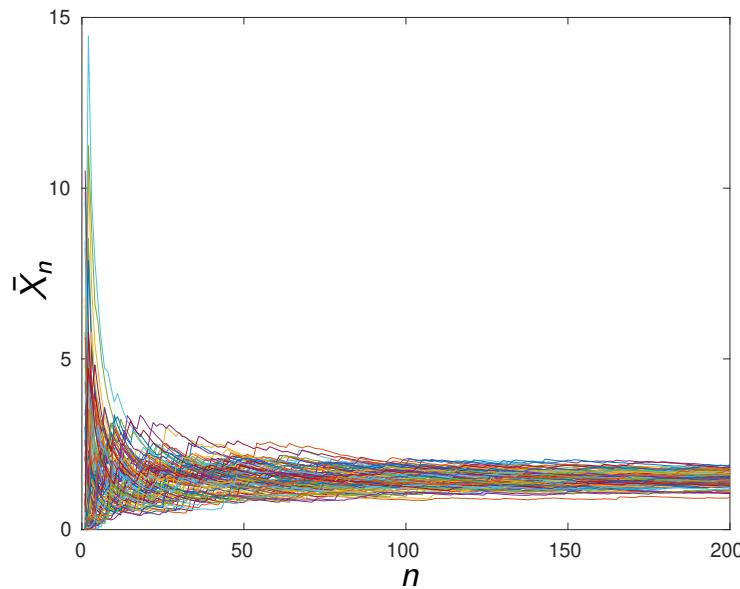
- ▶ i.e., $\mathbb{E}_{\theta}(\|X_1\|) < +\infty$, for all $\theta \in \Theta$.
- ▶ \bar{X}_n is a **strongly consistent** estimator of $\eta = \mathbb{E}_{\theta}(X_1)$.
- ▶ Nothing can be said about the quadratic risk without additional assumptions.

B) IID n -sample with finite second order moment

- ▶ i.e., $\mathbb{E}_{\theta}(\|X_1\|^2) < +\infty$, for all $\theta \in \Theta$.
- ▶ \bar{X}_n is **strongly consistent** and **consistent in the mean-square sense** for $\eta = \mathbb{E}_{\theta}(X_1)$.

39/55

Consistency: examples (cont'd)



Convergence of \bar{X}_n to the true mean
(for a Gamma n -sample with true mean $\mu = 1.5$)

40/55

Consistency: examples (cont'd)

C) IID n -sample (with any distribution)

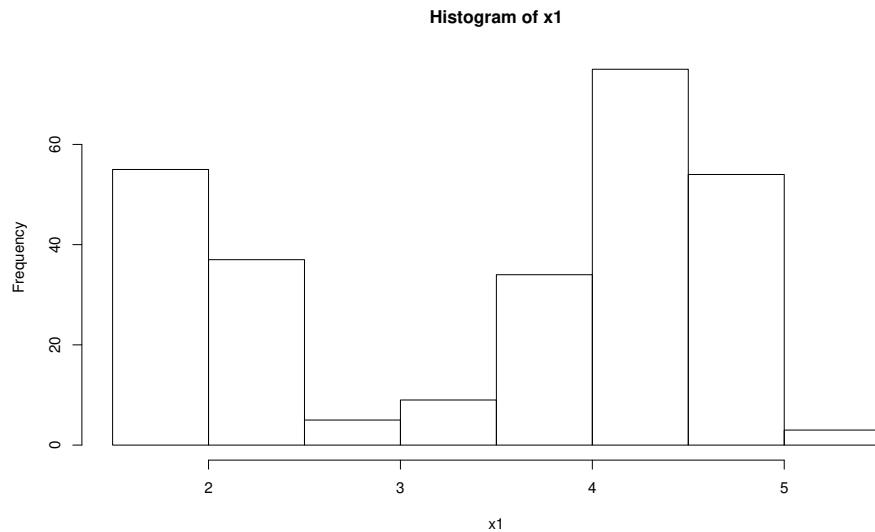
- ▶ Let $A \in \mathcal{A}$ and $\eta = g(\theta) = \mathbb{P}_\theta(X_1 \in A)$.
- ▶ Relative frequency: $\hat{\eta}_n = \frac{1}{n} \text{ card } \{i \leq n \mid X_i \in A\}$
- ▶ $\hat{\eta}_n$ is a **strongly** and **mean-square consistent** estimator of η .

Application: histograms

- ▶ Let $\mathcal{X} = \cup_{k=1}^K A_k$ denote a partition of \mathcal{X}
- ▶ vector-valued $\hat{\eta}_n$: $\hat{\eta}_n^{(k)} = \frac{1}{n} \text{ card } \{i \leq n \mid X_i \in A_k\}$
- ▶ $\hat{\eta}_n$ is a **strongly** and **mean-square consistent** estimator of $\eta = (\mathbb{P}_\theta(X_1 \in A_k))_{1 \leq k \leq K}$.

41/55

Consistency: examples (cont'd)



Example of a (un-normalized) histogram

42/55

Consistency: examples (cont'd)

D) Maximum of a uniform IID n -sample

- ▶ $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{U}_{[0, \theta]}$
- ▶ We estimate $\eta = \theta$ with $\hat{\eta}_n = \max_{i \leq n} X_i$.
- ▶ Exercise (TD 1): show that $\hat{\eta}_n$ is consistent, both strongly and in the mean-square sense.

E) Maximum likelihood estimator

- ▶ see below

43/55

Asymptotically unbiased estimator

Recall that $b_\theta(\hat{\eta}) = \mathbb{E}_\theta(\hat{\eta}) - g(\theta)$.

Definition: asymptotically unbiased

We will say that an estimator $\hat{\eta}_n$ is **asymptotically unbiased** if

$$b_\theta(\hat{\eta}) \xrightarrow{n \rightarrow +\infty} 0, \quad \forall \theta \in \Theta.$$

Proposition

$\hat{\eta}_n$ is **consistent in the mean-square sense** if, and only if, the two following conditions met:

- i $\hat{\eta}_n$ is **asymptotically unbiased**,
- ii $\text{var}_\theta(\hat{\eta}_n) \rightarrow 0$, for all $\theta \in \Theta$. ($\text{tr}(\text{var}_\theta(\hat{\eta})) \rightarrow 0$ in the vector case)

Proof: Use the bias-variance decomposition!

□

44/55

Asymptotically unbiased estimator: example

$X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{U}_{[0, \theta]}$, and we want to estimate θ .

Let us prove that $\hat{\theta}_n = \max_{i \leq n} X_i$ is **asymptotically unbiased**.

Method 1: direct computation

- Compute the expectation: $\mathbb{E}_\theta(\hat{\theta}_n) = \frac{n}{n+1} \theta$ (cf. TD),
- hence the bias: $b_\theta(\hat{\theta}) = -\frac{\theta}{n+1} \rightarrow 0$.

Method 2: dominated convergence theorem

- We already know that $\hat{\theta}_n$ is **strongly consistent**;
- besides $|\hat{\theta}_n| \leq \theta$, \mathbb{P}_θ – a.s.;
- therefore $\mathbb{E}_\theta(\hat{\theta}_n) \rightarrow \theta$ by the dominated convergence theorem.

45/55

Consistency of the MLE

The MLE minimizes the following criterion:

$$\gamma_n(\theta) = -\frac{1}{n} \ln f_\theta(\underline{X}) = -\frac{1}{n} \sum_{k=1}^n \ln f_\theta(X_i).$$

Let $\theta \in \Theta$, and set $c = \text{cov}_\theta(S_\theta, \hat{\eta}) \in \mathbb{R}^P$. Then, $\forall \theta \in \Theta$,

$$\gamma_n(\theta) - \gamma_n(\theta_*) = \frac{1}{n} \sum_{k=1}^n \ln \frac{f_{\theta_*}(X_i)}{f_\theta(X_i)} \xrightarrow[n \rightarrow +\infty]{\text{ps}} \int_{S_1} \ln \frac{f_{\theta_*}(x)}{f_\theta(x)} f_{\theta_*}(x) \nu_1(dx).$$

(assuming that $Z_i = \frac{f_{\theta_*}(X_i)}{f_\theta(X_i)}$ has a finite first order moment).

Definition / property: Kullback-Leibler divergence

$$D_{\text{KL}}(f_{\theta_*} || f_\theta) = \int_{S_1} \ln \frac{f_{\theta_*}(x)}{f_\theta(x)} f_{\theta_*}(x) \nu_1(dx) \geq 0$$

Consistency of the MLE (cont'd)

Set $\Delta_n(\theta_*, \theta) = \frac{1}{n} \sum_{k=1}^n \ln \frac{f_{\theta_*}(X_i)}{f_\theta(X_i)}$ and $\Delta(\theta_*, \theta) = D_{\text{KL}}(f_{\theta_*} || f_\theta)$.

We have $\Delta_n(\theta_*, \theta) \xrightarrow[n \rightarrow +\infty]{\mathbb{P}_{\theta_*} - \text{ps}} \Delta(\theta_*, \theta)$ for all θ , and $\Delta(\theta_*, \theta_*) = 0$.

Theorem: Consistency of the MLE

Assume that, for all $\theta_* \in \Theta$,

i) $\sup_{\theta \in \Theta} |\Delta_n(\theta_*, \theta) - \Delta(\theta_*, \theta)| \xrightarrow[n \rightarrow +\infty]{\mathbb{P}_{\theta_*}} 0$

ii) and, for all $\epsilon > 0$,

$$\inf_{\theta \in \Theta, \|\theta - \theta_*\| \geq \epsilon} \Delta(\theta_*, \theta) > 0.$$

Then the MLE is (weakly) consistent.

Lecture outline

1 – Point estimation: definition and notations

2 – Quadratic risk of an estimator

3 – A lower bound on the quadratic risk

4 – Asymptotic properties

5 – Warming up exercises

Exercise 1 (quadratic risk)

Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ with $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+^+$.

We want to estimate $g(\theta) = \mu$. We consider the estimators

$$\hat{\mu}_1 = \bar{X}_n, \quad \hat{\mu}_2 = \mu_0, \quad \hat{\mu}_3 = \frac{1}{2}\mu_0 + \frac{1}{2}\bar{X}_n, \quad \hat{\mu}_4 = \bar{X}_n + c,$$

where μ_0 and c are given real numbers.

Questions

- ① Prove the bias-variance decomposition formula in the scalar case (see slide 18)
- ② Compute the quadratic risk of each of these estimators
- ③ Prove that $\hat{\mu}_2$ and $\hat{\mu}_3$ are not comparable.
- ④ Prove that $\hat{\mu}_4$ is not admissible.

Exercise 2 (efficiency of an estimator)

Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{B}(\theta)$ with $\theta \in \Theta =]0, 1[$.

Recall that (see Exercises in Lecture 1):

- ▶ the log-likelihood of the n -sample is

$$\ln \mathcal{L}(\theta; \underline{x}) = \ln f_\theta(\underline{x}) = n \ln(1 - \theta) - \ln \left(\frac{\theta}{1 - \theta} \right) \sum_{i=1}^n x_i,$$

- ▶ the MLE is $\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n X_i$.

Questions

- ① Check that the model satisfies the hypotheses for Cramér-Rao's inequality, and compute Cramér-Rao's bound.
- ② Is the MLE $\hat{\theta}_n$ efficient?

47/55

Solution of Exercise 1

- ① Bias-variance decomposition

$$R_\theta(\hat{\eta}) = \mathbb{E}_\theta((\hat{\eta} - g(\theta))^2) = \text{var}_\theta(\hat{\eta}) + b_\theta(\hat{\eta})^2.$$

Proof

$$\begin{aligned} R_\theta(\hat{\eta}) &= \mathbb{E}_\theta((\hat{\eta} - g(\theta))^2) \\ &= \mathbb{E}_\theta((\hat{\eta} - \mathbb{E}_\theta(\hat{\eta}) + b_\theta(\hat{\eta}))^2) \\ &= \underbrace{\mathbb{E}_\theta((\hat{\eta} - \mathbb{E}_\theta(\hat{\eta}))^2)}_{\text{var}_\theta(\hat{\eta})} + b_\theta(\hat{\eta})^2 + 2 \underbrace{\mathbb{E}_\theta(\hat{\eta} - \mathbb{E}_\theta(\hat{\eta}))}_{=0} b_\theta(\hat{\eta}) \\ &= \text{var}_\theta(\hat{\eta}) + b_\theta(\hat{\eta})^2. \end{aligned}$$

□

48/55

Solution of Exercise 1 (cont'd)

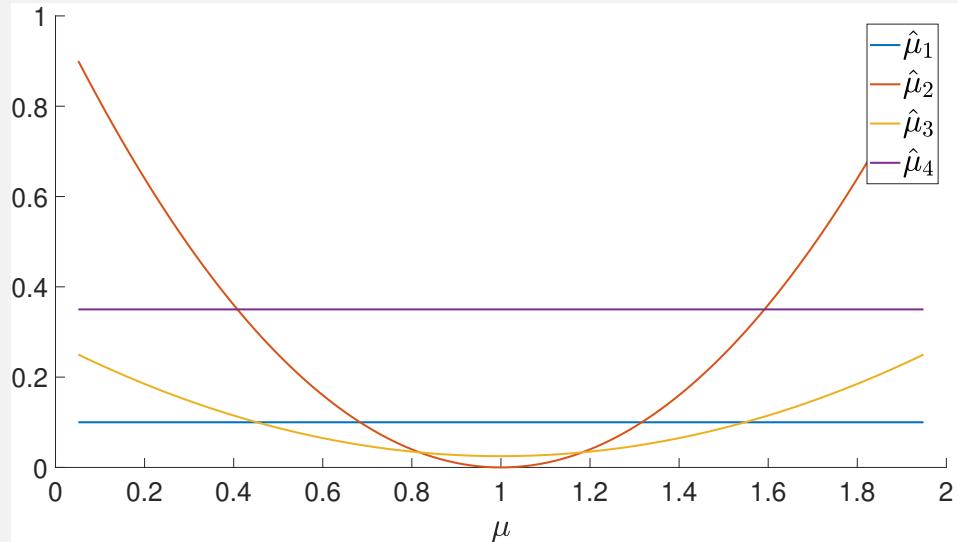
② Compute the bias and variance of each estimator, and then conclude using the bias-variance decomposition.

	expectation	bias	variance	quadratic risk
\bar{X}_n	μ	0	$\frac{\sigma^2}{n}$	$\frac{\sigma^2}{n}$
μ_0	μ_0	$\mu_0 - \mu$	0	$(\mu_0 - \mu)^2$
$\frac{1}{2} (\mu_0 + \bar{X}_n)$	$\frac{1}{2} (\mu_0 + \mu)$	$\frac{1}{2} (\mu_0 - \mu)$	$\frac{1}{4} \frac{\sigma^2}{n}$	$\frac{1}{4} \frac{\sigma^2}{n} + \frac{1}{4} (\mu_0 - \mu)^2$
$\bar{X}_n + c$	$\mu + c$	c	$\frac{\sigma^2}{n}$	$\frac{\sigma^2}{n} + c^2$

Reminder: $\text{var}_\theta(\alpha X + \beta) = \alpha^2 \text{var}_\theta(X)$.

49/55

Solution of Exercise 1 (cont'd)



Draw the four risks for $\sigma^2 = 1$, $n = 10$, $\mu_0 = 1$ and $c = 0.5$.

50/55

Solution of Exercise 1 (cont'd)

③ Let us compute the risk two well-chosen points.

For $\theta = (\mu_0, 1)$ we have

$$R_\theta(\hat{\mu}_2) = 0, \quad R_\theta(\hat{\mu}_3) = \frac{1}{4n}, \quad \text{therefore } R_\theta(\hat{\mu}_2) < R_\theta(\hat{\mu}_3).$$

For $\theta = \left(\mu_0 + \frac{1}{\sqrt{n}}, 1\right)$ we have

$$R_\theta(\hat{\mu}_2) = \frac{1}{n}, \quad R_\theta(\hat{\mu}_3) = \frac{1}{2n}, \quad \text{therefore } R_\theta(\hat{\mu}_2) > R_\theta(\hat{\mu}_3).$$

Therefore the estimators $\hat{\mu}_2$ and $\hat{\mu}_3$ are not comparable.

51/55

Solution of Exercise 1 (cont'd)

④ We have:

$$\begin{cases} R_\theta(\hat{\mu}_4) = \frac{\sigma^2}{n} + c^2 \\ R_\theta(\hat{\mu}_1) = \frac{\sigma^2}{n} \end{cases}$$

Therefore, $\forall \theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_*^+$, $R_\theta(\hat{\mu}_4) > R_\theta(\hat{\mu}_1)$

Thus $\hat{\mu}_4$ is not admissible.

52/55

Solution of Exercise 2

① Let us check that the model satisfies the regularity conditions C₁ and C₂, and that Fisher's information does not vanish.

⇒ C₁: since $\Theta =]0, 1[$, the densities

$$f_\theta(\underline{x}) = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i}$$

are all supported on $\mathcal{S} = \{0, 1\}^n$.

⇒ C₂: $\Theta =]0, 1[$ is an open subset of \mathbb{R} , $\theta \mapsto f_\theta(\underline{x})$ is differentiable on Θ for all \underline{x} , and the score

$$S_\theta(\underline{X}) = \frac{\partial(\ln f_\theta)}{\partial \theta}(X_i) = \frac{n}{\theta(1-\theta)} (\bar{X}_n - \theta)$$

53/55

Solution of Exercise 2 (cont'd)

is centered: $\mathbb{E}_\theta(S_\theta(\underline{X})) = \frac{n}{\theta(1-\theta)} (\mathbb{E}_\theta(\bar{X}_n) - \theta) = 0$.

⇒ Finally, we check that the Fisher information does not vanish:

$$I(\theta) = \text{var}_\theta(S_\theta(\underline{X})) = \left(\frac{n}{\theta(1-\theta)} \right)^2 \text{var}_\theta(\bar{X}_n) = \frac{n}{\theta(1-\theta)} > 0.$$

⇒ The Cramér-Rao bound for θ is

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

54/55

Solution of Exercise 2 (cont'd)

② The estimator $\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n X_i$ is unbiased:

$$\mathbb{E}_{\theta}(\hat{\theta}_n) = \mathbb{E}_{\theta}(X_1) = \theta,$$

and its variance is

$$\text{var}(\hat{\theta}) = \frac{1}{n} \text{var}(X_1) = \frac{\theta(1-\theta)}{n} = I(\theta)^{-1}.$$

Therefore it is efficient. □

Remark: it is easy to check that $\hat{\theta}_n$ is a regular estimator (see definition on slide 30), since

- a the density f_{θ} is differentiable with respect to θ ,
- b the integrals boil down to finite sums over $\{0, 1\}^n$.

Chapter 3

Asymptotic distributions Confidence intervals



CentraleSupélec

Statistics and Learning

Arthur Tenenhaus[†], Julien Bect & Laurent Le Brusquet

(firstname.lastname@centralesupelec.fr)

Teaching: CentraleSupélec / Department of Mathematics

Research: Laboratory of signals and systems (L2S)

[†]: Course coordinator

1/48

Lecture 3/10

Asymptotic distributions and confidence intervals

In this lecture you will learn how to...

- ▶ Take the asymptotic approach one step further, introducing **asymptotic distributions**.
- ▶ Learn what **confidence intervals** are and show how to construct them (using, again, asymptotic arguments if needed)

2/48

Lecture outline

1 – Convergence rate and asymptotic distribution

- 1.1 – Definitions and examples
- 1.2 – Theoretical tools
- 1.3 – Asymptotic efficiency

2 – Confidence regions and confidence intervals

- 2.1 – Definition and example
- 2.2 – Exact confidence intervals
- 2.3 – Asymptotic confidence intervals

3 – Warming up exercises

3/48

Mathematical framework

In this section:

- we consider a **statistical model**

$$\left(\mathcal{X}, \mathcal{A}, \left\{ \mathbb{P}_\theta^X, \theta \in \Theta \right\} \right),$$

assumed (most of the time) to be **parametric** ($\Theta \subset \mathbb{R}^\rho$);

- $X_1, X_2, \dots \stackrel{\text{iid}}{\sim} P_\theta$, defined on a common $(\Omega, \mathcal{F}, \mathbb{P}_\theta)$
- we want to estimate a “**quantity of interest**”:
 - either θ itself (we assume in this case that $\Theta \subset \mathbb{R}^\rho$),
 - or, more generally, $\eta = g(\theta) \in \mathbb{R}^q$.

4/48

Lecture outline

1 – Convergence rate and asymptotic distribution

- 1.1 – Definitions and examples
- 1.2 – Theoretical tools
- 1.3 – Asymptotic efficiency

2 – Confidence regions and confidence intervals

- 2.1 – Definition and example
- 2.2 – Exact confidence intervals
- 2.3 – Asymptotic confidence intervals

3 – Warming up exercises

Lecture outline

1 – Convergence rate and asymptotic distribution

- 1.1 – Definitions and examples
- 1.2 – Theoretical tools
- 1.3 – Asymptotic efficiency

2 – Confidence regions and confidence intervals

- 2.1 – Definition and example
- 2.2 – Exact confidence intervals
- 2.3 – Asymptotic confidence intervals

3 – Warming up exercises

Convergence rate

Let $\hat{\eta}_n = \hat{\eta}_n(X_1, \dots, X_n)$ be a consistent estimator of $\eta = g(\theta)$.

Definition

If there exists a sequence $(a_n)_{n \in \mathbb{N}^*}$ of positive numbers such that:

- ▶ $\lim_{n \rightarrow \infty} a_n = \infty$,
- ▶ $a_n (\hat{\eta}_n - \eta) \xrightarrow[n \rightarrow \infty]{d} Z$,
- ▶ where Z is a **non-degenerate*** random variable (or vector),
then $\hat{\eta}_n$ converges to η at the rate $\frac{1}{a_n}$.

* We say that Z is **degenerate** if:

- ▶ scalar case: $\exists c \in \mathbb{R}, Z = c$ a.s.;
- ▶ vector case: $\exists a \in \mathbb{R}^q \setminus \{0\}, \exists c \in \mathbb{R}, \sum_{j=1}^q a_j Z^{(j)} = c$ a.s.;

Exercise. Let Z be a random vector with finite second order moments.

⇒ Prove that Z is non-degenerate iff its covariance matrix is invertible.

5/48

Asymptotic normality

Let $\hat{\eta}_n = \hat{\eta}_n(X_1, \dots, X_n)$ be a **consistent** estimator of $\eta = g(\theta)$.

Definition

If there exists

- ▶ a sequence $(a_n)_{n \in \mathbb{N}^*}$ of positive numbers s.t. $\lim_{n \rightarrow \infty} a_n = \infty$,
- ▶ a symmetric positive-definite matrix $\Sigma(\theta)$,

such that

$$a_n (\hat{\eta}_n - \eta) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \Sigma(\theta)), \quad (1)$$

then we say that $\hat{\eta}_n$ is **asymptotically normal**.

Vocabulary. $\Sigma(\theta)$ is called the **asymptotic covariance matrix** (asymptotic variance, in the scalar case).

Note: it can be proved that (1) with $a_n \rightarrow +\infty$ implies consistency.

6/48

Relation between convergence in distribution and in probability.

We already know that convergence in probability implies convergence in distribution. Let $(Y_n)_{n \in \mathbb{N}^*}$ be a sequence of RV with values in \mathbb{R}^d .

Proposition

If $Y_n \xrightarrow{d} c$, with $c \in \mathbb{R}^d$ a constant, then $Y_n \xrightarrow{\mathbb{P}} c$.

Corollary

If there exists $c \in \mathbb{R}^d$,

- ▶ a RV Z with values in \mathbb{R}^d ,
- ▶ a sequence $(a_n)_{n \in \mathbb{N}^*}$ of real numbers such that $\lim_{n \rightarrow \infty} a_n = \infty$,

such that

$$a_n (Y_n - c) \xrightarrow[n \rightarrow \infty]{d} Z$$

then

$$Y_n \xrightarrow[n \rightarrow \infty]{\mathbb{P}} c.$$

Proof (exercise): use above proposition and Slutsky's theorem (see below). □

Probability refresher: the Central Limit Theorem (CLT)

Theorem

Let

- ▶ a sequence $(X_n)_{n \in \mathbb{N}^*}$ of IID RV taking values in \mathbb{R}^d , with finite second order moments.
- ▶ $\mu = \mathbb{E}(X_1)$ and $\Sigma = \text{var}(X_1) \in \mathbb{R}^{d \times d}$.

Then : $\sqrt{n} (\bar{X}_n - \mu) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \Sigma),$

with $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ the sample mean.

⇒ The sample mean \bar{X}_n

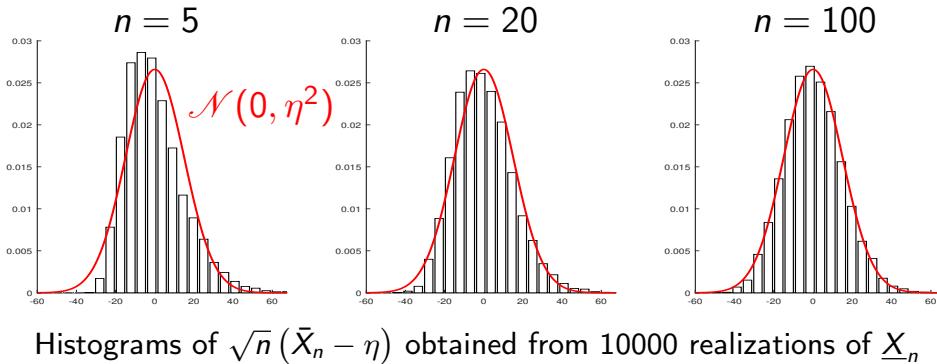
- ▶ is an **asymptotically Gaussian** estimator of $\mu = \mathbb{E}(X_1)$
- ▶ with **convergence rate** $\frac{1}{\sqrt{n}}$.

Example: component reliability

Recall that

- ▶ $X_i \stackrel{\text{iid}}{\sim} \mathcal{E}(\theta)$, $\theta > 0$, and $\eta = \mathbb{E}_\theta(X_1) = \frac{1}{\theta}$.
- ▶ $\hat{\eta}_n = \bar{X}_n$ is obtained by ML and the method of moments.

⇒ Direct application of the CLT: $\sqrt{n}(\bar{X}_n - \eta) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \eta^2)$.



8/48

Another example: indicator function

Let $(X_n)_{n \geq 1}$ be a sequence of IID RV with values in $(\mathcal{X}, \mathcal{A})$.

For a given $A \in \mathcal{A}$, we estimate $\eta = \mathbb{P}(X_1 \in A)$ by

$$\hat{\eta}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{X_i \in A}.$$

Direct application of the CLT: $\Rightarrow Y_i = \mathbf{1}_{X_i \in A} \stackrel{\text{iid}}{\sim} \text{Ber}(\eta)$

$$\sqrt{n}(\hat{\eta}_n - \eta) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \eta(1 - \eta)).$$

Concl.: if $0 < \eta < 1$, then $\hat{\eta}_n$ is **asymptotically Gaussian**, with

- ▶ convergence rate: $\frac{1}{\sqrt{n}}$,
- ▶ asymptotic variance: $\eta(1 - \eta)$.

9/48

Lecture outline

1 – Convergence rate and asymptotic distribution

1.1 – Definitions and examples

1.2 – Theoretical tools

1.3 – Asymptotic efficiency

2 – Confidence regions and confidence intervals

2.1 – Definition and example

2.2 – Exact confidence intervals

2.3 – Asymptotic confidence intervals

3 – Warming up exercises

The continuous mapping theorem

Theorem (Mann-Wald)

Let

- ▶ $h : \mathbb{R}^d \rightarrow \mathbb{R}^q$ a measurable function
- ▶ Y a random vector, taking values in \mathbb{R}^d ,

such that

h is continuous at the point Y , almost surely.

Then, for any sequence $(Y_n)_{n \in \mathbb{N}^*}$ of RV with values in \mathbb{R}^d ,

- (i) $Y_n \xrightarrow{\text{as}} Y \Rightarrow h(Y_n) \xrightarrow{\text{as}} h(Y),$
- (ii) $Y_n \xrightarrow{\mathbb{P}} Y \Rightarrow h(Y_n) \xrightarrow{\mathbb{P}} h(Y),$
- (iii) $Y_n \xrightarrow{d} Y \Rightarrow h(Y_n) \xrightarrow{d} h(Y).$

Proof: see CIP for the case where h is continuous. General case: admit.

10/48

Example: component reliability (cont'd)

Recall that

- $X_i \stackrel{\text{iid}}{\sim} \mathcal{E}(\theta)$, $\theta > 0$, and $\eta = \mathbb{E}_\theta(X_1) = \frac{1}{\theta}$.
- $\hat{\eta}_n = \bar{X}_n$ is obtained by ML and the method of moments.

Law of large numbers (strong and in L^2):

$$\hat{\eta}_n = \bar{X}_n \xrightarrow{\text{as, } L^2} \eta.$$

By the continuous mapping theorem:

$$\hat{\theta}_n = \frac{1}{\hat{\eta}_n} \xrightarrow{\text{as}} \frac{1}{\eta} = \theta,$$

therefore $\hat{\theta}_n$ is **strongly consistent**.

Exercise: prove that $\hat{\theta}_n$ is also consistent the L^2 sense.

11/48

Slutsky's theorem

Theorem

Let

- $(X_n)_{n \in \mathbb{N}^*}$ a sequence of random vectors that converges in distribution to a RV X :

$$X_n \xrightarrow[n \rightarrow \infty]{d} X,$$

- $(Y_n)_{n \in \mathbb{N}^*}$ a sequence of random vectors that converges in distribution (therefore in probability) to a **constant** c :

$$Y_n \xrightarrow[n \rightarrow \infty]{d} c,$$

Then

$$(X_n, Y_n) \xrightarrow[n \rightarrow \infty]{d} (X, c).$$

Remark: $Y_n \xrightarrow[n \rightarrow \infty]{d} c$ implies $Y_n \xrightarrow[n \rightarrow \infty]{\mathbb{P}} c$ (constant limit).

12/48

Example: component reliability (cont'd)

Recall that (CLT) $\sqrt{n}(\bar{X}_n - \eta) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \eta^2)$.

Since $\bar{X}_n \xrightarrow[n \rightarrow \infty]{\text{as}} \eta$ (constant), we have by Slutsky's theorem:

$$(\sqrt{n}(\bar{X}_n - \eta), \bar{X}_n) \xrightarrow[n \rightarrow \infty]{d} (Z, \eta) \quad \text{with } Z \sim \mathcal{N}(0, \eta^2).$$

Therefore, by the continuous mapping theorem,

$$\sqrt{n} \frac{(\bar{X}_n - \eta)}{\bar{X}_n} \xrightarrow[n \rightarrow \infty]{d} \frac{Z}{\eta} \sim \mathcal{N}(0, 1),$$

since $(z, y) \mapsto \frac{z}{y}$ is continuous at any point where $y \neq 0$.

13/48

Linearization method ("delta method")

Theorem ("delta theorem")

Let $(Y_n)_{n \in \mathbb{N}^*}$ be a sequence of RV with values in \mathbb{R}^d , s.t.

$$\sqrt{n}(Y_n - m) \xrightarrow[n \rightarrow \infty]{d} Z,$$

Y a random vector, taking values in \mathbb{R}^d and $m \in \mathbb{R}^d$.

Then, for any $h : \mathbb{R}^d \rightarrow \mathbb{R}^q$ that is differentiable at m ,

$$\sqrt{n}(h(Y_n) - h(m)) \xrightarrow[n \rightarrow \infty]{d} (Dh)(m)Z,$$

where $(Dh)(m)$ is the Jacobian matrix of h at m :

$$(Dh)(m) = \left((\partial_j h_i)(m) \right)_{1 \leq i \leq q, 1 \leq j \leq d}.$$

Intuition: $h(y) - h(m) \approx (Dh)(m)(y - m)$.

14/48

Special cases

Gaussian case

If $\sqrt{n}(Y_n - m) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \Sigma)$, then

$$\sqrt{n}(h(Y_n) - h(m)) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}\left(0, (Dh)(m) \Sigma (Dh)(m)^\top\right).$$

Scalar case

If $d = q = 1$ and $\sqrt{n}(Y_n - m) \xrightarrow[n \rightarrow \infty]{d} Z$, then

$$\sqrt{n}(h(Y_n) - h(m)) \xrightarrow[n \rightarrow \infty]{d} h'(m) Z.$$

Remark: if $h'(m) = 0$, and if h is twice differentiable at m , show that

$$n(h(Y_n) - h(m)) \xrightarrow[n \rightarrow \infty]{d} \frac{1}{2} h''(m) Z^2.$$

15/48

Proof (scalar case)

Consider the function ψ defined by :

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

ψ is continuous at m because h est differentiable at m . Since $Y_n \xrightarrow[n \rightarrow \infty]{d} m$,

$$\psi(Y_n) \xrightarrow[n \rightarrow \infty]{d} \psi(m) = h'(m),$$

and thus (Slutsky)

$$(\sqrt{n}(Y_n - m), \psi(Y_n)) \xrightarrow[n \rightarrow \infty]{d} (Z, h'(m)).$$

Finally, we have

$$\sqrt{n}(h(Y_n) - h(m)) = \sqrt{n}(Y_n - m) \psi(Y_n) \xrightarrow[n \rightarrow \infty]{d} h'(m) Z. \quad \square$$

Example: component reliability (cont'd)

We already saw that

- $\hat{\theta}_n = 1/\bar{X}_n$ is a consistent estimator of θ ,
- $\sqrt{n}(\bar{X}_n - \eta) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \eta^2)$, where $\eta = \frac{1}{\theta}$.

Using the delta method with $h(\eta) = \frac{1}{\eta}$, we get

$$\sqrt{n}\left(\frac{1}{\bar{X}_n} - \theta\right) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}\left(0, \eta^2 (h'(\eta))^2\right),$$

hence, since $h'(\eta) = -\frac{1}{\eta^2}$,

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \theta^2).$$

⇒ the estimator $\hat{\theta}_n$ is **asymptotically Gaussian**.

16/48

Example: component reliability (cont'd)

Another application: comparing estimators of $\eta = \mathbb{E}_\theta(X_1)$.

1) For $\hat{\eta}^{(1)} = \bar{X}_n$, we have (CLT): $\sqrt{n}(\hat{\eta}^{(1)} - \eta) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \eta^2)$.

2) For $\hat{\eta}^{(2)} = \sqrt{\frac{1}{2n} \sum_{i=1}^n X_i^2}$ (see lecture #1) ?

- Since $\mathbb{E}(X_1^2) = 2\eta^2$ et $\mathbb{E}(X_1^4) = 24\eta^4$, we have (CLT):

$$\sqrt{n}\left(\frac{1}{n} \sum_{i=1}^n X_i^2 - 2\eta^2\right) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, 20\eta^4).$$

- Hence, using the delta method with $h(z) = \sqrt{\frac{1}{2}z}$,

$$\sqrt{n}\left(\hat{\eta}^{(2)} - \eta\right) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}\left(0, \frac{5}{4}\eta^2\right).$$

Conclusion: $\hat{\eta}^{(1)}$ is "asymptotically preferable" to $\hat{\eta}^{(2)}$.

(Actually, it can be proved that $\hat{\eta}^{(1)}$ is efficient; see comput. of the FIM below).

17/48

Asymptotic comparison of (scalar) estimators

Let $\hat{\eta}_n$ and $\tilde{\eta}_n$ be two estimators of $\eta = g(\theta) \in \mathbb{R}$,

- ▶ asymptotically Gaussian.
- ▶ with asymptotic variances $\sigma^2(\theta)$ and $\tilde{\sigma}^2(\theta)$.

Definition: asymptotically preferable

If

- ▶ the two estimators have the same convergence rate,
- ▶ $\sigma^2(\theta) \leq \tilde{\sigma}^2(\theta) \quad \forall \theta \in \Theta$,

then we say that

$\hat{\eta}_n$ is asymptotically preferable to $\tilde{\eta}_n$

("strictly" if $\exists \theta \in \Theta$ such that $\sigma^2(\theta) < \tilde{\sigma}^2(\theta)$).

Note: comparing vector-valued estimators \Rightarrow compare matrices...

Lecture outline

1 – Convergence rate and asymptotic distribution

- 1.1 – Definitions and examples
- 1.2 – Theoretical tools
- 1.3 – Asymptotic efficiency

2 – Confidence regions and confidence intervals

- 2.1 – Definition and example
- 2.2 – Exact confidence intervals
- 2.3 – Asymptotic confidence intervals

3 – Warming up exercises

Asymptotic efficiency

Recall the Cramér-Rao lower bound (scalar parameter)

$\forall \hat{\theta}$ regular UE of θ , $\forall \theta \in \Theta$,

$$R_\theta(\hat{\theta}) = \text{var}_\theta(\hat{\theta}) \geq \frac{1}{n} I_1^{-1}(\theta),$$

with $I_1(\theta) = \text{var}_\theta(S_\theta(X_1))$.

► When equality holds for all θ , the estimator is called **efficient**.

Asymptotic efficiency

Definition. An estimator is called **asymptotically efficient** if

- ▶ it is asymptotically normal at the rate $\frac{1}{\sqrt{n}}$,
- ▶ with asymptotic variance $I_1^{-1}(\theta)$.

Remark: this definition is valid for the vector-valued case as well, replacing the variance by the covariance matrix

18/48

Asymptotic efficiency of the MLE

Context: $X_1, X_2, \dots \stackrel{\text{iid}}{\sim} P_\theta$ and, $\forall \theta \in \Theta$, P_θ admits a pdf f_θ .

Definition: regular model

The statistical model is called **regular** if

- ▶ **conditions C₁–C₄ hold**, (C₃ and C₄ defined below)
- ▶ $\forall \theta \in \Theta$, the Fisher information matrix $I_1(\theta)$ is positive definite.

Theorem

If the statistical model is **regular** and if the MLE $\hat{\theta}_n$ is **consistent**, then it is **asymptotically efficient** :

$$\sqrt{n} (\hat{\theta}_n - \theta) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, I_1^{-1}(\theta)).$$

19/48

Fisher information in regular models

Reminder. The **Fisher information** brought by \underline{X} is the matrix

$$I_{\underline{X}}(\theta) = \text{var}_{\theta}(S_{\theta}(\underline{X})) = \mathbb{E}_{\theta} \left(S_{\theta}(\underline{X}) S_{\theta}(\underline{X})^{\top} \right).$$

Proposition: another expression for the FIM

In a regular model, we have

$$I_{\underline{X}}(\theta) = -\mathbb{E}_{\theta} \left(\nabla_{\theta} \left(S_{\theta}(\underline{X})^{\top} \right) \right), \quad (\star)$$

In other words : $\forall \theta \in \Theta, \forall j \leq p, \forall k \leq p,$

$$(I_{\underline{X}}(\theta))_{j,k} = -\mathbb{E}_{\theta} \left(\frac{\partial}{\partial \theta_j} S_{\theta}^{(k)}(\underline{X}) \right) = -\mathbb{E}_{\theta} \left(\frac{\partial^2}{\partial \theta_j \partial \theta_k} \ln f_{\theta}(\underline{X}) \right).$$

Remark: actually, if C₁–C₃ hold, then C₄ and (\star) are equivalent.

20/48

Example: component reliability (cont'd)

Question: is $\hat{\theta}_n = 1/\bar{X}_n$ asymptotically efficient?

We have already computed the score: $S_{\theta}(X_1) = \frac{1}{\theta} - X_1.$

Computation of **Fisher's information** (two approaches):

Comput. of $\mathbb{E}_{\theta} (S_{\theta}(X_1)^2)$

$$I_1(\theta) = \text{var}_{\theta}(X_1) = \eta^2 = \frac{1}{\theta^2}$$

Comput. of $-\mathbb{E}_{\theta} \left(\frac{\partial S_{\theta}}{\partial \theta}(X_1) \right)$

$$I_1(\theta) = -\mathbb{E}_{\theta} \left(-\frac{1}{\theta^2} \right) = \frac{1}{\theta^2}$$

Conclusion: since $\sqrt{n} \left(\frac{1}{\bar{X}_n} - \theta \right) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \theta^2),$

$\hat{\theta}_n = \frac{1}{\bar{X}_n}$ is **asymptotically efficient**.

⇒ We recover the conclusions of the theorem (C₁–C₄ hold indeed).

21/48

Regular models: regularity conditions C₃ and C₄

Reminder: C₁ and C₂ were defined in Lecture #2.

Regularity condition C₃

$\theta \mapsto f_\theta(\underline{x})$ is twice continuously differentiable for ν -almost all \underline{x} .

Regularity condition C₄

At any point $\theta \in \Theta$, we have

$$\int_{\mathcal{S}} \nabla_{\theta} \nabla_{\theta}^{\top} f_{\theta}(\underline{x}) \nu(d\underline{x}) = \nabla_{\theta} \int_{\mathcal{S}} \nabla_{\theta}^{\top} f_{\theta}(\underline{x}) \nu(d\underline{x}).$$

In other words: $\forall \theta \in \Theta, \forall k \leq p, \forall j \leq p,$

$$\int_{\mathcal{S}} \frac{\partial^2 f_{\theta}(\underline{x})}{\partial \theta_k \partial \theta_j} \nu(d\underline{x}) = \frac{\partial}{\partial \theta_k} \int_{\mathcal{S}} \frac{\partial f_{\theta}(\underline{x})}{\partial \theta_j} \nu(d\underline{x}).$$

Example: an MLE that is not asymptotically Gaussian

Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{U}_{[0, \theta]}$, with $\theta > 0$ unknown.

⚠ This model is not regular (why?).

It can be proved that (cf. TD1, exercise 1.3)

- ▶ $\hat{\theta}_n = \max_{i \leq n} X_i$ is the MLE of θ , and
- ▶ $n(\hat{\theta}_n - \theta) \xrightarrow[n \rightarrow \infty]{d} -Z$ with $Z \sim \mathcal{E}\left(\lambda = \frac{1}{\theta}\right)$.

In this particular case

- ⇒ the MLE is **not asymptotically Gaussian**;
- ⇒ the **convergence rate** is $\frac{1}{n}$: faster than $\frac{1}{\sqrt{n}}$.

Lecture outline

1 – Convergence rate and asymptotic distribution

- 1.1 – Definitions and examples
- 1.2 – Theoretical tools
- 1.3 – Asymptotic efficiency

2 – Confidence regions and confidence intervals

- 2.1 – Definition and example
- 2.2 – Exact confidence intervals
- 2.3 – Asymptotic confidence intervals

3 – Warming up exercises

Lecture outline

1 – Convergence rate and asymptotic distribution

- 1.1 – Definitions and examples
- 1.2 – Theoretical tools
- 1.3 – Asymptotic efficiency

2 – Confidence regions and confidence intervals

- 2.1 – Definition and example
- 2.2 – Exact confidence intervals
- 2.3 – Asymptotic confidence intervals

3 – Warming up exercises

Motivation

Problem

A point estimator necessarily makes some estimation error.
How can we “report” this error?

Two approaches:

- ▶ provide, in addition to the estimated value,
 - ▶ the distribution of the estimator $\hat{\eta}$, exact or approximate,
 - ▶ or at least some “measure of dispersion”
(e.g., its standard deviation);
- ▶ give, instead of a point estimation $\hat{\eta}$,
 a confidence interval for η .

22/48

Confidence regions and confidence intervals

Recall that $\eta = g(\theta)$. We denote by $\mathcal{P}(N)$ the subsets of $N = g(\Theta)$.

Definition: confidence region

Let $\alpha \in]0, 1[$. A confidence region with level (at least) $1 - \alpha$ for η is a statistics $I_\alpha(\underline{X})$ taking values in $\mathcal{P}(N)$, such that:

$$\forall \theta \in \Theta, \quad \mathbb{P}_\theta(g(\theta) \in I_\alpha(\underline{X})) \geq 1 - \alpha.$$

We say that $I_\alpha(\underline{X})$ is a confidence region with level exactly $1 - \alpha$ if

$$\forall \theta \in \Theta, \quad \mathbb{P}_\theta(g(\theta) \in I_\alpha(\underline{X})) = 1 - \alpha.$$

(Some authors also write: of “size” $1 - \alpha$.)

Scalar case: if $I_\alpha(\underline{X})$ is an interval, it is called a confidence interval.

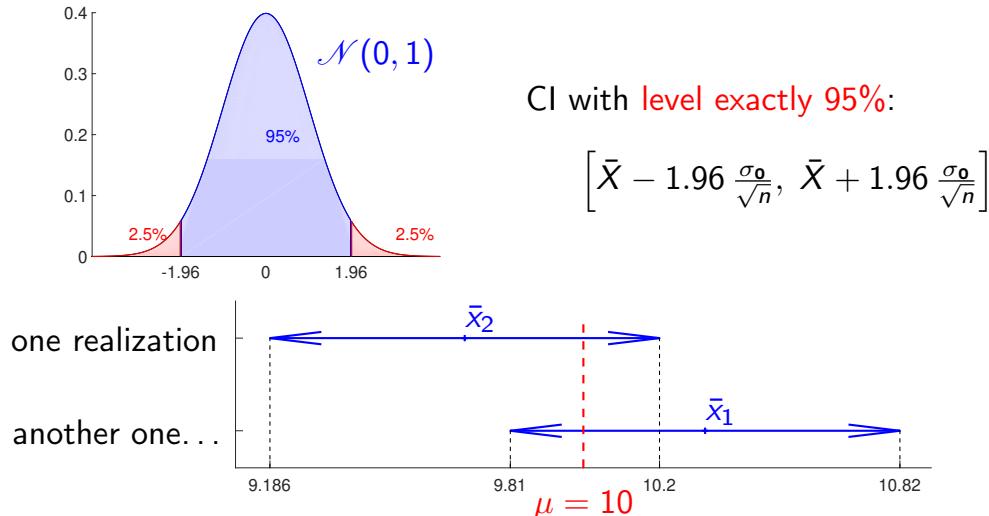
23/48

Example: $\mathcal{N}(\mu, \sigma_0^2)$ n -sample, with known σ_0^2

Since $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma_0^2}{n}\right)$, $T = \sqrt{n} \frac{\bar{X} - \mu}{\sigma_0} \sim \mathcal{N}(0, 1)$, therefore

$$\mathbb{P}_\mu \left(\sqrt{n} \frac{\bar{X} - \mu}{\sigma_0} \in [q_{\frac{\alpha}{2}}, q_{1-\frac{\alpha}{2}}] \right) = 1 - \alpha,$$

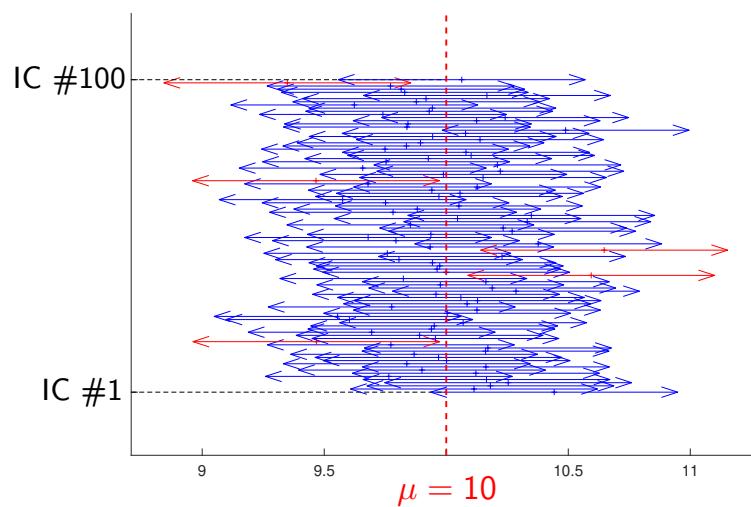
with q_r the quantile of order r of the $\mathcal{N}(0, 1)$ distribution.



24/48

Interpretation: simulations

We simulate 100 realizations with $\mu = 10$ and $\sigma_0 = 1$.



In red: realizations where the IC does not contain $\mu = 10$.

► The proportion of cases where the CI does not contain μ is (approx.) α .

25/48

Lecture outline

1 – Convergence rate and asymptotic distribution

- 1.1 – Definitions and examples
- 1.2 – Theoretical tools
- 1.3 – Asymptotic efficiency

2 – Confidence regions and confidence intervals

- 2.1 – Definition and example
- 2.2 – Exact confidence intervals
- 2.3 – Asymptotic confidence intervals

3 – Warming up exercises

Pivotal functions

The method can be formalized using **pivotal functions**.

Definitions

A function

$$T : \underline{\mathcal{X}} \times N \rightarrow \mathbb{R}$$

is called **pivotal** if the distribution of the RV $T = T(\underline{X}, \eta)$ **does not depend on θ** . We say that the distribution of $T(\underline{X}, \eta)$ is **free** from the parameter.

Back to the **example**: $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma_0^2)$ with known σ_0 .

Then $T = \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma_0}$ is pivotal since

$$\sqrt{n} \frac{\bar{X}_n - \mu}{\sigma_0} \sim \mathcal{N}(0, 1).$$

Remark: we can also choose $T = \sqrt{n} (\bar{X}_n - \mu) \sim \mathcal{N}(0, \sigma_0^2)$.

Probability refresher: quantiles

Definition: quantile of order r

Let $F(x)$ be the cdf of a probability distribution on \mathbb{R} .

For $0 < r < 1$, the **quantile of order r** of the distribution is defined as:

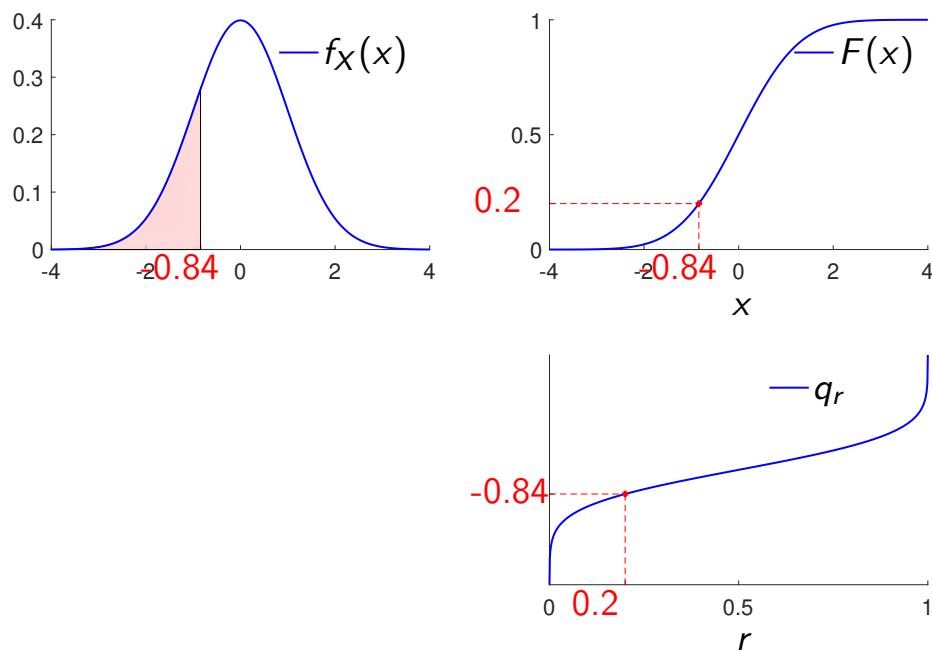
$$q_r = \inf \{x \in \mathbb{R}, F(x) \geq r\}.$$

Properties:

- ▶ If F is continuous, then $F(q_r) = r$.
- ▶ If, in addition, F is strictly increasing, then $q_r = F^{-1}(r)$.

