Statistical exploitation of measurements

Lecture Booklet

MASTER

FUNDAMENTAL PHYSICS AND APPLICATIONS

Pr Fabrice DEVAUX

fabrice.devaux@univ-fcomte.fr

UNIVERSITÉ DE FRANCHE-COMTÉ

UBFC UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

TABLE OF CONTENTS

CHAPTER 1. INTRODUCTION TO INVERSE PROBLEM 4

1. INTRODUCTION 4
2. DIRECT AND INVERSE PROBLEM 4
3. THE BAYES THEOREM 5
   3.1 example 1 5
   3.2 Example 2 6
   3.3 To summarize 7

CHAPTER 2. ESTIMATION 8

1. INTRODUCTION 8
2. GENERAL PROPERTIES 9
   2.1 Consistent (asymptotically unbiased) estimator 9
   2.2 Unbiased estimator 9
3. ESTIMATORS OF MEAN 10
4. ESTIMATORS OF VARIANCE 10
5. WHY THE MEASUREMENT ERRORS ARE OFTEN (NOT ALWAYS!) MODELED BY A GAUSSIAN LAW? 10
6. CONFIDENCE INTERVALS 12
7. RECURSIVE ESTIMATORS 13
8. WIENER FILTERING 13

CHAPTER 3. KARHUNEN-LOÈVE TRANSFORM OR PRINCIPAL COMPONENT ANALYSIS 16

1. PURPOSE AND GENERAL DESCRIPTION 16
   1.1 Example 1 (KLT): multispectral image compression 16
   1.2 Example 2 (KLT): Encoding bits reduction of an image 16
   1.3 Example 3 (PCA): Correlated student marks 16
2. MATHEMATICAL DESCRIPTION OF THE KARHUNEN-LOÈVE TRANSFORM 16
3. APPLICATION OF KLT 19
   3.1 Encoding bits reduction of an image 19
   3.2 PCA of students' marks 22

CHAPTER 4. HYPOTHESIS TESTING 24

1. INTRODUCTION 24
   1.1 Definition 24
   1.2 Types of tests 24
2. STATISTICAL CONTROL OF QUALITY 24
   2.1 Reception quality control 24
   2.2 Statistical mastering of production 25
3. FIT TEST: $\chi 2$ TEST 26
   3.1 Definition of a $\chi k2$ distribution law of k degrees of freedom 26
   3.2 $\chi 2$ test 26

CHAPTER 5. INVERSE PROBLEM AND LEAST-SQUARES FITTING 28

1. GENERAL FRAME 28
2. WHY USING LEAST SQUARES? 29
3. LINEAR LEAST SQUARES WITH TWO PARAMETERS a AND b : $yi = a\,xi + b$ 29
4. LINEAR LEAST SQUARES: MATRIX FORMULATION 30
5. LINEARIZED LEAST SQUARES 31
6. NONLINEAR LEAST SQUARES: THE GAUSS-NEWTON METHOD 31
   6.1 Newton method 31
   6.2 The Gauss-Newton method 31
   6.3 Uncertainties on the parameters 32
   6.3 Example: fitting of Gaussian distribution of data 32

ANNEX 35

A.1 Arithmetic mean (or arithmetic average) *35*
A.2 Expected value *35*
A.3 Standard deviation *35*
A.4 sample standard deviation *36*
A.5 Bayes theorem *36*

Chapter 1. Introduction to inverse problem

1. Introduction

An inverse problem, in science, is the process of calculating from a set of observations the causal factors that produced them: for example, calculating an image in X-ray computed tomography, source reconstruction in acoustics, or calculating the density of the Earth from measurements of its gravity field. It is called an inverse problem because it starts with the effects and then calculates the causes. It is the inverse of a forward problem (or direct problem), which starts with the causes and then calculates the effects.

2. Direct and inverse problem

Let be an unknown quantity (a temperature, an angle, a resistance, …) whose precise value, named θ, will remain unknown because of some uncertainties in the measurement process. The objective of this chapter can be defined as: How to use at best all the available knowledges to give the most precise information on θ. These available knowledges consist in:

- The measurements constituting a set of N data named $\boldsymbol{D} = \{d_1, d_2, d_3, \dots, d_N\}$
- A model of the measurement process (for example in optics, we know how an image is formed with a microscope if we have determined its impulse response)
- Any information about θ known before the measurement process, called *a priori* information.

In this context, we are leaded to distinguish:

- *The direct problem:* Making a model of the measurement process. Example: how is formed by a microscope the image of an object?
- *The inverse problem:* What can we infer on θ (the object) from the measurements (the numbers forming the recorded image)?

The general answer to the inverse problem can be formulated as:

> *Build a probability law on θ, from all the available informations (measurements, model of measurement, a priori information). This probability law is called a posteriori (after the measurements).*

3. The Bayes theorem

To start, let us first resolve small exercises to understand the above notions.

3.1 example 1

> *It is known that at a given date, 3% of a population is infected with hepatitis: If the person is sick, then the test is positive with a 95% probability. If the person is not sick, then the test is positive with a 10% probability.*
> *Question: A person is randomly tested in the population and the test is positive. How likely is the person tested to be sick?*

- In that case we have a priori information: *3% of a population is infected* that is noted:
$$\mathrm{P}_{\mathrm{prior}}(\mathrm{sick}) = 3\%$$
- Model of the measurement: $\mathrm{P}(+|\mathrm{sick}) = 95\%$, $\mathrm{P}(+|\mathrm{health}) = 10\%$†

The answer to the inverse problem "*How likely is the person tested to be sick?*" corresponds to the conditional probability $\mathrm{P}(\mathrm{sick}|+)$. Of course, the conditional probability $\mathrm{P}(\mathrm{sick}|-)$ is also very important. The solution of the inverse problem uses the Bayes theorem, which is given in the annex (see A.5).
In the context of the inverse problem, by naming $\boldsymbol{D}$ the data $\{d_1, d_2, d_3, \dots, d_N\}$ the Bayes theorem is defined as:

$$P(\theta|\boldsymbol{D}) = \frac{P(\boldsymbol{D}|\theta)P_{prior}(\theta)}{P(\boldsymbol{D})} \quad , \qquad [1.1]$$

For the example it yields to: $P(sick|+) = \frac{P(+|sick)P_{prior}(sick)}{P(+)} = \frac{0.95\times0.03}{0.95\times0.03