27/48

Quantile function of the $\mathcal{N}(0, 1)$ distribution



28/48

How to use pivotal functions

Let $T(\underline{X}, \eta)$ be a pivotal function and $\alpha \in]0, 1[$.

Proposition

Assume that the cdf F of $T(\underline{X}, \eta)$ is continuous and strictly increasing, and denote by $q_r = F^{-1}(r)$ the quantile of order r .

Then, for all $\gamma \in [0, \alpha]$:

$$\begin{aligned} I_\alpha^\gamma(\underline{X}) &= \{\eta \in N \text{ such that } q_\gamma \leq T(\underline{X}, \eta) \leq q_{\gamma+1-\alpha}\} \\ &= T^{-1}(\underline{X}, [q_\gamma, q_{\gamma+1-\alpha}]) \end{aligned}$$

is a confidence interval for η with level exactly $1 - \alpha$.

Proof. $\mathbb{P}_\theta(g(\theta) \in I_\alpha^\gamma(\underline{X})) = \mathbb{P}_\theta(q_\gamma \leq T(\underline{X}, \eta) \leq q_{\gamma+1-\alpha}) = F(q_{\gamma+1-\alpha}) - F(q_\gamma) = 1 - \alpha$

□

Example: $\mathcal{N}(\mu, \sigma_0^2)$ n -sample, with known σ_0^2

Consider once more the pivotal function

$$T(\underline{X}, \mu) = \sqrt{n} \frac{(\bar{X} - \mu)}{\sigma_0} \sim \mathcal{N}(0, 1).$$

For all $\gamma \leq \alpha$, we obtain a CI with level (exactly) $1 - \alpha$:

$$I_\alpha^\gamma = \left[\bar{X} - \frac{\sigma_0}{\sqrt{n}} q_{1-\alpha+\gamma}, \quad \bar{X} - \frac{\sigma_0}{\sqrt{n}} q_\gamma \right],$$

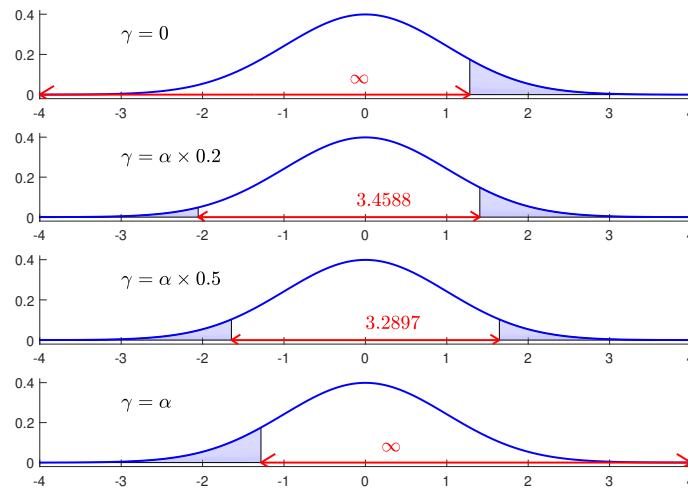
with q_r the quantile of order r of the $\mathcal{N}(0, 1)$ distribution.

For instance, with $\gamma = \frac{\alpha}{2}$ and $\alpha = 0.05$:

$$\begin{aligned} -q_{1-\alpha+\gamma} &= -q_{0.975} \approx -1.96 \\ -q_\gamma &= -q_{0.025} \approx +1.96 \end{aligned}$$

30 / 48

How to choose γ ?



Density of the $\mathcal{N}(0, 1)$ distribution and corresponding quantiles for $\alpha = 0.1$ and several values of γ (in red: $q_{\gamma+1-\alpha} - q_\gamma$).

Usual criterion: value s.t. the CI has minimal length (here $\gamma = \frac{\alpha}{2}$).

31/48

Example: component reliability (cont'd)

It can be proved that:

$$T(\underline{X}, \eta) = \frac{\bar{X}}{\eta} \sim \Gamma(n, n).$$

Thus, a CI with level exactly $1 - \alpha$ is :

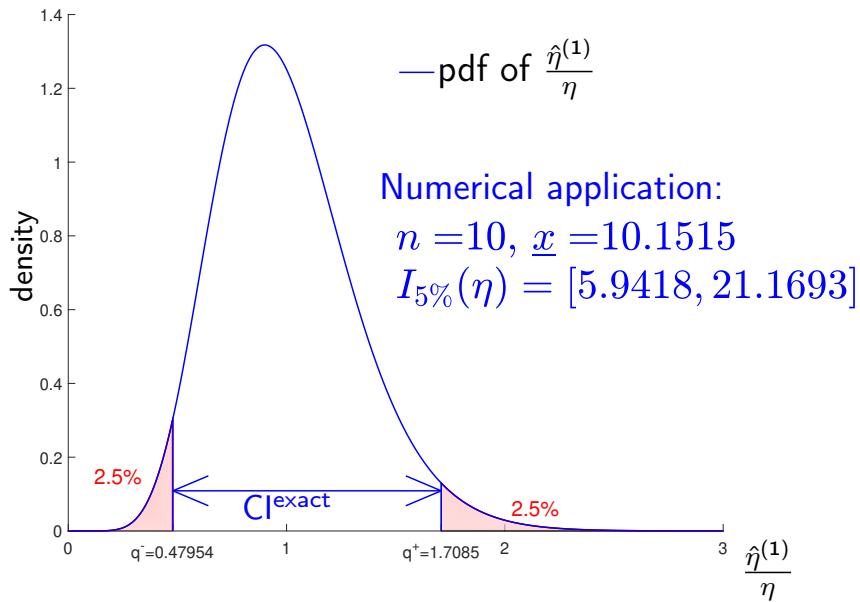
$$I_\alpha^\gamma = \left[\frac{\bar{X}}{q_{\gamma+1-\alpha}}, \frac{\bar{X}}{q_\gamma} \right],$$

with q_r the quantile of order r of the $\Gamma(n, n)$ distribution.

Choice of γ : we can take $\gamma = \frac{\alpha}{2}$ for simplicity, or search numerically for the value γ such that the length $1/q_\gamma - 1/q_{1+\gamma-\alpha}$ is minimal.

32/48

Example: component reliability (cont'd)



Probability density function of the pivotal distribution $\Gamma(n, n)$ and corresponding quantiles for $\alpha = 0.05$ and $\gamma = \frac{\alpha}{2}$.

33/48

Lecture outline

1 – Convergence rate and asymptotic distribution

- 1.1 – Definitions and examples
- 1.2 – Theoretical tools
- 1.3 – Asymptotic efficiency

2 – Confidence regions and confidence intervals

- 2.1 – Definition and example
- 2.2 – Exact confidence intervals
- 2.3 – Asymptotic confidence intervals

3 – Warming up exercises

Motivation and goal

Problem

It is sometimes (often) difficult to find a pivotal function.

Solution: use once again an asymptotic approach.

- ▶ Intervals with “approximate guarantees” will be obtained.
- ▶ Comput. become easier with the tools that we already have (CLT, Slutsky, delta method...).



Any analysis carried out in an asymptotic setting is
approximate when n is finite.

- ⇒ The results can be poor for small n ...

34 / 48

Asymptotic confidence regions (intervals)

We set $\underline{X}_n = (X_1, \dots, X_n)$. Recall that $\eta = g(\theta)$ and $N = g(\Theta)$.

Definition: asymptotic confidence region

An asymptotic confidence region with level (at least) $1 - \alpha$ is a statistic $I_{n,\alpha}(\underline{X}_n)$, with values in $\mathcal{P}(N)$, such that

$$\forall \theta \in \Theta, \quad \lim_{n \rightarrow \infty} \mathbb{P}_{\theta}(g(\theta) \in I_{n,\alpha}(\underline{X}_n)) \geq 1 - \alpha.$$

(variant: “exactly” if equality holds for all θ .)

Recall that for an “exact” CR with level (at least) $1 - \alpha$,

$$\forall \theta \in \Theta, \quad \mathbb{P}_{\theta}(g(\theta) \in I_{n,\alpha}(\underline{X}_n)) \geq 1 - \alpha$$

(here, “exact” means “non asymptotic”).

35 / 48

Asymptotic pivotal function

Definition

A (sequence of) function(s)

$$T_n : \mathcal{X}^n \times N \rightarrow \mathbb{R}$$

is an **asymptotic pivotal function** if the **limit** distribution of $T_n(\underline{X}_n, \eta)$ does not depend on θ :

$$T_n(\underline{X}_n, \eta) \xrightarrow[n \rightarrow \infty]{d} T_\infty.$$

where T_∞ is a RV whose distribution is free of θ .

How to use asymptotic pivotal functions:

⇒ exactly as we used the non-asymptotic ones !

36/48

Example: component reliability (cont'd)

We already saw that (Slutsky + continuity theorem)

$$\sqrt{n} \frac{(\bar{X}_n - \eta)}{\bar{X}_n} \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, 1).$$

⇒ Asymptotic pivotal function :

$$T_n(\underline{X}_n, \eta) = \sqrt{n} \frac{\bar{X}_n - \eta}{\bar{X}_n}.$$

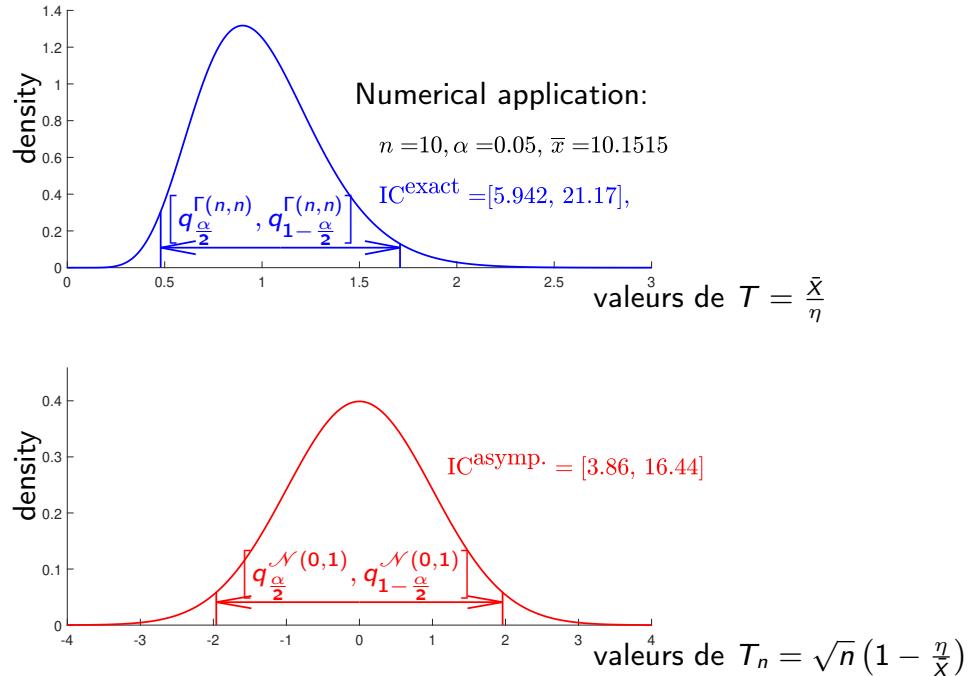
⇒ Asymptotic CI with level (exactly) $1 - \alpha$ for η :

$$I_{n,\alpha} = \left[\left(1 - \frac{1}{\sqrt{n}} q_{1-\frac{\alpha}{2}} \right) \bar{X}_n, \left(1 + \frac{1}{\sqrt{n}} q_{1-\frac{\alpha}{2}} \right) \bar{X}_n \right]$$

with q_r the quantile of order r of the $\mathcal{N}(0, 1)$ distribution.

37/48

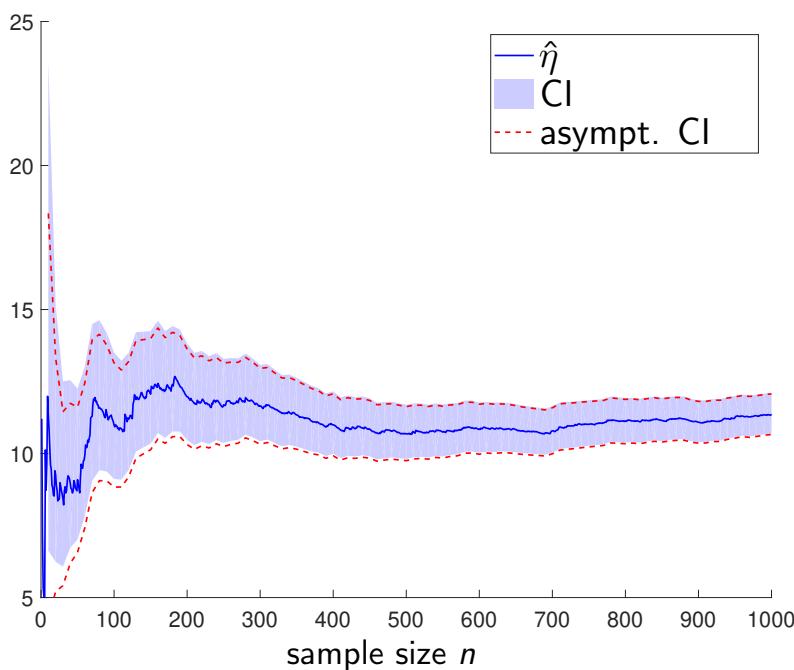
Example: component reliability (cont'd)



⚠ Do not confuse intervals on pivotal functions $[q_{\frac{\alpha}{2}}, q_{1-\frac{\alpha}{2}}]$ and confidence interval for η .

38/48

Example: component reliability (cont'd)



Comparison of exact and asymptotic CIs, as a function of n

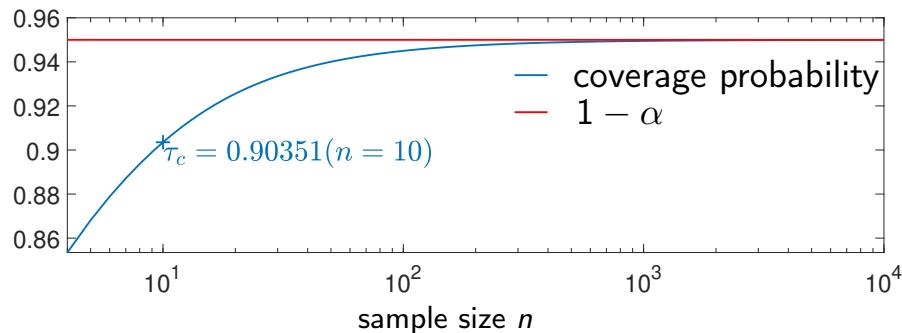
39/48

Coverage probability of a confidence interval

Definition

For $\theta \in \Theta$, the **coverage probability** of $I_{n,\alpha}(\underline{X}_n)$ is defined by

$$\tau_{n,\theta}^c(I_{n,\alpha}(\underline{X}_n)) = \mathbb{P}_\theta(\eta \in I_{n,\alpha}(\underline{X}_n))$$



Ex. "component reliability": $\tau_{n,\theta}^c$ for the asympt. CI with level 95%

Remark. If $I_{n,\alpha}(\underline{X}_n)$ is an asympt. CI with level $1 - \alpha$, then :

$$\forall \theta, \lim_{n \rightarrow \infty} \tau_\theta^c(I_{n,\alpha}(\underline{X}_n)) \geq 1 - \alpha.$$

40/48

Lecture outline

1 – Convergence rate and asymptotic distribution

- 1.1 – Definitions and examples
- 1.2 – Theoretical tools
- 1.3 – Asymptotic efficiency

2 – Confidence regions and confidence intervals

- 2.1 – Definition and example
- 2.2 – Exact confidence intervals
- 2.3 – Asymptotic confidence intervals

3 – Warming up exercises

Exercise 1 (asymptotic distribution)

Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{E}(\theta)$, with $\theta > 0$.

Let η denote the probability of exceeding a given threshold $x_0 > 0$:

$$\eta = \mathbb{P}_\theta(X \geq x_0) = \exp(-\theta x_0).$$

Questions

- ① Study the asymptotic behaviour of the sample mean \bar{X}_n .
- ② Propose an estimator $\hat{\eta}_n^{(1)}$ as a function of \bar{X}_n , using the substitution method.
- ③ Study the asymptotic behaviour of $\hat{\eta}_n^{(1)}$.
- ④ Let $\hat{\eta}_n^{(2)} = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{X_i \geq x_0}$. Is one of the estimators asymptotically preferable to the other?

41/48

Exercise 2 (exact confidence interval)

Definition: Rayleigh distribution with parameter σ^2

$X \sim \mathcal{R}(\sigma^2)$ if X admits the pdf $f(x) = \frac{x}{\sigma^2} \exp\left(-\frac{x^2}{2\sigma^2}\right)$, $x \geq 0$.

Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{R}(\sigma^2)$, with $\sigma^2 > 0$.

Questions

- ① Find a pivotal function.

Hint: if $X \sim \mathcal{R}(\sigma^2)$ then $Y = X^2 \sim \mathcal{E}\left(\frac{1}{2\sigma^2}\right)$.

- ② Deduce a confidence interval for σ^2 with level 95%.

42/48

Solution of Exercise 1

❶ Appliquant le TCL :

$$\sqrt{n} \left(\bar{X}_n - \frac{1}{\theta} \right) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N} \left(0, \frac{1}{\theta^2} \right)$$

❷ $\eta = \exp \left(-\frac{x_0}{\frac{1}{\theta}} \right) = h \left(\frac{1}{\theta} \right)$

avec $h : u \mapsto \exp \left(-\frac{x_0}{u} \right)$ continue sur \mathbb{R}_+^* .

Utilisant la méthode de substitution à \bar{X}_n estimateur de $\frac{1}{\theta}$:

$$\hat{\eta}_n^{(1)} = h(\bar{X}_n) = \exp \left(-\frac{x_0}{\bar{X}_n} \right)$$

43 / 48

Solution of Exercise 1

❸ h est dérivable sur \mathbb{R}_+^* avec $h'(u) = \frac{x_0}{u^2} \exp \left(-\frac{x_0}{u} \right)$.

Appliquant le Delta théorème dans le contexte gaussien :

$$\sqrt{n} \left(h(\bar{X}_n) - h \left(\frac{1}{\theta} \right) \right) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N} \left(h' \left(\frac{1}{\theta} \right)^2 \frac{1}{\theta^2} \right)$$

Soit :

$$\sqrt{n} \left(\hat{\eta}_n^{(1)} - \eta \right) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N} \left((x_0 \theta \exp(-\theta x_0))^2 \right)$$

La variance asymptotique de $\hat{\eta}_n^{(1)}$ est $\sigma_1^2(\theta) = (x_0 \theta \exp(-\theta x_0))^2$

44 / 48

Solution of Exercise 1

❸ $\hat{\eta}_n^{(2)} = \frac{1}{n} \sum_{i=1}^n Z_i$ avec $Z_i = \mathbb{1}_{X_i \geq x_0}$ avec $\begin{cases} Z_1, \dots, Z_n \text{ IID} \\ Z_1 \sim \mathcal{B}(\eta) \end{cases}$

Appliquant le TCL, $\hat{\eta}_n^{(2)}$ est asymptotiquement gaussien :

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n Z_i - \mathbb{E}(Z_1) \right) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \text{var}(Z_1))$$

soit

$$\begin{aligned} \sqrt{n} \left(\hat{\eta}_n^{(2)} - \eta \right) &\xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \eta(1-\eta)) \\ &\xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(\exp(-\theta x_0)(1 - \exp(-\theta x_0))) \end{aligned}$$

La var. asympt. de $\hat{\eta}_n^{(2)}$ est $\sigma_2^2(\theta) = \exp(-\theta x_0)(1 - \exp(-\theta x_0))$

45/48

Solution of Exercise 1

Soit $\Delta(\theta) = \sigma_2^2(\theta) - \sigma_1^2(\theta)$.

$$\begin{aligned} \Delta(\theta) &= \exp(-\theta x_0)(1 - \exp(-\theta x_0) - x_0^2 \theta^2 \exp(-\theta x_0)) \\ &= \exp(-\theta x_0) \varphi(\theta x_0) \end{aligned}$$

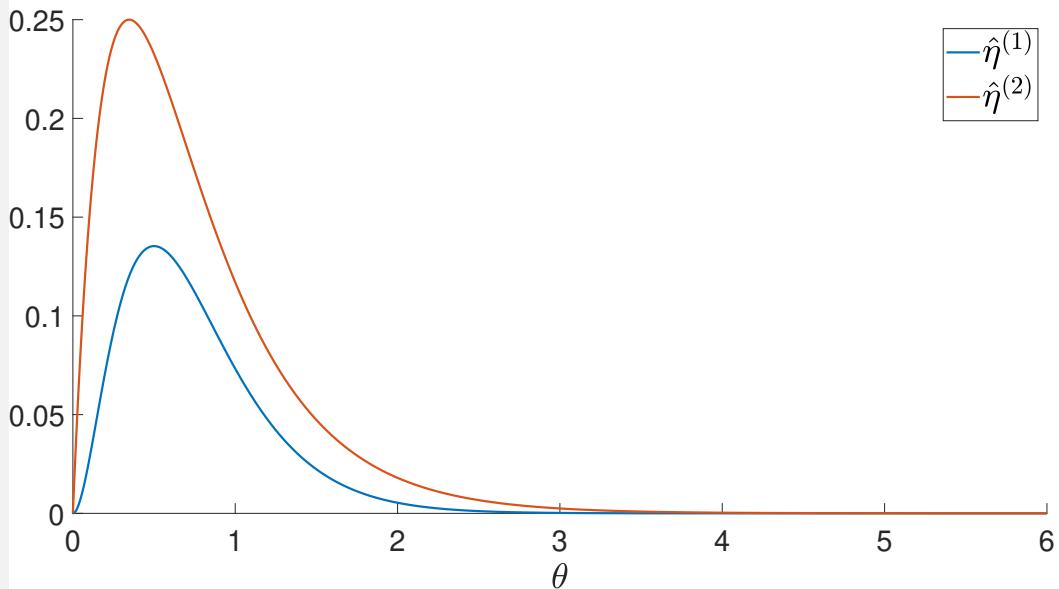
avec $\varphi(u) = 1 - \exp(-u)(1 + u^2)$.

Un tableau de variation de φ montre que $\varphi > 0$ sur \mathbb{R}_+ .

$\hat{\eta}_n^{(1)}$ est donc asymptotiquement préférable à $\hat{\eta}_n^{(2)}$.

46/48

Solution of Exercise 1



Tracés des 2 variances asymptotiques pour $x_0 = 2.0$.

47/48

Corrigé de l'exercice 2

En utilisant l'indication : $X_i^2 \sim \mathcal{E}\left(\frac{1}{2\sigma^2}\right)$

Les X_i étant indépendants :

$$\sum_{i=1}^n X_i^2 \sim \Gamma\left(n, \frac{1}{2\sigma^2}\right) \quad (\text{rappel : } \mathcal{E}(\lambda) = \Gamma(1, \lambda))$$

■ $T(\underline{X}, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^n X_i^2 \sim \Gamma\left(n, \frac{1}{2}\right)$ est pivotale pour σ^2 .

On en déduit un IC pour σ^2 de niveau (exactement) $1 - \alpha$:

$$I_{\alpha}^{\gamma=\frac{\alpha}{2}} = \left[\frac{1}{q_{0.975}} \sum_{i=1}^n X_i^2, \frac{1}{q_{0.025}} \sum_{i=1}^n X_i^2 \right].$$

où q_r est le quantile d'ordre r de la loi $\Gamma\left(n, \frac{1}{2}\right)$

Remarque : en prenant la racine carré, on obtient un IC pour σ

48/48

Chapter 4

Bayesian estimation



CentraleSupélec

Statistics and Learning

Arthur Tenenhaus[†], Julien Bect & Laurent Le Brusquet

(firstname.lastname@centralesupelec.fr)

Teaching: CentraleSupélec / Department of Mathematics

Research: Laboratory of signals and systems (L2S)

[†]: Course coordinator

1/37

Lecture 4/10

Bayesian estimation

In this lecture you will learn how to...

- ▶ Introduce the concept of prior information.
- ▶ Present the basics of the Bayesian approach.
- ▶ Explain how to construct estimators using prior information.

2/37

Lecture outline

- 1 – Introduction: the Bayes risk
- 2 – Bayesian statistics: prior / posterior distribution
- 3 – Choosing a prior distribution
- 4 – Bayes estimators
- 5 – Warming up exercise

3/37

Lecture outline

- 1 – Introduction: the Bayes risk
- 2 – Bayesian statistics: prior / posterior distribution
- 3 – Choosing a prior distribution
- 4 – Bayes estimators
- 5 – Warming up exercise

Recap: comparing estimators

Quadratic risk: $R_\theta(\hat{\eta}) = \mathbb{E}_\theta (\|\hat{\eta} - g(\theta)\|^2)$.

Definition

We will say that $\hat{\eta}'$ is (weakly) **preferable** to $\hat{\eta}$ if

- ▶ $\forall \theta \in \Theta, R_\theta(\hat{\eta}') \leq R_\theta(\hat{\eta}),$

We will say that it is **strictly preferable** to $\hat{\eta}$ if, in addition,

- ▶ $\exists \theta \in \Theta, R_\theta(\hat{\eta}') < R_\theta(\hat{\eta}),$

Remarks

- ▶ The relation “is preferable to” is a **partial order** on risk functions.
- ▶ **In general there is no optimal estimator**, i.e., no estimator that is preferable to all the others (unless we restrict the class of estimators that is considered).

4/37

Comparing (all) estimators: two approaches

Two approaches make it possible to refine the comparison for the cases where the risk functions R_θ cannot be compared:

- ① the **minimax** (or “worst case”) approach:

$$R_{\max}(\hat{\eta}) = \sup_{\theta \in \Theta} R_\theta(\hat{\eta}),$$

⇒ not discussed in this class;

- ② the **Bayesian** (or “average case”) approach:

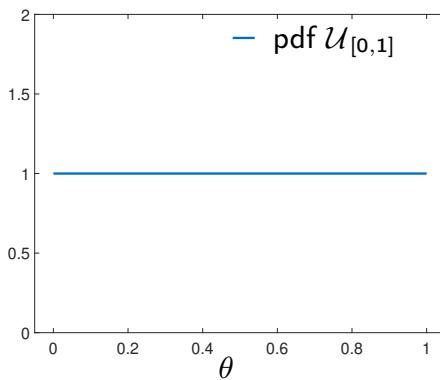
$$R_{\text{Bayes}, \pi}(\hat{\eta}) = \int_{\Theta} R_\theta(\hat{\eta}) \pi(d\theta),$$

where π is a probability measure on Θ , to be chosen.

⇒ this is the topic of this lecture.

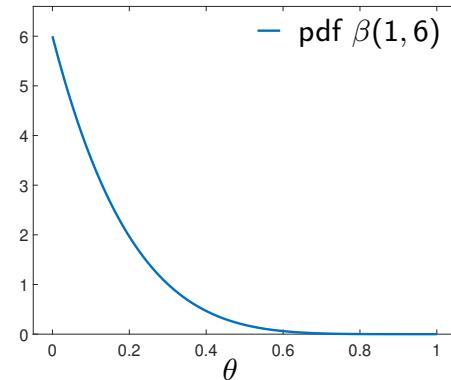
5/37

Example: white balls / red balls (see lecture #1)



Measure π : uniform over $[0, 1]$

$$\hat{\theta}_a = \frac{\sum_{i=1}^n X_i + 1}{n + 2}$$



Measure π : $\beta(1, 6)$

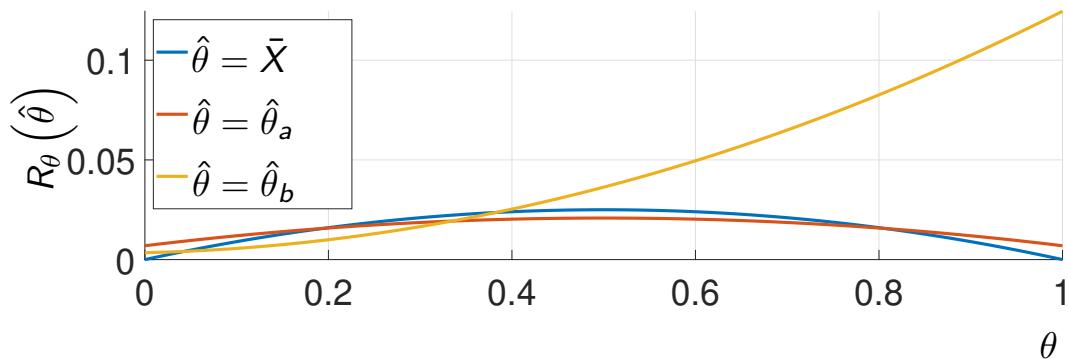
$$\hat{\theta}_b = \frac{\sum_{i=1}^n X_i + 1}{n + 7}$$

Observation: $\hat{\theta}_b = \frac{n+2}{n+7} \hat{\theta}_a$,

⇒ the second estimator provides smaller estimates

6/37

Example: white balls / red balls (with $n = 10$)



	$\hat{\theta} = \bar{X}$	$\hat{\theta} = \hat{\theta}_a$	$\hat{\theta} = \hat{\theta}_b$
$R_{\max}(\hat{\theta})$	0.025 $\frac{1}{4n}$	≈ 0.0208 $\frac{1}{4(n+2)}$	≈ 0.1246 $\frac{36}{(n+7)^2}$ (valid for $n \leq 77$)
$R_{\text{Bayes}, \pi}(\hat{\theta})$ with $\pi \sim \mathcal{U}_{[0,1]}$	≈ 0.0167 $\frac{1}{6n}$	≈ 0.0162 $\frac{n+4}{6(n+2)^2}$	≈ 0.0456 $\frac{n+69}{6(n+7)^2}$
$R_{\text{Bayes}, \pi}(\hat{\theta})$ with $\pi \sim \beta(1, 6)$	≈ 0.0107 $\frac{3}{28n}$	≈ 0.0129 $\frac{3n+22}{28(n+2)^2}$	≈ 0.0089 $\frac{3n+42}{28(n+7)^2}$

Exercise: prove the expressions of R_{\max} and $R_{\text{Bayes}, \pi}$ for $\hat{\theta} = \bar{X}$.

7/37

The beta family of distributions

Let $X \sim \beta(a, b)$ with $(a, b) = \theta \in (\mathbb{R}_*)^2$. Its pdf is :

$$f_\theta(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1} \mathbb{1}_{[0,1]}(x).$$

Moments

- ▶ expectation : $\mathbb{E}_\theta(X) = \frac{a}{a+b}$
- ▶ variance : $\text{var}_\theta(X) = \frac{ab}{(a+b)^2(a+b+1)}$

Special case

Properties

- ▶ If $X \sim \beta(a, 1)$, then $-\log(X) \sim \mathcal{E}\left(\frac{1}{a}\right)$.
- ▶ If $X \sim \Gamma(a, \lambda)$, $Y \sim \Gamma(b, \lambda)$, and $X \perp\!\!\!\perp Y$, then $\frac{X}{X+Y} \sim \beta(a, b)$.

Unknown parameter \rightarrow random variables

We will assume from now on a dominated model: pdf $f_\theta(\underline{x})$.

Consider the Bayesian risk (quadratic, in this case)

$$\begin{aligned} R_{\text{Bayes}, \pi}(\hat{\eta}) &= \int_{\Theta} R_\theta(\hat{\eta}) \pi(d\theta) \\ &= \int_{\Theta} \mathbb{E}_\theta (\|\hat{\eta} - g(\theta)\|^2) \pi(d\theta). \end{aligned}$$

It can be re-written as :

$$R_{\text{Bayes}, \pi}(\hat{\eta}) = \iint_{\mathcal{X} \times \Theta} \|\hat{\eta}(\underline{x}) - g(\theta)\|^2 \underbrace{f_\theta(\underline{x}) \nu(d\underline{x}) \pi(d\theta)}_{\text{Probability meas. on } \mathcal{X} \times \Theta}.$$

Unknown parameter → random variables (cont'd)

Let us introduce a new random variable ϑ , such that

$$(\underline{X}, \vartheta) \sim f_{\theta}(\underline{x}) \nu(d\underline{x}) \pi(d\vartheta). \quad (*)$$

Then the Bayesian risk can be re-written more simply as:

$$R_{\text{Bayes}, \pi} = \mathbb{E} (\|\hat{\eta} - g(\vartheta)\|^2),$$

where the expectation is, this time, over both \underline{X} and ϑ .

Bayesian approach

In Bayesian statistics, the **unknown parameter θ** is (also) **modeled as a random variable**.

(Technical remark: the introduction of a new random variable ϑ such that (*) holds is always possible, if we are willing to replace the underlying set Ω by $\tilde{\Omega} = \Omega \times \Theta$, provided that Θ is endowed with a σ -algebra \mathcal{F}_{Θ} such that $\theta \mapsto \mathbb{P}_{\theta}(E)$ is \mathcal{F}_{Θ} -measurable for all $E \in \mathcal{F}$.)

9/37

Lecture outline

1 – Introduction: the Bayes risk

2 – Bayesian statistics: prior / posterior distribution

3 – Choosing a prior distribution

4 – Bayes estimators

5 – Warming up exercise

Bayesian statistical models

Technical assumptions: we assume from now on that

- ▶ Θ is endowed with a σ -algebra \mathcal{F}_Θ . For inst.: if $\Theta \subset \mathbb{R}^p$, $\mathcal{F}_\Theta = \mathcal{B}(\Theta)$;
- ▶ $\theta \mapsto \mathbb{P}_\theta(E)$ is \mathcal{F}_Θ -measurable for all $E \in \mathcal{F}$ (σ -algebra on Ω).

Definition

A **Bayesian statistical model** consists of

- ▶ a statistical model as previously defined:

$$(\underline{\mathcal{X}}, \underline{\mathcal{A}}, \left\{ \mathbb{P}_\theta^X, \theta \in \Theta \right\}),$$

- ▶ a probability distrib. π , called **prior distribution**, on $(\Theta, \mathcal{F}_\Theta)$.

Dominated model \rightarrow makes it possible to define a **likelihood**.

10/37

Joint, prior and posterior distributions

Recall that we have introduced a new random variable ϑ , such that

$$(\underline{X}, \vartheta) \sim f_\theta(\underline{x}) \nu(d\underline{x}) \pi(d\vartheta). \quad (*)$$

Bayesian vocabulary

We call:

- ▶ **joint distribution** the distribution of \underline{X} and ϑ , that is, $(*)$,
- ▶ **prior distribution** the marginal distribution \mathbb{P}^ϑ of ϑ , that is, π ,
- ▶ **posterior distribution** the distribution $\mathbb{P}^{\vartheta|\underline{X}}$ of ϑ given the data.

Interpretation (“subjective Bayes”)

- ▶ prior distribution \rightarrow knowledge about θ before data acquisition
- ▶ posterior distribution $\rightarrow \dots$ after data acquisition

11/37

By the way... what is the conditional distribution $\mathbb{P}^{\vartheta|X}$?

General definition: beyond the scope of this lecture!

(\Rightarrow uses the notion of kernel)

Assume that (ϑ, X) has a density with respect to $\nu \otimes \nu_\Theta$, for some measure ν_Θ sur $(\Theta, \mathcal{F}_\Theta)$.

We will *define* $\mathbb{P}^{\vartheta|X=x}$ as the measure with density

$$f^{\vartheta|X}(\theta | x) = \frac{f^{\vartheta,X}(\theta, x)}{f^X(x)}$$

with respect to ν_Θ , for all x such that $f^X(x) > 0$.

Then we have, for any measurable function φ s.t. $\varphi(\vartheta, X) \in L^1$,

$$\mathbb{E}(\varphi(\vartheta, X) | X) \stackrel{\text{a.s.}}{=} \int_{\Theta} \varphi(\theta, X) f^{\vartheta|X}(\theta | X) \nu_\Theta(d\theta).$$

12/37

Joint and marginal densities

We will assume[†] from now on that π admits a pdf

- ▶ wrt a measure ν_Θ on $(\Theta, \mathcal{F}_\Theta)$, e.g., Lebesgue's measure,
- ▶ we will write (abusively): $\pi(d\theta) = \pi(\theta) d\theta$.

Proposition

The joint distribution admits the **joint pdf**

$$f^{(X,\vartheta)}(x, \theta) = f_\theta(x) \pi(\theta),$$

and the corresponding **marginal densities** are

$$\begin{aligned} f^\vartheta(\theta) &= \pi(\theta), \\ f^X(x) &= \int f_\theta(x) \pi(\theta) d\theta. \end{aligned}$$

[†]: This is not actually an assumption, since we can always use $\nu_\Theta = \pi$ (with the pdf equal to 1).

13/37

Proof

Joint pdf (informal proof)

$$\begin{aligned}\mathbb{P}^{(\underline{X}, \vartheta)}(d\underline{x}, d\theta) &= f_\theta(\underline{x}) \nu(d\underline{x}) \quad \pi(\theta) d\theta \\ &= \underbrace{f_\theta(\underline{x}) \pi(\theta)}_{\text{joint pdf}} \quad \nu(d\underline{x}) d\theta\end{aligned}$$

Marginal densities \rightarrow we just need to integrate:

$$\begin{aligned}f^\vartheta(\theta) &= \int f_\theta(\underline{x}) \pi(\theta) \nu(d\underline{x}) = \pi(\theta), \\ f^{\underline{X}}(\underline{x}) &= \int f_\theta(\underline{x}) \pi(\theta) d\theta.\end{aligned}$$

□

14/37

Likelihood and Bayes' formula

Recall the **conditional density**:

$$f^{Y|Z}(y | z) = \frac{f^{(Y,Z)}(y, z)}{f^Z(z)}, \quad \forall z \text{ s.t. } f^Z(z) \neq 0. \quad (*)$$

Proposition

i) The conditional distribution of \underline{X} given ϑ admits the pdf

$$f^{\underline{X}|\vartheta}(\underline{x} | \theta) = f_\theta(\underline{x}) \quad (\text{"likelihood"}).$$

ii) The posterior distribution (ϑ given \underline{X}) admits the pdf :

$$f^{\vartheta|\underline{X}}(\theta | \underline{x}) = \frac{f_\theta(\underline{x}) \pi(\theta)}{f^{\underline{X}}(\underline{x})} \quad (\text{Bayes' formula}).$$

Proof. Simply apply (*) to the joint pdf. □

15/37

Remark: proportionality

The term $\frac{1}{f^X(\underline{x})}$ plays the role of a **normalizing constant**:

$$f^{\vartheta|X}(\theta | \underline{x}) = \frac{f_\theta(\underline{x}) \pi(\theta)}{f^X(\underline{x})}.$$

Notation. The symbol “ \propto ” indicates **proportionality**. Thus,

$$f^{\vartheta|X}(\theta | \underline{x}) \propto f_\theta(\underline{x}) \pi(\theta),$$

or, less formally,

posterior pdf \propto likelihood \times prior pdf.

The “constant” $f^X(\underline{x})$ is often difficult to compute, but in some situations the computation can be avoided (MAP estimator, MCMC numerical methods...).

16/37

Example: white balls / red balls (cont'd)

Reminder: we want to estimate $\theta = \frac{W}{W+R}$ from $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Ber}(\theta)$.

Density of the observations:

$$f_\theta(\underline{x}) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} = \theta^{N(\underline{x})} (1 - \theta)^{n-N(\underline{x})}.$$

with $N(\underline{x}) = \sum_{i=1}^n x_i$.

Let us choose a $\beta(a_0, b_0)$ prior:

$$\pi(\theta) \propto \theta^{a_0-1} (1 - \theta)^{b_0-1}.$$

(The choice of the prior distribution will be discussed later.)

17/37

Example: white balls / red balls (cont'd)

Then we have:

$$\begin{aligned} f^{\vartheta | \underline{X}}(\theta | \underline{x}) &\propto f_{\theta}(\underline{x}) \pi(\theta) \\ &\propto \theta^{N(\underline{x})} (1-\theta)^{n-N(\underline{x})} \cdot \theta^{a_0-1} (1-\theta)^{b_0-1} \\ &= \theta^{a_0+N(\underline{x})-1} (1-\theta)^{b_0+n-N(\underline{x})-1}. \end{aligned}$$

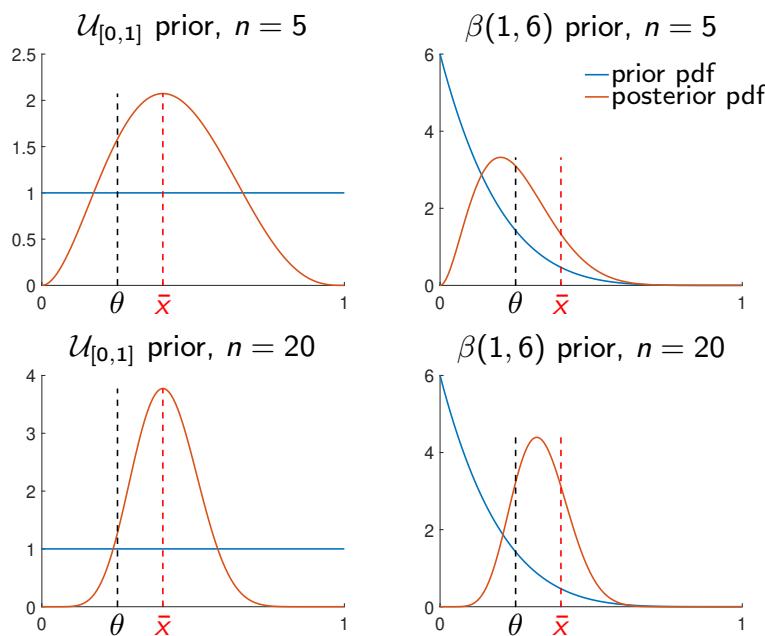
We recognize (up to a cst) the pdf of the $\beta(a_n, b_n)$ distrib., with

$$\begin{cases} a_n = a_0 + N, \\ b_n = b_0 + n - N. \end{cases}$$

Conclusion. Posterior distribution: $\vartheta | \underline{X} \sim \beta(a_n, b_n)$.

18/37

Example: white balls / red balls (cont'd)



Remark: for $n \rightarrow \infty$, we have a $\mathbb{E}(\vartheta | \underline{X}_n) = \bar{X}_n + O(\frac{1}{n})$ with $\text{var}(\vartheta | \underline{X}_n) \simeq \frac{\theta(1-\theta)}{n}$.

19/37

Example: component reliability

Reminder: $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{E}(\theta) = \mathcal{E}\left(\frac{1}{\eta}\right)$, hence the likelihood:

$$\begin{aligned}\mathcal{L}(\eta, \underline{x}_n) &= f(\underline{x}_n | \eta) = \prod_{i=1}^n \frac{1}{\eta} \exp\left(-\frac{1}{\eta} x_i\right) \\ &= \eta^{-n} \exp\left(-\frac{1}{\eta} \sum_{i=1}^n x_i\right).\end{aligned}$$

(Here we directly use η as our unknown parameter.)

We choose (see below) a truncated $\mathcal{N}(\eta_0, \sigma_0^2)$ prior for η :

$$\pi(\eta) \propto \exp\left(-\frac{(\eta - \eta_0)^2}{2\sigma_0^2}\right) \mathbb{1}_{\eta \geq 0}.$$

20/37

Example: component reliability (cont'd)

Posterior distribution of η . From Bayes' formula we get:

$$p(\eta | \underline{x}_n) \propto \underbrace{\eta^{-n} \exp\left(-\frac{1}{\eta} \sum_{i=1}^n x_i\right)}_{\text{likelihood}} \cdot \underbrace{\exp\left(-\frac{(\eta - \eta_0)^2}{2\sigma_0^2}\right)}_{\text{prior pdf}}.$$



This time we fail to recognize a “familiar” density
 ➔ numerical evaluation of the integrals

$$\begin{aligned}f(\underline{x}_n) &= \int \eta^{-n} e^{-\frac{1}{\eta} \sum_{i=1}^n x_i} e^{-\frac{(\eta - \eta_0)^2}{2\sigma_0^2}} d\eta \\ \mathbb{E}(\eta | \underline{X}_n = \underline{x}_n) &= \frac{1}{f(\underline{x}_n)} \int \eta \cdot \eta^{-n} e^{-\frac{1}{\eta} \sum_{i=1}^n x_i} e^{-\frac{(\eta - \eta_0)^2}{2\sigma_0^2}} d\eta\end{aligned}$$

21/37

Example: component reliability (cont'd)

Numerical application. $\eta_0 = 14.0$, $\sigma_0 = 1.0$ and the true value is $\eta_* = 11.4$.

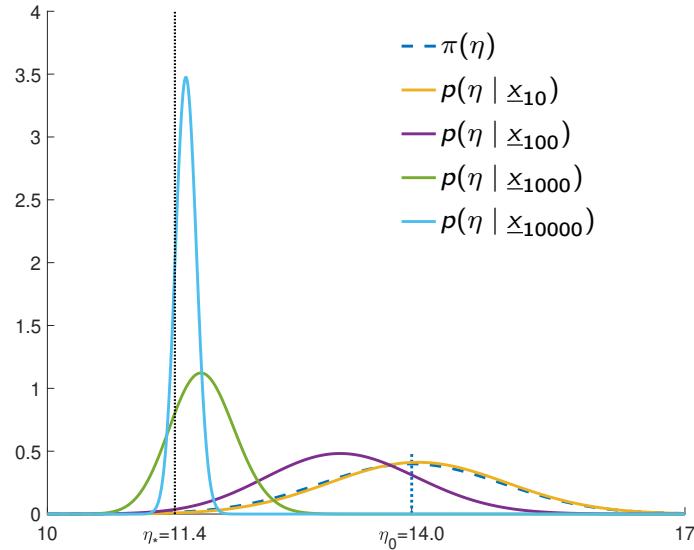


Figure – Prior and posterior densities of η , for four values of n .

22/37

Lecture outline

1 – Introduction: the Bayes risk

2 – Bayesian statistics: prior / posterior distribution

3 – Choosing a prior distribution

4 – Bayes estimators

5 – Warming up exercise

Several approaches

Two kinds of sources of prior information:

- ▶ “historical” **data**,
- ▶ **experts**: subjective knowledge, field expertise, etc.

Advanced topics (not covered in this course):

- ▶ merging several sources of prior information,
- ▶ “weakly informative” or “objective” priors,
- ▶ least favorable priors (cf. minimax),
- ▶ ...

23/37

Example: white balls / red balls (cont'd)

Assume that we have data from a past experiment:

- ▶ sample of $n_0 = 20$ draws,
- ▶ $N_0 = 15$ white balls drawn.

Choice of a prior distribution

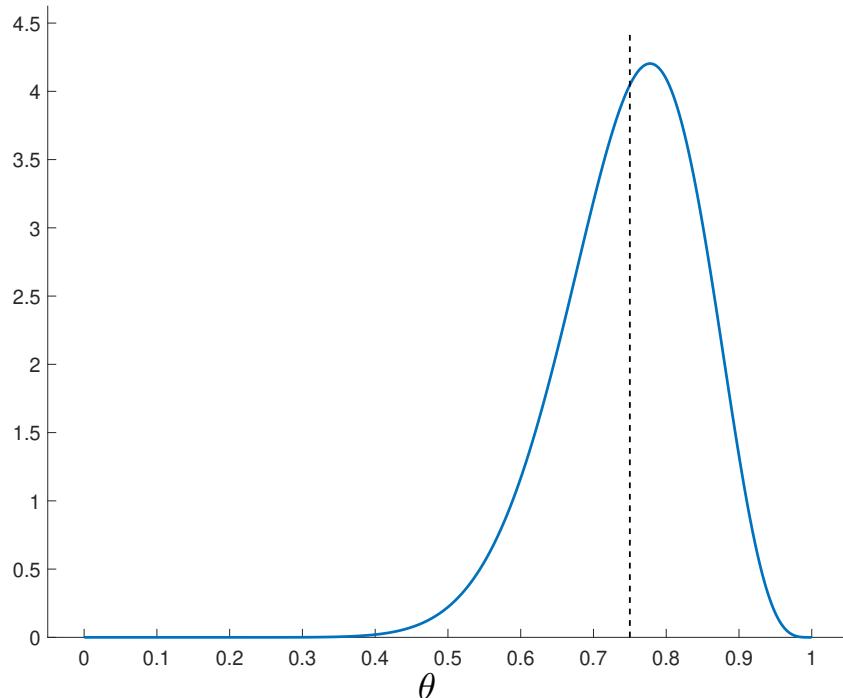
We can decide, e.g., to choose a $\beta(a_0, b_0)$ prior,
with $a_0 = N_0 = 15$ and $b_0 = n_0 - N_0 = 5$.

Arguments in favour of this choice:

- ▶ the shape of the distrib. makes computations easier (see below);
- ▶ expectation : $\frac{a_0}{a_0+b_0} = p_0$, with $p_0 = \frac{N_0}{n_0}$;
- ▶ variance: $\frac{a_0 b_0}{(a_0+b_0)^2(a_0+b_0+1)} \approx \frac{p_0(1-p_0)}{n_0}$ \Rightarrow variance of \bar{X}_{n_0} .

24/37

Example: white balls / red balls (cont'd)



25/37

Example: component reliability

We have the following pieces of information:

- ▶ The manufacturer claims that the lifetime of its components is approximately $\eta_0 = 6$ months.
- ▶ A field expert estimates that the accuracy of the manufacturer's data is roughly $\varepsilon_0 = 10\%$.

Choice of a prior distribution (elicitation)

We can decide, e.g., to choose a $\mathcal{N}(\eta_0, \sigma_0)$ prior, truncated to $[0, +\infty)$, with $\sigma_0 = \varepsilon_0 \eta_0 / 1.96$.

Arguments in favour of this choice:

- ▶ The prior is (approx.) centered on the manufacturer's value η_0 .
- ▶ $\approx 95\%$ of the prior probability is supported by the interval $[0.9\eta_0, 1.1\eta_0]$.
- ▶ The choice of a Gaussian shape and the value 95% are arbitrary.

26/37

Conjugate priors \Rightarrow easier computations !

Families of conjugate prior distributions

A **family of distributions** (densities) is called **conjugate** for a given statistical model if, for any prior π in this family, the posterior $f^{\vartheta|X}$ remains inside the family.

Examples.

- ▶ $\text{Ber}(\theta)$ sample + β prior,
- ▶ $\mathcal{N}(\mu, \sigma^2)$ sample with known σ^2 + \mathcal{N} prior on μ ,
- ▶ $\mathcal{N}(\mu, \sigma^2)$ sample with known μ + \mathcal{IG}^\dagger prior on σ^2 ,
- ▶ $\mathcal{E}(\theta)$ sample + gamma prior,
- ▶ ...

\dagger : inverse gamma. $Z \sim \mathcal{IG}$ if $1/Z$ has a gamma distribution.

27/37

Lecture outline

1 – Introduction: the Bayes risk

2 – Bayesian statistics: prior / posterior distribution

3 – Choosing a prior distribution

4 – Bayes estimators

5 – Warming up exercise

Bayes estimators

Goal

We want to construct estimators of $\eta = g(\theta)$ taking into account

- ▶ the data \underline{x} ,
- ▶ and the prior distribution π .

28/37

Bayes estimators

Let $L : N \times N \rightarrow \mathbb{R}$ be a loss function.

- ▶ Reminder: we “lose” $L(\eta, \tilde{\eta})$ if we estimate $\tilde{\eta}$ when the true value is η .

Definition: Bayesian estimator

A **Bayesian estimator** is an estimator that minimizes the **posterior expected loss**:

$$\hat{\eta} = \arg \min_{\tilde{\eta} \in N} J(\tilde{\eta}, \underline{X})$$

with

$$\begin{aligned} J(\tilde{\eta}, \underline{x}) &= \mathbb{E}(L(g(\vartheta), \tilde{\eta}) \mid \underline{X} = \underline{x}) \\ &= \int_{\Theta} L(g(\theta), \tilde{\eta}) f^{\vartheta|\underline{X}}(\theta \mid \underline{x}) d\theta. \end{aligned}$$

(☞ J is well defined for $\mathbb{P}^{\underline{X}}$ -almost all \underline{x} .)

Remark: equivalently, a Bayesian estimator minimizes the Bayes risk R_{π} .

29/37

Quadratic loss

Consider the quadratic loss function $L(\eta, \tilde{\eta}) = \|\eta - \tilde{\eta}\|^2$:

$$J(\tilde{\eta}, \underline{x}) = \int_{\Theta} \|g(\theta) - \tilde{\eta}\|^2 f^{\vartheta|\underline{X}}(\theta | \underline{x}) d\theta.$$

Proposition

In this case the Bayesian estimator is

$$\hat{\eta} = \mathbb{E}(g(\vartheta) | \underline{X}) = \int_{\Theta} g(\theta) f^{\vartheta|\underline{X}}(\theta | \underline{X}) d\theta.$$

⇒ $\hat{\eta}$ is the posterior mean of ϑ

Remark: it can also be written as

$$\hat{\eta}(\underline{x}) = \frac{\int_{\Theta} g(\theta) f_{\theta}(\underline{x}) \pi(\theta) d\theta}{f^{\underline{X}}(\underline{x})} = \frac{\int_{\Theta} g(\theta) f_{\theta}(\underline{x}) \pi(\theta) d\theta}{\int_{\Theta} f_{\theta}(\underline{x}) \pi(\theta) d\theta}.$$

30/37

Example: white balls / red balls (cont'd)

With a $\beta(a_0, b_0)$ prior on ϑ , we have seen that:

$$\vartheta | \underline{X} \sim \beta(N + a_0, n - N + b_0)$$

with $N = \sum_{i=1}^n X_i$.

The expectation of the $\beta(a, b)$ distribution is $\frac{a}{a+b}$, thus:

$$\hat{\theta} = \mathbb{E}(\vartheta | \underline{X}) = \frac{N + a_0}{n + a_0 + b_0}.$$

Remark: we recover the expressions of $\hat{\theta}_a$ and $\hat{\theta}_b$.

31/37

Another example: Gaussian n -sample (with known σ^2)

It can be proved (see PC #4) that $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma_0^2)$

- ▶ with $\theta \in \mathbb{R}$ (unknown), $\sigma_0 > 0$ (known),
- ▶ and $\vartheta \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2)$,

then

$$\vartheta | \underline{X} \sim \mathcal{N}\left(\frac{\sigma_\theta^2 \sum_{i=1}^n X_i + \sigma_0^2 \mu_\theta}{n\sigma_\theta^2 + \sigma_0^2}, \frac{\sigma_\theta^2 \sigma_0^2}{n\sigma_\theta^2 + \sigma_0^2}\right)$$

Hence the Bayesian estimator (for the quadratic loss):

$$\hat{\theta} = \lambda \bar{X} + (1 - \lambda) \mu_\theta \quad \text{with } \lambda = \frac{n\sigma_\theta^2}{n\sigma_\theta^2 + \sigma_0^2}$$

Interpretation

- ▶ when $n \rightarrow \infty$, $\hat{\theta} \approx \bar{X}$ (the prior no longer has influence)
- ▶ with finite n , when $\frac{\sigma_0}{\sigma_\theta} \gg 1$, $\hat{\theta} \approx \mu_\theta$ (the data is ignored).

32/37

L^1 loss

Assume for simplicity that $\eta = \theta \in \mathbb{R}$.

Consider the loss function $L(\theta, \tilde{\theta}) = |\theta - \tilde{\theta}|$:

$$J(\tilde{\theta}, \underline{x}) = \int_{\Theta} |\theta - \tilde{\theta}| f^{\vartheta | \underline{X}}(\theta | \underline{x}) d\theta.$$

Proposition

In this case the Bayesian estimator $\hat{\theta}$ is such that

$$\int_{-\infty}^{\hat{\theta}} f^{\vartheta | \underline{X}}(\theta | \underline{X}) d\theta = \int_{\hat{\theta}}^{\infty} f^{\vartheta | \underline{X}}(\theta | \underline{X}) d\theta = \frac{1}{2} \quad \mathbb{P}^{\underline{X}}\text{-a.s.}$$

⇒ $\hat{\theta}$ is a **median** of the posterior density of ϑ

Remark: when ϑ has a symmetric posterior density, the two Bayesian estimators (L^1 and L^2 loss) coincide.

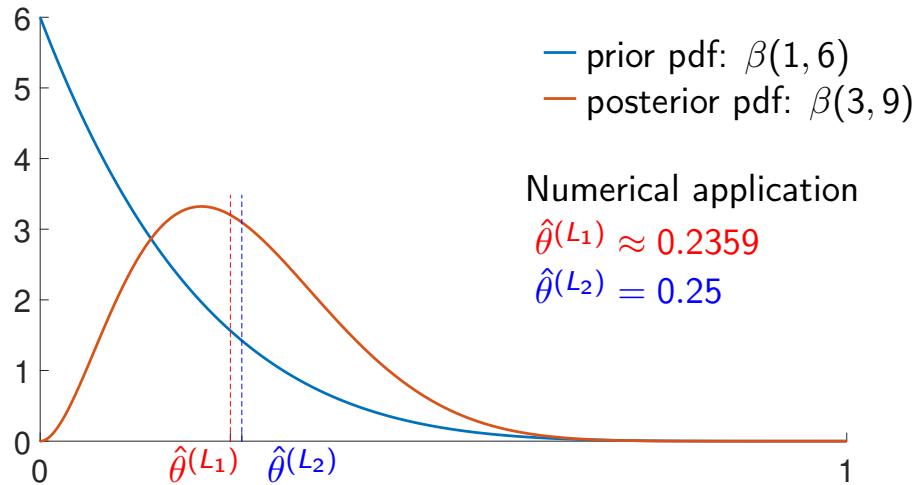
Example: mean of a Gaussian n -sample, with a Gaussian prior.

33/37

Example: white balls / red balls (cont'd)

Observed sample ($n = 5$): $\underline{x} = (W, R, R, W, R)$.

Prior on η : $\vartheta \sim \beta(1, 6)$, with $\theta = \mathbb{P}(X_1 = W)$.



34/37

Lecture outline

1 – Introduction: the Bayes risk

2 – Bayesian statistics: prior / posterior distribution

3 – Choosing a prior distribution

4 – Bayes estimators

5 – Warming up exercise

Exercise (exponential likelihood + gamma prior)

Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{E}(\theta)$ with $\theta \in \Theta = (0, +\infty)$.

We endow θ with a $\text{Gamma}(\alpha_0, \beta_0)$ prior.

Questions

- ❶ Show that the gamma prior is conjugate, and find the parameters α_n and β_n of the posterior distribution.
- ❷ Give the Bayesian estimator of θ , for the quadratic loss.
- ❸ prove that this estimator tends to the MLE when the parameters α_0 and β_0 tend to a certain limit to be specified.

35/37

Solution of exercise 1

Preliminary remark: in this solution we use the same notation, as often done in practice, for the “deterministic” parameter θ and the corresponding random variable, denoted by ϑ in the lecture.

i) First write the likelihood:

$$L(\theta; \underline{x}) = f(\underline{x} | \theta) = \prod_{i=1}^n \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_{i=1}^n x_i},$$

and the prior density:

$$\pi(\theta) = \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} \theta^{\alpha_0-1} e^{-\beta_0 \theta} \propto \theta^{\alpha_0-1} e^{-\beta_0 \theta}.$$

The posterior density then follow from the Bayes formula:

$$f(\theta | \underline{x}) \propto L(\theta; \underline{x}) \pi(\theta) \propto \theta^{\alpha_0+n} e^{-\theta(\beta_0 + \sum_{i=1}^n x_i)}$$

36/37

Solution of exercise 1 (cont'd)

The distribution of θ given \underline{X} , aka posterior distribution, is therefore a gamma distribution with parameters

- ▶ $\alpha_n = \alpha_0 + n$,
- ▶ $\beta_n = \beta_0 + \sum_{i=1}^n X_i$.

ii) The Bayesian estimator for the quadratic loss is given by the posterior expectation of θ given the data:

$$\mathbb{E}(\theta | \underline{X}) = \frac{\alpha_n}{\beta_n} = \frac{\alpha_0 + n}{\beta_0 + \sum_{i=1}^n X_i}.$$

iii) This estimator tends to the MLE $1/\bar{X}_n$ when both α_0 and β_0 tend to zero.

Chapter 5

Hypothesis testing



CentraleSupélec

Statistics and Learning

Arthur Tenenhaus[†], Julien Bect & Laurent Le Brusquet

(firstname.lastname@centralesupelec.fr)

Teaching: CentraleSupélec / Department of Mathematics

Research: Laboratory of signals and systems (L2S)

[†]: Course coordinator

1/43

Lecture 5/10

Hypothesis testing

In this lecture you will learn how to...

- ▶ make (binary) decisions through hypothesis testing,
- ▶ choose and construct a test,
- ▶ define and compute risks of error of the first and second kind.

2/43

Lecture outline

1 – Examples and first definitions

- 1.1 – Two introductory examples
- 1.2 – Risks associated to a test

2 – Parametric tests

- 2.1 – Simple null vs simple alternative
- 2.2 – Composite hypotheses
- 2.3 – Asymptotic tests

3 – Testing for goodness of fit

- 3.1 – Pearson's χ^2 test
- 3.2 – BONUS: Kolmogorov-Smirnov test

5 – Warming up exercise

3/43

Lecture outline

1 – Examples and first definitions

- 1.1 – Two introductory examples
- 1.2 – Risks associated to a test

2 – Parametric tests

- 2.1 – Simple null vs simple alternative
- 2.2 – Composite hypotheses
- 2.3 – Asymptotic tests

3 – Testing for goodness of fit

- 3.1 – Pearson's χ^2 test
- 3.2 – BONUS: Kolmogorov-Smirnov test

5 – Warming up exercise

Lecture outline

1 – Examples and first definitions

- 1.1 – Two introductory examples
- 1.2 – Risks associated to a test

2 – Parametric tests

- 2.1 – Simple null vs simple alternative
- 2.2 – Composite hypotheses
- 2.3 – Asymptotic tests

3 – Testing for goodness of fit

- 3.1 – Pearson's χ^2 test
- 3.2 – BONUS: Kolmogorov-Smirnov test

5 – Warming up exercise

Example: component reliability

Reminder: $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{E}(\theta)$, $\theta > 0$.

Problem

The manufacturer want to propose a one-year warranty...

⇒ is it a good idea ?

Formalization

The manufacturer considers that it is a "good idea" if:

the return rate is lower than 10%

\Updownarrow

$$\mathbb{P}_\theta(X_1 \leq 1) = 1 - \exp(-\theta) < 0.1$$

\Updownarrow

$$\theta < \theta_0 = -\ln(0.9)$$

Example: component reliability

Therefore, the manufacturer wants to know if $\theta < \theta_0$ or $\theta \geq \theta_0$.

- ➡ hypothesis to be tested: $H_0 : \theta \geq \theta_0$
(component quality is not sufficient)

Making (binary) decisions from data

We want to evaluate the “compatibility” between H_0 and \underline{x} :

- ▶ if a strong incompatibility is detected,
➡ H_0 is rejected (and the warranty proposed);
- ▶ otherwise, H_0 is accepted.

Note the asymmetry between the two scenarios
(H_0 = is retained by default)

Hypothesis tests make it possible to formalize this decision making.

5/43

Another example / construction of a first test

Goal: test the mean parameter of a Gaussian distribution.

- ▶ $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma_0^2)$ (σ_0 known; $n = 10$, $\sigma_0 = 2.5$)
- ▶ hypothesis to be tested $\rightarrow H_0 : \theta = \theta_0$ (fixed),
- ▶ alternative hypothesis $\rightarrow H_1 : \theta = \theta_1$ (fixed, and s.t. $\theta_0 < \theta_1$).

Approach. Making a decision about H_0 means estimating if it is

- ▶ either true ➡ $\delta = 0$,
- ▶ or false ➡ $\delta = 1$.

Constraint. We want δ to be such that, if $\theta = \theta_0$ (H_0 true),

$$\mathbb{P}_{\theta_0}(\delta = 1) = 5\% (= \alpha).$$

Intuitive construction of a test: $\delta = \mathbb{1}_{\bar{X} > t}$

- ▶ where t is such that $\mathbb{P}_{\theta_0}(\delta = 1) = \mathbb{P}_{\theta_0}(\bar{X} > t) = 5\%$.

6/43

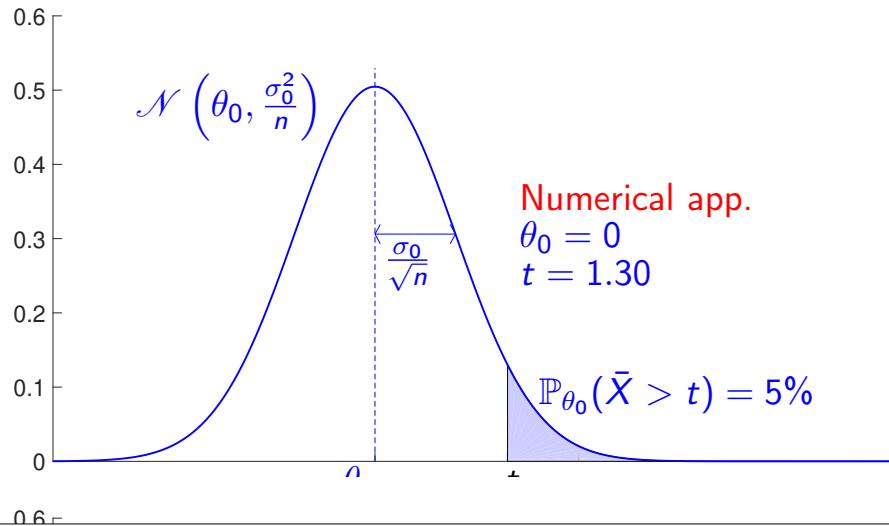
If H_0 is true ($\theta = \theta_0$) H_1 is true ($\theta = \theta_1$): $\bar{X} \sim \mathcal{N}\left(\theta_0, \frac{\sigma_0^2}{n}\right)$,
therefore

$$t = \theta_0 + q_{0.95} \frac{\sigma_0}{\sqrt{n}}$$

where q_r is the $\mathcal{N}(0, 1)$ quantile of order r .

$$\mathbb{P}_{\theta_1}(\delta = 0) = \mathbb{P}_{\theta_1}(\bar{X} \leq t) = \Phi\left(\frac{t - \theta_1}{\sigma_0/\sqrt{n}}\right)$$

where Φ is the cdf of the $\mathcal{N}(0, 1)$ distribution.



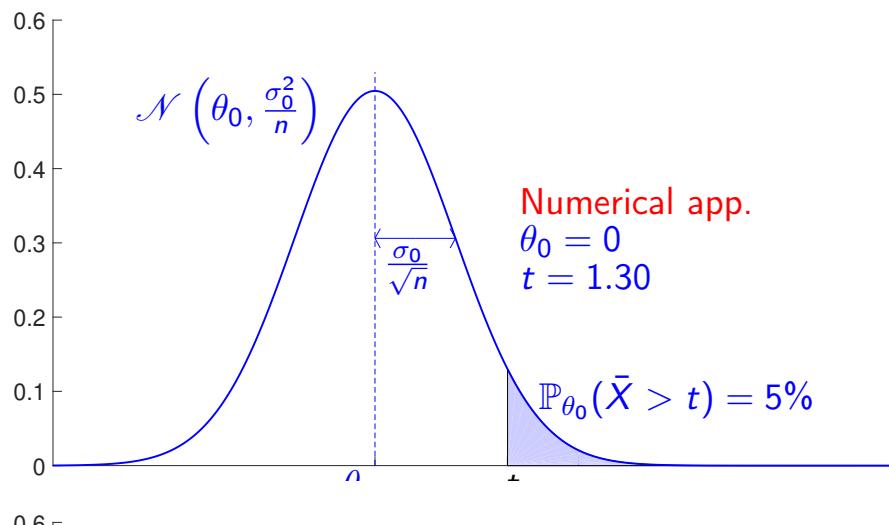
If H_0 is true ($\theta = \theta_0$) H_1 is true ($\theta = \theta_1$): $\bar{X} \sim \mathcal{N}\left(\theta_0, \frac{\sigma_0^2}{n}\right)$,
therefore

$$t = \theta_0 + q_{0.95} \frac{\sigma_0}{\sqrt{n}}$$

where q_r is the $\mathcal{N}(0, 1)$ quantile of order r .

$$\mathbb{P}_{\theta_1}(\delta = 0) = \mathbb{P}_{\theta_1}(\bar{X} \leq t) = \Phi\left(\frac{t - \theta_1}{\sigma_0/\sqrt{n}}\right)$$

where Φ is the cdf of the $\mathcal{N}(0, 1)$ distribution.



Vocabulary: a first overview

- ▶ hypothesis H_0 : **null hypothesis**
- ▶ hypothesis H_1 : **alternative hypothesis**

- ▶ α : **(significance) level** at which we want to test
- ▶ $\mathbb{P}_{\theta_0}(\delta = 1)$: **risk of the first kind** (or risk of type I error)
- ▶ $\mathbb{P}_{\theta_1}(\delta = 0)$: **risk of the second kind** (... of type II error)
- ▶ $\mathbb{P}_{\theta_1}(\delta = 1)$: **power** of the test

- ▶ $\mathcal{R}_\delta = \{\underline{x} \in \mathcal{X} \text{ tel que } \delta(\underline{x}) = 1\}$: **critical region**
(a.k.a. rejection region) of the test

- ▶ for a test written as: $\delta(\underline{x}) = 1 \iff T(\underline{x}) > t$,
 - ▶ T is the (scalar) **test statistic**,
 - ▶ $t \in \mathbb{R}$ is the **critical value** of this statistic.

9/43

Lecture outline

1 – Examples and first definitions

- 1.1 – Two introductory examples
- 1.2 – Risks associated to a test

2 – Parametric tests

- 2.1 – Simple null vs simple alternative
- 2.2 – Composite hypotheses
- 2.3 – Asymptotic tests

3 – Testing for goodness of fit

- 3.1 – Pearson's χ^2 test
- 3.2 – BONUS: Kolmogorov-Smirnov test

5 – Warming up exercise

How to formulate an hypothesis testing problem

Recall that we have a statistical model, which parameterized by θ :

$$\mathcal{P}^X = \left\{ \mathbb{P}_\theta^X, \theta \in \Theta \right\}.$$

Statistical hypothesis

A **statistical hypothesis** is represented by a subset of \mathcal{P}^X , and thus by a subset of Θ .

Notation. Let $\Theta_j \subset \Theta$ denote the subset representing H_j

$$\Rightarrow H_j : \theta \in \Theta_j$$

Parametric / non-parametric test

A testing problem is called parametric if Θ is finite-dimensional.

10/43

How to formulate an hypothesis testing problem (cont'd)

Null hypothesis

We call **null hypothesis** the hypothesis $H_0 : \theta \in \Theta_0$

- ▶ that we “want to test”, and
- ▶ that will be **retained “by default”** unless it is clearly at odds with the data.

Legal analogy: presumption of innocence

Alternative hypothesis

We call **alternative hypothesis** the hypothesis $H_1 : \theta \in \Theta_1$

- ▶ that will be **chosen if H_0 is rejected**.
- ▶ We assume that $\Theta_1 \cap \Theta_0 = \emptyset$.

Remark : we can assume wlog that $\Theta_0 \cup \Theta_1 = \Theta$.

11/43

Examples of parametric tests

Example 1.

- ▶ $X_1, X_2, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{E}(\theta)$, with $\theta \in \Theta = [0, +\infty[$,
- ▶ $\Theta_0 = \{\theta \geq \theta_0\}$; $\Theta_1 = \{\theta < \theta_0\}$ with $\theta_0 > 0$ a given threshold.
 ■■■ cf. component reliability example.

Example 2. Same example, with :

- ▶ $\Theta_0 = \{\theta_0\}$ (**singleton**) ; $\Theta_1 = \{\theta \neq \theta_0\}$,
- ▶ or $\Theta_0 = \{\theta_0\}$; $\Theta_1 = \{\theta < \theta_0\}$.

Definitions: simple / composite hypotheses

An hypothesis H_j is called **simple** if Θ_j is a singleton.
 It is called **composite** otherwise.

12/43

Other examples of (non-parametric) tests

Goodness-of-fit tests

for a distribution or family of distributions

- ▶ see Section 3

Other types of tests

- ▶ testing the independence of two variables
- ▶ testing the symmetry of a distribution
- ▶ ...

13/43

Test procedures

Definition: test (procedure)

A **test** is a statistic $\delta = \delta(\underline{X})$ with values in $\{0, 1\}$:

$$\begin{aligned}\delta : \quad \underline{\mathcal{X}} &\mapsto \{0, 1\}, \\ \underline{x} &\rightarrow \begin{cases} 0 & \text{if } H_0 \text{ is accepted,} \\ 1 & \text{if it is rejected (in favour of } H_1). \end{cases}\end{aligned}$$

Definition: critical region of a test

The **critical region** \mathcal{R}_δ of a test δ is the region of rejection

$$\mathcal{R}_\delta = \{\underline{x} \in \underline{\mathcal{X}} \text{ such that } \delta(\underline{x}) = 1\}.$$

14/43

Quantifying the risks of error

Definition: risk of the first kind

We call **risk of the first kind**, or **risk of type I error**, the probability to reject H_0 when it is true :

$$\mathbb{P}_\theta(\delta = 1) = \mathbb{E}_\theta(\delta), \quad \theta \in \Theta_0.$$

(Δ This risk depends on the value of θ , for $\theta \in \Theta_0$.)

Definition: risk of the second kind

We call **risk of the second kind**, or **risk of type II error**, the probability to accept H_0 when it is false :

$$\mathbb{P}_\theta(\delta = 0) = 1 - \mathbb{E}_\theta(\delta), \quad \theta \in \Theta_1.$$

(Note the asymmetry of terminology
 \rightarrow more emphasis is put on H_0 .)

15/43

Definition: power of a test

We call **power** the probability to reject H_0 when it is wrong:

$$\mathbb{P}_\theta(\delta = 1) = \mathbb{E}_\theta(\delta), \quad \theta \in \Theta_1.$$

Remark: equal to “1 - risk of type II error”.

Usual approach[†] for the construction of tests.

Let $0 < \alpha < 1$ be a level of risk. We will look for tests s.t.

- ▶ $\forall \theta \in \Theta_0, \mathbb{P}_\theta(\delta = 1) \leq \alpha$;
⇒ control of the risk of type I errors.
The test δ is said to have **level (at most) α** .
- ▶ $\forall \theta \in \Theta_1, \mathbb{P}_\theta(\delta = 1)$ “as large as possible”;
⇒ capacity to reject H_0 when it is false.

Typical values: $\alpha = 5\%, 1\%, 1\% \dots$

† a.k.a. Neyman's

16/43

Definition: size of a test

We say that δ has **level exactly α** , or **size α** , if

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\delta = 1) = \alpha.$$

Definition: comparing two tests

Let δ and δ' be two tests with level (at most) α . We say that δ' is **uniformly more powerful** than δ if

$$\forall \theta \in \Theta_1, \quad \mathbb{P}_\theta(\delta' = 1) \geq \mathbb{P}_\theta(\delta = 1).$$

(Some authors require a strict inequality at one or all $\theta \in \Theta_1$.)

Remarks :

- ▶ this is a **partial order** on power functions,
- ▶ whenever possible, we will look for the **uniformly most powerful test at level α** (i.e., a test with α , that is uniformly more powerful than all other tests with level α).

17/43

Lecture outline

- 1 – Examples and first definitions
 - 1.1 – Two introductory examples
 - 1.2 – Risks associated to a test
- 2 – Parametric tests
 - 2.1 – Simple null vs simple alternative
 - 2.2 – Composite hypotheses
 - 2.3 – Asymptotic tests
- 3 – Testing for goodness of fit
 - 3.1 – Pearson's χ^2 test
 - 3.2 – BONUS: Kolmogorov-Smirnov test
- 5 – Warming up exercise

Lecture outline

- 1 – Examples and first definitions
 - 1.1 – Two introductory examples
 - 1.2 – Risks associated to a test
- 2 – Parametric tests
 - 2.1 – Simple null vs simple alternative
 - 2.2 – Composite hypotheses
 - 2.3 – Asymptotic tests
- 3 – Testing for goodness of fit
 - 3.1 – Pearson's χ^2 test
 - 3.2 – BONUS: Kolmogorov-Smirnov test
- 5 – Warming up exercise

Likelihood ratio test

Assume **two simple hypotheses** : $\Theta_0 = \{\theta_0\}$ et $\Theta_1 = \{\theta_1\}$.

Denote by $\mathcal{L} : (\theta, \underline{x}) \mapsto \mathcal{L}(\theta, \underline{x})$ the **likelihood function**[†].

Definition: likelihood ratio test

We call **likelihood ratio test** the test

$$\delta^{\text{LR}} = \begin{cases} 1 & \text{if } T > t, \\ 0 & \text{otherwise,} \end{cases}$$

built using the **likelihood ratio statistic**:

$$T = \frac{\mathcal{L}(\theta_1, \underline{X})}{\mathcal{L}(\theta_0, \underline{X})}.$$

[†] It can be proved that the family $\{\mathbb{P}_{\theta_0}^X, \mathbb{P}_{\theta_1}^X\}$ is always dominated (ex. facultatif / Radon-Nikodym).

Fundamental result

Let $\alpha \in (0, 1)$.

Theorem: Neyman-Pearson “lemma”

Assume that there **exists**^{*} a threshold t_α such that

- ▶ the associated LR test δ^{LR} has **level exactly α** (i.e., has size α).

Then δ^{LR} is **most powerful**[†] at the level α :

- ▶ for any test $\tilde{\delta}$ with level (at most) α , δ^{LR} is more powerful than $\tilde{\delta}$.

► The LR test is **optimal** in this setting.

^{*} Always true if the cdf of T is continuous.

[†] No need to specify “uniformly” since H_1 is simple.

Back to the Gaussian example

Likelihood ratio test :

$$\begin{aligned} T &= \frac{\frac{1}{(\sqrt{2\pi}\sigma_0)^n} \exp\left(-\frac{\sum_{i=1}^n (X_i - \theta_1)^2}{2\sigma_0^2}\right)}{\frac{1}{(\sqrt{2\pi}\sigma_0)^n} \exp\left(-\frac{\sum_{i=1}^n (X_i - \theta_0)^2}{2\sigma_0^2}\right)} \\ &= \exp\left(-\frac{n(\theta_1^2 - \theta_0^2)}{2\sigma_0^2}\right) \exp\left(\frac{(\theta_1 - \theta_0)}{\sigma_0^2} \sum_{i=1}^n X_i\right). \end{aligned}$$

$$\theta_1 > \theta_0 \text{ therefore } \delta = 1 \iff T > t \iff \sum_{i=1}^n X_i > c$$

⇒ the test that was previously constructed is optimal.

20/43

Test statistic and p-value

The result of a test can be expressed using the concept of **p-value**.

Definition: p-value

Let T be the test statistic of a test of the form $\delta = \mathbb{1}_{T > t_\alpha}$.

Definition. We call **p-value** the **statistic**

$$\text{pval}(\underline{x}) = \mathbb{P}_{\theta_0}(T(\underline{X}) > T(\underline{x}))$$

taking values in $(0, 1)$.

⚠ Function of the data!

Let F_0 denote the cdf of T under H_0 . Then:

$$\text{pval}(\underline{x}) = 1 - F_0(T(\underline{x})).$$

21/43

Interpretation of the p-value

Assume that F_0 is continuous and strictly increasing:

$$\forall \alpha \in (0, 1), \quad \exists! t_\alpha \in \mathbb{R}, \quad \delta = \mathbb{1}_{T > t_\alpha} \text{ has level exactly } \alpha$$

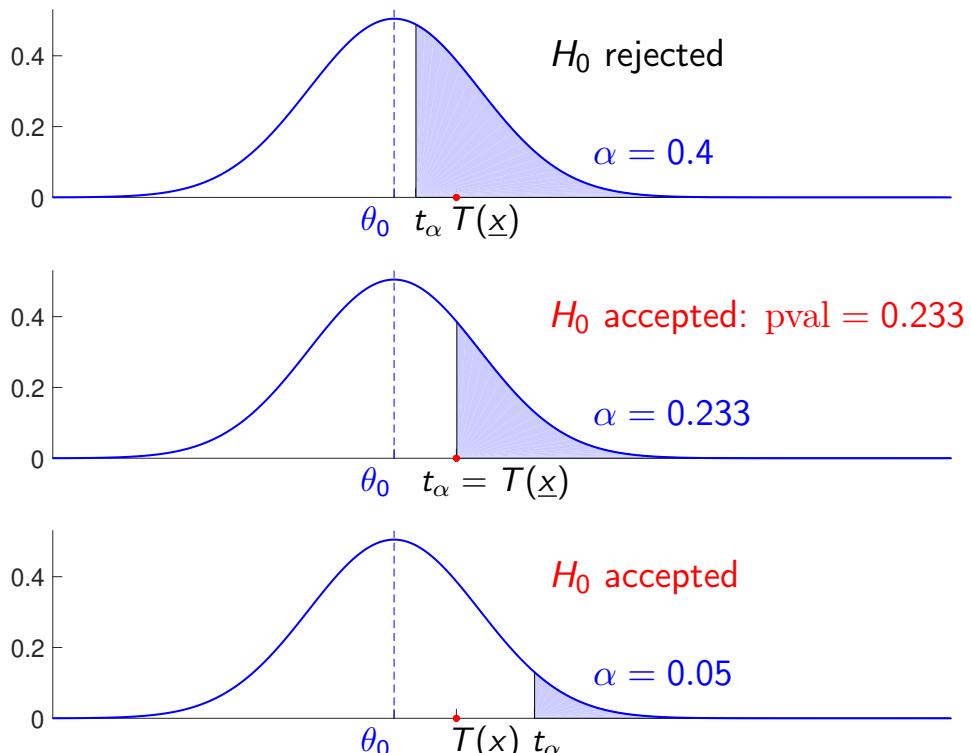
Proposition

$$H_0 \text{ is rejected at the level } \alpha \iff T > t_\alpha \iff \text{pval} < \alpha.$$

Interpretation: p-value = measure of the “evidence” against H_0 .

p-value	evidence against H_0
$\text{pval} < 0.01$	strong evidence
$0.01 \leq \text{pval} < 0.1$	weak evidence
$0.1 < \text{pval}$	no evidence

22/43



(pval is the maximal level α at which H_0 is accepted.)

23/43

Proof

Note that t_α is, by construction, such that

$$F_0(t_\alpha) = 1 - \alpha.$$

Thus we have

$$\begin{aligned}\delta = 1 &\iff T > t_\alpha \\ &\iff F_0(T) > F_0(t_\alpha) = 1 - \alpha \\ &\iff \text{pval} < \alpha\end{aligned}$$

□

Lecture outline

1 – Examples and first definitions
 1.1 – Two introductory examples
 1.2 – Risks associated to a test

2 – Parametric tests
 2.1 – Simple null vs simple alternative
 2.2 – Composite hypotheses
 2.3 – Asymptotic tests

3 – Testing for goodness of fit
 3.1 – Pearson's χ^2 test
 3.2 – BONUS: Kolmogorov-Smirnov test

5 – Warming up exercise

Examples of problems with composite hypotheses

Simple null / composite alternative

- ▶ $\Theta_0 = \{\theta_0\}$ / $\Theta_1 = \{\theta > \theta_0\}$ (one-sided test),
- ▶ $\Theta_0 = \{\theta_0\}$ / $\Theta_1 = \{\theta \neq \theta_0\}$ (two-sided test),
- ▶ ...

Composite null / composite alternative

- ▶ $\Theta_0 = \{\theta \leq \theta_0\}$ / $\Theta_1 = \{\theta > \theta_0\}$ (one-sided test),
- ▶ $\Theta_0 = \{\mu = \mu_0\}$ / $\Theta_1 = \{\mu = \mu_1\}$,
where $\theta = (\mu, \sigma^2)$ with unknown σ^2 (nuisance parameter),
- ▶ $\Theta_0 = \{\theta^{(1)} = \theta^{(2)}\}$ / $\Theta_1 = \{\theta^{(1)} \neq \theta^{(2)}\}$,
where $\theta \in \Theta = \mathbb{R}^2$ (equality of two parameters),
- ▶ ...

24/43

Differences with the case of simple hypotheses

- ▶ Test with level (at most) α , when Θ_0 is composite :

$$\forall \theta \in \Theta_0, \mathbb{P}_\theta(\delta = 1) \leq \alpha \iff \underbrace{\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\delta = 1)}_{\text{size of the test}} \leq \alpha.$$

- ▶ If Θ_1 is composite, the **power** is a function of $\theta \in \Theta_1$:

$$\begin{aligned} \Theta_1 &\rightarrow [0, 1] \\ \theta &\mapsto \mathbb{P}_\theta(\delta = 1). \end{aligned}$$

25/43

Differences with the case of simple hypotheses (cont'd)

- ▶ Generalized likelihood ratio test

- ▶ Test statistic :

$$T(\underline{X}) = \frac{\sup_{\theta \in \Theta_1} \mathcal{L}(\theta; \underline{X})}{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta; \underline{X})}.$$

- ▶ The test is not, in general, uniformly most powerful (UMP) at level α .

- ▶ p-value for a test of the form $\delta = \mathbb{1}_{T > t_\alpha}$:

$$\text{pval} = \sup_{\theta \in \Theta_0} (1 - F_\theta(T)).$$

where F_θ is the cdf of T under \mathbb{P}_θ .

26/43

Back to the Gaussian example / testing the mean

Case 1. $H_0 : \theta = \theta_0$ / $H_1 : \theta = \theta_1$, with $\theta_0 < \theta_1$

- ▶ Recall the optimal test:

$$\delta(\underline{X}) = 1 \iff \bar{X} > t_\alpha \text{ with } t_\alpha = \theta_0 + q_{1-\alpha} \frac{\sigma_0}{\sqrt{n}}$$

Analysis of the optimal test δ

- ▶ δ is the same for any $\theta_1 > \theta_0$ (it only depends on α and θ_0);
- ▶ $\theta \mapsto \mathbb{P}_\theta(\delta = 1) = 1 - \Phi\left(\frac{t_\alpha - \theta}{\sigma_0/\sqrt{n}}\right)$ is increasing, therefore $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\delta = 1)$ is attained at θ_0 in the following cases:

Case 2. $H_0 : \theta = \theta_0$ / $H_1 : \theta > \theta_0$

Case 3. $H_0 : \theta \leq \theta_0$ / $H_1 : \theta > \theta_0$

Conclusion on cases 2 and 3

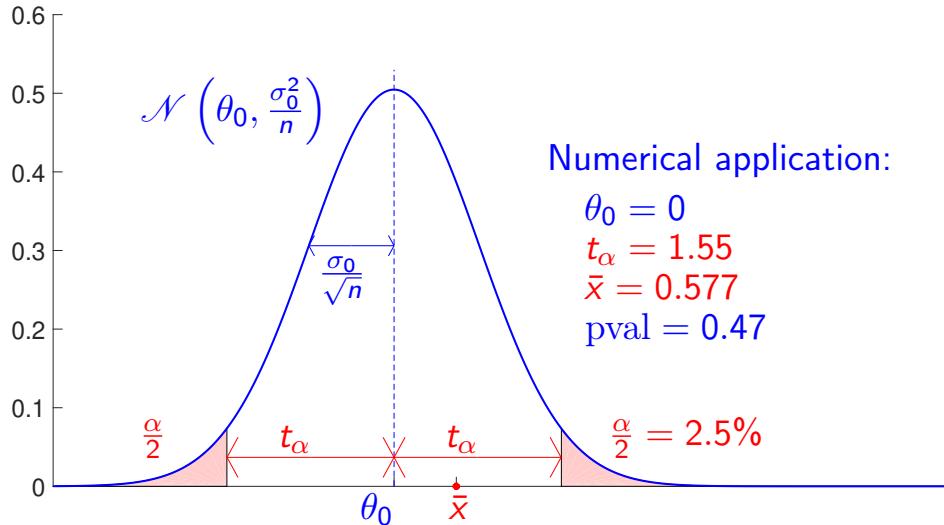
- ▶ δ has level exactly α ,
- ▶ δ is UMP at the level α .

27/43

Case 4 (two-sided test): $H_0 : \Theta_0 = \{\theta_0\} / H_1 : \Theta_1 = \{\theta \neq \theta_0\}$

Idea[†]: use $T(\underline{X}) = |\bar{X} - \theta_0|$

⇒ H_0 is rejected when $T(\underline{X}) > t_\alpha$, with $t_\alpha = \frac{\sigma_0}{\sqrt{n}} q_{1-\frac{\alpha}{2}}$.



[†] Exercise: Show that this is the generalized LR test when $\sigma^2 = \sigma_0^2$ is known.

28/43

Example: component reliability (cont'd)

Reminder: $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{E}(\theta)$

$H_0 : \Theta_0 = \{\theta \geq \theta_0\}$ (component is not reliable enough)
 $H_1 : \Theta_1 = \{\theta < \theta_0\}$ (component is reliable enough)

Likelihood ratio test.

$H_0 : \Theta_0 = \{\theta_0\} / H_1 : \Theta_1 = \{\theta_1\}$ with $\theta_1 < \theta_0$

$$\begin{aligned} T^{\text{LR}}(\underline{X}) &= \frac{\theta_1^n \exp(-\theta_1 \sum_{i=1}^n X_i)}{\theta_0^n \exp(-\theta_0 \sum_{i=1}^n X_i)} \\ &= \left(\frac{\theta_1}{\theta_0} \right)^n \exp((\theta_0 - \theta_1) \sum_{i=1}^n X_i) \end{aligned}$$

Example: component reliability (cont'd)

Critical (rejection) zone of the LR test at level α :

$$\mathcal{R}_\alpha = \left\{ \underline{x} \text{ tel que } T^{\text{LR}}(\underline{x}) > t_\alpha^{\text{LR}} \right\} = \left\{ \underline{x} \text{ tel que } T(\underline{x}) = \bar{x} > t_\alpha \right\}.$$

Reminder : if $\theta = \theta_0$, then $\theta_0 \bar{X} \sim \Gamma(p = n, \lambda = n)$.

⇒ $t_{\alpha,n} = \frac{1}{\theta_0} q_{1-\alpha}$

where q_r is the $\Gamma(p = n, \lambda = n)$ quantile of order r .

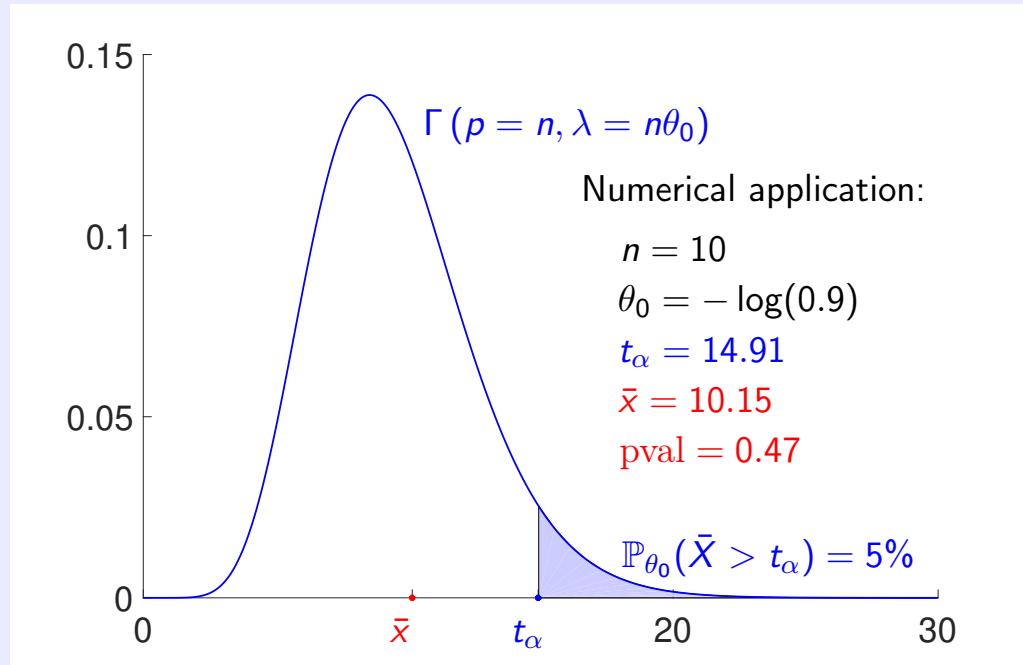
Analysis (similar to previous example)

- ▶ the LR test is **the same for any $\theta_1 < \theta_0$** ,
- ▶ the function $\theta \mapsto \mathbb{P}_\theta(\delta = 1)$ is strictly ↘.

Summary. The test that we have built is **UMP at the level α** .

Remark: same principle for any one-sided test on this model.

Example: component reliability (cont'd)



- ⇒ at the 5% level, H_0 is not rejected
- ⇒ out of precaution, the manufacturer will not propose a warranty

Lecture outline

1 – Examples and first definitions

- 1.1 – Two introductory examples
- 1.2 – Risks associated to a test

2 – Parametric tests

- 2.1 – Simple null vs simple alternative
- 2.2 – Composite hypotheses
- 2.3 – Asymptotic tests

3 – Testing for goodness of fit

- 3.1 – Pearson's χ^2 test
- 3.2 – BONUS: Kolmogorov-Smirnov test

5 – Warming up exercise

Context : $X_1, X_2, \dots \stackrel{\text{iid}}{\sim} P_\theta$

When distribution of $T_n(\underline{X}_n)$ is hard to determine

⇒ use of the limit distribution for $n \rightarrow \infty$.

Example: component reliability

$$\mathcal{R}_{\alpha,n} = \{\underline{x}_n \text{ such that } T_n(\underline{x}_n) = \bar{x}_n > t_{\alpha,n}^\infty\}.$$

with $t_{\alpha,n}^\infty$ chosen in such a way that :

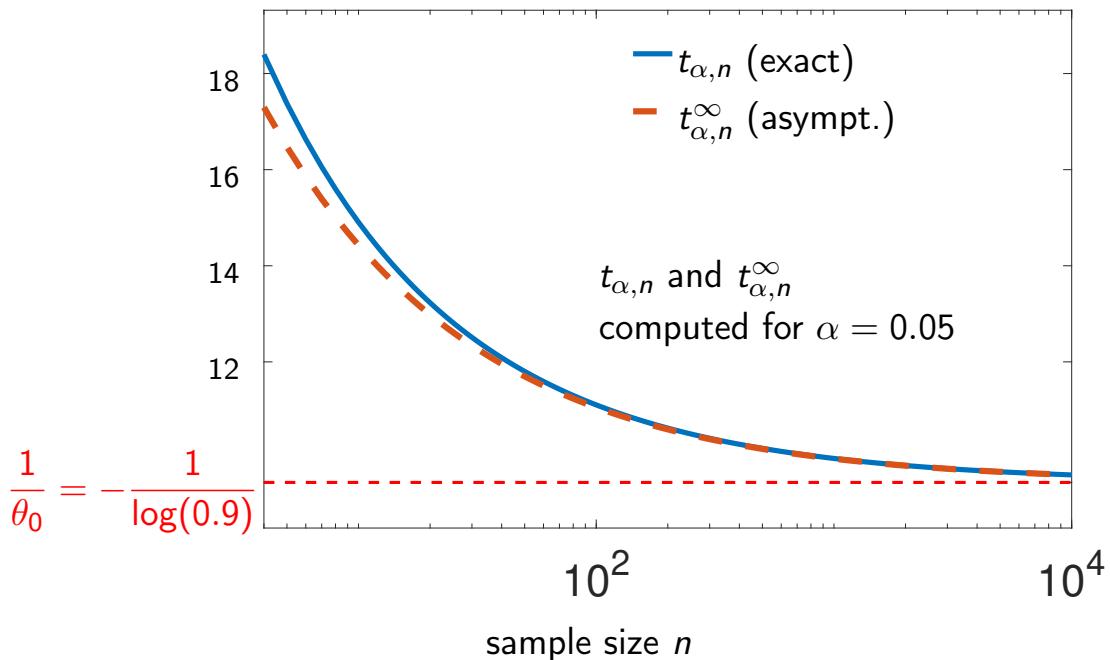
$$\lim_{n \rightarrow \infty} \mathbb{P}_{\theta_0} (T_n(\underline{X}_n) > t_{\alpha,n}^\infty) = \alpha.$$

By the CLT under H_0 : $\sqrt{n} \left(\bar{X}_n - \frac{1}{\theta_0} \right) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N} \left(0, \frac{1}{\theta_0^2} \right)$, therefore

$$t_{\alpha,n}^\infty = \frac{1}{\theta_0} + \frac{1}{\theta_0 \sqrt{n}} q_{1-\alpha}$$

where q_r is the $\mathcal{N}(0, 1)$ quantile of order r .

Example: component reliability (cont'd)



30 / 43

Lecture outline

- 1 – Examples and first definitions
 - 1.1 – Two introductory examples
 - 1.2 – Risks associated to a test
- 2 – Parametric tests
 - 2.1 – Simple null vs simple alternative
 - 2.2 – Composite hypotheses
 - 2.3 – Asymptotic tests
- 3 – Testing for goodness of fit
 - 3.1 – Pearson's χ^2 test
 - 3.2 – BONUS: Kolmogorov-Smirnov test
- 5 – Warming up exercise

Lecture outline

1 – Examples and first definitions

- 1.1 – Two introductory examples
- 1.2 – Risks associated to a test

2 – Parametric tests

- 2.1 – Simple null vs simple alternative
- 2.2 – Composite hypotheses
- 2.3 – Asymptotic tests

3 – Testing for goodness of fit

- 3.1 – Pearson's χ^2 test
- 3.2 – BONUS: Kolmogorov-Smirnov test

5 – Warming up exercise

Goodness-of-fit test for a single distribution

Context: $X_1, X_2, \dots \stackrel{\text{iid}}{\sim} P$ with **unknown** P (can be anything);

► $\theta = P$, $\Theta = \{ \text{probability distributions on } (\mathbb{R}, \mathcal{B}(\mathbb{R})) \}$.

Statistical hypotheses to be tested

For a given probability P_0 , we consider the hypotheses:

$$\begin{aligned} H_0 : P &= P_0 \\ H_1 : P &\neq P_0 \end{aligned}$$

Component reliability example:

- The component manufacturer knows, from past analyses, that the component lifetimes should follow a $\mathcal{E}(\theta_0)$ distribution.
- In order to check that the production line is still properly working, he wants to test if $H_0 : P = \mathcal{E}(\theta_0)$ is still true.

Pearson's χ^2 test statistic

Let (A_1, \dots, A_K) be a partition of P_0 's support, and

► $N = (N_1, \dots, N_K)$ with

$$N_k = \sum_{i=1}^n \mathbb{1}_{A_k}(X_i) \rightarrow \text{observed frequencies (counts)},$$

► $p = (p_1, \dots, p_K)$ with

$$p_k = P_0(X_1 \in A_k) \rightarrow np_k = \text{expected frequencies under } H_0.$$

Proposition

Under hypothesis H_0 , N follows a **multinomial** $\text{Multi}(n, p)$ distribution, and

$$T_n = \sum_{k=1}^K \frac{(N_k - np_k)^2}{np_k} \xrightarrow[n \rightarrow \infty]{d} \chi^2(K-1)$$

$(\chi^2 \text{ distribution with } K-1 \text{ degrees of freedom})$

32/43

Pearson's chi-squared test (χ^2)

Recall that we want to test $H_0 : P = P_0$ against $H_1 : P \neq P_0$.

Chi-square (χ^2) goodness-of-fit test

Let $0 < \alpha < 1$ and let T denote Pearson's statistic:

$$T = \sum_{k=1}^K \frac{(N_k - np_k)^2}{np_k}.$$

The chi-squared (χ^2) test is

$$\delta = \mathbb{1}_{T > t_\alpha},$$

where t_α is the $\chi^2(K-1)$ quantile of order $1-\alpha$.



In practice: choose A_1, \dots, A_K such that $np_k \geq 5, \forall k$.

33/43

The multinomial family of distributions

Parameters

- ▶ n integer, ≥ 1 ,
- ▶ K integer, ≥ 2 and $p \in (\mathbb{R}_+^*)^K$ such that $\sum_{k=1}^K p_k = 1$.

Let n_1, \dots, n_K entiers ≥ 0 such that $\sum_{k=1}^K n_k = n$:

$$\text{If } N \sim \text{Multi}(n, p), \mathbb{P}(N_1 = n_1, \dots, N_K = n_K) = \frac{n!}{n_1! \dots n_K!} p_1^{n_1} \dots p_K^{n_K}$$

Moments

- ▶ expectation : $\mathbb{E}_p(N) = np$
- ▶ covariance matrix : $\text{cov}_p(N_i, N_j) = n(p_i \delta_{ij} - p_i p_j)$

Marginal distributions

- ▶ Marginal distributions are binomial : $N_j \sim \mathcal{B}(n, p_j)$.

The χ^2 family of distributions

Parameters

- ▶ q integer, ≥ 1 : number of “degrees of freedom”.

Definition. If $Y_1, \dots, Y_q \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1)$ then

$$T = \sum_{k=1}^q Y_k^2 \sim \chi^2(q)$$

The χ^2 distribution is a **special case of the Γ distribution** :

$$\chi^2(q) = \Gamma\left(p = \frac{q}{2}, \lambda = \frac{1}{2}\right)$$

► The properties of the χ^2 follow from those of the Γ distribution.

Expectation

- ▶ $\mathbb{E}_q(T) = q$

Variance

- ▶ $\text{var}_q(T) = 2q$

Goodness-of-fit test to a family for distributions

Component reliability example

Goal: test if the lifetimes are exponentially distributed.

⇒ Null hypothesis $H_0: \exists \theta > 0, P = P_\theta = \mathcal{E}(\theta)$.

Two-step approach

- ① Estimate θ from the data → $\hat{\theta}$.
- ② Test the goodness of fit to $P_{\hat{\theta}}$.

Details

$$\hat{p}_k = P_{\hat{\theta}}(X_1 \in A_k)$$

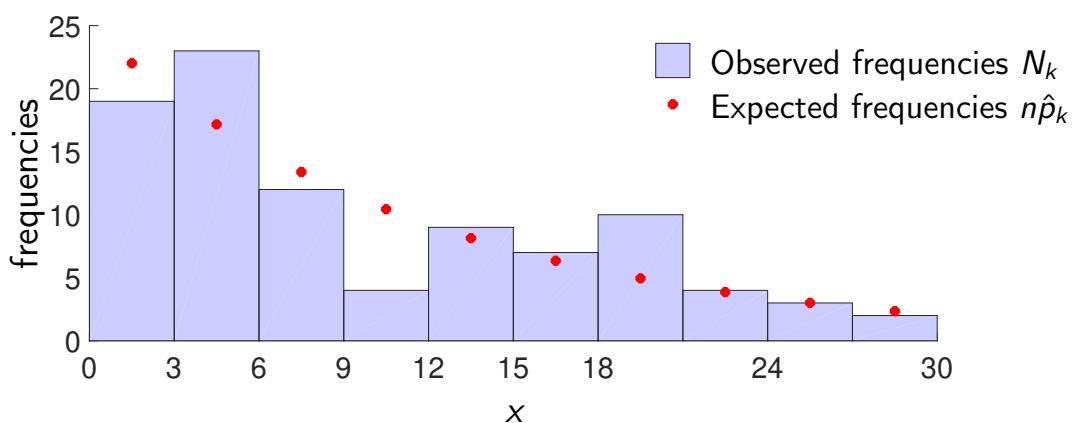
$$T(\underline{X}_n) = \sum_{k=1}^K \frac{(N_k - n\hat{p}_k)^2}{n\hat{p}_k} \xrightarrow[n \rightarrow \infty]{d} \chi^2(K - 1 - q) \text{ with } q = \text{card}(\theta)$$

H_0 is rejected if $T(\underline{X}_n) > t_\alpha$

with t_α the $\chi^2(K - 1 - q)$ quantile of order $1 - \alpha$.

34 / 43

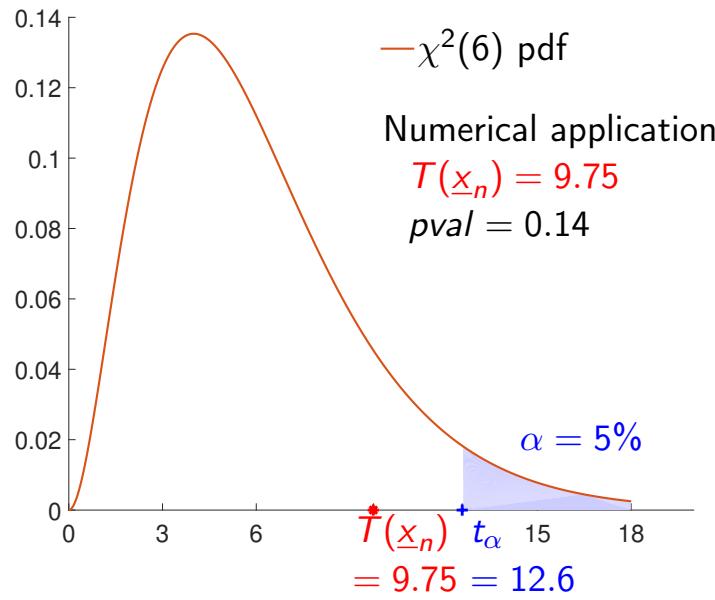
Example: component reliability



$$T(\underline{X}_n) = \sum_{k=1}^8 \frac{(N_k - n\hat{p}_k)^2}{n\hat{p}_k} \xrightarrow[n \rightarrow \infty]{d} \chi^2(8 - 1 - 1)$$

35 / 43

Numerical application. $n = 100$, $T(\underline{x}_n) = 9.75$



⇒ at the 5% level, H_0 is accepted

36 / 43

Lecture outline

- 1 – Examples and first definitions
 - 1.1 – Two introductory examples
 - 1.2 – Risks associated to a test
- 2 – Parametric tests
 - 2.1 – Simple null vs simple alternative
 - 2.2 – Composite hypotheses
 - 2.3 – Asymptotic tests
- 3 – Testing for goodness of fit
 - 3.1 – Pearson's χ^2 test
 - 3.2 – BONUS: Kolmogorov-Smirnov test
- 5 – Warming up exercise

Goodness-of-fit test for a single distribution : $H_0 : P = P_0$.

Kolmogorov-Smirnov distance

We call **Kolmogorov-Smirnov distance** the quantity

$$D_n = \sup_x |\hat{F}_n(x) - F_0(x)|,$$

with F_0 the cdf of P_0 and \hat{F}_n empirical cdf $\Rightarrow \hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}}$.

Kolmogorov-Smirnov test

Under the null hypothesis H_0 , if F_0 is continuous:

$$T(\underline{X}_n) = \sqrt{n} D_n \xrightarrow[n \rightarrow \infty]{d} \mathcal{K},$$

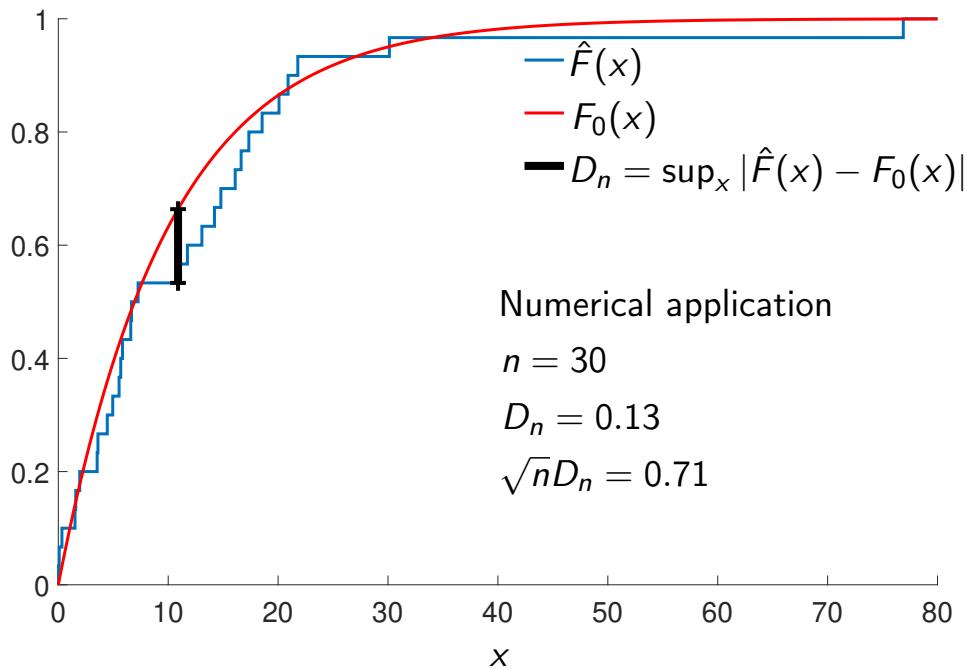
where \mathcal{K} is the **Kolmogorov-Smirnov distribution**.

$\Rightarrow H_0$ is rejected if $T_n > t_\alpha$, with t_α the $(1 - \alpha)$ -quantile of \mathcal{K} .

37/43

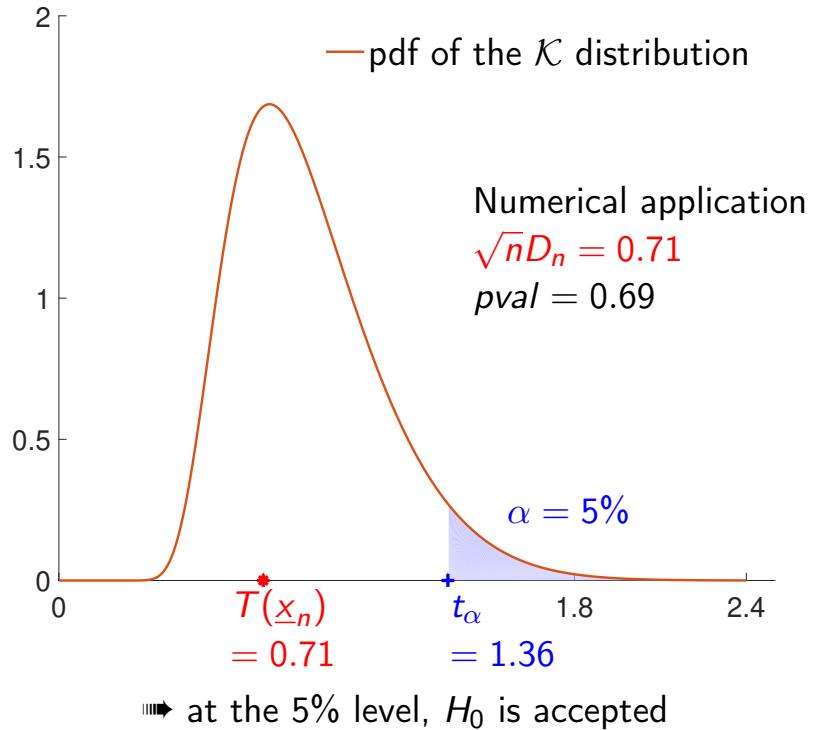
Example: component reliability

$$H_0 : P = \mathcal{E}(\theta_0) \text{ with } \theta_0 = 0.1$$



38/43

Example: component reliability (cont'd)



39/43

Lecture outline

1 – Examples and first definitions

- 1.1 – Two introductory examples
- 1.2 – Risks associated to a test

2 – Parametric tests

- 2.1 – Simple null vs simple alternative
- 2.2 – Composite hypotheses
- 2.3 – Asymptotic tests

3 – Testing for goodness of fit

- 3.1 – Pearson's χ^2 test
- 3.2 – BONUS: Kolmogorov-Smirnov test

5 – Warming up exercise

Exercise (Hypothesis test for a proportion)

In the context of a coin toss game, we want to test if the coin is balanced.

Questions

- ❶ Propose a statistical experiment to test this hypothesis.
Specify the underlying statistical model, and define the null and alternative hypotheses.
- ❷ Propose a test at the asymptotic level α .

40/43

Solution of Exercise 1

i) on réalise n expériences de "pile ou face" dont les issues sont modélisées par n variables aléatoires X_1, \dots, X_n indépendantes et identiquement distribuées selon une loi de $Ber(\theta)$.

We want to test if

$$H_0 : \theta = \frac{1}{2}, \quad \text{c'est-à-dire} \quad \Theta_0 = \left\{ \frac{1}{2} \right\} \quad (\text{hypothèse simple}),$$

vs.

$$H_1 : \theta \neq \frac{1}{2} \quad \text{donc} \quad \Theta_1 = \left] 0, \frac{1}{2} \right[\cup \left[\frac{1}{2}, 1 \right[\quad (\text{hypothèse bilatère}).$$

On parle de test bilatère.

41/43

Solution of Exercise 1 (suite)

ii) Posons $\hat{\theta}_n = \bar{X}_n$, la moyenne empirique de l'échantillon. Par application directe du TCL, il vient que :

$$\frac{\hat{\theta}_n - \theta}{\sqrt{\theta(1-\theta)/n}} \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, 1)$$

Pour construire un test asymptotique bilatéral de niveau α , on se place sous H_0 . Il vient la convergence en loi suivante:

$$2\sqrt{n} \left(\hat{\theta}_n - \frac{1}{2} \right) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, 1).$$

On considère une zone de rejet de la forme: $2\sqrt{n}|\hat{\theta}_n - \frac{1}{2}| > c_\alpha$. où c_α est choisi en fixant le risque de première espèce à α .

42/43

Solution of Exercise 1 (suite)

ii) Let

$$\lim_{n \rightarrow \infty} \mathbb{P}\left(2\sqrt{n}|\hat{\theta}_n - \frac{1}{2}| > c_\alpha\right) = \alpha.$$

On en déduit que $c_\alpha = q_{1-\frac{\alpha}{2}}$, quantile d'ordre $1 - \frac{\alpha}{2}$ d'une $\mathcal{N}(0, 1)$.

On rejette l'hypothèse H_0 au profit de H_1 au risque α de se tromper dès que:

$$|\hat{\theta}_n - \frac{1}{2}| > q_{1-\frac{\alpha}{2}} \frac{1}{2\sqrt{n}}.$$

Ainsi l'écart entre $\hat{\theta}_n$ et $1/2$ est considéré comme significatif au risque α dès qu'il est supérieur à $q_{1-\frac{\alpha}{2}} \frac{1}{2\sqrt{n}}$.

43/43

Chapter 6

Introduction to supervised learning Linear models for regression



CentraleSupélec

Statistics and Learning

Arthur Tenenhaus[†], Julien Bect & Laurent Le Brusquet

(firstname.lastname@centralesupelec.fr)

Teaching: CentraleSupélec / Department of Mathematics

Research: Laboratory of signals and systems (L2S)

[†]: Course coordinator

1/41

Lecture 6/10

Introduction to supervised learning Linear models for regression

In this lecture you will learn how to...

- ▶ explain the basic concepts of statistical learning
- ▶ set up the mathematical framework for regression and classification problems
- ▶ build & use linear regression models

2/41

Lecture outline

1 – Introduction to (supervised) statistical learning

1.1 – Statistical learning

1.2 – The mathematical framework of supervised learning

2 – Linear regression

2.1 – Introduction to regression models

2.2 – Linear model / quadratic loss

2.3 – Back to statistical inference

2.4 – Other loss functions

2.5 – Limitations of “ordinary least squares”

3/41

Lecture outline

1 – Introduction to (supervised) statistical learning

1.1 – Statistical learning

1.2 – The mathematical framework of supervised learning

2 – Linear regression

2.1 – Introduction to regression models

2.2 – Linear model / quadratic loss

2.3 – Back to statistical inference

2.4 – Other loss functions

2.5 – Limitations of “ordinary least squares”

Lecture outline

1 – Introduction to (supervised) statistical learning

1.1 – Statistical learning

1.2 – The mathematical framework of supervised learning

2 – Linear regression

2.1 – Introduction to regression models

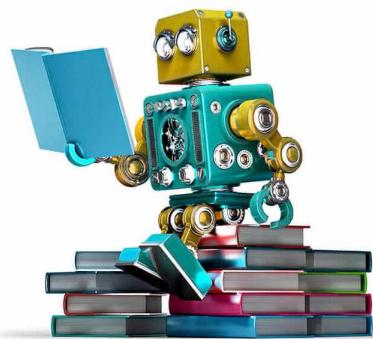
2.2 – Linear model / quadratic loss

2.3 – Back to statistical inference

2.4 – Other loss functions

2.5 – Limitations of “ordinary least squares”

Machine learning (*apprentissage automatique*)



One possible definition...

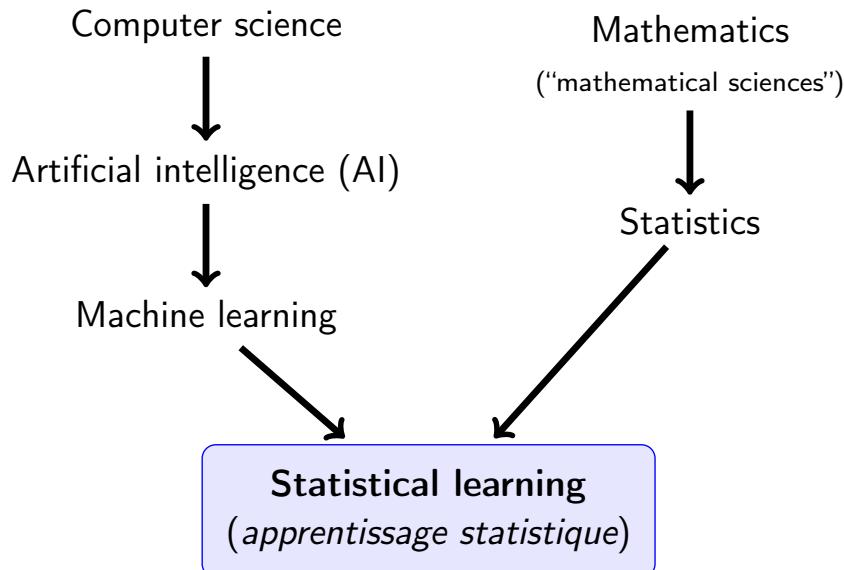
“*Machine learning is the study of computational methods for improving performance by mechanizing the acquisition of knowledge from experience.*”

→ data !

(P. Langley and H. A. Simon (1995). Comm. of the ACM, 38(11):54–64)

Image: J. Walsh (2016). Machine Learning: The Speed-of-Light Evolution of AI and Design.
<https://www.autodesk.com/redshift/machine-learning/>

Statistical learning: a “disciplinary” point of view



Remark: in practice, “machine learning” (*apprentissage automatique*) and “statistical learning” (*apprentissage statistique*) are often used interchangeably.

5/41

Example: handwritten character recognition



A subset of the MNIST database
containing 70 000 b&w images[†] of size 28×28 pixels

Supervised learning problems: examples are provided with a **label**.

⇒ Learn to **classify** a new image in one of the 10 classes.

[†] 60 000 training examples and 10 000 test examples → <http://yann.lecun.com/exdb/mnist/>
(to this day, the best error rates achieved on this problem are about 0.2%)

6/41

Example: real estate pricing in Ames (Iowa)



Data Description

- SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
- MSSubClass: The building class
- MSZoning: The general zoning classification
- LotFrontage: Linear feet of street connected to property
- LotArea: Lot size in square feet
- Street: Type of road access
- Alley: Type of alley access
- LotShape: General shape of property
- ...

Database of real estate transactions data
(sales price + 79 attributes; 1460 transactions)

Supervised learning problem: here, the price plays the role of a **label**.

⇒ Learn to **predict** the price of a house from its 79 attributes.

Source: Kaggle competition "House Prices: Advanced Regression Techniques"
(<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>)

7/41

Several forms of learning

- ▶ **Supervised** learning: examples with **labels**.
 - ▶ analogy: learning with a teacher.
 - ⇒ lectures #6 to #9
- ▶ **Unsupervised** learning: examples **without labels**
 - ▶ analogy: learning without a teacher, discovery
 - ⇒ lecture #10

and also... (not covered in this course)

- ▶ **Active** learning
 - ▶ the labels are queried sequentially;
 - ▶ example: detection of bank frauds
→ in-depth analysis of "suspicious" cases only.
- ▶ **Reinforcement** learning...

8/41

Numerous fields of application

- ▶ Computer vision
- ▶ Speech recognition
- ▶ Natural Language Processing (NLP)
- ▶ Fraud detection
- ▶ Personalized medicine
- ▶ Recommender systems & targeted marketing
- ▶ ...

9/41

Lecture outline

1 – Introduction to (supervised) statistical learning

- 1.1 – Statistical learning
- 1.2 – The mathematical framework of supervised learning

2 – Linear regression

- 2.1 – Introduction to regression models
- 2.2 – Linear model / quadratic loss
- 2.3 – Back to statistical inference
- 2.4 – Other loss functions
- 2.5 – Limitations of “ordinary least squares”

ML vocabulary: instance space and label space

Instance space: \mathcal{X}

- ▶ instances $x_1, \dots, x_n \in \mathcal{X}$

Label space: \mathcal{Y}

- ▶ labels $y_1, \dots, y_n \in \mathcal{Y}$

MNIST example:

```

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6
7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7
8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8
9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9

```

Class: zero, one, ... nine

$$\mathcal{X} = \{0, 1\}^{28 \times 28}$$

$$\mathcal{Y} = \{\text{"zero"}, \dots, \text{"nine"}\}$$

In this and the following lectures, we will always assume:

$$\mathcal{X} = \mathbb{R}^p$$

$\mathcal{Y} = \mathbb{R} \rightarrow$ regression, or

$\mathcal{Y} = \{0, 1\} \rightarrow$ classification[†].

[†] more precisely: *binary* classification. However, binary classification methods can also be useful for "multi-class" problems (such as MNIST)...

10/41

Statistical model

The statistical model of supervised learning

i) In supervised learning, we consider an **iid n -sample**:

$$(X_1, Y_1), \dots, (X_n, Y_n) \stackrel{\text{iid}}{\sim} P^{X,Y}$$

where $P^{X,Y}$ is an unknown probability measure on $\mathcal{X} \times \mathcal{Y}$.

ii) Unless explicitly mentioned, we make **no assumption on the distribution**: $\theta = P^{X,Y}$ and $\Theta = \{\text{probability measures on } \mathcal{X} \times \mathcal{Y}\}$.

Notation. We denote by (X, Y) another pair of RVs, which follows the same distribution $P^{X,Y}$ but is **not observed**.



change of notation (wrt previous lectures)

⇒ observations: $X_i \in \mathcal{X} \rightarrow (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$

11/41

Goal

Goal of supervised learning (informally)

We want to “learn” from data[†] a **prediction function**[‡]

$$\begin{aligned}\hat{h} : \mathcal{X} &\rightarrow \mathcal{Y} \\ x &\mapsto y = \hat{h}(x)\end{aligned}$$

such that the RVs Y and $\hat{h}(X)$ are as “close” as possible.

[†] We should write $\hat{h}(x) = \hat{h}(x; (X_1, Y_1), \dots, (X_n, Y_n)) \dots$

[‡] If \mathcal{Y} is finite, it is also called **classification function** or “classifier”.

To this end, let us consider a **loss function**:

$$\begin{aligned}L : \mathcal{Y} \times \mathcal{Y} &\rightarrow \mathbb{R}^+ \\ (y, \tilde{y}) &\mapsto L(y, \tilde{y}).\end{aligned}$$

⇒ $L(y, \hat{h}(x))$ quantifies the loss when y is predicted by $\hat{h}(x)$.

12/41

Goal (cont'd)

Definition: risk (generalization error)

Given a loss function L and a prediction function h , the **risk**, or **generalization error**, is defined as :

$$R(h) = \mathbb{E}(L(Y, h(X))),$$

where the expectation is with respect to (X, Y) .



This risk depends on the unknown distribution $\theta = P^{X,Y}$:

$$R_\theta(\textcolor{blue}{h}) = \iint_{\mathcal{X} \times \mathcal{Y}} L(y, \textcolor{blue}{h}(x)) P^{X,Y}(dx, dy).$$

⇒ From now on, we will simply write $R(h)$.

13/41

Goal (cont'd)

The optimal prediction function depends on the unknown distribution $P^{X,Y}$:

$$h_* = h_*(P^{X,Y}) = \operatorname{argmin}_h R(h).$$

(existence/uniqueness not guaranteed)

Goal of supervised learning

We want to construct, from the data $(X_1, Y_1), \dots, (X_n, Y_n)$, a prediction function

$$\begin{aligned} \hat{h} : \mathcal{X} &\rightarrow \mathcal{Y} \\ x &\mapsto y = \hat{h}(x) \end{aligned}$$

such that the risk $R(\hat{h})$ is as close as possible to the optimal risk

$$R_* = \inf_h R(h)$$

(also called “Bayes risk”).

14/41

Lecture outline

1 – Introduction to (supervised) statistical learning

- 1.1 – Statistical learning
- 1.2 – The mathematical framework of supervised learning

2 – Linear regression

- 2.1 – Introduction to regression models
- 2.2 – Linear model / quadratic loss
- 2.3 – Back to statistical inference
- 2.4 – Other loss functions
- 2.5 – Limitations of “ordinary least squares”

Lecture outline

1 – Introduction to (supervised) statistical learning

1.1 – Statistical learning

1.2 – The mathematical framework of supervised learning

2 – Linear regression

2.1 – Introduction to regression models

2.2 – Linear model / quadratic loss

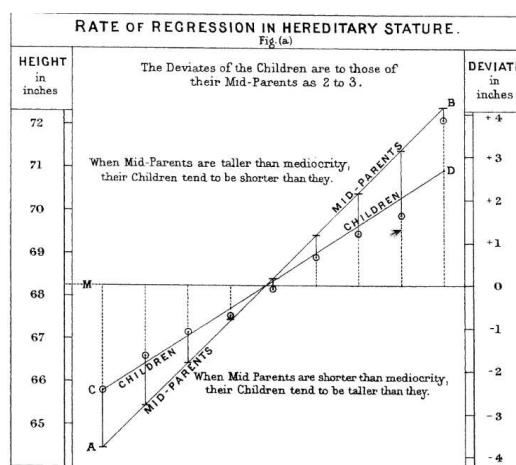
2.3 – Back to statistical inference

2.4 – Other loss functions

2.5 – Limitations of “ordinary least squares”

Regression

We consider in the rest of this lecture the **regression** case: $\mathcal{Y} = \mathbb{R}$.



Francis Galton (1886). "Regression Towards Mediocrity in Hereditary Stature", *Journal of the Anthropological Institute*, 15:246–263.

Stat. vocab.: Y = response variable / X = explanatory variables.

Quadratic loss

Consider for a start the quadratic loss:

$$L(y, \tilde{y}) = (y - \tilde{y})^2.$$

(this is the most commonly used in regression settings)

Proposition

For the quadratic loss, the optimal prediction function is

$$\forall x \in \mathcal{X}, \quad h_*(x) = \mathbb{E}(Y|X=x).$$

Vocabulary : $x \mapsto \mathbb{E}(Y|X=x)$ is sometimes called “regression function”.

We will consider this loss function **until further notice**.

16/41

Quadratic loss (cont'd)

Proof. By the law of total expectation, we get:

$$R(h) = \mathbb{E} \left(\underbrace{\mathbb{E}((Y-h(X))^2 | X)}_{\circledast} \right).$$

The term \circledast can be decomposed as :

$$\begin{aligned} & \mathbb{E}((Y-h(X))^2 | X) \\ &= \mathbb{E}((Y - \mathbb{E}(Y|X) + \mathbb{E}(Y|X) - h(X))^2 | X) \\ &= \text{var}(Y|X) + (\mathbb{E}(Y|X) - h(X))^2. \end{aligned}$$

The first term does not depend on h , and the second one is minimal when $h(X) = \mathbb{E}(Y|X)$ a.s.. □

17/41

Empirical risk

Recall that the joint distribution $P^{X,Y}$ is unknown

- ⇒ the risk $R(h)$ cannot be computed.

Definition: empirical risk

We call **empirical risk** the risk

$$\hat{R}_n(h) = \iint_{\mathcal{X} \times \mathcal{Y}} L(y, h(x)) \hat{P}_n(dx, dy) = \frac{1}{n} \sum_{i=1}^n L(Y_i, h(X_i))$$

associated to the empirical measure $\hat{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i, Y_i}$.

With the quadratic loss :

$$\hat{R}_n(h) = \frac{1}{n} \sum_{i=1}^n (Y_i - h(X_i))^2.$$

18/41

Empirical risk minimization

A general learning method:

- ① Choose a family \mathcal{H} of prediction functions.
- ② Select the function h which **minimizes the empirical risk**:

$$\hat{h}^{\text{ERM}} = \operatorname{argmin}_{h \in \mathcal{H}} \hat{R}_n(h).$$

Example: “linear” (affine) prediction functions

$$\begin{aligned} \mathcal{H} = & \left\{ h : \mathbb{R}^p \rightarrow \mathbb{R} \mid \exists \beta \in \mathbb{R}^{p+1}, \forall x \in \mathcal{X}, \right. \\ & \left. h(x) = \beta_0 + \beta_1 x^{(1)} + \dots + \beta_p x^{(p)} \right\} \end{aligned}$$



the ERM method is reasonable if \mathcal{H} is “not too large”

- ⇒ otherwise, complex models must be *penalized* (more on this later)

19/41

Other examples of families of prediction functions

- ▶ linear models with general basis functions

$$h(x) = \beta_1 h_1(x) + \dots + \beta_K h_K(x),$$

where the functions $h_k : \mathcal{X} \rightarrow \mathbb{R}$ are known;

- ▶ additive models

$$h(x) = h_1(x^{(1)}) + \dots + h_p(x^{(p)}),$$

where the h_k 's belong to a given family of $\mathbb{R} \rightarrow \mathbb{R}$ functions;

- ▶ neural networks,
- ▶ decision trees,
- ▶ generalized linear/additive models
- ▶ ...

20/41

Lecture outline

1 – Introduction to (supervised) statistical learning

- 1.1 – Statistical learning
- 1.2 – The mathematical framework of supervised learning

2 – Linear regression

- 2.1 – Introduction to regression models
- 2.2 – Linear model / quadratic loss
- 2.3 – Back to statistical inference
- 2.4 – Other loss functions
- 2.5 – Limitations of “ordinary least squares”

Residual sum of squares

We consider prediction functions h of the form :

$$h(x) = \beta_0 + \beta_1 x^{(1)} + \dots + \beta_p x^{(p)} = \beta^\top x$$

$$\text{with } \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \text{ and } x = \begin{pmatrix} 1 \\ x^{(1)} \\ \vdots \\ x^{(p)} \end{pmatrix}.$$

Definition: RSS / least squares criterion

Empirical risk: $\hat{R}(h) = \frac{1}{n} \sum_{i=1}^n (Y_i - \beta^\top X_i)^2$.

We define the **Residual Sum of Squares** (RSS):

$$\text{RSS}(\beta) = \sum_{i=1}^n (Y_i - \beta^\top X_i)^2$$

or **least squares criterion**.

21/41

Matrix-vector notations

$$\text{Let } \underline{X} = \begin{pmatrix} 1 & X_1^{(1)} & \dots & X_1^{(p)} \\ 1 & X_2^{(1)} & \dots & X_2^{(p)} \\ \vdots & & & \vdots \\ 1 & X_n^{(1)} & \dots & X_n^{(p)} \end{pmatrix} \text{ and } \underline{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}.$$

⇒ \underline{X} has size $n \times (p+1)$ and \underline{Y} has length n .

Matrix form of the criterion

$$\begin{aligned} \text{RSS}(\beta) &= \|\underline{Y} - \underline{X}\beta\|^2 \\ &= (\underline{Y} - \underline{X}\beta)^\top (\underline{Y} - \underline{X}\beta) \\ &= \beta^\top \underline{X}^\top \underline{X}\beta - 2\underline{Y}^\top \underline{X}\beta + \underline{Y}^\top \underline{Y} \end{aligned}$$

22/41

Minimization of the least squares criterion

Assumption

We assume $\underline{X}^\top \underline{X}$ almost surely invertible

implies $p + 1 \leq n$.

Let $\tilde{\beta} = (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top \underline{Y}$. Then:

$$\begin{aligned}\text{RSS}(\beta) &= \beta^\top \underline{X}^\top \underline{X} \beta - 2 \underline{Y}^\top \underline{X} \beta + \underline{Y}^\top \underline{Y} \\ &= (\beta - \tilde{\beta})^\top \underline{X}^\top \underline{X} (\beta - \tilde{\beta}) + c\end{aligned}$$

where c is a constant (i.e., does not depend on β).

Indeed: $\tilde{\beta}^\top \underline{X}^\top \underline{X} \beta = \underline{Y}^\top \underline{X} (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top \underline{X} \beta = \underline{Y}^\top \underline{X} \beta$.

23/41

Minimization of the least squares criterion

Reminder : $\text{RSS}(\beta) = (\beta - \tilde{\beta})^\top \underline{X}^\top \underline{X} (\beta - \tilde{\beta}) + c$.

We have:

- ❶ $\forall a \in \mathbb{R}^{p+1}$, $a^\top \underline{X}^\top \underline{X} a = \|\underline{X} a\|^2 \geq 0$,
- ❷ $\underline{X}^\top \underline{X}$ is invertible, hence positive definite.

- (i) implies that $\text{RSS}(\beta)$ is minimal at $\tilde{\beta}$;
- (ii) implies that the minimizer is unique ($a^\top \underline{X}^\top \underline{X} a = 0 \implies a = 0$).

Proposition: least squares estimator

When $\underline{X}^\top \underline{X}$ is invertible,

$$\hat{\beta} = (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top \underline{Y}$$

is the unique minimizer of the RSS (least squares criterion).

24/41

Matrix calculus

The result can also be found using matrix calculus.

Let $v \in \mathbb{R}^q$, $z \in \mathbb{R}^q$ and $M \in \mathbb{R}^{q \times q}$.

1) differentiation of $h(z) = v^\top z = \sum_{j=1}^q v_j z_j$

$$\nabla_z h(z) = \begin{pmatrix} \frac{\partial h}{\partial z_1} \\ \vdots \\ \frac{\partial h}{\partial z_q} \end{pmatrix} = \begin{pmatrix} v_1 \\ \vdots \\ v_q \end{pmatrix} = v \quad \text{therefore} \quad \nabla_z (v^\top z) = v.$$

2) differentiation of $h(z) = z^\top Mz = \sum_{i,j=1}^p z_i M_{i,j} z_j$

$$\nabla_z h(z) = \begin{pmatrix} \frac{\partial h}{\partial z_1} \\ \vdots \\ \frac{\partial h}{\partial z_q} \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^q M_{1,j} z_j + \sum_{i=1}^q M_{i,1} z_i \\ \vdots \\ \sum_{j=1}^q M_{1,j} z_j + \sum_{i=1}^q M_{i,1} z_i \end{pmatrix}$$

$$\text{therefore} \quad \nabla_z (z^\top Mz) = (M + M^\top)z.$$

Matrix calculus (cont'd)

Application to the minimization of the least squares criterion.

Recall that

$$\text{RSS}(\beta) = \beta^\top \underline{X}^\top \underline{X} \beta - 2 \underline{Y}^\top \underline{X} \beta + \underline{Y}^\top \underline{Y}$$

Thus we have

$$\nabla_\beta \text{RSS}(\beta) = 2 \underline{X}^\top \underline{X} \beta - 2 \underline{X}^\top \underline{Y} = 2 (\underline{X}^\top \underline{X} \beta - \underline{X}^\top \underline{Y}),$$

and finally:

$$\nabla_\beta \text{RSS}(\hat{\beta}) = 0 \implies \hat{\beta} = (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top \underline{Y}.$$

□

Goodness of fit

Without explanatory variables, we would have

$$\hat{h}(x) = \hat{\beta}_0, \quad \text{with} \quad \hat{\beta}_0 = \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i.$$

Let us set $\text{TSS} = \sum_{i=1}^n (Y_i - \bar{Y})^2 \rightarrow \text{Total Sum of Squares.}$

Definition: coefficient R^2 of determination

Reminder : $\text{RSS}(\beta) = \sum_{i=1}^n (Y_i - \beta^\top X_i)^2$. We set :

$$R^2 = 1 - \frac{\text{RSS}(\hat{\beta})}{\text{TSS}}.$$

Properties.

- ▶ $0 \leq R^2 \leq 1$,
- ▶ if $R^2 = 1$, then $\forall i, Y_i = \hat{\beta} X_i$.

25/41

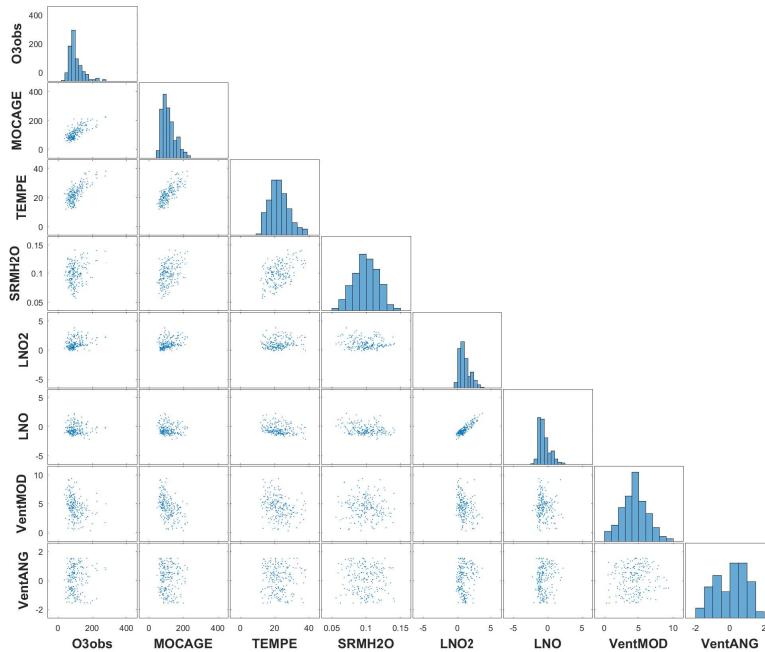
“Ozone” example: presentation of the data

variable	description
O3obs	concentration of ozone on day $t + 1$
MOCAGE	pollution prediction obtained by a deterministic computation fluid dynamics (CFD) model
TEMPE	MétéoFrance temperature forecast for day $t + 1$
RMH2O	humidity ratio at day t
NO2	nitrogen dioxide concentration on day t
NO	nitrogen monoxide concentration on day t
VentMOD	wind strength on day t
VentANG	wind orientation of day t

Learning task

- ▶ predict the ozone concentration on day $t + 1$ from data available on day t
- ▶ predict if the concentration will exceed $150 \mu\text{g}/\text{m}^3$ (classification task, cf. lecture #7).

“Ozone” example: data visualization



27/41

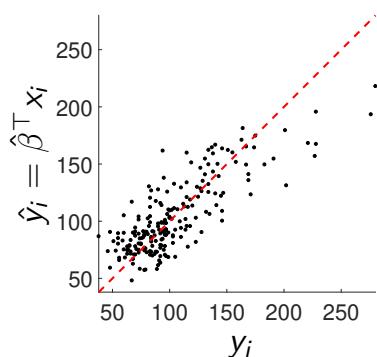
“Ozone” example: linear regression

Linear regression using $n = 210$ days of data.

Remark. All variables centered and normalized for the sake of interpretability.

β_0	MOCAGE	TEMPE	RMH2O	NO2	NO	VentMOD	VentANG
103.4	28.9	22.5	-3.2	-34.4	37.9	1.4	2.6

Coefficient of determination. $R^2 = 65.7\%$



Observations:

- ▶ the negative coefficient associated to NO2 is surprising (but NO2 is correlated with NO);
- ▶ RMH2O, VentMOD and VentANG appear to be of lesser importance;
- ▶ the model explains partly the data.

28/41

Lecture outline

1 – Introduction to (supervised) statistical learning

1.1 – Statistical learning

1.2 – The mathematical framework of supervised learning

2 – Linear regression

2.1 – Introduction to regression models

2.2 – Linear model / quadratic loss

2.3 – Back to statistical inference

2.4 – Other loss functions

2.5 – Limitations of “ordinary least squares”

Properties of the least squares estimator

Recall that, until now: $(X_1, Y_1), \dots, (X_n, Y_n) \stackrel{\text{iid}}{\sim} P^{X,Y}$.

⇒ in the section, we assume instead **deterministic X_i 's**

(equivalently, we work “conditionally on the X_i 's”).

Assume moreover that

$$(i) \quad \forall i, \quad Y_i = \beta^\top X_i + \epsilon_i$$

where the errors ϵ_i are

$$(ii) \quad \text{centered: } \mathbb{E}(\epsilon_i) = 0,$$

$$(iii) \quad \text{uncorrelated: } i \neq j \Rightarrow \text{cov}(\epsilon_i, \epsilon_j) = 0,$$

$$(iv) \quad \text{homoscedastic: } \text{var}(\epsilon_i) = \sigma^2 \text{ for some } \sigma^2 > 0.$$

Properties of the least squares estimator

Proposition

Under these assumptions, $\hat{\beta}$ is an **unbiased** estimator:

$$\mathbb{E}(\hat{\beta}) = \beta,$$

and its **covariance matrix** is:

$$\text{var}(\hat{\beta}) = \sigma^2 (\underline{X}^\top \underline{X})^{-1}.$$

30/41

Properties of the least squares estimator

Proof.

Recall that the X_i 's are assumed deterministic.

Let $\underline{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^\top$. Then:

$$(i) \quad \Rightarrow \quad \begin{cases} \underline{Y} &= \underline{X}\beta + \underline{\epsilon} \\ \hat{\beta} &= (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top \underline{Y} = \beta + (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top \underline{\epsilon} \end{cases}$$

$$(ii) \quad \Rightarrow \quad \mathbb{E}(\hat{\beta}) = \beta + (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top \mathbb{E}(\underline{\epsilon}) = \beta$$

$$(iii)+(iv) \quad \Rightarrow \quad \text{var}(\hat{\beta}) = (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top \text{var}(\underline{\epsilon}) \underline{X} (\underline{X}^\top \underline{X})^{-1} \\ = \sigma^2 (\underline{X}^\top \underline{X})^{-1}$$

□

31/41

Distribution of $(\hat{\beta}, \hat{\sigma}^2)$ under a normality assumption

Assume furthermore that $(v) \in \mathbb{R}$ is Gaussian:

$$\log \mathcal{L}(\beta, \sigma^2; Y) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - \beta^\top X_i)^2.$$

Proposition: MLE of (β, σ^2) (see PC)

The MLE is $\begin{cases} \hat{\beta} = \arg \min_{\beta} \sum_{i=1}^n (Y_i - \beta^\top X_i)^2, \\ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{\beta}^\top X_i)^2. \end{cases}$

⇒ We recover the least square estimator of β

Student's theorem: distribution of $(\hat{\beta}, \hat{\sigma}^2)$ (see PC)

- $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (\underline{X}^\top \underline{X})^{-1})$,
- $\hat{\beta}$ et $\hat{\sigma}^2$ are independent.
- $\hat{\sigma}^2 \sim \frac{\sigma^2}{n} \chi^2(n-p-1)$,

32/41

Tests / CI on the value of a component of β

We know that $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 v_j)$ with $v_j = [(\underline{X}^\top \underline{X})^{-1}]_{j,j}$.

Pivotal function

$$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\frac{n \hat{\sigma}^2 v_j}{n-p-1}}} \sim \mathcal{T}(n-p-1)$$

with $\mathcal{T}(n-p-1)$: Student distrib. with $n-p-1$ degrees of freedom
(⇒ defined on next page)

Remark:

$$\frac{n \hat{\sigma}^2}{n-p-1} = \frac{1}{n-p-1} \sum_{i=1}^n (Y_i - \hat{\beta}^\top X_i)^2$$

is an unbiased estimator of σ^2 (see PC).

33/41

The Student family of distributions

Definition of $\mathcal{T}(k)$, k integer ≥ 1

Let U and V be two RVs such that

- ▶ $U \sim \mathcal{N}(0, 1)$
- ▶ $V \sim \chi^2(k)$
- ▶ U and V are independent

then $T = \frac{U}{\sqrt{\frac{V}{k}}}$ follows a **Student distribution with k degrees of freedom**.

Properties

$$\mathcal{T}(k) \xrightarrow[k \rightarrow \infty]{d} \mathcal{N}(0, 1)$$

Exercise : prove it.

Probability density function

$$f(x) = \frac{1}{\sqrt{k\pi}} \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}$$

Mean

- ▶ for $k \geq 2$, $\mathbb{E}_k(T) = 0$

Variance

- ▶ for $k \geq 3$, $\text{var}_k(T) = \frac{k}{k-2}$

Proof

It follows from Student's theorem that

- ▶ $U = \frac{\hat{\beta}_j - \beta_j}{\sigma \sqrt{v_j}} \sim \mathcal{N}(0, 1)$
- ▶ $V = \frac{n \hat{\sigma}^2}{\sigma^2} \sim \chi^2(n - p - 1)$,
- ▶ and U and V are independent.

Thus

$$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\frac{n \hat{\sigma}^2 v_j}{n-p-1}}} = \frac{U}{\sqrt{\frac{V}{n-p-1}}} \sim \mathcal{T}(n - p - 1)$$

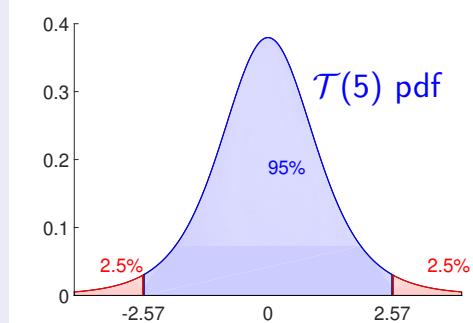
by definition of Student's distribution with $k = n - p - 1$ degrees of freedom. □

Test for $H_0 : \beta_j = 0$ / $H_1 : \beta_j \neq 0$

Let $0 < \alpha < 1$.

Take $\beta_j = 0$ in the def. of T (i.e. assume H_0) and

$$\delta = \mathbb{1}_{|T| > q_{1-\frac{\alpha}{2}}}$$



Exact confidence interval for β_j

$$\left[\hat{\beta}_j - \sqrt{\frac{n \hat{\sigma}^2 v_j}{n - p - 1}} q_{1-\frac{\alpha}{2}}, \hat{\beta}_j + \sqrt{\frac{n \hat{\sigma}^2 v_j}{n - p - 1}} q_{1-\frac{\alpha}{2}} \right]$$

q_r : quantile of order r of $\mathcal{T}(n - p - 1)$

35/41

“Ozone” example: CIs and p-values

	Cl _{95%}	t	pval
β_0	[100.1, 106.7]	62.9	< 10 ⁻⁶
MOCAGE	[21.1, 36.8]	7.4	< 10 ⁻⁶
TEMPE	[16.5, 28.5]	7.6	< 10 ⁻⁶
RMH2O	[-7.0, 0.6]	-1.7	0.095
NO2	[-53.0, -15.7]	-3.7	< 10 ⁻³
NO	[19.8, 55.4]	4.2	< 10 ⁻³
VentMOD	[-2.7, 5.4]	0.7	0.49
VentANG	[-0.8, 6.0]	1.6	0.12

with t : realized value of T for the corresponding coefficient

Remark: regression without RMH2O, VentMOD et VentANG

► the coefficient of determination drops from 65.7% to 64.5%.

36/41

Lecture outline

1 – Introduction to (supervised) statistical learning

1.1 – Statistical learning

1.2 – The mathematical framework of supervised learning

2 – Linear regression

2.1 – Introduction to regression models

2.2 – Linear model / quadratic loss

2.3 – Back to statistical inference

2.4 – Other loss functions

2.5 – Limitations of “ordinary least squares”

“Ozone” example: data corruption

Assume that 5 out of n measurements of ozone concentration ($n = 210$) are **corrupted** (approx. 2%).

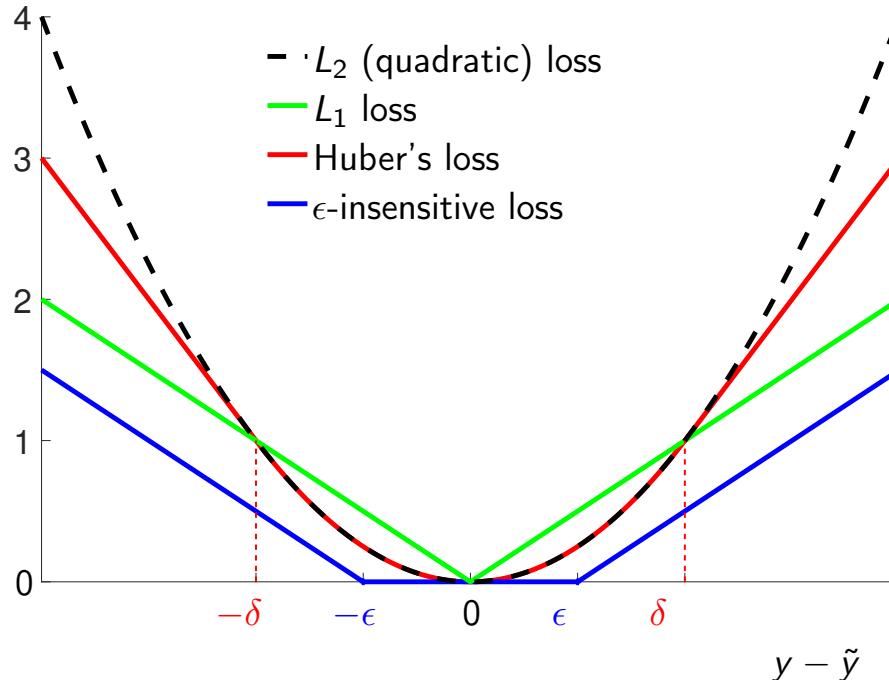
	β_0	MOCAGE	TEMPE	RMH2O	NO2	NO	VentMOD	VentANG
w/o	103.4	28.9	22.6	-3.2	-34.4	37.6	1.4	2.6
with	125.2	79.2	-15.6	24.2	-155.1	141.4	4.7	24.9

► Strong sensitivity of the coefficients to “outliers”.

Solution

Use a **loss function** that leads to a prediction function with better **robustness properties** than the quadratic loss.

Usual loss functions



38/41

L_1 loss

Loss function : $L(y, \tilde{y}) = |y - \tilde{y}|$.

Proposition

(see PC)

For the L_1 loss, the optimal prediction function is

$$\forall x \in \mathcal{X}, \quad h_*(x) = \text{med}(Y|X = x)$$

“Ozone” example

	β_0	MOCAGE	TEMPE	RMH2O	NO2	NO	VentMOD	VentANG
w/o	100.8	27.5	19.2	-3.3	-32.2	33.9	-1.0	3.9
with	101.4	28.3	18.6	-1.6	-35.1	37.5	0.5	3.2

⇒ better stability with respect to outliers.

39/41

Lecture outline

1 – Introduction to (supervised) statistical learning

1.1 – Statistical learning

1.2 – The mathematical framework of supervised learning

2 – Linear regression

2.1 – Introduction to regression models

2.2 – Linear model / quadratic loss

2.3 – Back to statistical inference

2.4 – Other loss functions

2.5 – Limitations of “ordinary least squares”

Limitations of “ordinary least squares”

Recall that \underline{X} has size #individuals \times #variables ($n \times (p + 1)$).

Critical cases for “ordinary least squares”

- ▶ when $\underline{X}^\top \underline{X}$ not invertible,
- ▶ or poorly conditioned.

Typical cases

- ▶ when the number of variables is large
($p + 1 > n$, sometimes $p \gg n$)

Example: genomics.

- ▶ when there are strong correlations between explanatory variables

Example: “ozone” data (cf. variables NO and NO2)

➡ lack of interpretability of the coefficients

One possible solution: penalized regression

A **penalty** term is added to the empirical risk:

$$\hat{\beta} = \arg \min_{\beta} \underbrace{\text{RSS}(\beta)}_{\text{data "fidelity"} \quad \text{+}} + \underbrace{\lambda}_{\text{hyperparameter}} \underbrace{\Omega(\beta)}_{\text{penalty}}.$$

⇒ see Lecture 8/10

Chapter 7

Classification: logistic regression Generalization error



Statistics and Learning

Arthur Tenenhaus[†], Julien Bect & Laurent Le Brusquet

(firstname.lastname@centralesupelec.fr)

Teaching: CentraleSupélec / Department of Mathematics

Research: Laboratory of signals and systems (L2S)

[†]: Course coordinator

1/34

Lecture 7/10

Classification: logistic regression.
Generalization error.

In this lecture you will learn how to...

- ▶ Classify using logistic regression
- ▶ Define relevant performance measures for classifiers
- ▶ Estimate the risk (generalization error)
in a regression or classification problem

2/34

Lecture outline

1 – Classification: logistic regression

- 1.1 – Introduction
- 1.2 – Linear models for classification
- 1.3 – Estimation of the parameter β
- 1.4 – Performance evaluation & choice of δ_0
- 1.5 – Extensions

2 – Estimation of the risk (generalization error)

- 2.1 – Problem
- 2.2 – Zoom on an illuminating special case
- 2.3 – Training set and test set

3/34

Lecture outline

1 – Classification: logistic regression

- 1.1 – Introduction
- 1.2 – Linear models for classification
- 1.3 – Estimation of the parameter β
- 1.4 – Performance evaluation & choice of δ_0
- 1.5 – Extensions

2 – Estimation of the risk (generalization error)

- 2.1 – Problem
- 2.2 – Zoom on an illuminating special case
- 2.3 – Training set and test set

Lecture outline

1 – Classification: logistic regression

1.1 – Introduction

1.2 – Linear models for classification

1.3 – Estimation of the parameter β

1.4 – Performance evaluation & choice of δ_0

1.5 – Extensions

2 – Estimation of the risk (generalization error)

2.1 – Problem

2.2 – Zoom on an illuminating special case

2.3 – Training set and test set

Mathematical framework and objectives

Notations

- ▶ $(X_1, Y_1), \dots, (X_n, Y_n) \stackrel{\text{iid}}{\sim} P^{X, Y}$
- ▶ $P^{X, Y}$: unknown distribution on $\mathcal{X} \times \mathcal{Y}$
- ▶ $\mathcal{X} \subset \mathbb{R}^p$, $\mathcal{Y} = \{0, 1, \dots, K - 1\}$
- ▶ unless otherwise stated: $K = 2$ (**binary classification**)

Objectives

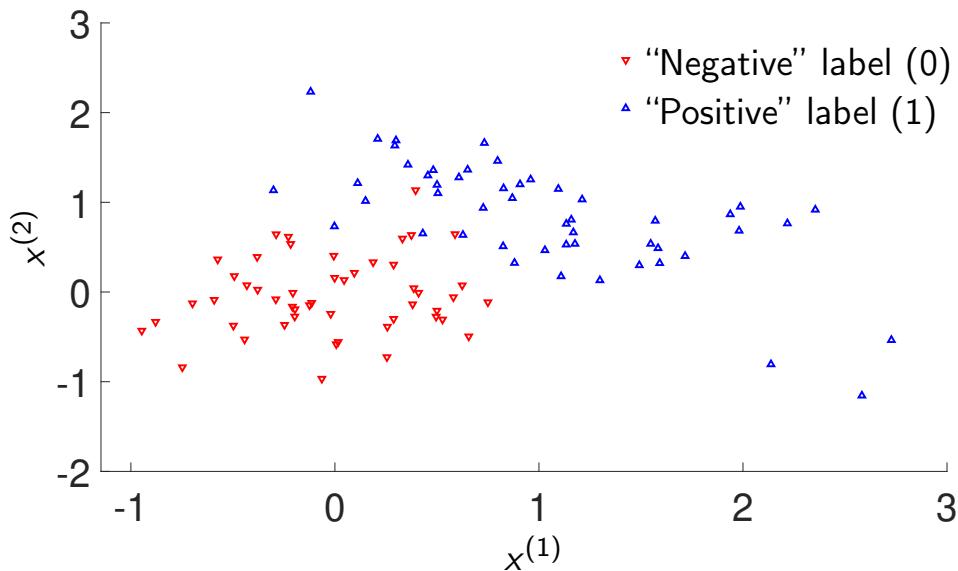
Construct a (good) prediction function $h : x \mapsto \{0, 1\}$.

Synonyms: **classification function**, or “classifier”.

Objectives of this section

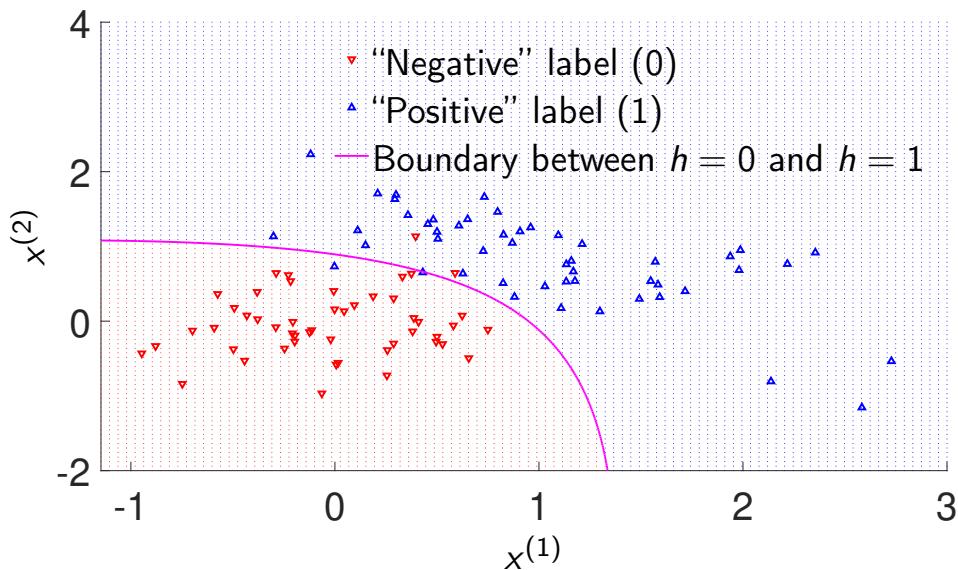
- ▶ present the **logistic regression** method
- ▶ define relevant **risk measures** for classification

Example with two explanatory variables ($p = 2$)



5/34

A taste of things to come: a possible classifier



6/34

Lecture outline

1 – Classification: logistic regression

1.1 – Introduction

1.2 – Linear models for classification

1.3 – Estimation of the parameter β

1.4 – Performance evaluation & choice of δ_0

1.5 – Extensions

2 – Estimation of the risk (generalization error)

2.1 – Problem

2.2 – Zoom on an illuminating special case

2.3 – Training set and test set

Logistic regression: a classification method !?!



Despite the name “**regression**”,
it is actually a **classification** method!

Explanation:

- ▶ it is indeed a regression method, since it focuses on the **regression** function $x \mapsto \mathbb{E}(Y | X = x)$,
- ▶ but the label Y is assumed binary, and thus the goal is actually to address (binary) **classification** problems.

Logistic regression: principle

Remark: if $P^{Y|X}$ were known, we could compute, for a given loss function, the **optimal classification function**:

$$h^* = \operatorname{argmin}_h \mathbb{E}(L(Y, h(X)))$$

$$\Leftrightarrow h^*(x) = \operatorname{argmin}_{t \in \mathcal{Y}} \mathbb{E}(L(Y, t) | X = x) \quad P^X\text{-pp.}$$

General principle

- ▶ approximate $P^{Y|X}$ using a parametric model $P_{\beta}^{Y|X}$,
- ▶ then deduce the classification function from the model.

Here $\mathcal{Y} = \{0, 1\}$, therefore

- ▶ $Y|X \sim \text{Bernoulli}(p(X))$ with $p(x) = \mathbb{P}(Y = 1|X = x)$,
- ▶ and thus we need to approximate $x \mapsto p(x)$.

8/34

Logistic regression: model

Model

Logistic regression assumes that $\exists \beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p$, such that [†]

$$\mathbb{P}(Y = 1|X = x) = \frac{\exp(\beta_0 + \beta^\top x)}{1 + \exp(\beta_0 + \beta^\top x)}$$

or, equivalently,

$$\text{logit}(\mathbb{P}(Y = 1|X = x)) = \beta_0 + \beta^\top x$$

with

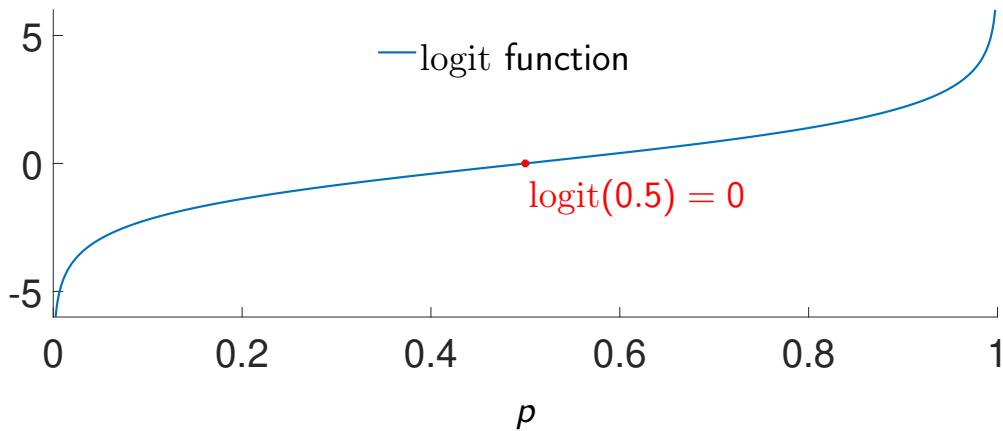
$$\begin{aligned} \text{logit} : (0, 1) &\rightarrow \mathbb{R} \\ p &\mapsto \ln\left(\frac{p}{1-p}\right) \end{aligned}$$

the **logit function**.

[†] and therefore $\mathbb{P}(Y = 0|X = x) = \frac{1}{1 + \exp(\beta_0 + \beta^\top x)}$

9/34

The logistic function



⇒ logit defines a correspondence: proba $p \in (0, 1) \longleftrightarrow \beta_0 + \beta^\top x \in \mathbb{R}$

10/34

Remark: generalized linear models (GLM)

The logistic regression model has the form

- ▶ $Y|X \sim \text{Bernoulli}(\mathbb{E}_\beta(Y|X))$,
- ▶ $\mathbf{g}(\mathbb{E}_\beta(Y|X)) = \beta_0 + \beta^\top X$, with $\mathbf{g} = \text{logit}$.

⇒ special case of the **generalized linear model (GLM)**
(\mathbf{g} is called link function)

Remark: we have already met another GLM model

- ▶ $Y|X \sim \mathcal{N}(\mathbb{E}_\beta(Y|X), \sigma^2)$
- ▶ $\mathbf{g}(\mathbb{E}_\beta(Y|X)) = \beta_0 + \beta^\top X$ with $\mathbf{g} = \text{Id}$

11/34

Generalized linear models

Definition

The GLM contains all statistical models such that

- $Y|X$ follows a distribution from an **exponential family**:

$$f^{Y|X}(y|x) = C(\eta)h(y) \exp(\eta y) \quad \text{with } \eta = \eta(x).$$

- $\mathbf{g}(\mathbb{E}_\beta(Y|X=x)) = \beta_0 + \beta^\top x.$

Vocabulary. The function g is called the **link function**.[†]

Example. Bernoulli distributions form an exponential family.

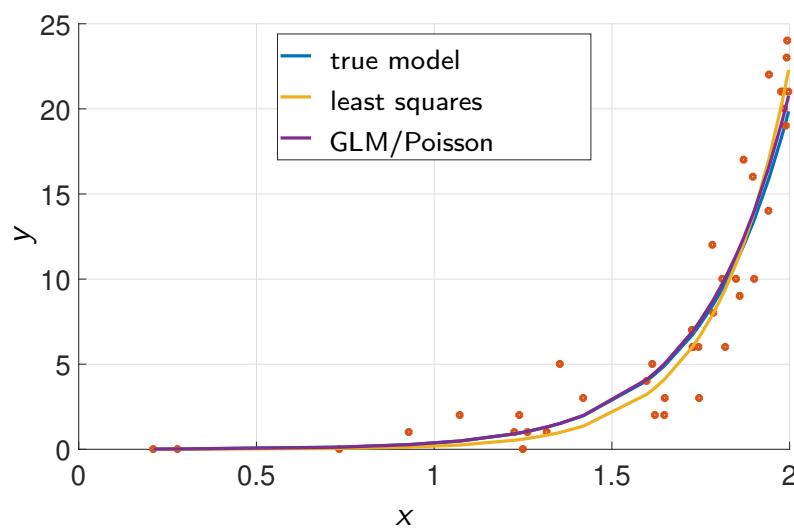
$$\begin{aligned} f(y) &= \theta^y (1-\theta)^{1-y} \\ &= (1-\theta) \exp\left(\ln\left(\frac{\theta}{1-\theta}\right)y\right) \quad \Rightarrow \eta = \ln\left(\frac{\theta}{1-\theta}\right) \end{aligned}$$

[†] Let N denote the set of admissible value for η : g is often chosen to be a bijection from N to \mathbb{R} .

Example: $Y_i|X_i \stackrel{\text{iid}}{\sim} \text{Poisson}(\theta_i)$, with $\ln \theta_i = \beta_0 + \beta_1 X_i$

Poisson distributions form an exponential family:

$$\begin{aligned} f(y) &= \exp(-\theta) \frac{\theta^y}{y!} \\ &= \frac{1}{y!} \exp(-\theta) \exp(\ln(\theta)y) \quad \Rightarrow \eta = \ln(\theta) \end{aligned}$$



Classification function

Logistic regression leads naturally to a “soft” classification

$$\blacktriangleright P_{\beta}^{Y|X}(Y = 1|X = x) \in [0, 1]$$

“Hard” classification (taking values in $\mathcal{Y} = \{0, 1\}$)

Let $\delta_0 \in [0, 1]$ (decision threshold).

A **classification function** can be constructed as follows:

$$\begin{aligned} h_{\delta_0} : \mathcal{X} &\rightarrow \{0, 1\} \\ x &\mapsto \begin{cases} 1 & \text{if } P_{\beta}^{Y|X}(Y = 1|X = x) \geq \delta_0 \\ 0 & \text{if } P_{\beta}^{Y|X}(Y = 1|X = x) < \delta_0 \end{cases} \end{aligned}$$

$$P_{\beta}^{Y|X}(Y = 1|X = x) \geq \delta_0 \iff \beta_0 + \beta^T x \geq \text{logit}(\delta_0)$$

⇒ separation by a hyperplane in \mathcal{X}

12/34

Minimization of the misclassification risk

Let us consider the loss function $L(y, \tilde{y}) = \mathbb{1}_{y \neq \tilde{y}}$.

The corresponding risk is the **probability of misclassification**:

$$R(h_{\delta_0}) = \mathbb{E}(L(Y, h_{\delta_0}(X))) = \mathbb{P}(Y \neq h_{\delta_0}(X)).$$

Proposition

(see PC)

The **minimum** of $\delta_0 \mapsto R(h_{\delta_0})$ is attained at $\delta_0 = 0.5$

⇒ With $\delta_0 = 0.5$, the separating hyperplane is $\beta_0 + \beta^T x = 0$.

Remark: a more general formula can be proved for an asymmetric loss ($L(0, 1) \neq L(1, 0)$). See PHC’s lecture notes.

13/34

Lecture outline

1 – Classification: logistic regression

- 1.1 – Introduction
- 1.2 – Linear models for classification
- 1.3 – Estimation of the parameter β
- 1.4 – Performance evaluation & choice of δ_0
- 1.5 – Extensions

2 – Estimation of the risk (generalization error)

- 2.1 – Problem
- 2.2 – Zoom on an illuminating special case
- 2.3 – Training set and test set

Maximum likelihood estimator

Simplification of notations: $x \rightarrow \begin{pmatrix} 1 \\ x \end{pmatrix}$ and $\beta \rightarrow \begin{pmatrix} \beta_0 \\ \beta \end{pmatrix}$

■ $\Rightarrow P_{\beta}^{Y|X}(Y = 1|X = x) = \frac{\exp(\beta^T x)}{1 + \exp(\beta^T x)}$

Log-likelihood

(see PC)

$$\begin{aligned}\ell(\beta) &= \ln \mathcal{L}(\beta; \underline{x}, \underline{y}) \\ &= \sum_{i=1}^n \left\{ y_i \beta^T x_i - \ln \left(1 + \exp(\beta^T x_i) \right) \right\}\end{aligned}$$

Maximization of ℓ

Carried out using a numerical optimization algorithm

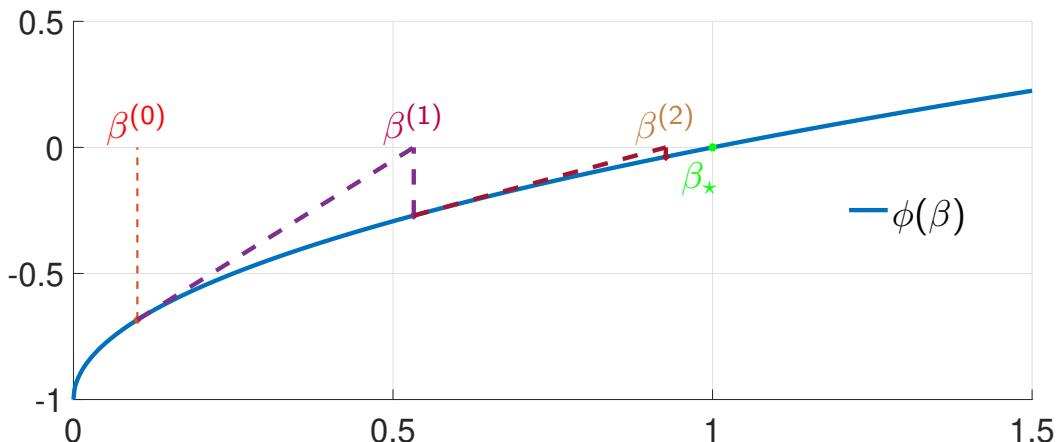
■ \Rightarrow for instance, the Newton-Raphson algorithm

Reminder: Newton-Raphson algorithm in one dimension

Let $\phi : \mathbb{R} \rightarrow \mathbb{R}$. We want β that satisfies $\phi(\beta) = 0$

Newton-Raphson algorithm is iterative:

- ▶ initialization: $\beta^{(0)}$
- ▶ iteration: $\beta^{(k+1)} = \beta^{(k)} - \frac{\phi(\beta^{(k)})}{\phi'(\beta^{(k)})}$



15/34

Maximization of ℓ using the Newton-Raphson method

Same algorithm but now in dimension $p + 1$, with:

- ▶ $\phi \rightarrow \nabla_{\beta} \ell$
- ▶ $\phi' \rightarrow \nabla_{\beta}^2 \ell$

The iteration follows:

$$\beta^{(k+1)} = \beta^{(k)} - [\nabla_{\beta}^2 \ell(\beta^{(k)})]^{-1} \nabla_{\beta} \ell(\beta^{(k)})$$

Under the following conditions:

- ▶ $\nabla_{\beta}^2 \ell(\cdot)$ is Lipschitz continuous,
- ▶ $\nabla_{\beta}^2 \ell(\beta^{(0)})$ is invertible
- ▶ $h_0 = [\nabla_{\beta}^2 \ell(\beta^{(0)})]^{-1} \nabla_{\beta} \ell(\beta^{(0)})$ small enough[†],

the algorithm converges to a point β^* such that $\nabla_{\beta} \ell(\beta^*) = 0$.

[†] cf. "Kantorovich theorem" on wikipedia for a more precise statement

16/34

Lecture outline

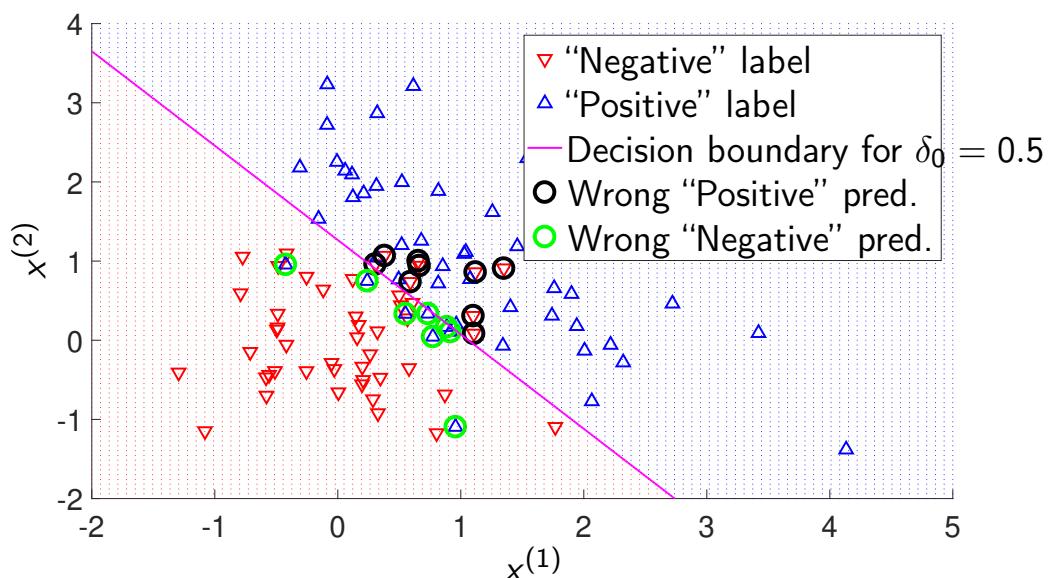
1 – Classification: logistic regression

- 1.1 – Introduction
- 1.2 – Linear models for classification
- 1.3 – Estimation of the parameter β
- 1.4 – Performance evaluation & choice of δ_0
- 1.5 – Extensions

2 – Estimation of the risk (generalization error)

- 2.1 – Problem
- 2.2 – Zoom on an illuminating special case
- 2.3 – Training set and test set

LR performed on the example with 2 explanatory variables



Prediction errors:

- ▶ “Negative” examples predicted as “Positive”
- ▶ “Positive” examples predicted as “Negative”

Confusion matrix & associated definitions

	Truth Negative (N)	Truth Positive (P)
Prediction Negative	True Negative (TN)	False Negative (FN)
Prediction Positive	False Positive (FP)	True Positive (TP)

True Positive Rate

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

(also called **sensitivity**)

True Negative Rate

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP}$$

(also called **specificity**)

18/34

Trade-off between True Negative Rate True Positive Rate

Alternative terminology, from the field of signal processing:

- ▶ $1 - TPR$ is the **miss rate** (false negative rate)
- ▶ $1 - TNR$ is the **false alarms rate** (false positive rate)

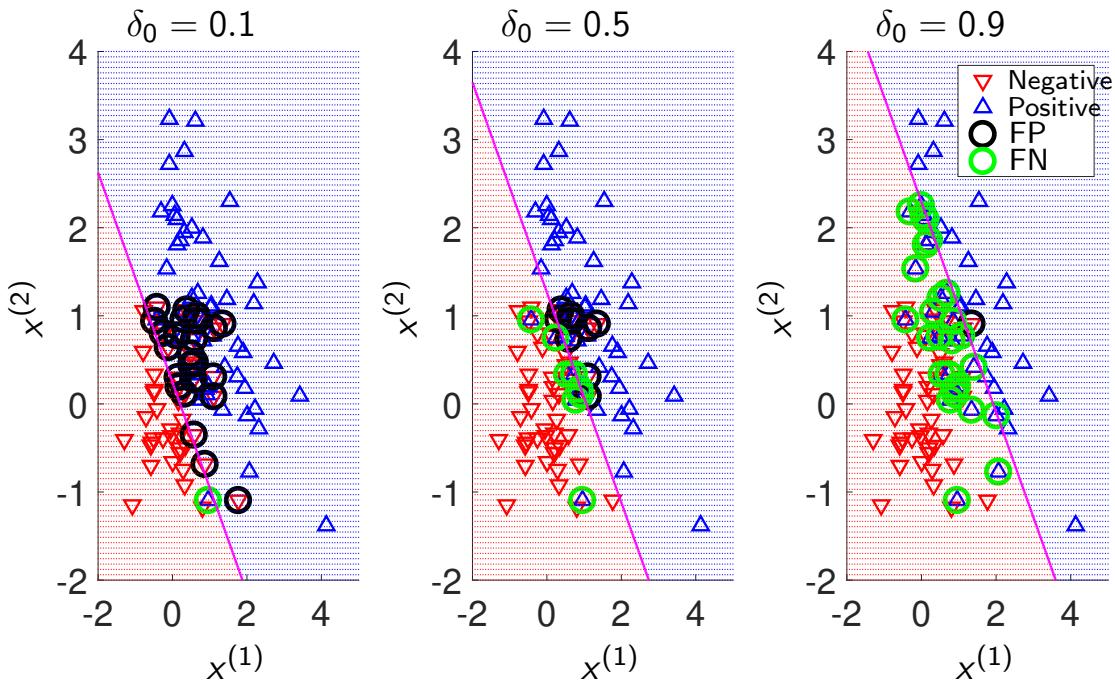
Trade-off.

The value of δ_0 impacts the trade-off TNR/TPR:

- ▶ reminder: $h_{\delta_0} = 1$ if $P_{\beta}^{Y|X}(Y = 1|X = x) \geq \delta_0$
- ▶ when $\delta_0 \nearrow$, $TNR \nearrow$, and $TPR \searrow$

19/34

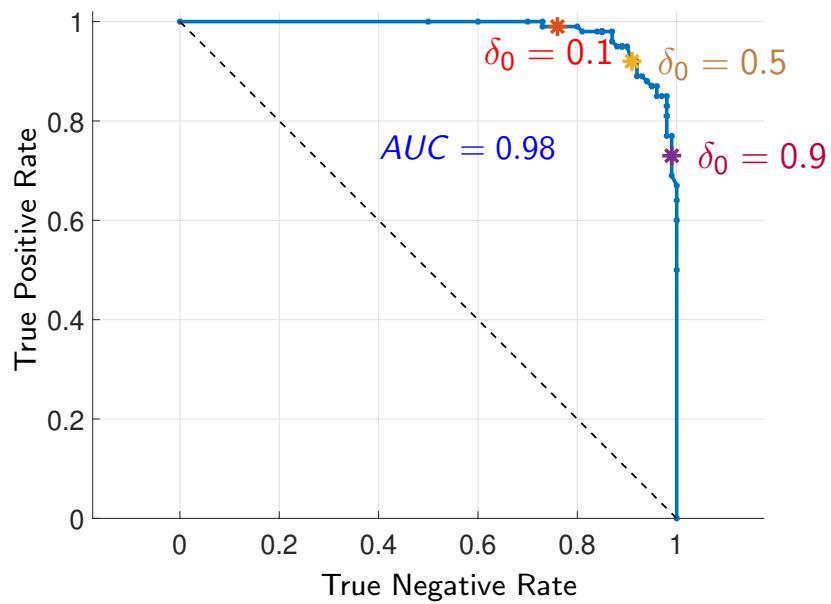
Influence of δ_0



20/34

ROC curve (Receiver Operating Characteristic)

- ▶ a tool for **decision support** (choice of δ_0)
- ▶ a tool useful for **classifier comparison**
- ▶ associated definition: **AUC** = Area Under Curve



21/34

Lecture outline

1 – Classification: logistic regression

- 1.1 – Introduction
- 1.2 – Linear models for classification
- 1.3 – Estimation of the parameter β
- 1.4 – Performance evaluation & choice of δ_0
- 1.5 – Extensions

2 – Estimation of the risk (generalization error)

- 2.1 – Problem
- 2.2 – Zoom on an illuminating special case
- 2.3 – Training set and test set

Extension: large number of variables

How to handle the case where p is large

The log-likelihood is **penalized**:

- $L_1 : \hat{\beta} = \arg \max_{\beta} (\ell(\beta) - \lambda \|\beta\|^2)$
- $L_2 : \hat{\beta} = \arg \max_{\beta} (\ell(\beta) - \lambda \|\beta\|_1)$

⇒ see Lecture 8/10

p is “large” if $p \gg n$, or even simply $p \approx n$

Extension: more than two classes

Multiclass classification

Let $\{0, 1, \dots, K - 1\}$ be the set of labels (classes), $K \geq 3$.

One class is chosen as the reference class and $K - 1$ binary logistic regressions are performed (here class “0” was chosen):

$$\left\{ \begin{array}{lcl} \ln \left(\frac{P(Y=1|X=x)}{P(Y=0|X=x)} \right) & = & \beta_{1,0} + \beta_1^\top x \\ \vdots \\ \ln \left(\frac{P(Y=K-1|X=x)}{P(Y=0|X=x)} \right) & = & \beta_{K-1,0} + \beta_{K-1}^\top x \end{array} \right.$$

23/34

Lecture outline

1 – Classification: logistic regression

- 1.1 – Introduction
- 1.2 – Linear models for classification
- 1.3 – Estimation of the parameter β
- 1.4 – Performance evaluation & choice of δ_0
- 1.5 – Extensions

2 – Estimation of the risk (generalization error)

- 2.1 – Problem
- 2.2 – Zoom on an illuminating special case
- 2.3 – Training set and test set

Lecture outline

1 – Classification: logistic regression

- 1.1 – Introduction
- 1.2 – Linear models for classification
- 1.3 – Estimation of the parameter β
- 1.4 – Performance evaluation & choice of δ_0
- 1.5 – Extensions

2 – Estimation of the risk (generalization error)

- 2.1 – Problem
- 2.2 – Zoom on an illuminating special case
- 2.3 – Training set and test set

Problem

Back to the **general setting** (regression/classification).

Let \hat{h} be a predictor $\mathcal{X} \rightarrow \mathcal{Y}$ learned from data:

$$\hat{h}(x) = \hat{h}(x; (X_1, Y_1), \dots, (X_n, Y_n)) = \hat{h}(x; \underline{X}, \underline{Y}).$$

Recall that, given a loss function L , we define the **risk**, or **generalisation error** :

$$\begin{aligned}\mathcal{R}(\hat{h}) &= \mathbb{E} \left(L(Y, \hat{h}(X)) \mid \underline{X}, \underline{Y} \right) \\ &= \iint_{\mathcal{X} \times \mathcal{Y}} L(y, \hat{h}(x)) P^{X,Y}(dx, dy).\end{aligned}$$

Examples. $L(y, \tilde{y}) = (y - \tilde{y})^2$, $L(y, \tilde{y}) = |y - \tilde{y}|$, $L(y, \tilde{y}) = \mathbf{1}_{y \neq \tilde{y}}$, ...

Problem

How can we **estimate this risk** (which depends on $P^{X,Y}$) ?

Refresher: empirical risk

We call **empirical risk** the risk

$$\widehat{\mathcal{R}}_n = \iint_{\mathcal{X} \times \mathcal{Y}} L(y, \hat{h}(x)) \hat{P}_n(dx, dy) = \frac{1}{n} \sum_{i=1}^n L(Y_i, \hat{h}(X_i))$$

computed with $P^{X,Y}$ equal to $\hat{P}_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i, Y_i}$.

Question

Is this empirical risk $\widehat{\mathcal{R}}_n$, in general, a “good” estimator of the true risk $\mathcal{R}(\hat{h})$?

 the data is used twice !

Intuition: It is “risky” to estimate the risk from the error observed on the same data already used to construct \hat{h} ...

25/34

Lecture outline

1 – Classification: logistic regression

- 1.1 – Introduction
- 1.2 – Linear models for classification
- 1.3 – Estimation of the parameter β
- 1.4 – Performance evaluation & choice of δ_0
- 1.5 – Extensions

2 – Estimation of the risk (generalization error)

- 2.1 – Problem
- 2.2 – Zoom on an illuminating special case
- 2.3 – Training set and test set

Zoom on an illuminating special case

Consider the case of “ordinary” linear regression:

- ▶ $h(x) = \beta_0 + \beta_1 x^{(1)} + \dots + \beta_p x^{(p)}$,
- ▶ quadratic loss: $L(y, \tilde{y}) = (y - \tilde{y})^2$,
- ▶ $p+1 \leq n$ and $\underline{X}^\top \underline{X}$ an a.s. invertible $(p+1) \times (p+1)$ matrix.

Empirical risk minimization : $\hat{\beta} = (\underline{X}^\top \underline{X})^{-1} \underline{X}^\top \underline{Y}$.

Remark: link between $\widehat{\mathcal{R}}_n$ and the coefficient R^2 of determination:

$$\begin{aligned} R^2 &= 1 - \frac{\text{RSS}(\hat{\beta})}{\text{TSS}} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{\beta}^\top \underline{X}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\ &= 1 - \frac{\widehat{\mathcal{R}}_n}{\widehat{\text{var}}_n(Y)} \quad \text{with } \widehat{\text{var}}_n(Y) = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2. \end{aligned}$$

26/34

Zoom on an illuminating special case (cont'd)

Consider the generalization error wrt responses only:

$$\widetilde{\mathcal{R}}_n = \mathbb{E} \left(\frac{1}{n} \sum_{i=1}^n (\widetilde{Y}_i - \hat{\beta}^\top \underline{X}_i)^2 \mid \underline{X}, \underline{Y} \right),$$

with, for all i , \widetilde{Y}_i and Y_i iid conditionally to \underline{X} .

Proposition

Assume that the unknown distribution $P^{X,Y}$ is such that $Y_i = \beta^\top \underline{X}_i + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, independent of \underline{X}_i . Then

$$\begin{aligned} \mathbb{E}(\widetilde{\mathcal{R}}_n) &= \sigma^2 \left(1 + \frac{p+1}{n} \right), \\ \mathbb{E}(\widehat{\mathcal{R}}_n) &= \sigma^2 \left(1 - \frac{p+1}{n} \right). \end{aligned}$$

27/34

Zoom on an illuminating special case (cont'd)

Interpretation. On average, the empirical risk under-estimates the generalization error:

$$\mathbb{E} \left(\tilde{\mathcal{R}}_n - \hat{\mathcal{R}}_n \right) = 2 \frac{p+1}{n} \sigma^2 > 0.$$

Another way of looking at this result. Set

$$\eta = \frac{p+1}{n} = \frac{\text{number of coefficients}}{\text{sample size}}.$$

Then

$$\frac{\mathbb{E} \left(\tilde{\mathcal{R}}_n \right)}{\mathbb{E} \left(\hat{\mathcal{R}}_n \right)} = \frac{1+\eta}{1-\eta} \xrightarrow{\eta \rightarrow 1} +\infty.$$

28/34

Zoom on an illuminating special case (cont'd)

Proof. Let us compute first $\mathbb{E} \left(\tilde{\mathcal{R}}_n \mid \underline{X} \right)$ with (reminder)

$$\tilde{\mathcal{R}}_n = \mathbb{E} \left(\frac{1}{n} \sum_{i=1}^n \left(\tilde{Y}_i - \hat{\beta}^\top X_i \right)^2 \mid \underline{X}, \underline{Y} \right).$$

We have $\mathbb{E} \left(\tilde{Y}_i \mid \underline{X} \right) = \mathbb{E} \left(\hat{\beta}^\top X_i \mid \underline{X} \right) = \beta^\top X_i$, therefore

$$\begin{aligned} \mathbb{E} \left(\tilde{\mathcal{R}}_n \mid \underline{X} \right) &= \frac{1}{n} \sum_{i=1}^n \text{var} \left(\tilde{Y}_i - \hat{\beta}^\top X_i \mid \underline{X} \right) \\ &= \frac{1}{n} \sum_{i=1}^n \left(\underbrace{\text{var} \left(\tilde{Y}_i \mid \underline{X} \right)}_{=\sigma^2} + \underbrace{\text{var} \left(\hat{\beta}^\top X_i \mid \underline{X} \right)}_{=0} \right). \end{aligned}$$

29/34

Zoom on an illuminating special case (cont'd)

We already know that $\text{var}(\hat{\beta} | \underline{X}) = \sigma^2 (\underline{X}^\top \underline{X})^{-1}$. Therefore:

$$\begin{aligned}\circledast &= \text{var}(\hat{\beta}^\top X_i | \underline{X}) \\ &= X_i^\top \text{var}(\hat{\beta} | \underline{X}) X_i \\ &= \sigma^2 X_i^\top (\underline{X}^\top \underline{X})^{-1} X_i \\ &= \sigma^2 \text{tr}((\underline{X}^\top \underline{X})^{-1} X_i X_i^\top).\end{aligned}$$

By noting that $\underline{X}^\top \underline{X} = \sum_i X_i X_i^\top$, we get:

$$\begin{aligned}\sum_i \text{var}(\hat{\beta}^\top X_i | \underline{X}) &= \sigma^2 \text{tr}((\underline{X}^\top \underline{X})^{-1} \sum_i X_i X_i^\top) \\ &= \sigma^2 \text{tr}(I_{p+1}) = \sigma^2(p+1).\end{aligned}$$

30/34

Zoom on an illuminating special case (cont'd)

Thus, we have:

$$\begin{aligned}\mathbb{E}(\tilde{\mathcal{R}}_n | \underline{X}) &= \frac{1}{n} \sum_{i=1}^n \left(\underbrace{\text{var}(\tilde{Y}_i | \underline{X})}_{=\sigma^2} + \underbrace{\text{var}(\hat{\beta}^\top X_i | \underline{X})}_{=\circledast} \right) \\ &= \sigma^2 + \sigma^2 \frac{p+1}{n} = \sigma^2 \left(1 + \frac{p+1}{n}\right).\end{aligned}$$

Hence the result: $\mathbb{E}(\tilde{\mathcal{R}}_n) = \sigma^2 \left(1 + \frac{p+1}{n}\right)$.

Exercise: prove the second inequality, i.e.,

$$\mathbb{E}(\widehat{\mathcal{R}}_n) = \sigma^2 \left(1 - \frac{p+1}{n}\right).$$

⇒ see PC



31/34

Lecture outline

1 – Classification: logistic regression

- 1.1 – Introduction
- 1.2 – Linear models for classification
- 1.3 – Estimation of the parameter β
- 1.4 – Performance evaluation & choice of δ_0
- 1.5 – Extensions

2 – Estimation of the risk (generalization error)

- 2.1 – Problem
- 2.2 – Zoom on an illuminating special case
- 2.3 – Training set and test set

Training set and test set

Conclusion/extrapolation. The empirical risk is in general

- ▶ a **negatively biased estimator** of the risk,
- ▶ with a bias that is increasing when $p \nearrow$.

Solution: split the data in two sets

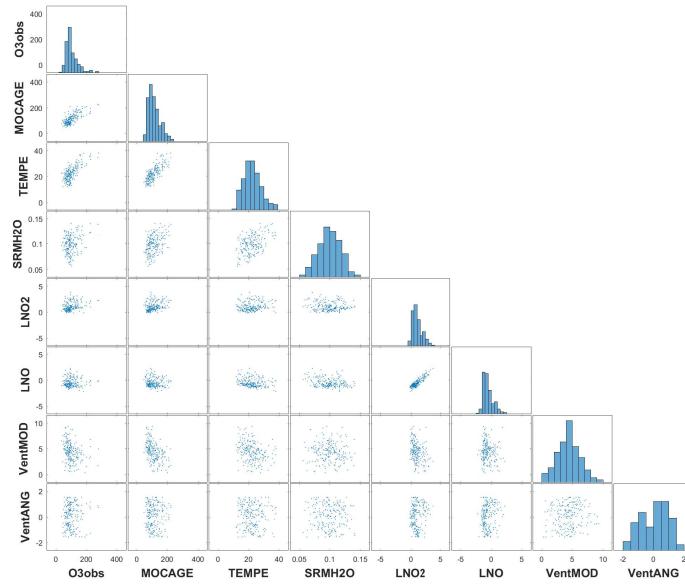
- ▶ **training** data: used to construct \hat{h} ,
- ▶ **test** data: used to estimate the generalization error.

Example:

training
(e.g., 80%)

test
(20%)

Exemple “Ozone” (cont’d from lecture #6)



Goal: predict the ozone concentration on day $t + 1$
from data available on day t

33 / 34

“Ozone” example: 70/30

All 7 explanatory variables and their 21 interactions are used.

Result from 10 random splits, 70% / 30%:

R^2	\hat{R}_n	\hat{R}_n^{test}
0.77185	345.1	573.32
0.76831	371.41	496.03
0.77292	343.96	608.62
0.76093	350.53	606.14
0.78584	345.45	669.66
0.75459	399.9	476.61
0.71367	343.72	643.72
0.77689	377.32	524.74
0.8176	317.83	695.86
0.79784	373.18	554.25

34 / 34

Chapter 8

Regularization and model selection



CentraleSupélec

Statistics and Learning

Arthur Tenenhaus[†], Julien Bect & Laurent Le Brusquet

(firstname.lastname@centralesupelec.fr)

Teaching: CentraleSupélec / Department of Mathematics

Research: Laboratory of signals and systems (L2S)

[†]: Course coordinator

1/37

Lecture 8/10

Regularization and model selection

In this lecture you will learn how to...

- ▶ Construct a regularized regression/classification model
- ▶ Include non-linearities in linear models
- ▶ Choose the value of hyper-parameters, select a model

2/37

Lecture outline

1 – Regularized regression (or classification): penalization

- 1.1 – Limitations of “ordinary least squares”
- 1.2 – Ridge regression
- 1.3 – LASSO regression

2 – Building models: feature engineering

- 2.1 – Non-linearities in linear models...
- 2.2 – Expansion in a basis

3 – Hyper-parameters, model selection

- 3.1 – Problem
- 3.2 – Cross validation
- 3.2 – AIC criterion

3/37

Lecture outline

1 – Regularized regression (or classification): penalization

- 1.1 – Limitations of “ordinary least squares”
- 1.2 – Ridge regression
- 1.3 – LASSO regression

2 – Building models: feature engineering

- 2.1 – Non-linearities in linear models...
- 2.2 – Expansion in a basis

3 – Hyper-parameters, model selection

- 3.1 – Problem
- 3.2 – Cross validation
- 3.2 – AIC criterion

Lecture outline

1 – Regularized regression (or classification): penalization

- 1.1 – Limitations of “ordinary least squares”
- 1.2 – Ridge regression
- 1.3 – LASSO regression

2 – Building models: feature engineering

- 2.1 – Non-linearities in linear models . . .
- 2.2 – Expansion in a basis

3 – Hyper-parameters, model selection

- 3.1 – Problem
- 3.2 – Cross validation
- 3.2 – AIC criterion

Limitations of “ordinary least squares”

Recall that \underline{X} has size #individuals \times #variables ($n \times (p + 1)$).

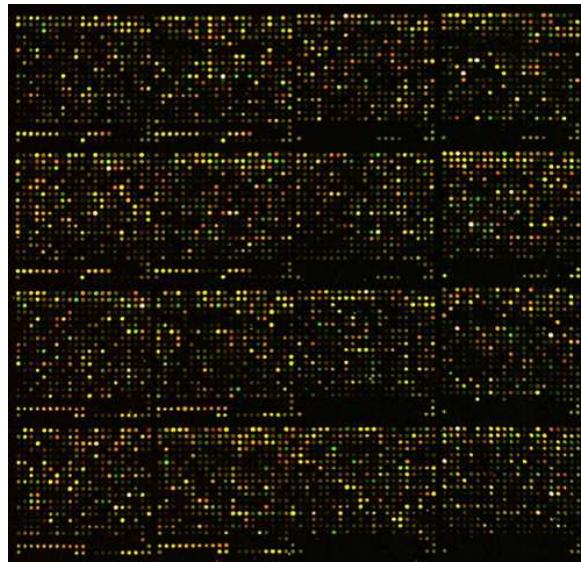
Critical situations for (ordinary) linear regression:

- ▶ when $\underline{X}^\top \underline{X}$ not invertible
- ▶ or poorly conditioned

Typical cases

- ① when the number of variables is large
- ② when there are strong correlations between explanatory variables

Example: $p \gg n$

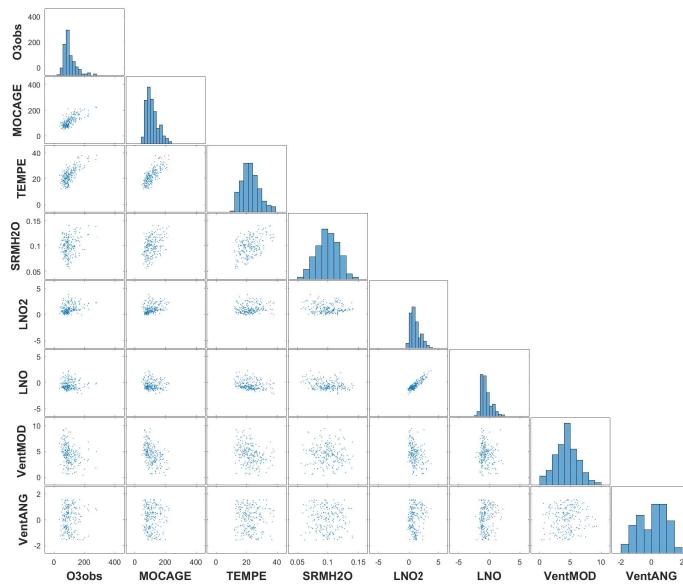


*Subset of a microarray for transcriptome analysis,
 $p \approx 25000$ for one patient*

Typically, $n \approx 10$ or 100 !

5/37

Example: strong correlation between explanatory variables



“Ozone” example → correlation between variables NO and NO2

6/37

Example: strong correlation... (cont'd)

Vector $\hat{\beta}$ obtained by OLS regression:

β_0	MOCAGE	TEMPE	RMH2O	NO2	NO	VentMOD	VentANG
103.4	28.9	22.5	-3.2	-34.4	37.9	1.4	2.6

Observations:

- ▶ The negative coefficient associated to NO2 is surprising
➡ hazardous interpretation of the coefficients
- ▶ The least influential variables (small coefficients) could perhaps be removed from the model?

7/37

One possible solution: penalized regression

A **penalty** term is added to the empirical risk :

$$\hat{\beta} = \arg \min_{\beta} \underbrace{\text{RSS}(\beta)}_{\text{data "fidelity"}}, + \underbrace{\lambda}_{\text{hyperparameter}} \underbrace{\Omega(\beta)}_{\text{penalty}}. \quad (*)$$

Expected benefits of penalization:

- ▶ make the solution of (*) **unique**,
- ▶ take **prior information** into account (this is related to the Bayesian approach),
- ▶ **avoid over-fitting** when the family of predictor functions is “large” (for linear models: $p \gg n$),
- ▶ make it **easier to interpret** the resulting model.

8/37

Lecture outline

1 – Regularized regression (or classification): penalization

1.1 – Limitations of “ordinary least squares”

1.2 – Ridge regression

1.3 – LASSO regression

2 – Building models: feature engineering

2.1 – Non-linearities in linear models...

2.2 – Expansion in a basis

3 – Hyper-parameters, model selection

3.1 – Problem

3.2 – Cross validation

3.2 – AIC criterion

Ridge regression

Penalty

$$\Omega(\beta) = \|\beta\|^2$$

$$\hat{\beta}^{\text{RIDGE}} = \arg \min_{\beta} \|\underline{Y} - \underline{X}\beta\|^2 + \lambda \|\beta\|^2$$

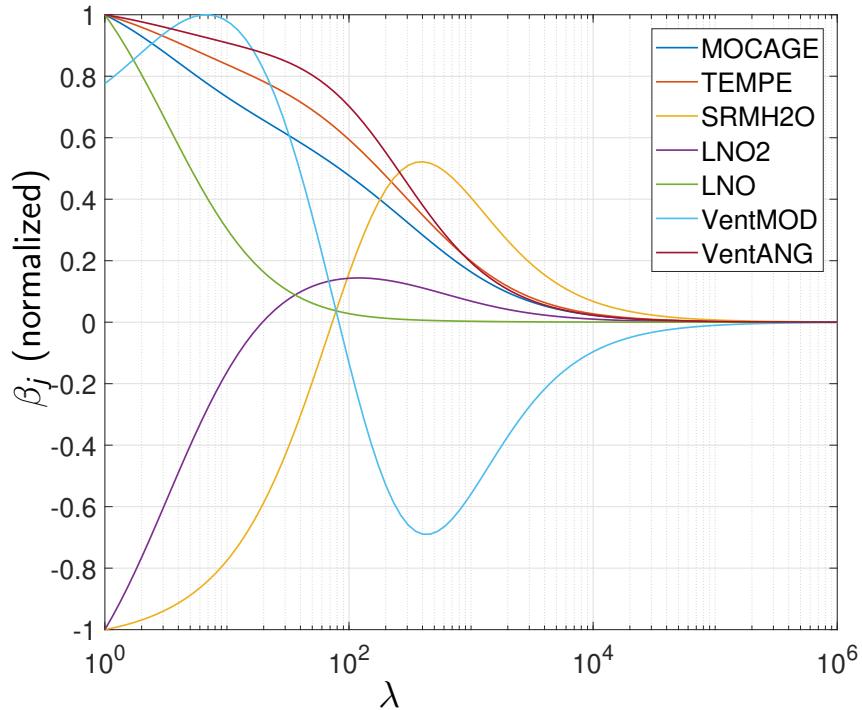
Exercise. Prove that:

$$\hat{\beta}^{\text{RIDGE}} = (\underline{X}^\top \underline{X} + \lambda I_{p+1})^{-1} \underline{X}^\top \underline{Y}.$$

⇒ When $\lambda \nearrow$, the conditioning of $(\underline{X}^\top \underline{X} + \lambda I_{p+1})$ improves.

Remark: $\hat{\beta}^{\text{RIDGE}}$ has a Bayesian interpretation (see PC).

“Ozone” example: β^{RIDGE} plot in function of λ



10/37

Lecture outline

- 1 – Regularized regression (or classification): penalization
 - 1.1 – Limitations of “ordinary least squares”
 - 1.2 – Ridge regression
 - 1.3 – LASSO regression

- 2 – Building models: feature engineering
 - 2.1 – Non-linearities in linear models...
 - 2.2 – Expansion in a basis

- 3 – Hyper-parameters, model selection
 - 3.1 – Problem
 - 3.2 – Cross validation
 - 3.2 – AIC criterion

LASSO regression

Penalty

$$\Omega(\beta) = \|\beta\|_1 = \sum_{j=1}^n |\beta_j|$$

$$\hat{\beta}^{\text{LASSO}} = \arg \min_{\beta} \|\underline{Y} - \underline{X}\beta\|^2 + \lambda \|\beta\|_1 \quad (*)$$

Minimization of the criterion

- ▶ no explicit expression for $\hat{\beta}^{\text{LASSO}}$
- ⇒ dedicated algorithms

11/37

LASSO regression: reformulation

$$\hat{\beta}^{\text{LASSO}} = \arg \min_{\beta} \|\underline{Y} - \underline{X}\beta\|^2 + \lambda \|\beta\|_1 \quad (*)$$

- ▶ Let $\hat{\beta}^{\text{OLS}}$ denote the OLS estimator of β :

$$\hat{\beta}^{\text{LASSO}} = \hat{\beta}^{\text{OLS}} \quad \text{for } \lambda = 0$$

- ▶ Since $\|\underline{Y} - \underline{X}\beta\|^2 = \|\underline{X}(\beta - \hat{\beta}^{\text{OLS}})\|^2 + c$, we have:

$$\hat{\beta}^{\text{LASSO}} = \arg \min_{\beta} \|\underline{X}(\beta - \hat{\beta}^{\text{OLS}})\|^2 + \lambda \|\beta\|_1$$

- ▶ Reformulation with a constraint: it can be proved that there exists $c_\lambda \in \mathbb{R}^+$ such that

$$\hat{\beta}^{\text{LASSO}} = \arg \min_{\beta} \|\underline{X}(\beta - \hat{\beta}^{\text{OLS}})\|^2 \text{ such that } \|\beta\|_1 \leq c_\lambda$$

(and similarly for $\hat{\beta}^{\text{RIDGE}}$)

12/37

LASSO regression: intuitive interpretation

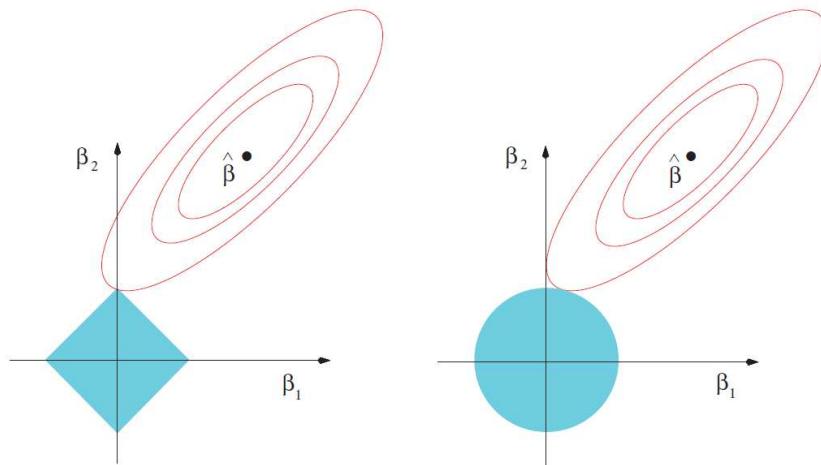
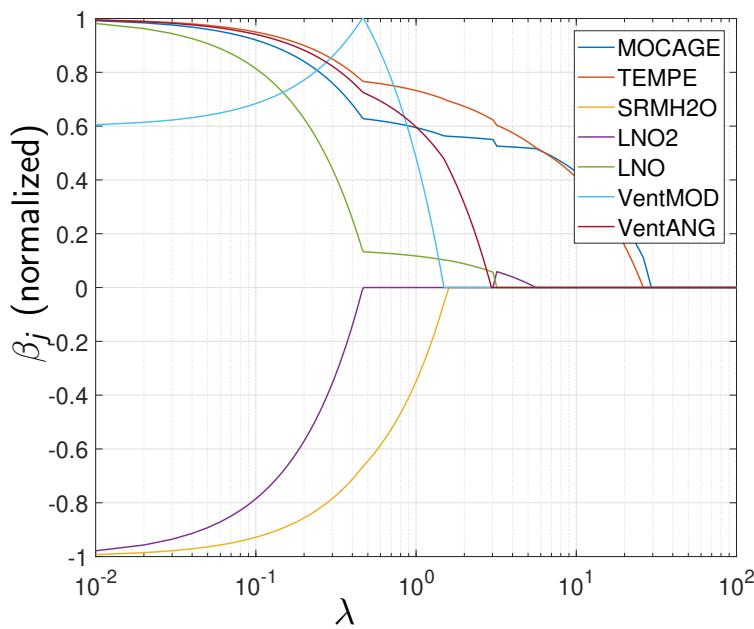


FIGURE 3.11. Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function.

Elements of Statistical Learning (2nd Ed.) ©Hastie, Tibshirani & Friedman 2009 Chap 3

13/37

“Ozone” example: $\hat{\beta}^{\text{LASSO}}$ versus λ



When $\lambda \nearrow$, the number of coefficients equal to zero ↗

14/37

“Ozone” example: $\hat{\beta}^{\text{LASSO}}$ for several λ

With $\lambda = 0$ (OLS)

β_0	MOCAGE	TEMPE	RMH2O	NO2	NO	VentMOD	VentANG
103.4	28.9	22.5	-3.2	-34.4	37.9	1.4	2.6

⇒ The coefficient for NO2 may seem surprising

With $\lambda = 0.5$

β_0	MOCAGE	TEMPE	RMH2O	NO2	NO	VentMOD	VentANG
103.4	18.1	17.2	-2.1	0	4.9	2.2	1.9

⇒ One of the two correlated variables is discarded,
makes it easier to **interpret the coefficients**

With $\lambda = 3$

β_0	MOCAGE	TEMPE	RMH2O	NO2	NO	VentMOD	VentANG
103.4	15.9	14.1	0	0	2.2	0	0

⇒ The remaining variables are progressively discarded

Choice of the hyper-parameter λ ?

15 / 37

Lecture outline

1 – Regularized regression (or classification): penalization

- 1.1 – Limitations of “ordinary least squares”
- 1.2 – Ridge regression
- 1.3 – LASSO regression

2 – Building models: feature engineering

- 2.1 – Non-linearities in linear models...
- 2.2 – Expansion in a basis

3 – Hyper-parameters, model selection

- 3.1 – Problem
- 3.2 – Cross validation
- 3.2 – AIC criterion

Lecture outline

1 – Regularized regression (or classification): penalization

- 1.1 – Limitations of “ordinary least squares”
- 1.2 – Ridge regression
- 1.3 – LASSO regression

2 – Building models: feature engineering

- 2.1 – Non-linearities in linear models...
- 2.2 – Expansion in a basis

3 – Hyper-parameters, model selection

- 3.1 – Problem
- 3.2 – Cross validation
- 3.2 – AIC criterion

Non-linearities in linear models...

If the empirical risk $\hat{\mathcal{R}}(\hat{h})$ is high, several possible causes:

- ▶ **noise**: intrinsic difficulty in predicting Y
 - ⇒ irreducible **statistical error**.
- ▶ **non-linearity** of the optimal predictor wrt the $X^{(j)}$'s
 - ⇒ reducible **approximation error**.

Possible workaround: $x^{(1)}, \dots, x^{(p)} \mapsto \tilde{x}^{(1)}, \dots, \tilde{x}^{(q)}$

- ▶ with $\tilde{x}^{(j)}$ function of $x^{(1)}, \dots, x^{(p)}$.
- ▶ The model is still **linear with respect to β** .

Examples

A few examples:

- ▶ **scalar transformations**: $\ln(x^{(j)})$, $\sqrt{x^{(j)}}$, $(x^{(j)})^k \dots$
- ▶ **interactions** (here, of order two): $x^{(j)}x^{(k)}$, $j \neq k$,
- ▶ higher-order interactions,
- ▶ (truncated) expansion in a basis...

 if $q \gg p$, **risk of over-fitting**.

Remarks: *feature engineering*

- ▶ Proposing new relevant variables
 - ⇒ **domain expertise** (or model selection...?)
- ▶ The same principle can be used to *reduce dimension*
 - ⇒ **features extraction**.

17/37

Lecture outline

1 – Regularized regression (or classification): penalization

- 1.1 – Limitations of “ordinary least squares”
- 1.2 – Ridge regression
- 1.3 – LASSO regression

2 – Building models: *feature engineering*

- 2.1 – Non-linearities in linear models...
- 2.2 – Expansion in a basis

3 – Hyper-parameters, model selection

- 3.1 – Problem
- 3.2 – Cross validation
- 3.2 – AIC criterion

Expansion in a basis

Principle

Let $\{\psi_m\}_{m>0}$ be a function basis of $L^2(\mathcal{X})^\dagger$.

Consider $\tilde{X}^{(m)} = \psi_m(X)$, $m = 1, \dots, M$

⇒ truncated expansion in the basis $\{\psi_m\}$.

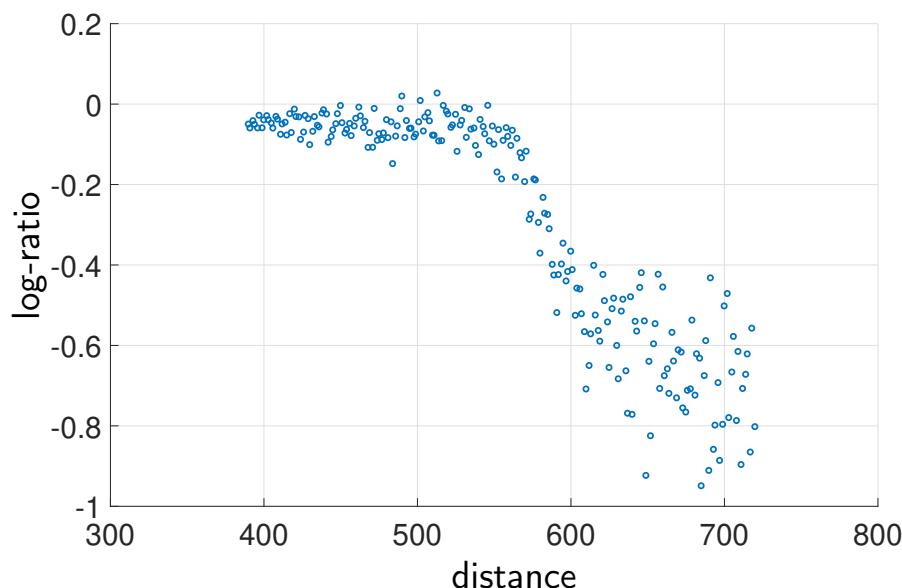
Examples of bases (preferably orthogonal):

- ▶ polynomial bases,
- ▶ wavelet bases,
- ▶ Fourier bases...

[†] or any other function space assumed to contain the optimal predictor h^* .

18/37

Example: LIDAR data



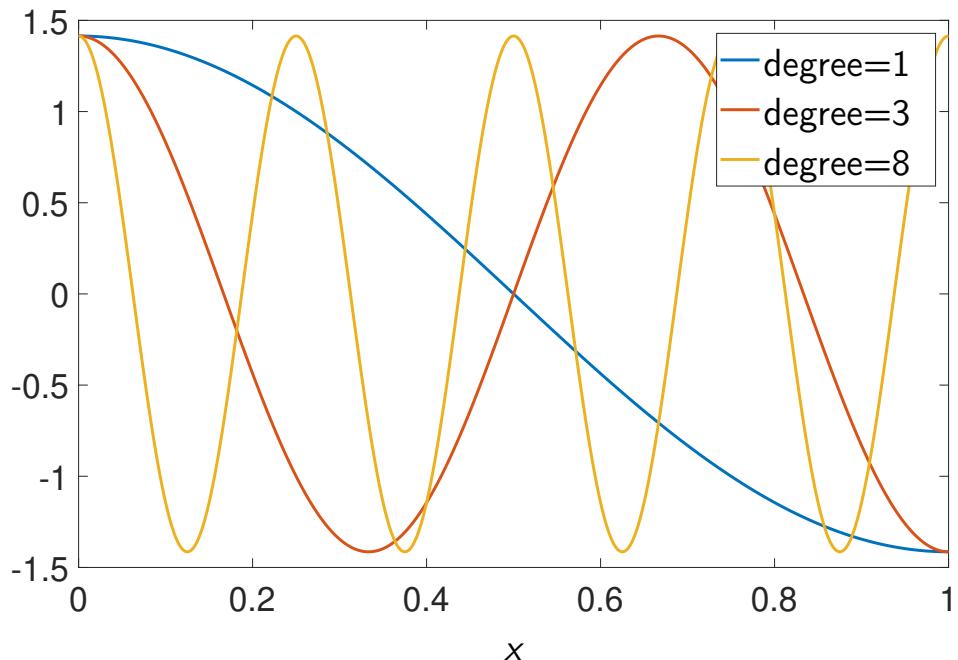
x-axis: distance travelled before the light is reflected back to its source
y-axis: logarithm of the ratio of received light from two laser sources

Data obtained from <http://matt-wand.utsacademics.info/webspr/lidar.html>

LIDAR: Light Detection And Ranging

19/37

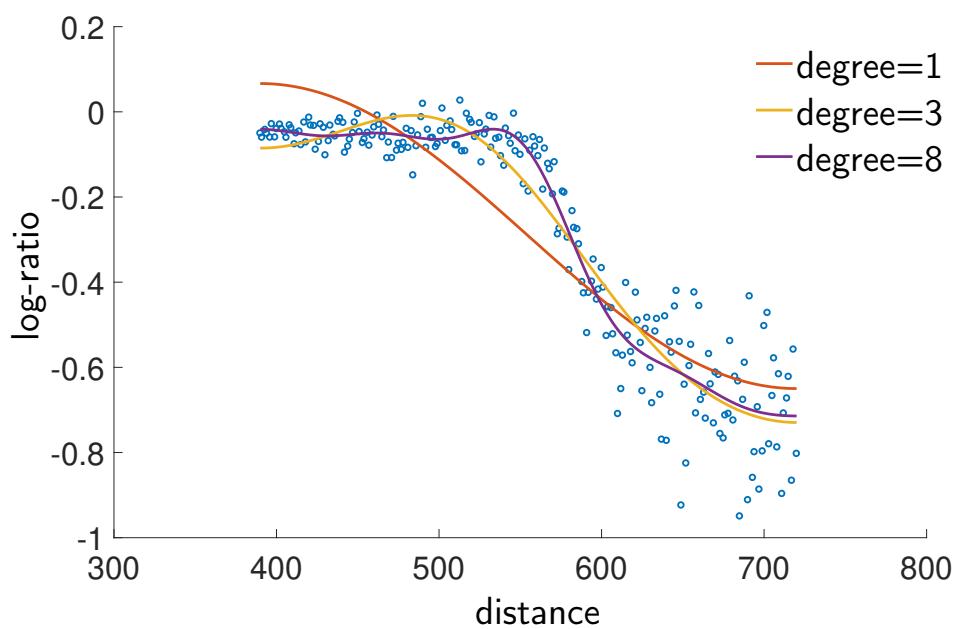
Basis of orthogonal cosines (basis of $L^2([0, 1])$)



20/37

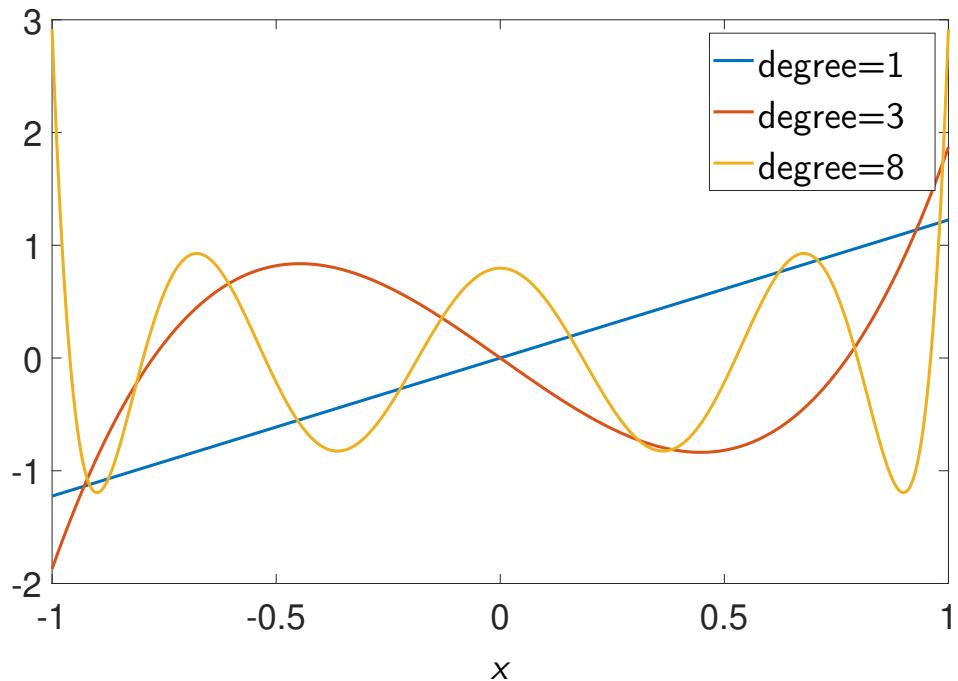
Example: LIDAR data (cont'd)

Quadratic loss + basis of cosines



21/37

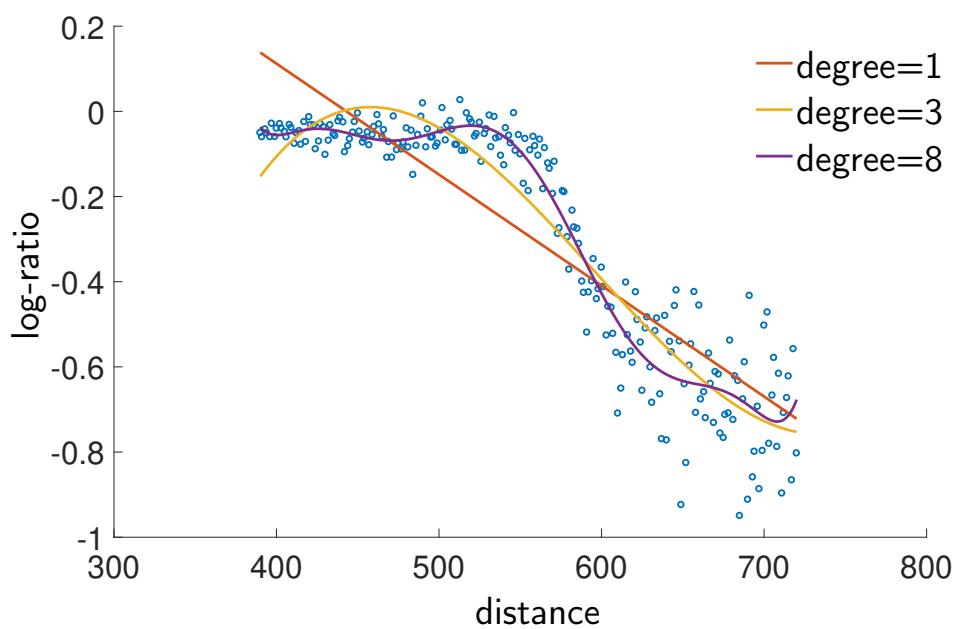
Legendre polynomials (orthonormal basis of $L^2([-1, 1])$)



22/37

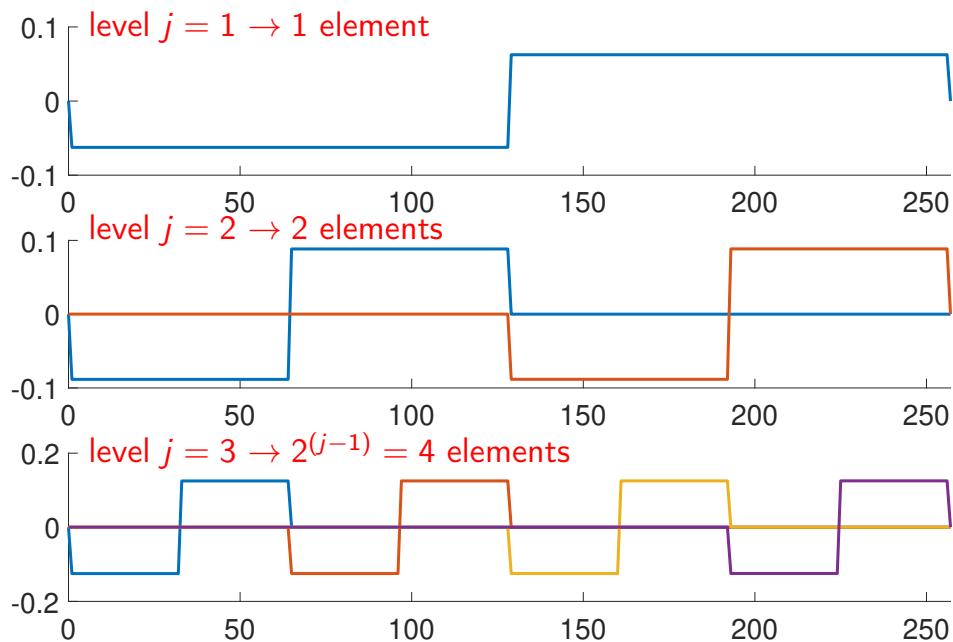
Example: LIDAR data (cont'd)

Quadratic loss + Legendre polynomials



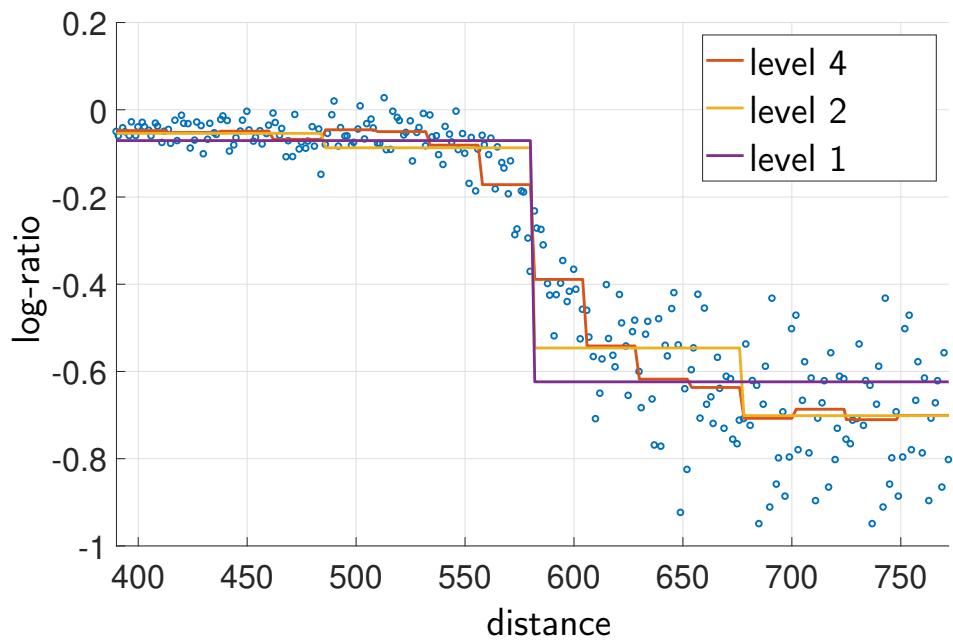
23/37

Haar wavelet basis



Example: LIDAR data (cont'd)

Quadratic loss + Haar wavelets



Lecture outline

1 – Regularized regression (or classification): penalization

- 1.1 – Limitations of “ordinary least squares”
- 1.2 – Ridge regression
- 1.3 – LASSO regression

2 – Building models: feature engineering

- 2.1 – Non-linearities in linear models...
- 2.2 – Expansion in a basis

3 – Hyper-parameters, model selection

- 3.1 – Problem
- 3.2 – Cross validation
- 3.2 – AIC criterion

Lecture outline

1 – Regularized regression (or classification): penalization

- 1.1 – Limitations of “ordinary least squares”
- 1.2 – Ridge regression
- 1.3 – LASSO regression

2 – Building models: feature engineering

- 2.1 – Non-linearities in linear models...
- 2.2 – Expansion in a basis

3 – Hyper-parameters, model selection

- 3.1 – Problem
- 3.2 – Cross validation
- 3.2 – AIC criterion

Problem #1: choosing a “good” family \mathcal{H}

Example. Selection of k variables among p . Let $J \subset \{1, \dots, p\}$:

$$h(x) = \beta_0 + \sum_{j \in J} \beta_j x^{(j)}.$$

⇒ Defines a family \mathcal{H}_J with $k_J = \text{card}(J) + 1$ parameters.

Example. Expansion in a basis, truncated at rank J :

$$h(x) = \sum_{k=0}^J \beta_k \psi_k(x).$$

⇒ Defines a family \mathcal{H}_J with $k_J = J + 1$ parameters.

Problem: model selection

How to choose the family \mathcal{H}_J (and, in particular, its “size” k_J) ?

Remark: replace $h(x)$ with $\ln \frac{h(x)}{1-h(x)}$ for logistic regression.

24/37

Problem #2: choosing a regularization hyper-parameter

Most methods require some “tuning”...

► Ridge/LASSO regression : $\hat{\beta} = \operatorname{argmin}_{\beta} \hat{\mathcal{R}}_{n,\lambda}^{\text{pen}}$, avec

$$\hat{\mathcal{R}}_{n,\lambda}^{\text{pen}}(\beta) = \hat{\mathcal{R}}_n(\beta) + \lambda \sum_j |\beta_j|^q, \quad q \in \{1, 2\},$$

► Choosing the number k of neighbours in a k -NN model:

$$h(x) = \frac{1}{k} \sum_{i \in \mathcal{V}_{n,k}(x)} y_i,$$

with $\mathcal{V}_{n,k}(x)$ the indices of the k nearest neighbours of x .

Problem: calibration

How to “tune” the value of such hyperparameters ?

25/37

Over-fitting: beware!

Idea

Choose the family \mathcal{H}_J , or the hyperparameter λ , in order to minimize (an estimation of) the generalization error.

⚠ again, the empirical risk $\hat{\mathcal{R}}_n$, estimated on the training data, is not appropriate !

Example. Polynomial regression in $x \in \mathbb{R}$, $\text{degree} \leq J$:

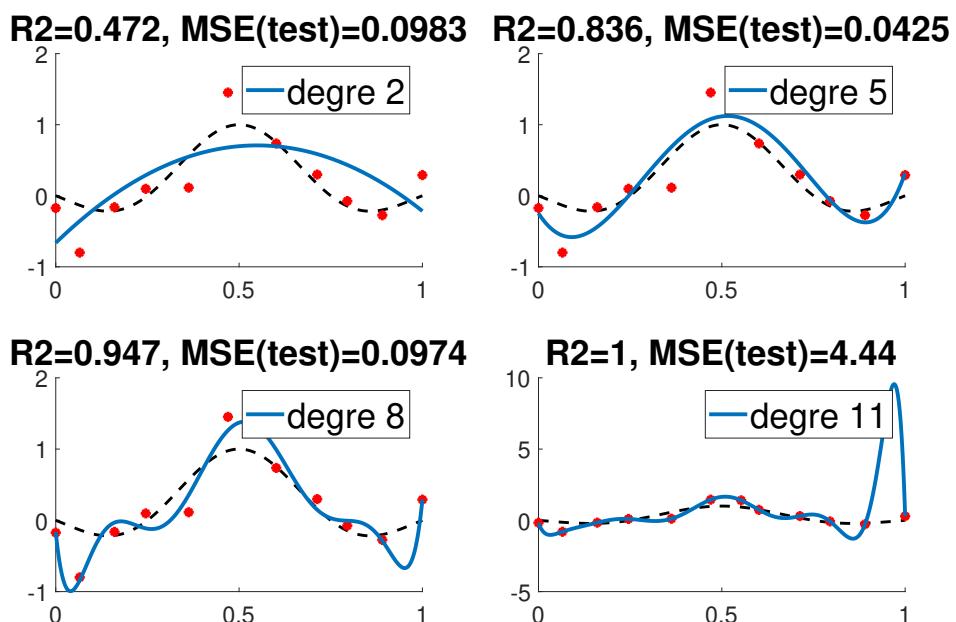
$$h(x) = \beta_0 + \beta_1 x + \dots + \beta_J x^J,$$

with $J = 2, 5, 8, 11$.

Recall that, in linear regression, the empirical risk has a downward bias proportional to the number of parameters in the model.

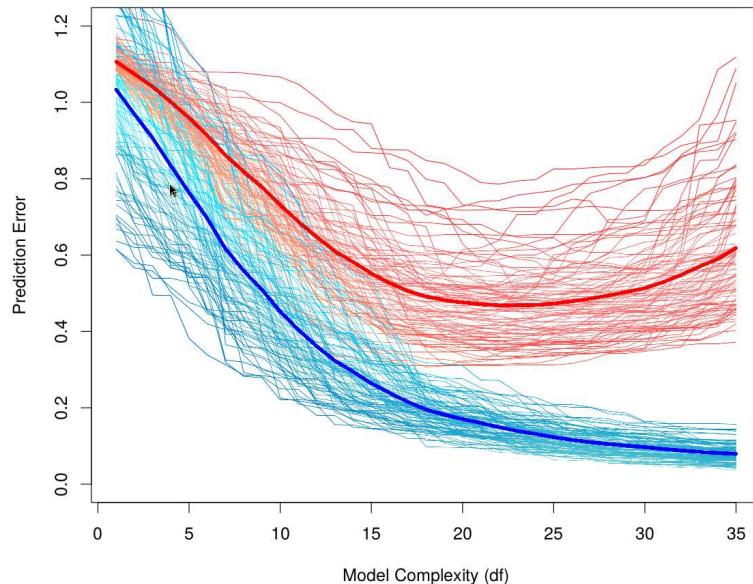
26/37

Example: polynomial regression



27/37

Understanding over-fitting: simulations



Blue: empirical risk $\hat{\mathcal{R}}_n$ / Red: error on the test set

Figure from Hastie, Tibshirani & Friedman (2017).
The Elements of Statistical Learning (12th edition), Springer.

28/37

Let's recapitulate...

Problem. We want to estimate the error to choose \mathcal{H} or λ but...

- ▶ it should be done neither on the **training data**
 (⇒ **over-fitting** problem),
- ▶ nor on the **test data**
 (⇒ **bias** in the final estimation of the generalization error).



29/37

Lecture outline

1 – Regularized regression (or classification): penalization

- 1.1 – Limitations of “ordinary least squares”
- 1.2 – Ridge regression
- 1.3 – LASSO regression

2 – Building models: feature engineering

- 2.1 – Non-linearities in linear models...
- 2.2 – Expansion in a basis

3 – Hyper-parameters, model selection

- 3.1 – Problem
- 3.2 – Cross validation
- 3.2 – AIC criterion

Solution: validation set

Idea: split the data in three sets

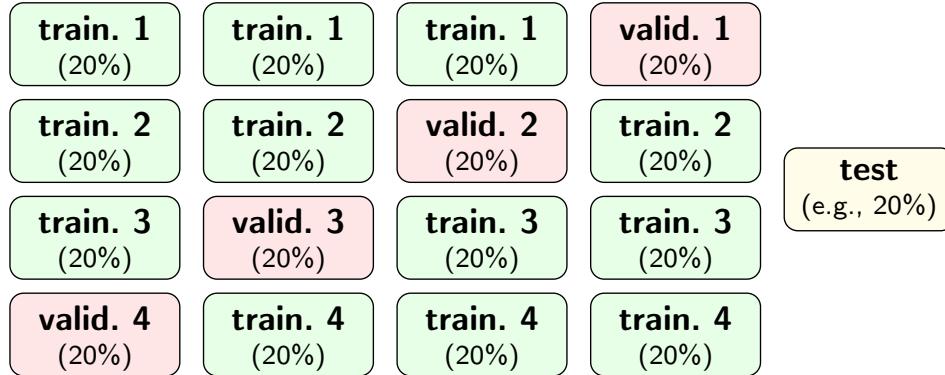
- ▶ **training** data: construct \hat{h} with given \mathcal{H}/λ ,
- ▶ **validation** set: choose \mathcal{H} , λ , etc.
- ▶ **test** data: estimate the generalization error.

Simple validation (hold-out)



Better validation: the cross validation method

k-fold cross-validation, here with $k = 4$:



► the error is averaged over the k validation sets.

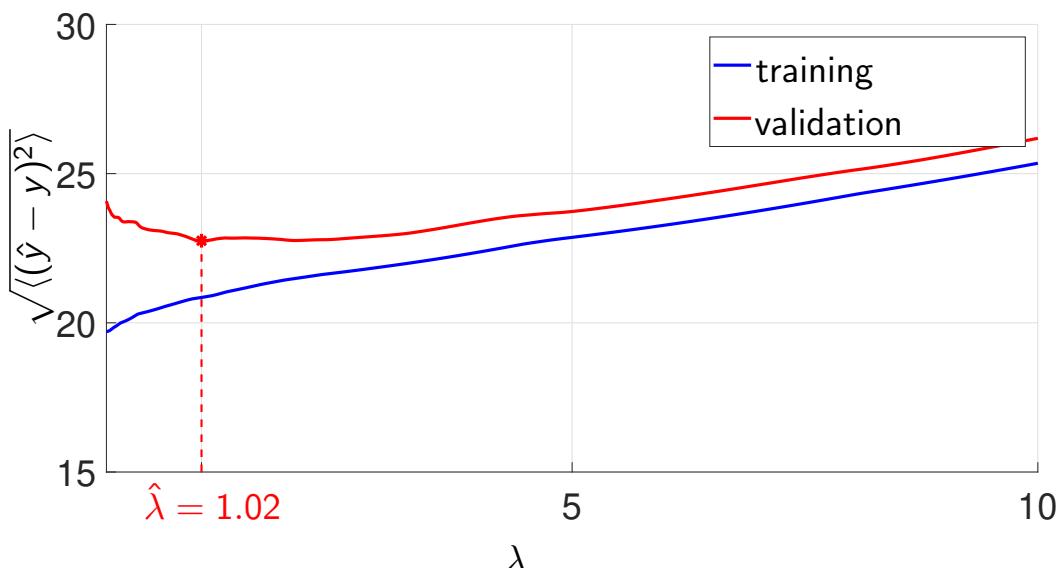
Special case: **leave-one-out** cross validation

- $k = n$ blocks (of size $n/k = 1$).

31/37

“Ozone” example: LASSO / choice of λ

- Predictor: LASSO regression using all variables and their interactions
- $\hat{\lambda}$ obtained by CV (LOO)



32/37

“Ozone” example: interactions

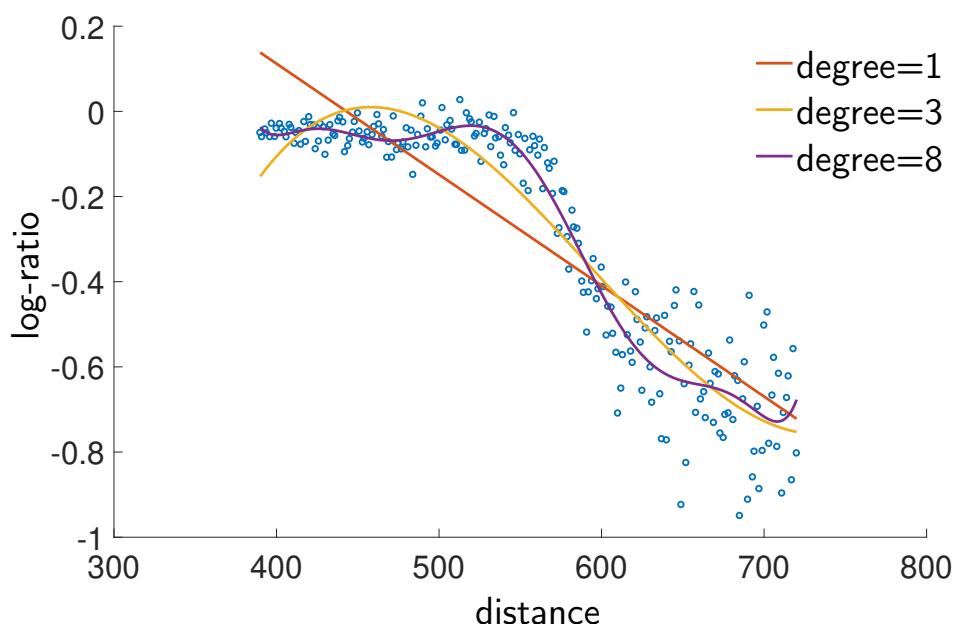
- We add variables of the form $X^{(j)}X^{(j')}$ and $X^{(j)}X^{(j')}X^{(j'')}$.
- LASSO regression (L^1 penalty).
- Hyper-parameter λ estimated through 10-fold CV.

model	$X^{(j)}$	$X^{(j)}X^{(j')}$	$X^{(j)}X^{(j')}X^{(j'')}$
total number of variables (q)	7	28	119
number of selected variables ($\beta_j \neq 0$)	4	9	8
\sqrt{MSE} CV (10-fold)	49.1	41.5	33.0
selected variables	MOCAGE TEMPE NO VentANG	MOCAGE TEMPE NO2 MOCAGE/TEMPE TEMPE/TEMPE TEMPE/MH2O TEMPE/NO2 NO2/VentANG VentANG/VentANG	MOCAGE TEMPE NO2 MOCAGE/TEMPE TEMPE/TEMPE TEMPE/RMH2O TEMPE²/MOCAGE VentANG ² /TEMPE

33/37

Example: LIDAR data (cont'd)

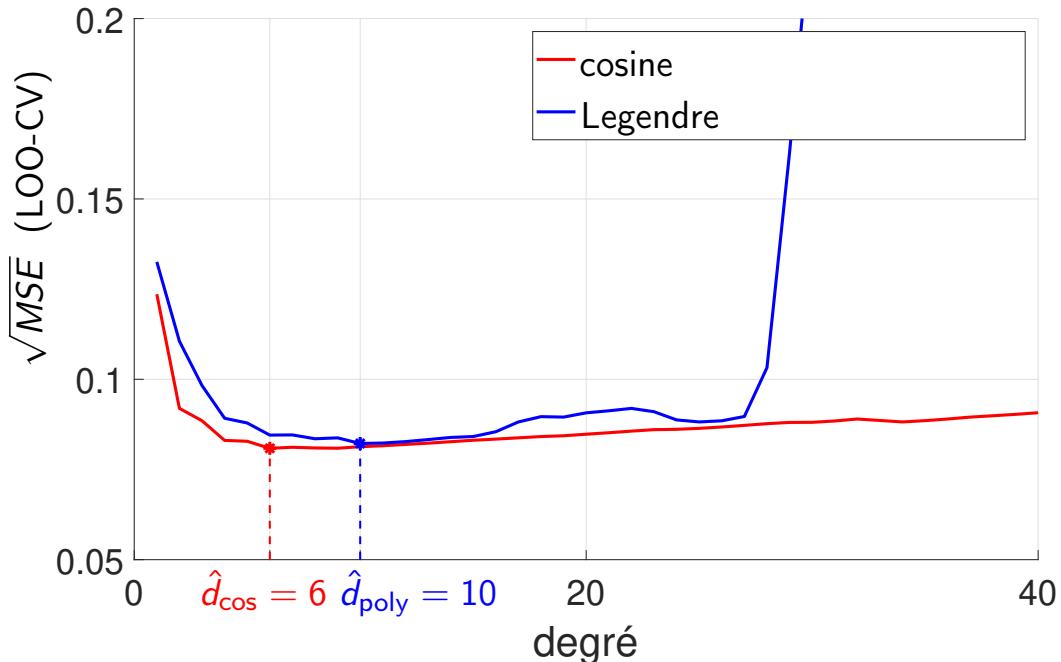
Quadratic loss + Legendre polynomials



34/37

Example: LIDAR data (cont'd)

Model selection



35 / 37

Lecture outline

1 – Regularized regression (or classification): penalization

- 1.1 – Limitations of “ordinary least squares”
- 1.2 – Ridge regression
- 1.3 – LASSO regression

2 – Building models: feature engineering

- 2.1 – Non-linearities in linear models...
- 2.2 – Expansion in a basis

3 – Hyper-parameters, model selection

- 3.1 – Problem
- 3.2 – Cross validation
- 3.2 – AIC criterion

Another approach to model selection: the AIC criterion

Assumption: parametric statistical models \mathcal{M}_j for $P^{Y|X}$.

Denote by $\hat{\theta}_j^{\text{MLE}}$ the MLE of θ in model \mathcal{M}_j .

Then the AIC criterion can also be used for model selection:

$$\hat{j} = \operatorname{argmin}_j \text{AIC}(j), \quad \text{AIC}(j) = -2 \ln \mathcal{L}(\hat{\theta}_j^{\text{MLE}}; \underline{X}, \underline{Y}) + 2k_j,$$

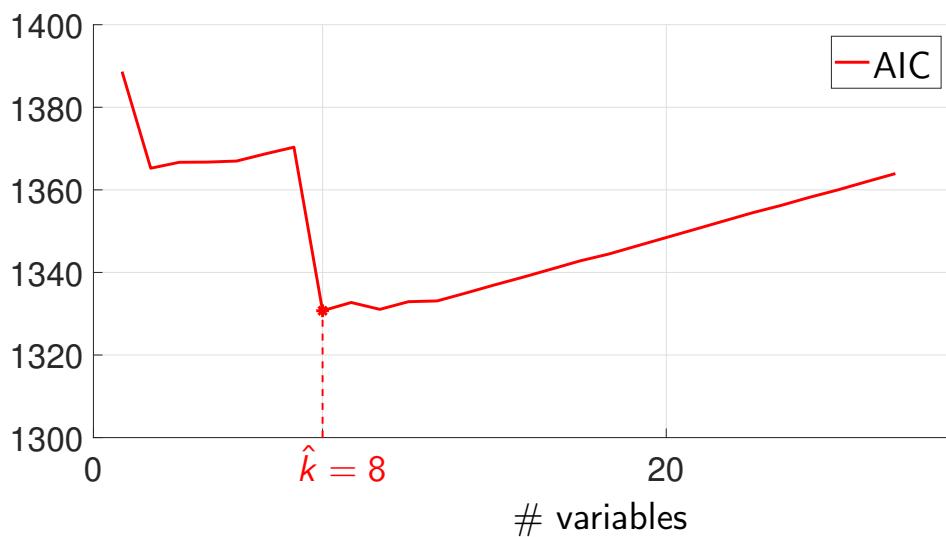
with k_j the number of parameters in model \mathcal{M}_j .

⇒ see PC for a partial justification (OLS linear regression)

36 / 37

“Ozone” example: AIC

- Predictor obtained by the ordinary least squares method, on an increasing number of variables
(linear terms first, then interactions)



37 / 37

Chapter 9

Some models for supervised learning



CentraleSupélec

Statistics and Learning

Arthur Tenenhaus[†], Julien Bect & Laurent Le Brusquet

(firstname.lastname@centralesupelec.fr)

Teaching: CentraleSupélec / Department of Mathematics

Research: Laboratory of signals and systems (L2S)

[†]: Course coordinator

1/33

Lecture 9/10

Some models for supervised learning

In this lecture you will learn how to...

- ▶ Predict with decision trees
- ▶ Predict with neural networks

2/33

Lecture outline

0 – Preliminary: classification with the log loss

1 – Decision trees

- 1.1 – Two introductory examples
- 1.2 – Recursive partitioning
- 1.3 – Prediction function

2 – Neural networks

- 2.1 – Neurons
- 2.2 – Multi-layer perceptrons
- 2.3 – Example
- 2.4 – Other architectures

3/33

Lecture outline

0 – Preliminary: classification with the log loss

1 – Decision trees

- 1.1 – Two introductory examples
- 1.2 – Recursive partitioning
- 1.3 – Prediction function

2 – Neural networks

- 2.1 – Neurons
- 2.2 – Multi-layer perceptrons
- 2.3 – Example
- 2.4 – Other architectures

Soft classification with the log loss

Back to logistic regression

- “soft” classifier: $h(x) = P_{\beta}^{Y|X}(Y = 1|X = x) \in [0, 1]$.

Definition: log loss for soft classification

$$L(y, h(x)) = \begin{cases} -\ln(h(x)) & \text{if } y = 1, \\ -\ln(1 - h(x)) & \text{if } y = 0. \end{cases}$$

Remark: $L(y, h(x)) \geq 0$ and $L(y, h(x)) = 0 \Leftrightarrow h(x) = y$.

Equivalence between MLE and empirical risk minimization

$$\begin{aligned} \hat{\mathcal{R}}(h) &= \sum_{i=1}^n L(y_i, h(x_i)) \\ &= -\ln \left(\underbrace{\prod_{i=1}^n (h(x_i))^{y_i} (1 - h(x_i))^{1-y_i}}_{\text{likelihood for } Y_i|X_i \stackrel{iid}{\sim} \text{Ber}(h(X_i))} \right) \end{aligned}$$

4/33

Lecture outline

0 – Preliminary: classification with the log loss

1 – Decision trees

- 1.1 – Two introductory examples
- 1.2 – Recursive partitioning
- 1.3 – Prediction function

2 – Neural networks

- 2.1 – Neurons
- 2.2 – Multi-layer perceptrons
- 2.3 – Example
- 2.4 – Other architectures

Lecture outline

0 – Preliminary: classification with the log loss

1 – Decision trees

- 1.1 – Two introductory examples
- 1.2 – Recursive partitioning
- 1.3 – Prediction function

2 – Neural networks

- 2.1 – Neurons
- 2.2 – Multi-layer perceptrons
- 2.3 – Example
- 2.4 – Other architectures

Binary classification: spam detection

Data collected over 4601 e-mails

- ▶ explanatory variables: relative freq. of 57 of the most used words
- ▶ variable to be explained: label “Spam” or “Email”
 - ➡ **categorical** variable (binary in this example)

TABLE 1.1. Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between spam and email.

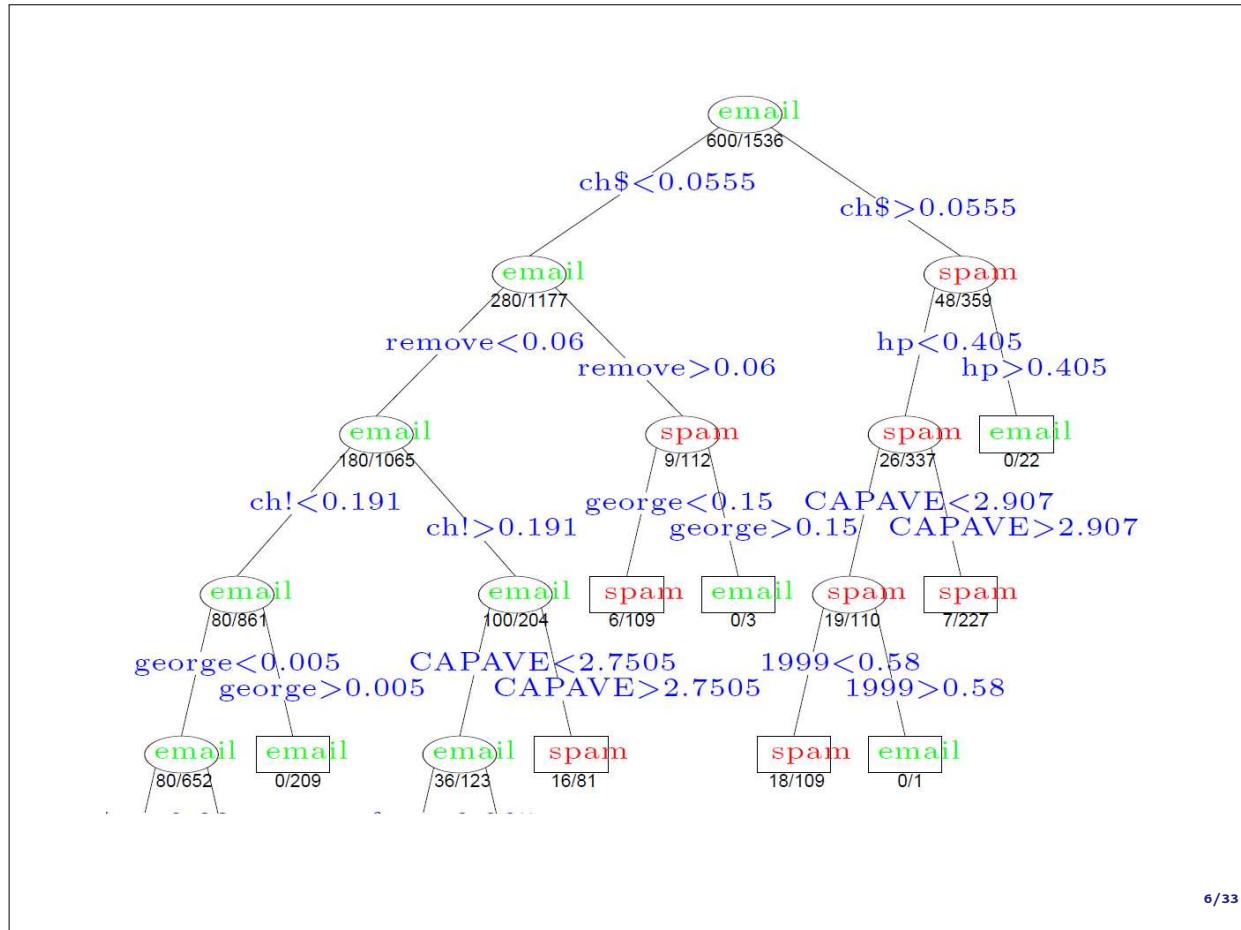
	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

TABLE 9.3. Spam data: confusion rates for the 17-node tree (chosen by cross-validation) on the test data. Overall error rate is 9.3%.

	Predicted	
	email	spam
True		
email	57.3%	4.0%
spam	5.3%	33.4%

Source: The Elements of Statistical Learning, Springer (for next slide also)

5/33



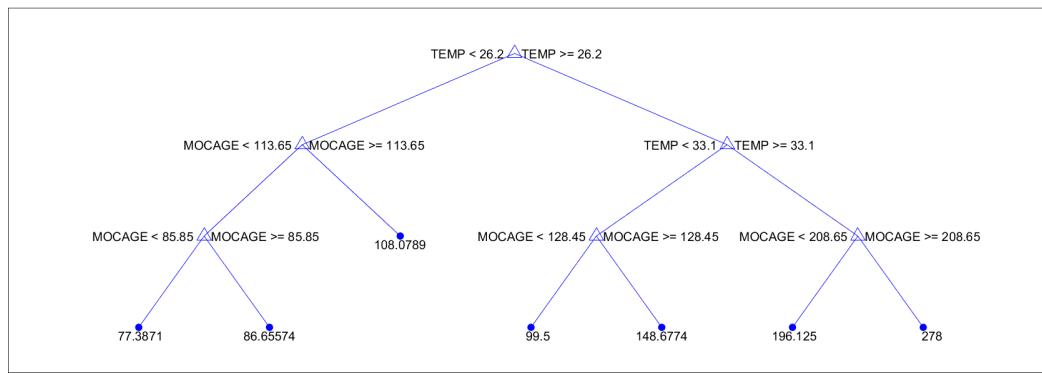
Regression tree: “Ozone” example

Simplified (for the sake of visualization)

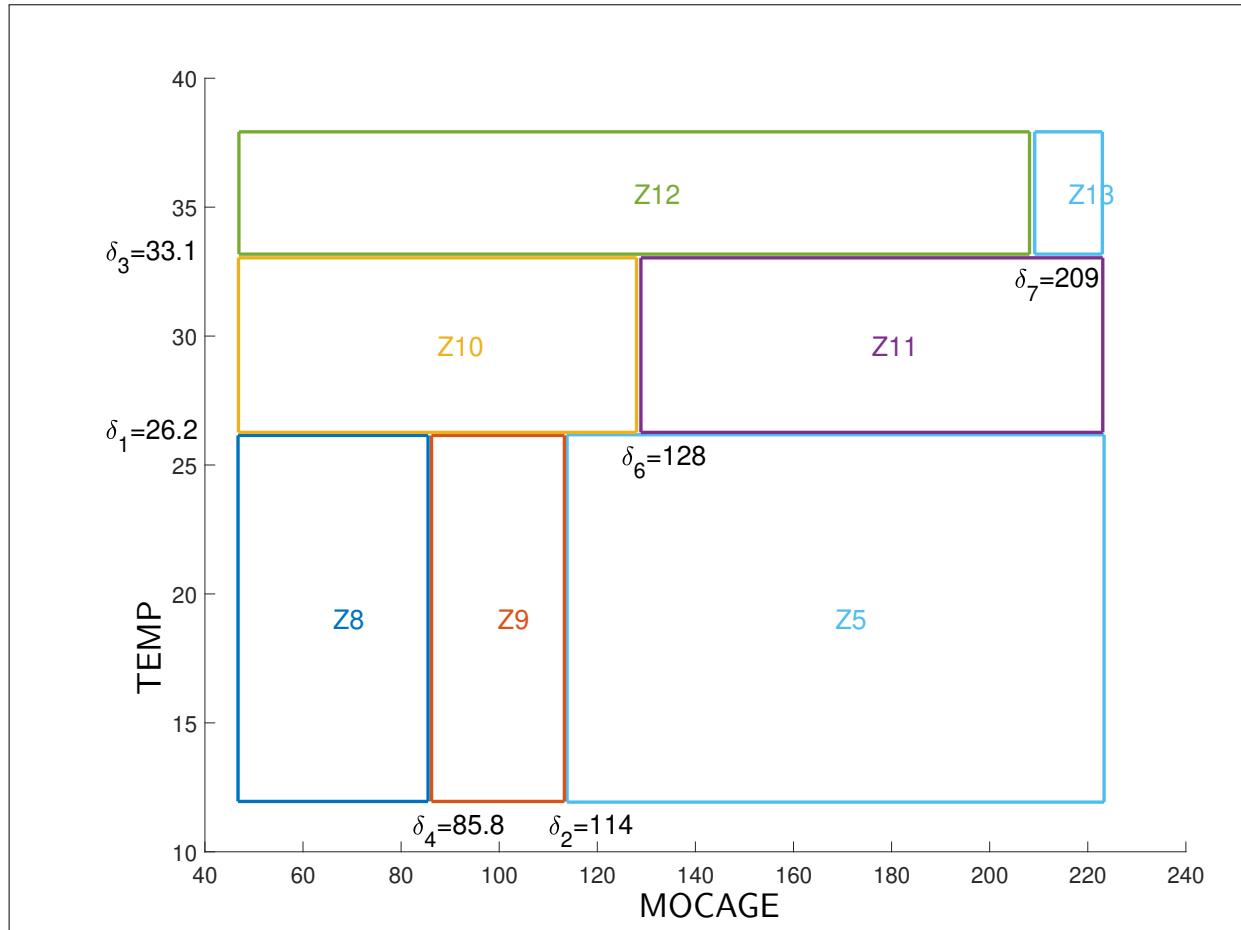
- ▶ predict variable 03 (**quantitative** variable)
- ▶ from variables MOCAGE and TEMP

Vocabulary. When the variable to be explained is

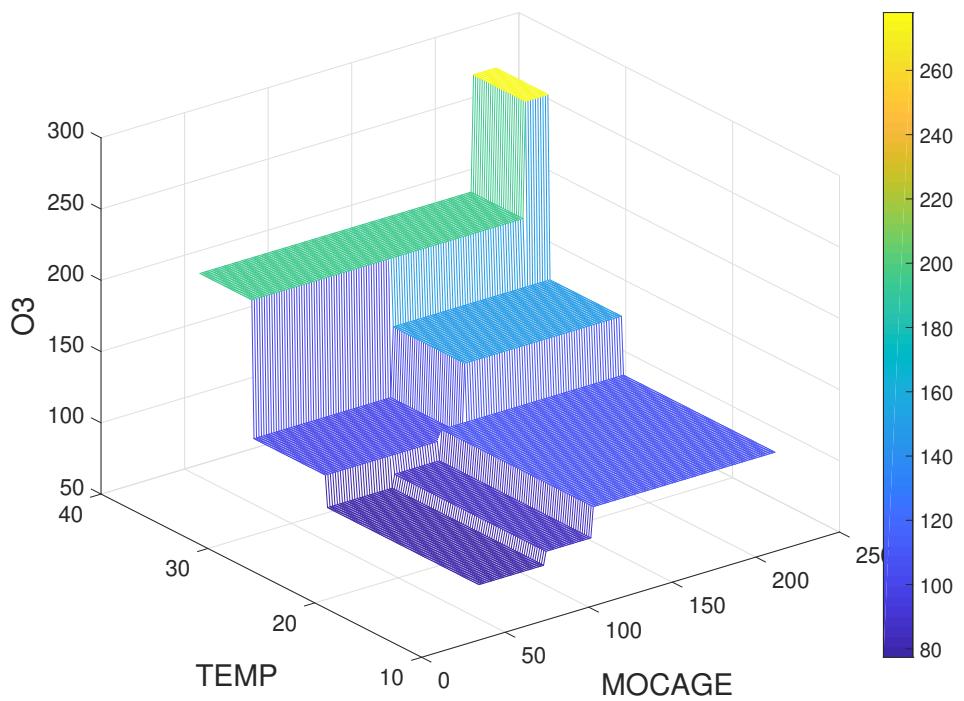
- ▶ **quantitative** → **regression tree**
- ▶ **categorical** → **classification tree**



7/33



Regression tree: “Ozone” example



Lecture outline

0 – Preliminary: classification with the log loss

1 – Decision trees

1.1 – Two introductory examples

1.2 – Recursive partitioning

1.3 – Prediction function

2 – Neural networks

2.1 – Neurons

2.2 – Multi-layer perceptrons

2.3 – Example

2.4 – Other architectures

Recursive partitioning: general principle

Goal

Construct a partition of \mathcal{X} from the data $(\underline{X}, \underline{Y})$.

Principle: **iterative** construction of a sequ. $(\mathcal{P}_m)_{m \geq 1}$ of partitions,

► $\mathcal{P}_m = \{Z_1^{(m)}, \dots, Z_m^{(m)}\}$, where partition \mathcal{P}_m contains m subsets.

Initialization: $\mathcal{P}_1 = \{\mathcal{X}\}$.

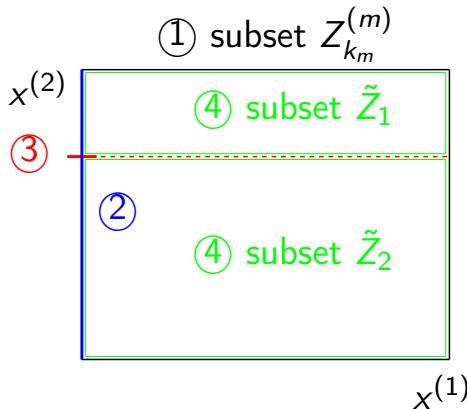
$\mathcal{P}_m \rightarrow \mathcal{P}_{m+1}$: **split** a subset $Z_{k_m}^{(m)}$ along one of the variables:

► $\tilde{Z}_1 = Z_{k_m}^{(m)} \cap \{x \text{ such that } x^{(j_m)} \leq \delta_m\}$

► $\tilde{Z}_2 = Z_{k_m}^{(m)} \cap \{x \text{ such that } x^{(j_m)} > \delta_m\}$

(the index j_m and the threshold δ_m still have to be specified)

An example with $p = 2$



Iteration $\mathcal{P}_m \rightarrow \mathcal{P}_{m+1}$:

- ▶ ① subset $Z_{k_m}^{(m)} \in \mathcal{P}_m$
- ▶ ② variable $x^{(j_m)}$ (here $j_m = 2$)
- ▶ ③ threshold δ_m
- ▶ ④ construction of \tilde{Z}_1 and \tilde{Z}_2

After splitting $Z_{k_m}^{(m)}$, we get:

$$\mathcal{P}_{m+1} = \mathcal{P}_m \cup \{\tilde{Z}_1, \tilde{Z}_2\} \setminus \{Z_{k_m}^{(m)}\}$$

11/33

Choice of k_m , j_m and δ_m

Let $D(Z)$ be a measure of the heterogeneity of a subset Z .

Example (for a quantitative label y)

$$D(Z) = \sum_{i \in Z} (y_i - \bar{y}_Z)^2$$

where \bar{y}_Z is the empirical mean computed over Z .

k_m , j_m and δ_m are jointly chosen in such a way that

$D(Z_{k_m}^{(m)}) - D(\tilde{Z}_1) - D(\tilde{Z}_2)$ is as large as possible

⇒ largest reduction of heterogeneity

(Recall that \tilde{Z}_1 and \tilde{Z}_2 are the subsets obtained by splitting $Z_{k_m}^{(m)}$)

12/33

Lecture outline

0 – Preliminary: classification with the log loss

1 – Decision trees

1.1 – Two introductory examples

1.2 – Recursive partitioning

1.3 – Prediction function

2 – Neural networks

2.1 – Neurons

2.2 – Multi-layer perceptrons

2.3 – Example

2.4 – Other architectures

Piecewise constant prediction function

Decision trees define a **piecewise constant prediction function** on the elements of the partition:

$$h_{\beta}(x) = \sum_{k=1}^m \beta_k \mathbb{1}_{Z_k^{(m)}}(x).$$

Remark: for a given partition, this is a **linear model** with respect to the m variables $\mathbb{1}_{Z_k^{(m)}}(x)$.

Estimation of the coefficients

Principle: to estimate $\beta^{(m)} = (\beta_1^{(m)}, \dots, \beta_m^{(m)})$,

- ▶ choose a **loss function** $L(y, h_\beta(x))$,
- ▶ then **minimize the empirical risk**.

Simplification:

$$\begin{aligned}\min_{\beta} \hat{\mathcal{R}}(h_\beta) &= \min_{\beta} \sum_{i=1}^n L(y_i, h_\beta(x_i)) \\ &= \min_{\beta} \sum_{k=1}^m \sum_{i \in Z_k^{(m)}} L(y_i, \beta_k) \\ &= \sum_{k=1}^m \min_{\beta_k} \sum_{i \in Z_k^{(m)}} L(y_i, \beta_k)\end{aligned}$$

Consequence: $\forall k, \hat{\beta}_k^{(m)} = \arg \min_{\beta_k} \sum_{i \in Z_k^{(m)}} L(y_i, \beta_k)$.

14/33

Two important special cases

Regression with the quadratic loss

$$\hat{\beta}_k^{(m)} = \operatorname{argmin}_{\beta_k} \sum_{i \in Z_k^{(m)}} (y_i - \beta_k)^2 = \bar{y}_{Z_k^{(m)}}$$

Binary classification with the logarithmic loss

Soft classification:

$$\begin{aligned}\hat{\beta}_k^{(m)} &= \operatorname{argmin}_{\beta_k \in [0,1]} \sum_{i \in Z_k^{(m)}} (-y_i \ln(\beta_k) - (1 - y_i) \ln(1 - \beta_k)) \\ &= \frac{1}{\operatorname{card}(Z_k^{(m)})} \cdot \operatorname{card} \left(i \in Z_k^{(m)} \text{ such that } y_i = 1 \right)\end{aligned}$$

Hard classification: threshold at $\delta_0 = \frac{1}{2}$ (cf. logistic regression).

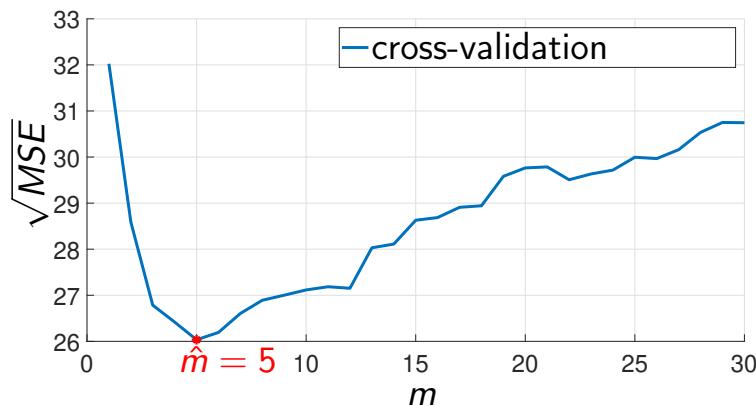
15/33

Choosing the size m of the partition

- ▶ m can either be given beforehand (\sim prior knowledge)
- ▶ or estimated by **cross-validation**.

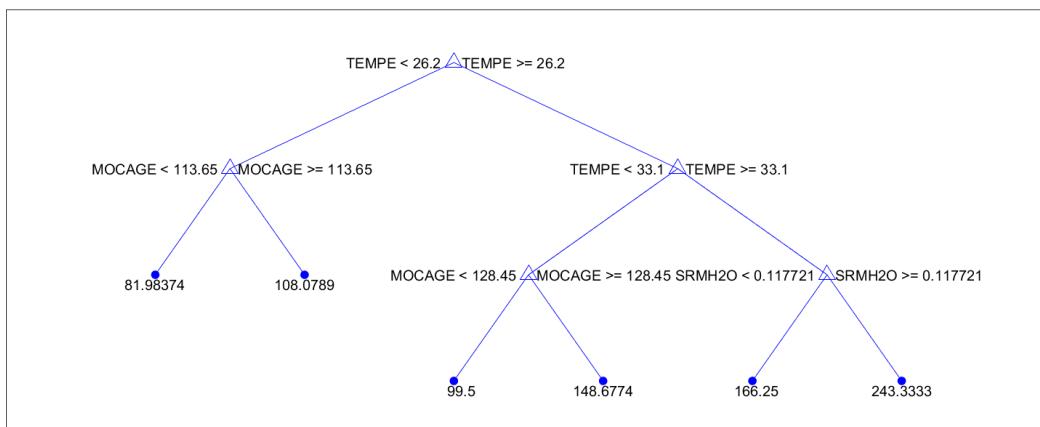
“Ozone” example

- ▶ Regression of O₃ with $p = 7$ explanatory variables
- ▶ m is chosen by leave-one-out cross-validation



16/33

Regression tree: “Ozone” example



```

1 if TEMPE<26.2 then node 2 elseif TEMPE>=26.2 then node 3 else 103.433
2 if MOCAGE<113.65 then node 4 elseif MOCAGE>=113.65 then node 5 else 88.1429
3 if TEMPE<33.1 then node 6 elseif TEMPE>=33.1 then node 7 else 153.673
4 fit = 81.9837
5 fit = 108.079
6 if MOCAGE<128.45 then node 8 elseif MOCAGE>=128.45 then node 9 else 138.59
7 if SRMH2O<0.117721 then node 10 elseif SRMH2O>=0.117721 then node 11 else 212.5
8 fit = 99.5
9 fit = 148.677
10 fit = 166.25
11 fit = 243.3333
  
```

17/33

More trees...

Disadvantages of decision trees

- ▶ high sensitivity to the sample ($\underline{x}, \underline{y}$)
- ▶ piecewise constant prediction on each subset (by construct.)
(not satisfactory if the optimal prediction function is smooth)

Extensions

- ▶ aggregation of decisions tree models
 - ⇒ **Random forests**
- ▶ weighted sum of weak classifiers
 - ⇒ **Boosting** (AdaBoost)

18/33

Lecture outline

0 – Preliminary: classification with the log loss

1 – Decision trees

- 1.1 – Two introductory examples
- 1.2 – Recursive partitioning
- 1.3 – Prediction function

2 – Neural networks

- 2.1 – Neurons
- 2.2 – Multi-layer perceptrons
- 2.3 – Example
- 2.4 – Other architectures

Lecture outline

0 – Preliminary: classification with the log loss

1 – Decision trees

- 1.1 – Two introductory examples
- 1.2 – Recursive partitioning
- 1.3 – Prediction function

2 – Neural networks

- 2.1 – Neurons
- 2.2 – Multi-layer perceptrons
- 2.3 – Example
- 2.4 – Other architectures

The (multipolar) biological neuron: axons, dendrites...

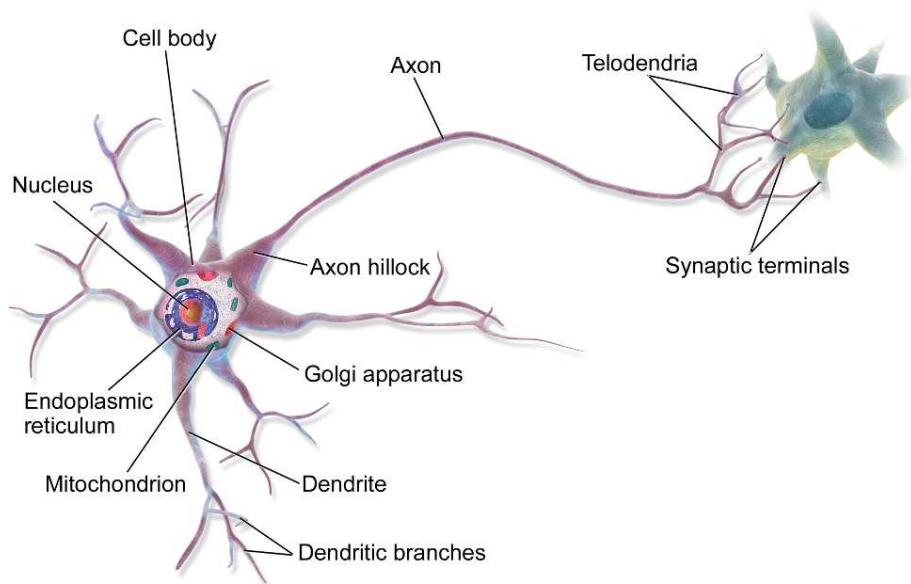


Image: Bruce Blaus, <https://commons.wikimedia.org>, CC BY 3.0

"A multipolar neuron is a type of neuron that possesses a single axon and many dendrites (and dendritic branches), allowing for the integration of a great deal of information from other neurons." (https://fr.wikipedia.org/wiki/Neurone_multipolaire)

19/33

The artificial neuron

Definition: neuron (McCulloch and Pitts, 1943)[†]

In statistical learning, a **neuron** with p variables (inputs) is a function, generally non-linear[‡], of the form

$$h(x) = \varphi(wx + b), \quad x \in \mathbb{R}^p,$$

where

- ▶ φ is an increasing $\mathbb{R} \rightarrow \mathbb{R}$ function;
- ▶ $w \in \mathbb{R}^{1 \times p}$, and $b \in \mathbb{R}$.

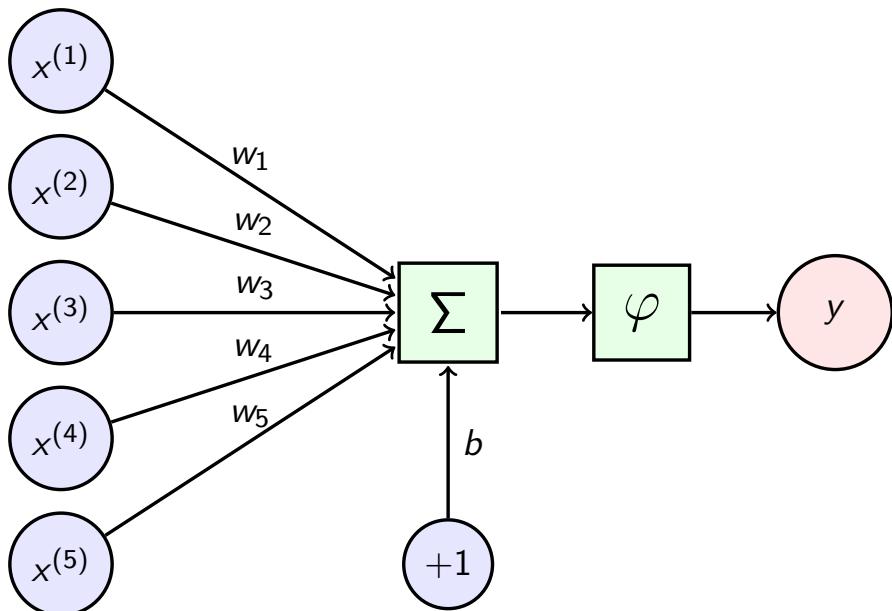
Vocabulary

- ▶ φ : **activation function**,
- ▶ w_1, \dots, w_p : **weights**,
- ▶ b : **bias** (nothing to do with the bias of an estimator).

[†] The original neuron of McCulloch & Pitts (1943) specifically used $\varphi = \text{sgn}$ as an activation function.

[‡] We will see later a situation where a linear neuron ($\varphi = \text{Id}$) is used.

The artificial neuron: illustration ($p = 5$)



Activation functions

Discontinuous activation functions (not recommended[†]):

- ▶ **Heaviside** function: $\varphi(v) = \mathbb{1}_{v \geq 0}$, or
- ▶ **sign** function: $\varphi(v) = \text{sgn}(v) = \mathbb{1}_{v > 0} - \mathbb{1}_{v < 0}$.

“S-shaped” functions, a.k.a. **sigmoids**:

- ▶ **logistic**[‡] : $\varphi(v) = \frac{1}{1+e^{-v}} = \frac{1}{2} + \frac{1}{2} \tanh\left(\frac{v}{2}\right)$, or
- ▶ **tanh** : $\varphi(v) = \tanh(v) = \frac{e^v - e^{-v}}{e^v + e^{-v}}$.

The **ReLU** (*Rectified Linear Unit*) function:

- ▶ $\varphi(v) = \max(0, v)$.

[†] Used in the oldest models, most notably the Rosenblatt's perceptron (1957), but abandoned since then because of their almost-everywhere zero gradient, which creates problems for optimization procedures.

[‡] The word “sigmoid” sometimes refers to this particular function.

22/33

Activation functions (cont'd)

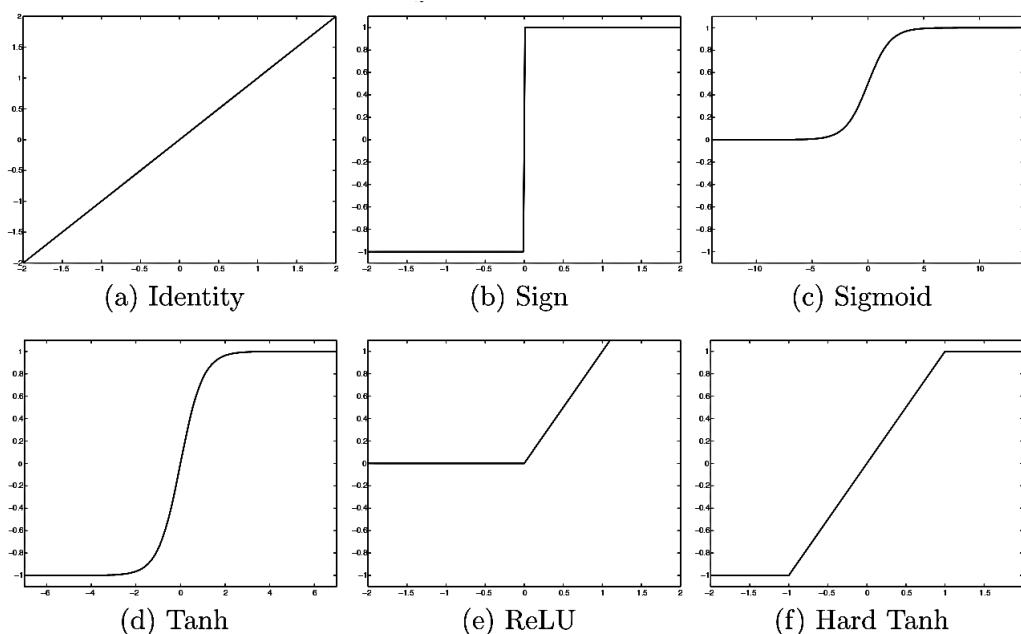


Image: C. C. Aggarwal (2018). *Neural networks and Deep Learning*, Springer.

23/33

Remark: relation with logistic regression

Remark. With the **logistic activation function** (sigmoid),

$$y = \varphi(v) = \frac{1}{1 + e^{-v}} \quad \Leftrightarrow \quad v = \ln\left(\frac{y}{1 - y}\right).$$

Since $v = wx + b$, we recover for $h(x)$ the form of the **logistic regression predictor**.

24/33

Lecture outline

0 – Preliminary: classification with the log loss

1 – Decision trees

- 1.1 – Two introductory examples
- 1.2 – Recursive partitioning
- 1.3 – Prediction function

2 – Neural networks

- 2.1 – Neurons
- 2.2 – Multi-layer perceptrons
- 2.3 – Example
- 2.4 – Other architectures

Multi-layer perceptron: definition

Let p, K be non-zero integers.

Definition: multi-layer perceptron[†] (MLP)

We call **multi-layer perceptron** with $M + 1$ layers, p variables (input) and K responses (output), any function $\mathbb{R}^p \rightarrow \mathbb{R}^K$ of the form

$$h = (\underline{\varphi}_M \circ g_M) \circ \cdots \circ (\underline{\varphi}_j \circ g_j) \circ \cdots \circ (\underline{\varphi}_1 \circ g_1),$$

where[‡]

- ▶ $g_k : \mathbb{R}^{m_{k-1}} \rightarrow \mathbb{R}^{m_k}$ is **affine**,
- ▶ $\underline{\varphi}_k : \mathbb{R}^{m_k} \rightarrow \mathbb{R}^{m_k}$ represents the action coordinate by coordinate of an **increasing function** $\varphi_k : \mathbb{R} \rightarrow \mathbb{R}$.
- ▶ m_0, m_1, \dots, m_M : non-zero integers, $m_0 = p, m_M = K$.

[†] Rosenblatt's original perceptron (1957) did not include hidden layers ($M = 1$). It was using the activation function $h(x) = \text{sgn}(x)$ as McCulloch and Pitts (1943), and weights $w_j \in \{-1, +1, -\infty\}$.

[‡] there will be one exception this rule later ("softmax" layer)

Multi-layer perceptron: definition (cont'd)

Vocabulary: **layers** of variables

- ▶ $z[0] = x$: input layer,
- ▶ $z[k] = (\underline{\varphi}_k \circ g_k)(z[k-1]), 1 \leq k < M$: **hidden layers**,
- ▶ $z[M] = y = (\underline{\varphi}_M \circ g_M)(z[M-1])$: output layer.

Remark. Let us write

$$g_k(z[k-1]) = W_k z[k-1] + b_k.$$

Then, for all $j \in \{1, \dots, m_k\}$ we recognize a **neuron**:

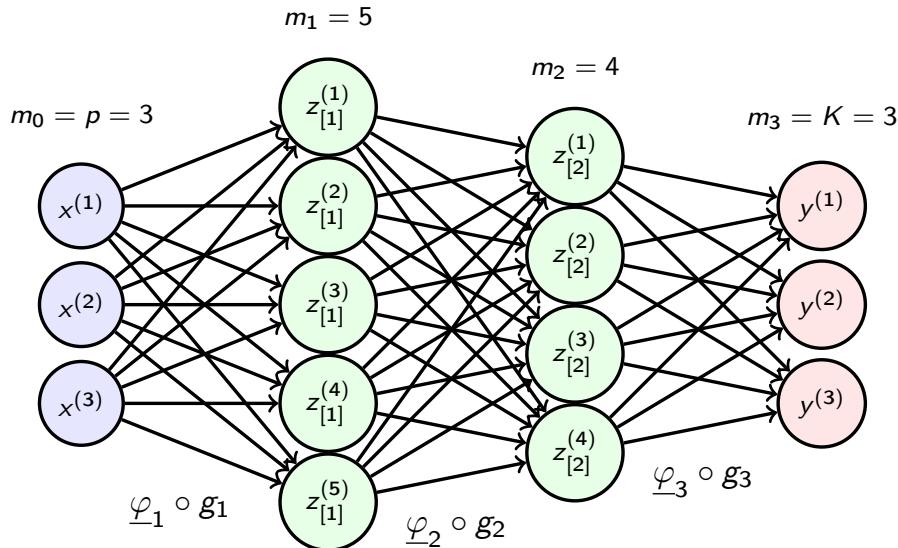
$$z^{(j)}_k = \varphi_k(w_{k,j} z[k-1] + b_k^{(j)}),$$

where $w_{k,j} = e_j^\top W_k$ is the j -th row of W_k .

► Vocabulary: weights, bias, activation function.

Multi-layer perceptron: illustration

Example of a multi-layer perceptron with $p = 3$ inputs, $K = 3$ outputs, and two hidden layers of sizes $m_1 = 5$ and $m_2 = 4$.



Vocabulary: fully connected, feed-forward neural network

27/33

Output layer: activation function

The output layer must be **adapted to the problem** at hand...

Regression. $\mathcal{Y} \subset \mathbb{R}$, or more generally \mathbb{R}^K .

- ▶ Perceptron with K outputs
- ▶ Activation function: $\varphi_M = \text{Id}$.
- ▶ Thus the last transformation $(\varphi_M \circ g_M)$ is **linear** (affine).

Classification. K classes, $\mathcal{Y} = [0, 1]^K$ ("soft" classification).

- ▶ Perceptron with K outputs, with $m_{M-1} = m_M = K$.
- ▶ Exception to the definition \Rightarrow the "**softmax**" layer:

$$z_{[M]}^{(j)} = \frac{\exp(z_{[M-1]}^{(j)})}{\sum_{j'=1}^p \exp(z_{[M-1]}^{(j')})}, \quad \sum_{j=1}^K z_{[M]}^{(j)} = 1.$$

Remark: alternatively, for **binary classification**, we can use a single output ($K = 1$ instead of $K = 2$) with the **logistic** function used as the activation function on the last layer.

28/33

Training: loss functions and regularization

The most commonly used loss functions[†] are

- ▶ **regression:** the **quadratic loss**
 - ▶ $L(y, \tilde{y}) = (y - \tilde{y})^2$ for the single-output case,
 - ▶ $L(y, \tilde{y}) = \|y - \tilde{y}\|^2$ if $K > 1$.
- ▶ **(soft) classification:** the **logarithmic loss**
 - ▶ For all $j \in \{1, \dots, K\}$, we have $y^{(j)} \in \{0, 1\}$ and $\tilde{y}^{(j)} \in [0, 1]$.
 - ▶ $L(y, \tilde{y}) = -\sum_{j=1}^K y^{(j)} \ln(\tilde{y}^{(j)})$.

Nb parameters is high \Rightarrow **regularize** to avoid **over-fitting**

- ▶ **penalization**, for instance L^1 (LASSO) or L^2 (ridge);
- ▶ other (not covered): early stopping, drop out...

[†] for instance https://scikit-learn.org/stable/modules/neural_networks_supervised.html

Training: numerical optimization

We want to **minimize the empirical risk** (possibly penalized)

$$\hat{\mathcal{R}}_n(\theta) = \frac{1}{n} \sum_{i=1}^n L(Y_i, h_\theta(X_i)),$$

where θ denotes the parameters of the model (weights, biases).

- ▶ Numerical methods are used to this end.

These methods use the **gradient of the criterion**. Two remarks:

- ▶ computational burden when n is large: random “**mini-batches**”
 - ▶ **stochastic gradient** method (not covered);
- ▶ recursive computation of the **gradient of a composition of fcts**
 - ▶ **back-propagation** method (not covered).

Lecture outline

0 – Preliminary: classification with the log loss

1 – Decision trees

- 1.1 – Two introductory examples
- 1.2 – Recursive partitioning
- 1.3 – Prediction function

2 – Neural networks

- 2.1 – Neurons
- 2.2 – Multi-layer perceptrons
- 2.3 – Example
- 2.4 – Other architectures

Example: MNIST

A 10x10 grid of handwritten digits, each digit being a 28x28 pixel image. The digits are arranged in rows corresponding to their class: 0s, 1s, 2s, 3s, 4s, 5s, 6s, 7s, 8s, and 9s. The digits are drawn in black on a white background.

70 000 images[†] of size 28×28 pixels (256 gray levels)

Problem: multi-class classification (10 classes);
training: 60 000 images / test: 10 000 images

Source: <http://yann.lecun.com/exdb/mnist/>

31/33

Example: MNIST

➡ see Jupyter / Python / Scikit-Learn notebook

32/33

Lecture outline

0 – Preliminary: classification with the log loss

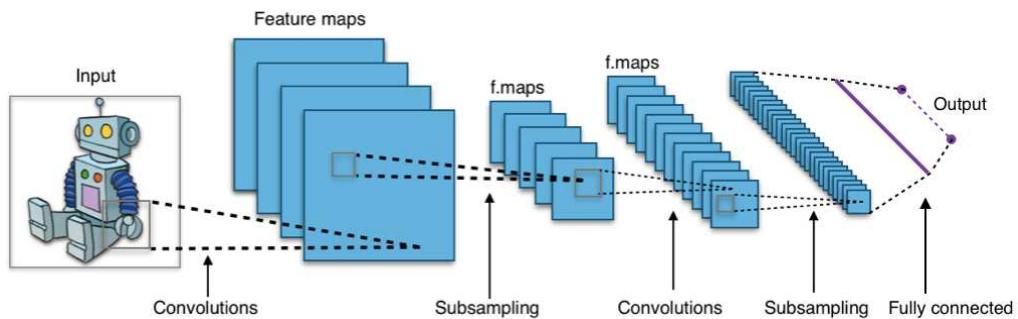
1 – Decision trees

- 1.1 – Two introductory examples
- 1.2 – Recursive partitioning
- 1.3 – Prediction function

2 – Neural networks

- 2.1 – Neurons
- 2.2 – Multi-layer perceptrons
- 2.3 – Example
- 2.4 – Other architectures

Convolutional neural networks (CNNs)



Schematic diagram of a typical CNN

Image: Aphex34, <https://commons.wikimedia.org>, CC BY-SA 4.0

Chapter 10

Unsupervised learning: two examples



CentraleSupélec

Statistics and Learning

Arthur Tenenhaus[†], Julien Bect & Laurent Le Brusquet

(firstname.lastname@centralesupelec.fr)

Teaching: CentraleSupélec / Department of Mathematics

Research: Laboratory of signals and systems (L2S)

[†]: Course coordinator

1/50

Lecture 10/10

Unsupervised learning: two examples

In this lecture you will...

- ▶ Understand the main ideas of unsupervised learning through two examples of unsupervised learning tasks.
- ▶ Learn how to reduce the dimension of a dataset using **principal component analysis**.
- ▶ Learn how to partition the data into clusters of similar examples (*clustering*) using the **K-means algorithm**.

2/50

Lecture outline

1 – Introduction to unsupervised learning

2 – Principal components analysis

2.1 – Low rank approximation

2.2 – Finding the optimal subspace: SVD

2.3 – Sample variance and covariance of PCA components

3 – Clustering

3.1 – Dissimilarity

3.2 – K -means algorithm

3.3 – Choice of the number of clusters

4 – A taste of some (more) advanced methods

3/50

Lecture outline

1 – Introduction to unsupervised learning

2 – Principal components analysis

2.1 – Low rank approximation

2.2 – Finding the optimal subspace: SVD

2.3 – Sample variance and covariance of PCA components

3 – Clustering

3.1 – Dissimilarity

3.2 – K -means algorithm

3.3 – Choice of the number of clusters

4 – A taste of some (more) advanced methods

Recap: supervised learning

- We observe **pairs** (X_i, Y_i) :

$$(X_1, Y_1), \dots, (X_n, Y_n) \stackrel{\text{iid}}{\sim} P^{X,Y},$$

with $X_i \in \mathcal{X}$: **instance** and $Y_i \in \mathcal{Y}$: **label**.

- We want to approach the **optimal predictor**

$$h^* = \operatorname{argmin}_h \mathbb{E}(L(Y, h(X))),$$

which is a **property of the conditional distribution** $P^{Y|X}$:

$$\begin{aligned} h^*(x) &= \operatorname{argmin}_{\tilde{y} \in \mathcal{Y}} \mathbb{E}(L(Y, \tilde{y}) \mid X = x) \\ &= \operatorname{argmin}_{\tilde{y} \in \mathcal{Y}} \int L(y, \tilde{y}) P^{Y|X=x}(dy). \end{aligned}$$

4/50

Unsupervised learning

Learning without a “teacher”:

- we observe **instances only**,

$$X_1, \dots, X_n \stackrel{\text{iid}}{\sim} P^X,$$

and we are interested in the distribution P^X .

Assume that $\mathcal{X} \subset \mathbb{R}^p$ and that P^X has a pdf f^X .

Problem: curse of dimensionality

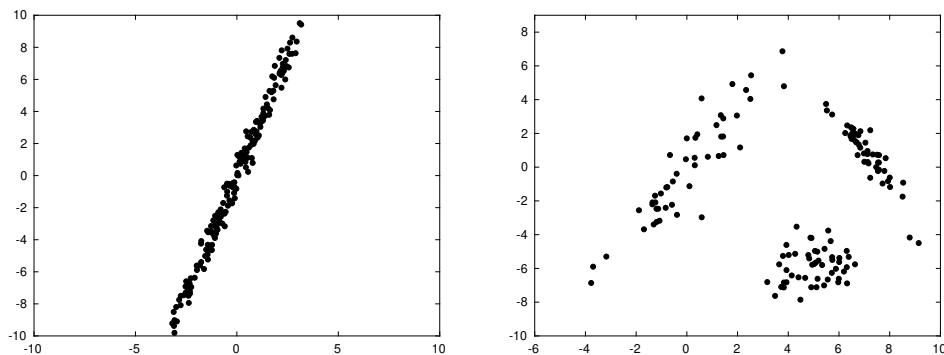
Estimating a “general” pdf f^X has a cost (sample size required to achieve a certain accuracy) that **scales exponentially with the dimension p .**[†]

[†] *Non-parametric statistics*, a branch of statistics which studies among other things density estimation under weak assumptions, provides theoretical results (not covered) that support this claim.

5/50

Goals in unsupervised learning

- ① Ideally, estimate the pdf f^X of the data distribution.
 - ➡ unless p is small enough (say, $p \lesssim 5$, rare in learning problems), this problem is in general **too difficult**[†].
- ② Reveal underlying “structures” in the distribution
(without explicitly constructing a density estimator)



[†] In low dimension, one can use, e.g., *kernel density estimators* (not covered).

6/50

Lecture outline

1 – Introduction to unsupervised learning

2 – Principal components analysis

- 2.1 – Low rank approximation
- 2.2 – Finding the optimal subspace: SVD
- 2.3 – Sample variance and covariance of PCA components

3 – Clustering

- 3.1 – Dissimilarity
- 3.2 – K -means algorithm
- 3.3 – Choice of the number of clusters

4 – A taste of some (more) advanced methods

Goal: dimension reduction

We are looking for a mapping

$$\begin{aligned} T : \mathcal{X} &\rightarrow \mathcal{Z} \subset \mathbb{R}^q \quad \text{with } q \ll p \\ x &\mapsto z = T(x) \end{aligned}$$

together with a **reconstruction** mapping

$$\begin{aligned} \tilde{T} : \mathcal{Z} &\rightarrow \mathcal{X} \\ z &\mapsto \hat{x} = \tilde{T}(z) \end{aligned}$$

such that

$$\frac{1}{n} \sum_{i=1}^n L(x_i, \hat{x}_i) = \frac{1}{n} \sum_{i=1}^n L\left(x_i, \tilde{T}\left(\underbrace{T(x_i)}_{z_i}\right)\right)$$

is as small as possible (where $L(x, \hat{x})$ denotes a loss function).

Remark: more generally, \mathcal{Z} could be a q -dimensional *manifold*, which is an abstract generalization of the concepts of curve ($q = 1$) and surface ($q = 2$); cf. differential geometry.

7/50

Lecture outline

1 – Introduction to unsupervised learning

2 – Principal components analysis

2.1 – Low rank approximation

2.2 – Finding the optimal subspace: SVD

2.3 – Sample variance and covariance of PCA components

3 – Clustering

3.1 – Dissimilarity

3.2 – K -means algorithm

3.3 – Choice of the number of clusters

4 – A taste of some (more) advanced methods

“Linear” dimension reduction

Let $x_1, \dots, x_n \in \mathbb{R}^p$ be an observed sample. Let $q < p$.

Definition: affine subspace

$\mathcal{A}_q \subset \mathbb{R}^p$ is an **affine subspace** of **dimension q** if there exists

- ▶ $\mu \in \mathbb{R}^p$,
- ▶ a matrix A of size $p \times q$ with **rank q** ,

such that $\mathcal{A}_q = \text{Aff}_{\mu, A} = \{y \in \mathbb{R}^p \text{ such that } y = \mu + Az, z \in \mathbb{R}^q\}$.

Definition: principal components analysis (PCA)

PCA consist in finding the **best approximation** of the data, for the **quadratic loss**, by an **affine subspace** \mathcal{A}_q .

The dimension q is either given or chosen automatically.

8/50

“Linear” dimension reduction (cont'd)

Thus, we are looking for $\mathcal{A}_q = \text{Aff}_{\mu, A}$ and (z_i) such that

$$\mu, A, (z_i) \in \underset{i=1}{\text{argmin}} \sum_{i=1}^n \|x_i - (\mu + Az_i)\|^2. \quad (*)$$



The solution is **not unique**.

☞ If \tilde{A} has the same range as A , then
there exists \tilde{z}_i 's such that $Az_i = \tilde{A}\tilde{z}_i$ for all i .

⇒ We will assume wlog that the columns of A are **orthonormal**:

$$A^\top A = \text{Id}_q.$$

Remark: the orthonormality assumption still does not make A unique...

9/50

“Linear” dimension reduction (cont’d)

☞ Fix some μ , A and (z_i) , and set $\tilde{z}_i = z_i - \bar{z}$. Then

$$\begin{aligned}\mu + Az_i &= \mu + A(\tilde{z}_i + \bar{z}) \\ &= \underbrace{\mu + A\bar{z}}_{\tilde{\mu}} + A\tilde{z}_i.\end{aligned}$$

⇒ We can constrain the z_i 's, wlog, to be such that $\bar{z} = 0$.

10/50

Partial result

Proposition

Minimizing the criterion for a given matrix A leads to:

$$\begin{aligned}\mu &= \bar{x}, \\ z_i &= A^\top(x_i - \bar{x}),\end{aligned}$$

and we have the geometric interpretation:

⇒ $\hat{x}_i = \mu + Az_i$ is the **orthogonal projection** of x_i on $\text{Aff}_{\mu, A}$.

Consequence. Plugging this result into $(*)$, we get

$$A = \underset{\text{argmin}}{} \sum_{i=1}^n \left\| (\text{Id}_p - AA^\top)(x_i - \bar{x}) \right\|^2.$$

11/50

Partial result: proof

Fix some A and (z_i) , with $\bar{z} = 0$, and set $v_i = x_i - Az_i$. Then

$$\begin{aligned}\sum_i \|x_i - (\mu + Az_i)\|^2 &= \sum_i \|v_i - \mu\|^2 \\ &= n \|\mu - \frac{1}{n} \sum_i v_i\|^2 + c\end{aligned}$$

where c does not depend on μ . Therefore, the optimal μ is

$$\mu = \frac{1}{n} \sum_i v_i = \bar{x} - A\bar{z} = \bar{x}.$$

Thus we set $\mu = \bar{x}$, and proceed similarly to determine each of the z_i 's.
For all i the minimum is attained (exercise) at

$$z_i = A^\top(x_i - \bar{x}),$$

and we check that $\bar{z} = \frac{1}{n} \sum_i z_i = A^\top(\bar{x} - \bar{x}) = 0$. □

Remark: the expressions can also be obtained quickly by setting the gradient of the criterion to zero.

12/50

Partial result: geometric interpretation

Assume temporarily, wlog, that $\bar{x} = 0$. Then

- ▶ $\mu = 0$,
- ▶ $\mathcal{A}_q = \text{Aff}_{0,A}$ is a linear subspace of \mathbb{R}^p ,
- ▶ $z_i = A^\top x_i$ and $\hat{x}_i = Az_i = AA^\top x_i$.

Proposition

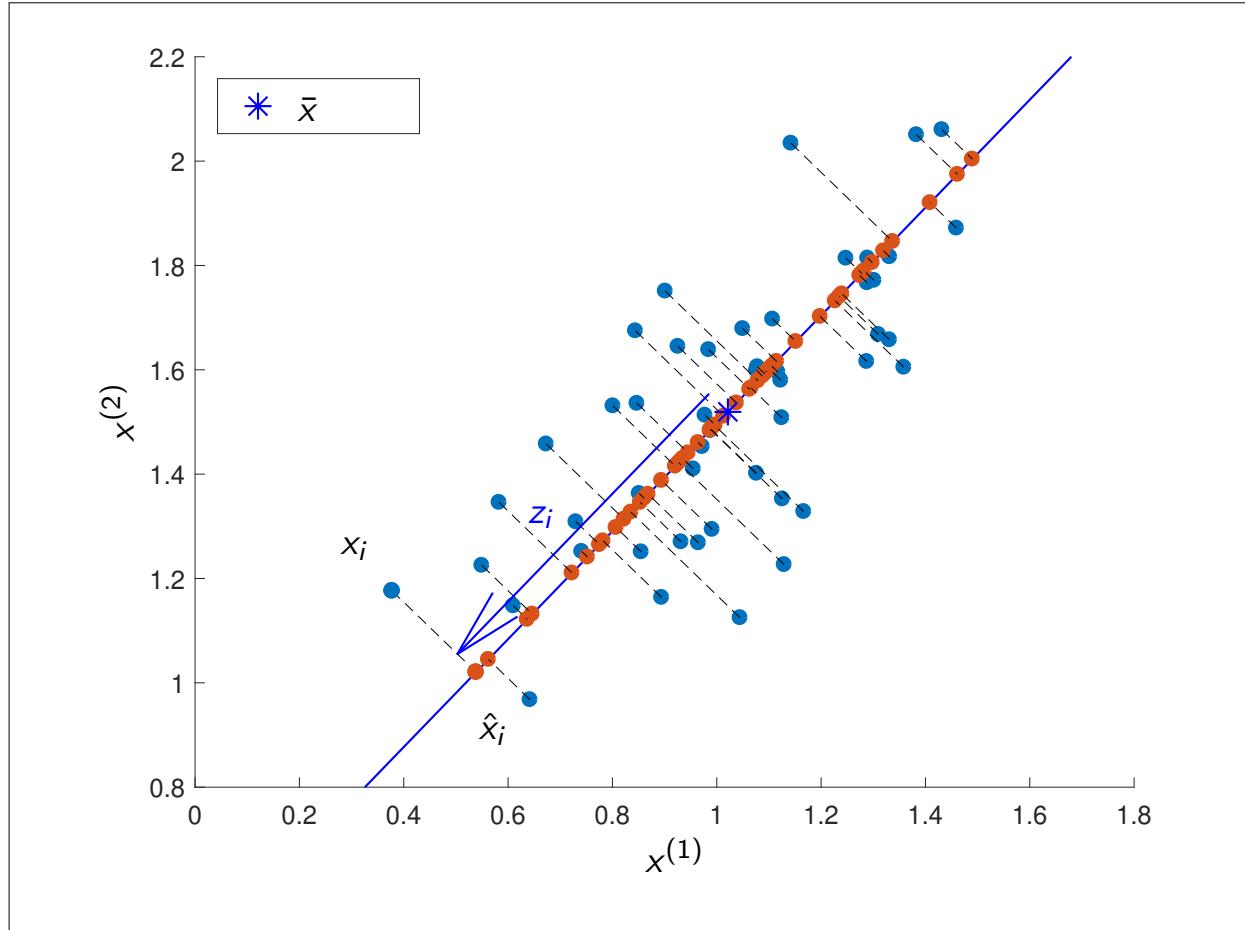
The $p \times p$ matrix AA^\top is the orthogonal projection matrix onto the linear subspace $\text{Aff}_{0,A}$.

Proof. Let v_1, \dots, v_q be the (orthonormal) columns of A .

Then, for all $x \in \mathbb{R}^p$ and $z = A^\top x$,

- ▶ $z^{(j)} = e_j^\top A^\top x = v_j^\top x$ is the scalar product between x and v_j ,
- ▶ $\hat{x} = AA^\top x = Az = \sum_j z^{(j)} v_j$ is the orthonormal projection of x onto $\text{Aff}_{0,A} = \text{span}\{v_1, \dots, v_q\}$.

13/50



Lecture outline

1 – Introduction to unsupervised learning

2 – Principal components analysis

2.1 – Low rank approximation

2.2 – Finding the optimal subspace: SVD

2.3 – Sample variance and covariance of PCA components

3 – Clustering

3.1 – Dissimilarity

3.2 – K -means algorithm

3.3 – Choice of the number of clusters

4 – A taste of some (more) advanced methods

Notations

Let X be the matrix of observations:

$$X = \begin{pmatrix} (x_1)^\top \\ \vdots \\ (x_n)^\top \end{pmatrix} = \begin{pmatrix} x_1^{(1)} & \dots & x_1^{(p)} \\ \vdots & & \vdots \\ x_n^{(1)} & \dots & x_n^{(p)} \end{pmatrix}$$

We will assume, wlog, that $\bar{x} = 0$.

We are looking for a matrix A such that

$$\begin{aligned} A &= \operatorname{argmin} \sum_{i=1}^n \left\| (\text{Id}_p - AA^\top) x_i \right\|^2 \\ &= \operatorname{argmin} \| (\text{Id}_p - AA^\top) X^\top \|_F^2 \end{aligned}$$

where $\|\cdot\|_F$ denotes the **Frobenius norm**:

$$\|M\|_F^2 = \sum_{i,j} M_{ij}^2 = \operatorname{tr}(M^\top M) = \operatorname{tr}(MM^\top).$$

15/50

Singular value decomposition (SVD)

Theorem

Let M be an $n \times p$ real matrix. There exist matrices

- ▶ U , orthogonal with size $n \times n$ ($U^\top U = \text{Id}_n$),
- ▶ V , orthogonal with size $p \times p$ ($V^\top V = \text{Id}_p$),
- ▶ $D = \operatorname{diag}(d_1, \dots, d_r, 0, \dots, 0)$ with size $n \times p$,
with $d_1 \geq d_2 \geq \dots \geq d_r > 0$

such that :

$$M = UDV^\top,$$

and r is the rank of both D et M .

The scalars $d_1, \dots, d_r, 0, \dots, 0$ are the **singular values of M** .

- ▶ d_1^2, \dots, d_r^2 are the non-zero eigenvalues of MM^\top and $M^\top M$.

Proof. See PC #8, bonus exercise. □

16/50

Solution of the optimization problem

Let U , D and V be the matrices obtained from the SVD of X :

$$X = UDV^\top.$$

Theorem

Let

- ▶ v_1, v_2, \dots, v_p the columns of V ,
- ▶ $V_q = (v_1 \mid \dots \mid v_q)$ the submatrix with the first q columns.

Then

$$V_q \in \operatorname{argmin}_A \left\| (\operatorname{Id}_p - AA^\top) X^\top \right\|_F^2.$$

17/50

Proof.

$$\left\| (\operatorname{Id}_p - AA^\top) X^\top \right\|_F^2 = \| VD^\top U^\top - AA^\top VD^\top U^\top \|_F^2$$

Properties of the Frobenius norm: if U and V are orthogonal,

$$\| VMU^\top \|_F^2 = \| M \|_F^2.$$

$$\text{Hence : } \left\| (\operatorname{Id}_p - AA^\top) X \right\|_F^2 = \| D^\top - V^\top AA^\top VD^\top \|_F^2.$$

Let $\mathcal{M}_{n,p,q}$ denote the set of all rank q matrices of size $n \times p$. Then

$$D_q = \operatorname{diag}(d_1, \dots, d_q, 0, \dots, 0) \in \operatorname{argmin}_{M \in \mathcal{M}_{n,p,q}} \| D^\top - M^\top \|_F^2$$

(diagonal matrix with the q largest singular values).

We obtain the result by checking that $V^\top V_q V_q^\top VD^\top = D_q^\top$. □

Recap: PCA

Algorithm: Principal components analysis (PCA)

Computing the PCA of a sample (x_1, \dots, x_n) consists in :

- ① Computing the mean \bar{x} and **centering the data**: $x_i \leftarrow x_i - \bar{x}$.
- ② Constructing the matrix X of centered data.
- ③ Computing the matrix V from the **SVD of X**
(the singular values are useful too, cf. next section)
- ④ **Reducing the dimension**: $z_i = V_q^\top x_i$.

Reconstruction. $\hat{x}_i = \bar{x} + V_q z_i$.

Vocabulary.

- ▶ v_1, \dots, v_q (columns of V_q): **principal axes**.
- ▶ $z_i^{(1)}, \dots, z_i^{(q)}$: **principal component**.

18/50

Example: handwritten digits (not MNIST, another one!)

Data: $n = 658$ images 16×16 of the digit "3" $\rightarrow p = 256$

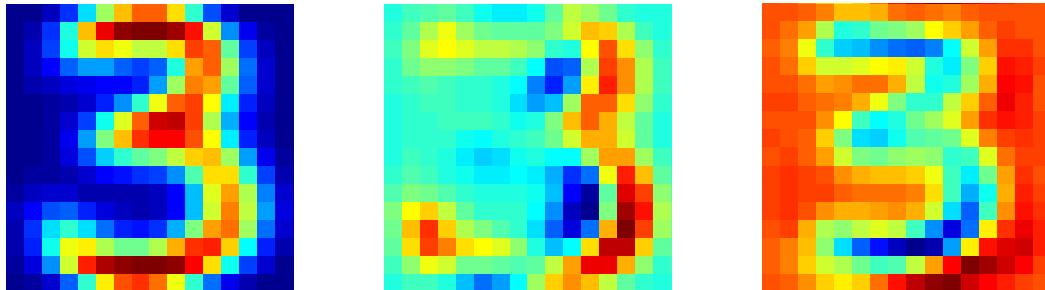


Source : The Elements of Statistical Learning, Springer

19/50

Example: handwritten digits (cont'd)

Visualization of the first two principal axes



mean \bar{x}

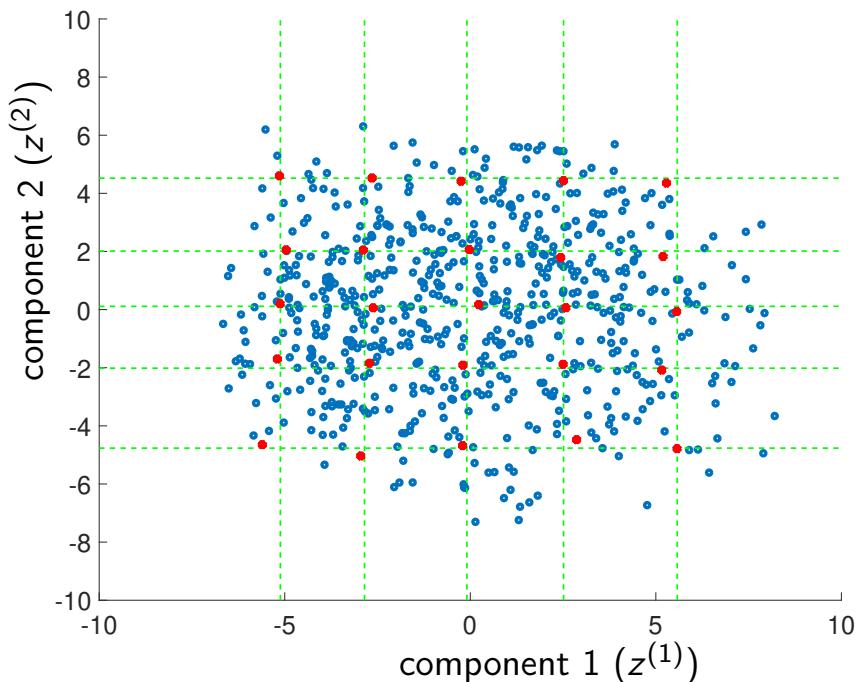
principal axis v_1

principal axis v_2

$$\forall i, \hat{x}_i = \bar{x} + z_i^{(1)} v_1 + z_i^{(2)} v_2$$

20/50

Principal plane ($z^{(1)}, z^{(2)}$)

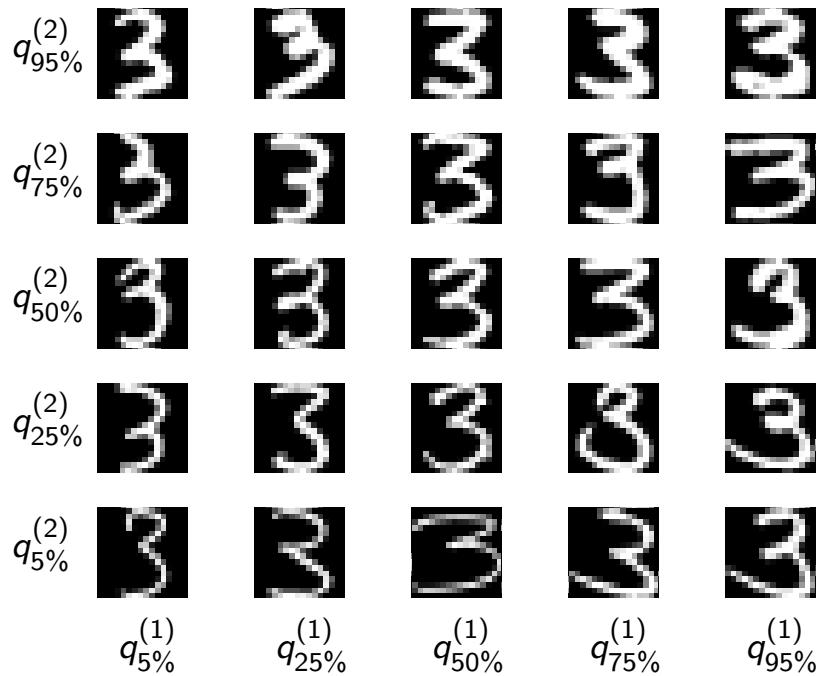


Dashed lines: 5%, 25%, 50%, 75%, 95% quantiles.

Red dots: examples shown on the next slide.

21/50

Interpretation of the components $(z^{(1)}, z^{(2)})$ based on the 25 examples selected on the previous slide.



22/50

Lecture outline

1 – Introduction to unsupervised learning

2 – Principal components analysis

2.1 – Low rank approximation

2.2 – Finding the optimal subspace: SVD

2.3 – Sample variance and covariance of PCA components

3 – Clustering

3.1 – Dissimilarity

3.2 – K-means algorithm

3.3 – Choice of the number of clusters

4 – A taste of some (more) advanced methods

Sample covariance matrix of the components

Let $\hat{\Sigma}_Z$ denote the sample covariance matrix of the q components

$$\begin{aligned}\hat{\Sigma}_Z &= \frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})(z_i - \bar{z})^\top \\ &= \frac{1}{n} \sum_{i=1}^n z_i z_i^\top \quad (\text{car } \bar{z} = \frac{1}{n} \sum_{i=1}^n z_i = 0) \\ &= \frac{1}{n} Z^\top Z\end{aligned}$$

with $Z = \begin{pmatrix} z_1^\top \\ \vdots \\ z_n^\top \end{pmatrix}$. Recall that $z_i = V_q^\top x_i$, and thus $Z = X V_q$.

Using $X = UDV^\top$, we get

$$\begin{aligned}\hat{\Sigma}_Z &= \frac{1}{n} V_q^\top V D^\top D V^\top V_q \\ &= \frac{1}{n} \text{diag}(d_1^2, \dots, d_q^2).\end{aligned}$$

23/50

Sample covariance matrix of the components (cont'd)

Conclusions.

- ▶ The (sample) variance of component $z^{(j)}$ is $\frac{d_j^2}{n}$.
 - ⇒ Components sorted by decreasing variance.
- ▶ The (sample) covariances are equal to zero.
 - ⇒ The components are uncorrelated.

24/50

Total variance of a sample

Definition / Proposition

The **total variance** of the p -variate sample (x_1, \dots, x_n) is

$$VT(x_1, \dots, x_n) = \sum_{j=1}^p \text{var}(x_1^{(j)}, \dots, x_n^{(j)}).$$

With centered x_i 's, we have

$$VT(x_1, \dots, x_n) = \frac{1}{n} \text{tr}(X^\top X) = \frac{1}{n} \sum_{j=1}^r d_j^2.$$

Proof. Using that the x_i 's are centered, we have

$$VT(x_1, \dots, x_n) = \sum_{j=1}^p \left(\frac{1}{n} \sum_{i=1}^n (x_i^{(j)})^2 \right) = \frac{1}{n} \|X\|_F^2 = \frac{1}{n} \text{tr}(X^\top X).$$

Then, using $X = UDV^\top$, with $U^\top U = \text{Id}_n$ and $V^\top V = \text{Id}_p$, we obtain

$$VT(x_1, \dots, x_n) = \frac{1}{n} \text{tr}(D^\top D) = \frac{1}{n} \sum_{j=1}^r d_j^2.$$

25/50

Proportion of explained variance

Total variance of the **reconstructed sample** $(\hat{x}_1, \dots, \hat{x}_n)$:

$$VT(\hat{x}_1, \dots, \hat{x}_n) = \frac{1}{n} \text{tr}(\hat{X}^\top \hat{X}) = ?.$$

Using $\hat{X} = ZV_q^\top$, we get:

$$VT(\hat{x}_1, \dots, \hat{x}_n) = \text{tr}(V_q \hat{\Sigma}_Z V_q^\top) = \frac{1}{n} \sum_{j=1}^q d_j^2.$$

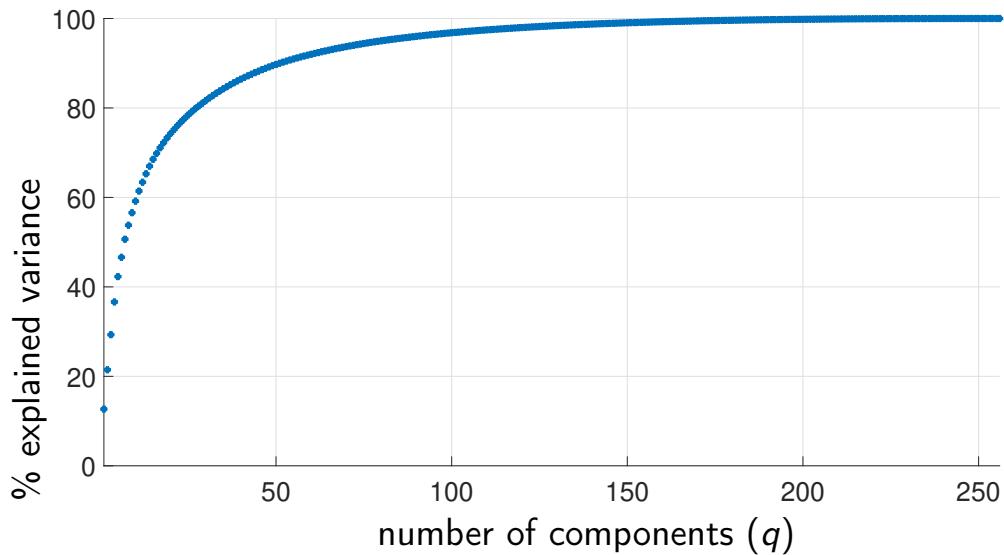
Proportion of explained variance

The **proportion of explained variance** is defined as

$$\frac{VT(\hat{x}_1, \dots, \hat{x}_n)}{VT(x_1, \dots, x_n)} = \frac{\sum_{j=1}^q d_j^2}{\sum_{j=1}^r d_j^2}.$$

26/50

Example: handwritten digits (MNIST, $p = 28^2 = 784$)



Remark: similarity with the coefficient of determination (R^2) in regression.

27/50

Lecture outline

1 – Introduction to unsupervised learning

2 – Principal components analysis

- 2.1 – Low rank approximation
- 2.2 – Finding the optimal subspace: SVD
- 2.3 – Sample variance and covariance of PCA components

3 – Clustering

- 3.1 – Dissimilarity
- 3.2 – K -means algorithm
- 3.3 – Choice of the number of clusters

4 – A taste of some (more) advanced methods

Definition : clustering, clusters

Let $E = \{x_1, \dots, x_n\}$ be a sample of n observations $x_i \in \mathcal{X}$.

- We assume that $\mathcal{X} \subset \mathbb{R}^p$, thus $E \subset \mathbb{R}^p$.

Definitions

Clustering[†] consists in **partitioning** the set E in K non-empty parts $E_k \subset E$, $1 \leq k \leq K$, that contain “similar” observations.

The number K is either given or chosen automatically.

The sets E_k are called **groups** or **clusters**.

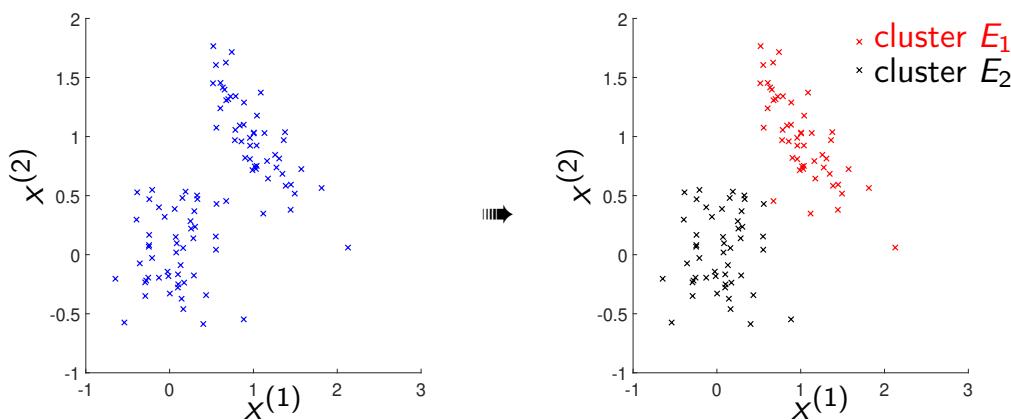
Notations.

- Denote by $\pi^{(k)} = \{i \leq n \mid x_i \in E^{(k)}\}$ the indices in E_k .
- $\Pi = \{\pi_1, \dots, \pi_K\}$ is a partition of $\{1, \dots, n\}$.

[†] also called *data partitioning*.

Example of clustering result

Example with $p = 2$ and $K = 2$



Lecture outline

1 – Introduction to unsupervised learning

2 – Principal components analysis

2.1 – Low rank approximation

2.2 – Finding the optimal subspace: SVD

2.3 – Sample variance and covariance of PCA components

3 – Clustering

3.1 – Dissimilarity

3.2 – K -means algorithm

3.3 – Choice of the number of clusters

4 – A taste of some (more) advanced methods

Dissimilarity: definition

We are looking for a partition such that, for all k ,

- ▶ the instances[†] in cluster E_k are “similar” to each other,
- ▶ and as dissimilar as possible to those in other clusters.

Definition

In clustering algorithms, we call **dissimilarity** the function $D : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ that is used to measure the “distance” between examples.

Remark: **not always a distance** but satisfies in general

- ▶ the **symmetry** property: $D(x, y) = D(y, x)$,
- ▶ the **positivity** property: $D(x, y) \geq 0$.

[†] a.k.a. “examples”, “observations”, “data”, “individuals”…

Dissimilarity: examples

- ▶ General form: $D(x_i, x_{i'}) = \sum_{j=1}^p d_j(x_i^{(j)}, x_{i'}^{(j)})$
- ▶ Quantitative variable: $d_j(x_i^{(j)}, x_{i'}^{(j)}) = f(|x_i^{(j)} - x_{i'}^{(j)}|)$.
Example: $d_j(x_i^{(j)}, x_{i'}^{(j)}) = (x_i^{(j)} - x_{i'}^{(j)})^2$.
Remark: it is often beneficial to normalize the variables:
 $x_i^{(j)} \rightarrow \frac{x_i^{(j)}}{s_j}$, (usual choice for s_j : sample standard deviation)
- ▶ Qualitative variable: $d_j(x_i^{(j)}, x_{i'}^{(j)}) = \text{cste if } x_i^{(j)} \neq x_{i'}^{(j)} (0 \text{ otherwise})$

31/50

Within-cluster and between-cluster inertia

Let us write $d_{ii'} = D(x_i, x_{i'})$.

Within-cluster inertia

It is defined as $W(\Pi)$ (W=Within) :

$$W(\Pi) = \frac{1}{2} \sum_{k=1}^K \sum_{i,i' \in \pi_k} d_{ii'}$$

Between-cluster inertia

It is defined as $B(\Pi)$ (B=Between) :

$$B(\Pi) = \frac{1}{2} \sum_{k,k' \neq k} \sum_{i \in \pi_k} \sum_{i' \in \pi_{k'}} d_{ii'}$$

Property: $W(\Pi) + B(\Pi) = \sum_{i,i'} d_{ii'}$

Definition: $T = \frac{1}{2} \sum_{i,i'} d_{ii'}$ is the total inertia.

- ▶ Does not depend on the partition.

32/50

Proof of the property

$$\begin{aligned}
 T &= \frac{1}{2} \sum_{i,i'} d_{ii'} \\
 &= \frac{1}{2} \sum_{k,k'} \sum_{i \in \pi_k} \sum_{i' \in \pi_{k'}} d_{ii'} \\
 &= \underbrace{\frac{1}{2} \sum_k \sum_{i,i' \in \pi_k} d_{ii'}}_{W(\Pi)} + \underbrace{\frac{1}{2} \sum_{k,k' \neq k} \sum_{i \in \pi_k} \sum_{i' \in \pi_{k'}} d_{ii'}}_{B(\Pi)}
 \end{aligned}$$

Optimal partition :

$$\Pi_* = \arg \min_{\Pi} W(\Pi)$$

Remark: since $W(\Pi) + B(\Pi) = T$, $\Pi_* = \arg \max_{\Pi} B(\Pi)$.

Problem : this is a combinatorial optimization problem

- ▶ 34105 partitions for $n = 10$ and $K = 4$,
- ▶ $\approx 7.5 \cdot 10^{11}$ partitions for $n = 20$ and $K = 5$.

Solution : look for a sub-optimal solution

⇒ K-means algorithm

Lecture outline

1 – Introduction to unsupervised learning

2 – Principal components analysis

2.1 – Low rank approximation

2.2 – Finding the optimal subspace: SVD

2.3 – Sample variance and covariance of PCA components

3 – Clustering

3.1 – Dissimilarity

3.2 – K -means algorithm

3.3 – Choice of the number of clusters

4 – A taste of some (more) advanced methods

Dissimilarity considered here : $d_{ii'} = \|x_i - x_{i'}\|^2$.

With this choice of dissimilarity :

$$W(\Pi) = \sum_{k=1}^K \sum_{i \in \pi_k} \|x_i - \bar{x}_k\|^2$$

where $\bar{x}_k = \frac{1}{|\pi_k|} \sum_{i \in \pi_k} x_i$ is the barycenter of the cluster.

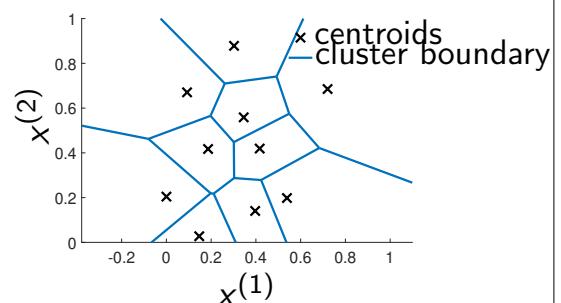
⇒ \bar{x}_k is called the **centroid** of cluster k .

Principle of the K -means algorithm

Iteratively,

- ▶ Given a partition Π , compute the centroids \bar{x}_k .
- ▶ Modify Π in such a way that each x_i is associated to the cluster π_k whose (current) centroid \bar{x}_k is the closest.

⇒ Voronoï diagram



Expressions of T , $W(\Pi)$ and $B(\Pi)$ for $d_{ii'} = \|x_i - x_{i'}\|^2$

$$\begin{aligned} T &= \frac{1}{2} \sum_{i,i'} \|x_i - x_{i'}\|^2 \\ &= \frac{1}{2} \sum_{i,i'} \|(x_i - \bar{x}) - (x_{i'} - \bar{x})\|^2 \\ &= \sum_i \|x_i - \bar{x}\|^2 - \sum_{i,i'} (x_i - \bar{x})^\top (x_{i'} - \bar{x}) \\ &= \sum_i \|x_i - \bar{x}\|^2 \end{aligned}$$

$$\begin{aligned} W(\Pi) &= \frac{1}{2} \sum_k \sum_{i,i' \in \pi_k} \|x_i - x_{i'}\|^2 \\ &= \frac{1}{2} \sum_k \sum_{i,i' \in \pi_k} \|(x_i - \bar{x}_k) - (x_{i'} - \bar{x}_k)\|^2 \\ &= \sum_k \sum_{i \in \pi_k} \|x_i - \bar{x}_k\|^2 \end{aligned}$$

$$\begin{aligned} B(\Pi) &= T - W(\Pi) \\ &= \sum_k \sum_{i \in \pi_k} \|x_i - \bar{x}\|^2 - \sum_k \sum_{i \in \pi_k} \|x_i - \bar{x}_k\|^2 \\ &= \sum_k \sum_{i \in \pi_k} \|\bar{x}_k - \bar{x}\|^2 \\ &= \sum_k |\pi_k| \|\bar{x}_k - \bar{x}\|^2 \end{aligned}$$

K-means algorithm

Require: $K > 0$ {number of clusters}
Require: $(\bar{x}_{1,0}, \dots, \bar{x}_{K,0})$ {centroids initialization}
 $t \leftarrow 0$
repeat

Step 1 {construction of Π_t from the centroids}
for all k **do**
 $\pi_{k,t} = \{i \text{ s.t. } k = \arg \min_{k'} \|x_i - \bar{x}_{k',t}\|\}$

end for

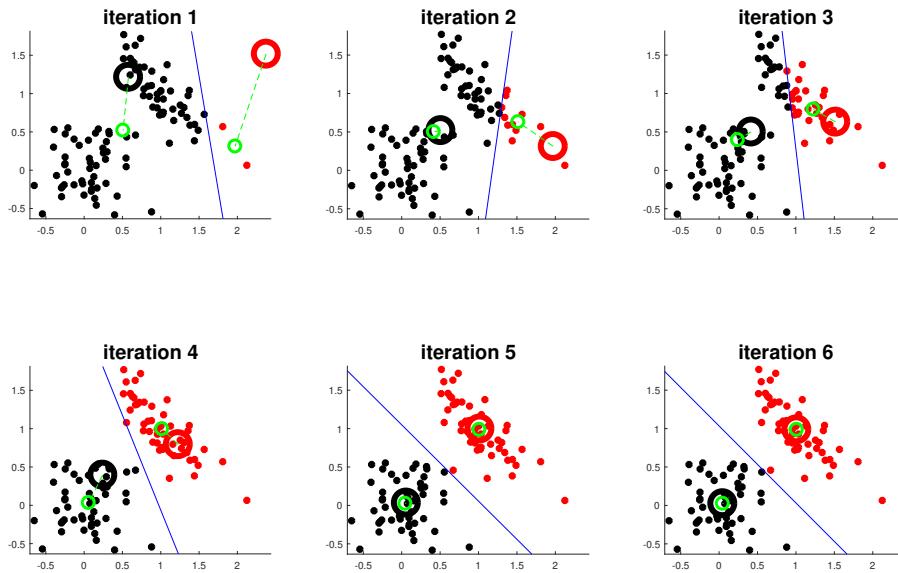
Step 2 {centroids update}
for all k **do**
 $\bar{x}_{k,t} = \frac{1}{|\pi_{k,t}|} \sum_{i \in \pi_{k,t}} x_i$

end for

$t \leftarrow t + 1$

until $W(\Pi_{t-1}) = W(\Pi_{t-2})$

return Π_{t-1}



36 / 50

Properties of the K -means algorithm

Proposition

Let $(\Pi_t)_{t \geq 0}$ denote the sequence of partitions constructed by the algorithm.

Then, there exists T such that :

- ① $\forall t \leq T, W(\Pi_t) < W(\Pi_{t-1})$,
- ② $W(\Pi_{T+1}) = W(\Pi_T)$.

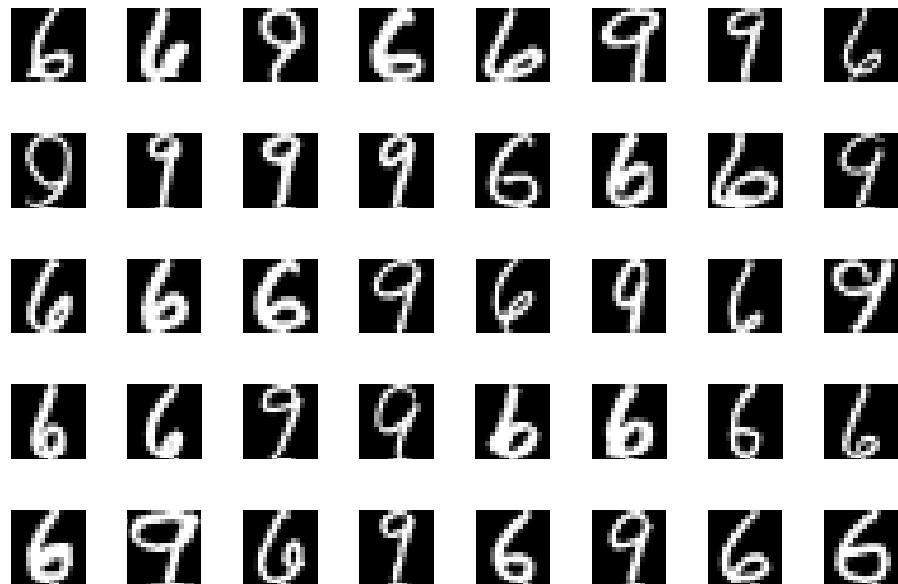


- The algorithm terminates in a finite number of iterations, but
- ▶ the partition Π_T is **not, in general, the optimal partition**;
 - ▶ it depends on the starting point $(\bar{x}_{1,0}, \dots, \bar{x}_{K,0})$.
 - ▶ Recommended: several trials with random starting points.

37 / 50

Example: handwritten digits

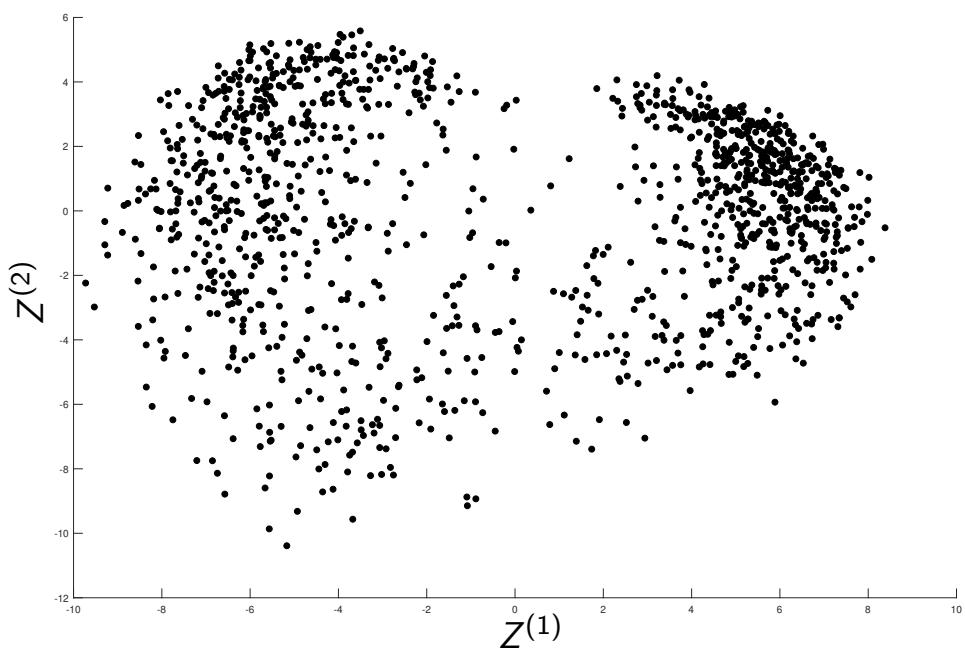
Consider the digits “6” and “9” (644 images each).



38/50

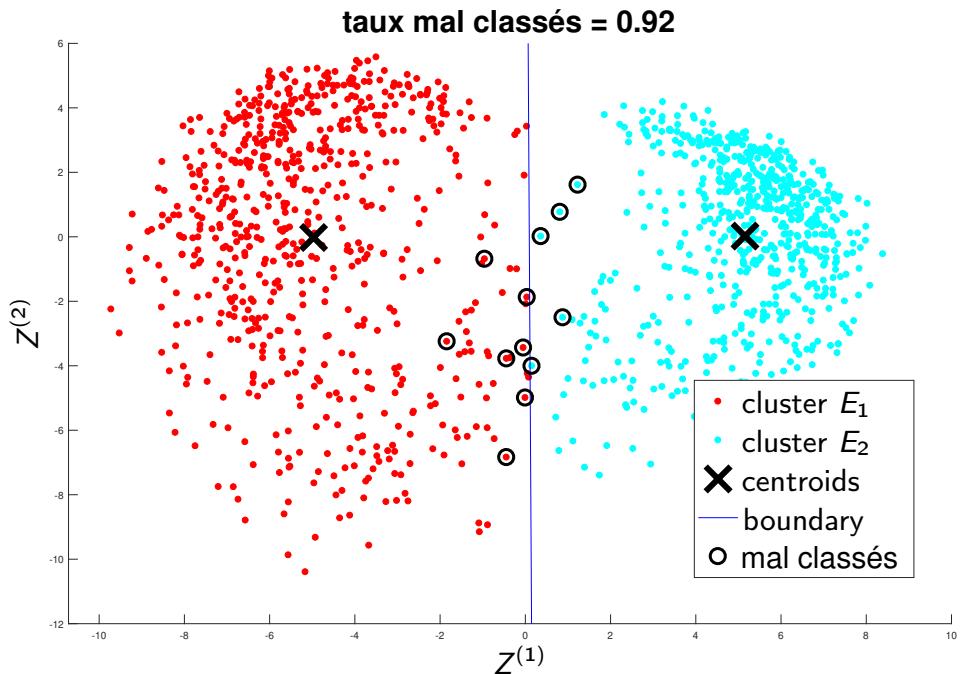
Example: handwritten digits

Represent each image by its first two principal components.



39/50

Example: handwritten digits



Note: here we use the labels, which are assumed unavailable in the non-supervised setting, to the sole purpose of evaluating the quality of the partition that we have obtained.

40/50

Lecture outline

1 – Introduction to unsupervised learning

2 – Principal components analysis

- 2.1 – Low rank approximation
- 2.2 – Finding the optimal subspace: SVD
- 2.3 – Sample variance and covariance of PCA components

3 – Clustering

- 3.1 – Dissimilarity
- 3.2 – K -means algorithm
- 3.3 – Choice of the number of clusters

4 – A taste of some (more) advanced methods

Homogeneity / dispersion

Reminder. We are looking for a partition such that, for all k ,

- ▶ the instances[†] in cluster E_k are “similar” to each other,
- ▶ and as dissimilar as possible to those in other clusters.

Definition: dispersion measure

The dispersion of cluster E_k is (often) measured by

$$S_k = \left(\frac{1}{|\pi_k|} \sum_{i \in \pi_k} \|x_i - \bar{x}_k\|^q \right)^{\frac{1}{q}},$$

with q a positive real number, to be chosen[†].

Interpretation. The smaller S_k , the more homogeneous the cluster.

[†] P.-H. Cournède's lecture notes and scikit-learn use $q = 1$.

Davies-Bouldin index

Definition: similarity of clusters E_k and $E_{k'}$

$$R_{k,k'} = \frac{S_k + S_{k'}}{\|\bar{x}_k - \bar{x}_{k'}\|}, \quad 1 \leq k, k' \leq K, \quad k \neq k'.$$

Interpretation. The clusters are more similar when their dispersion is large with respect to the distance between their centroids.

Definition: Davies-Bouldin index of a partition

$$\text{DB} = \frac{1}{K} \sum_{k=1}^K \max_{k' \neq k} R_{k,k'}$$

► Use: choose K in order to minimize DB.

Alternative method: silhouette values

Another indicator of the quality of a partition Π .

Let $i \in \pi_k$. For each x_i , define

- ▶ $a(x_i)$: average distance to other points in the same cluster
- ▶ $b(x_i)$: minimum average distance to points in another cluster

$$a(x_i) = \frac{1}{|\pi_k|} \sum_{i' \in \pi_k} \|x_{i'} - x_i\|$$

$$b(x_i) = \min_{k' \neq k} \left(\frac{1}{|\pi_{k'}|} \sum_{i' \in \pi_{k'}} \|x_{i'} - x_i\| \right)$$

Interpretation : $a(x_i) \ll b(x_i)$ if the clusters are homogeneous and well separated.

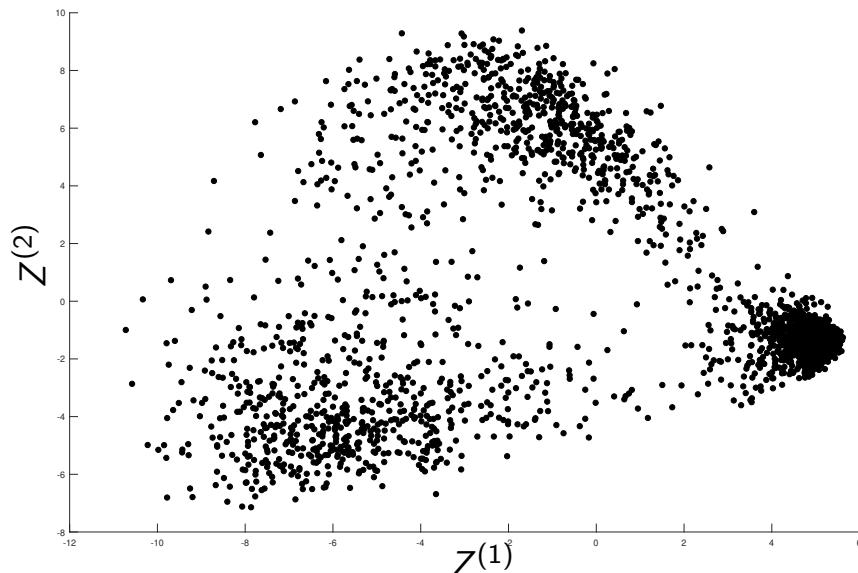
Silhouette value of partition Π

$$s(\Pi) = \frac{1}{n} \sum_{i=1}^n \frac{b(x_i) - a(x_i)}{\max(a(x_i), b(x_i))}$$

Choice of the number K of clusters:

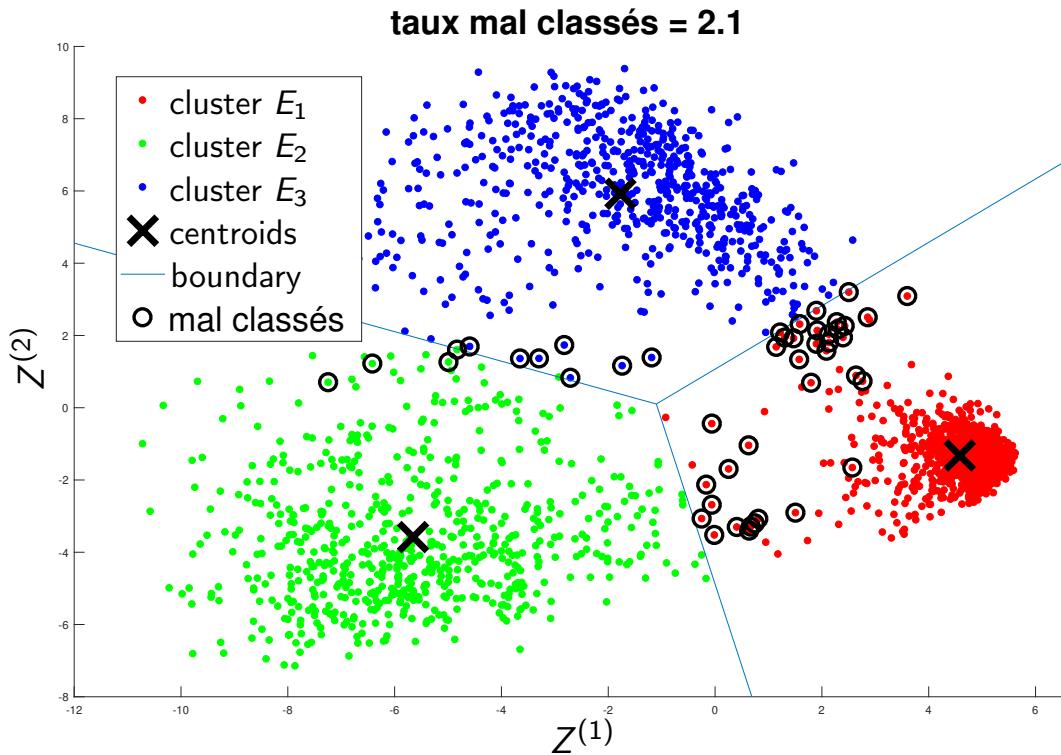
$\forall \Pi$, $s(\Pi) \leq 1$ and we choose the partition such that $s(\Pi)$ is maximal.

Example: handwritten digits with digits 1, 6 and 9

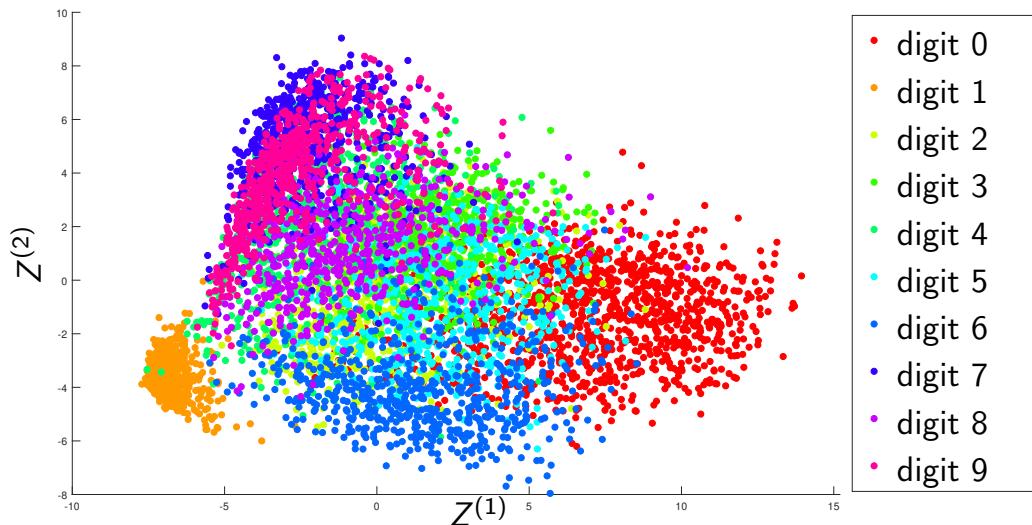


K	2	3	4	5	6	7	8
$DB(K)$	0.76	0.42	0.77	0.89	0.76	0.77	0.79
$s(K)$	0.55	0.73	0.65	0.58	0.60	0.59	0.58

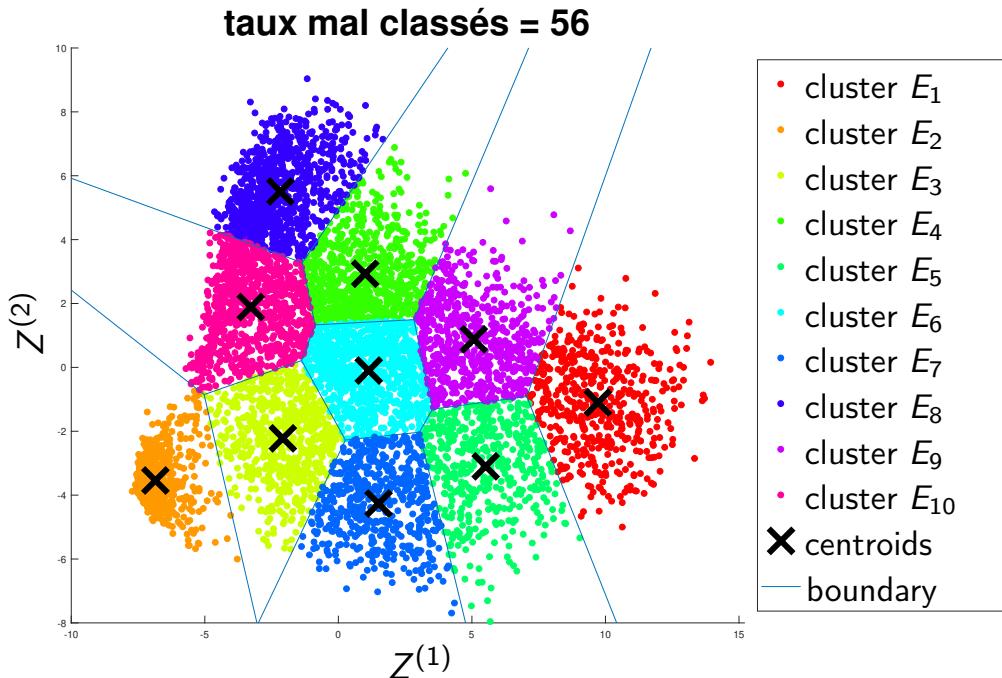
Example: handwritten digits with digits 1, 6 and 9



Example: handwritten digits with all digits



Example: handwritten digits with all digits



46/50

Example: handwritten digits with all digits

	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}	total
"0"	498	0	22	6	260	82	64	0	262	0	1194
"1"	0	1000	4	0	0	0	0	0	0	1	1005
"2"	3	1	234	122	12	202	54	3	60	40	731
"3"	1	0	29	230	4	211	5	5	131	42	658
"4"	0	21	70	112	2	42	3	144	19	239	652
"5"	2	0	61	37	66	171	88	1	119	11	556
"6"	3	6	135	0	128	43	335	0	10	4	664
"7"	0	2	2	49	0	6	0	458	1	127	645
"8"	2	7	82	138	1	93	1	17	41	160	542
"9"	0	10	0	64	0	3	0	303	7	257	644
total	509	1047	639	758	473	853	550	931	650	881	7291

Poor result \rightarrow need for a better dissimilarity measure !

(and, in particular, for a **better representation**)

47/50

Lecture outline

1 – Introduction to unsupervised learning

2 – Principal components analysis

2.1 – Low rank approximation

2.2 – Finding the optimal subspace: SVD

2.3 – Sample variance and covariance of PCA components

3 – Clustering

3.1 – Dissimilarity

3.2 – K -means algorithm

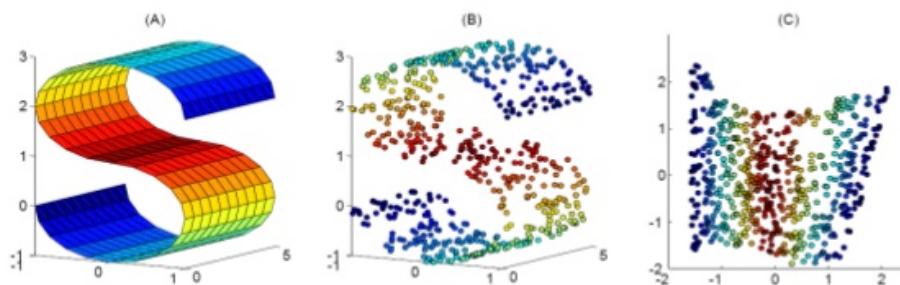
3.3 – Choice of the number of clusters

4 – A taste of some (more) advanced methods

Non-linear dimension reduction

Nonlinear Dimensionality Reduction

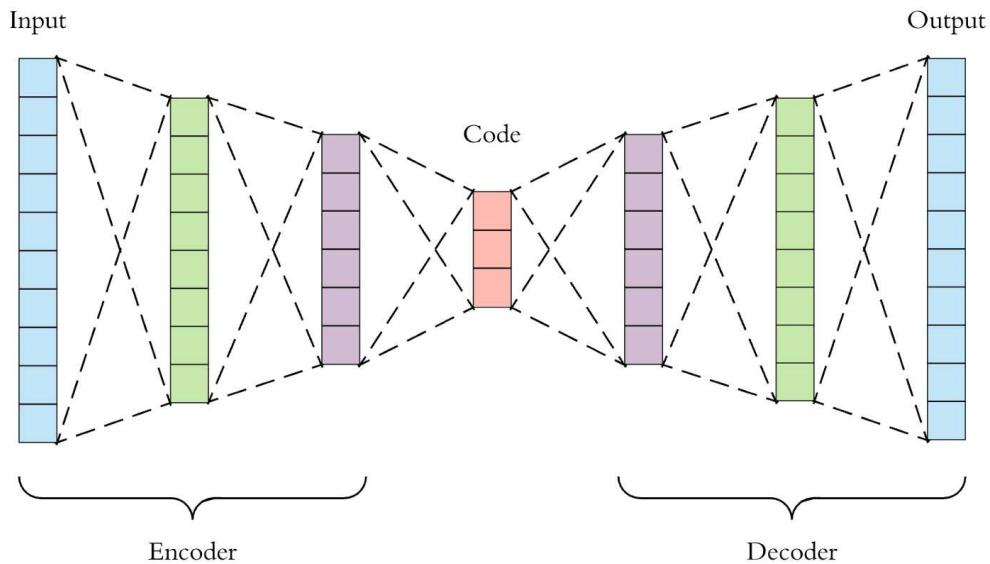
- Many data sets contain essential nonlinear structures that are invisible to PCA.



source: Yan Xu, Houston Machine Learning Meetup, 2017

48/50

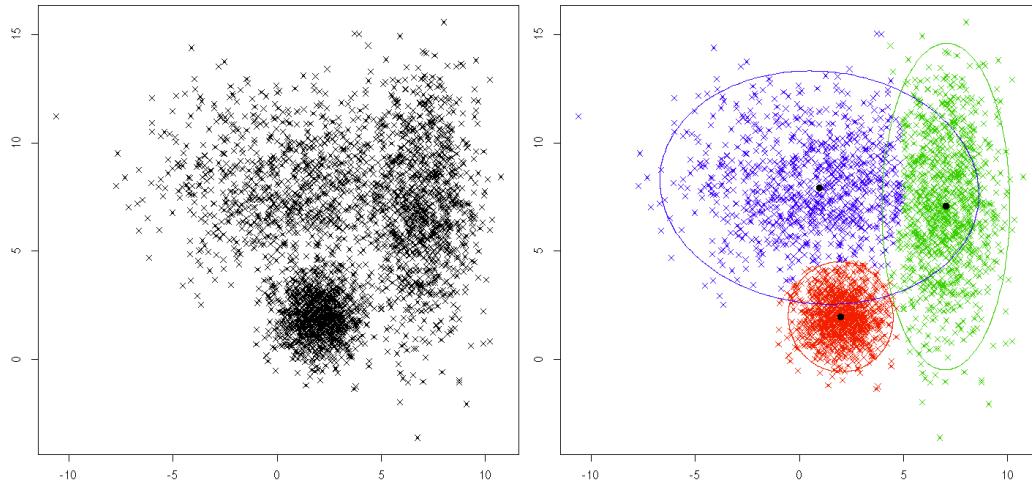
Example: auto-encoder



source: <https://towardsdatascience.com/Applied-Data-Deep-Learning-Part-3>

49/50

Clustering based on mixture models (see PC)



source: bioinfo-fr.net

50/50

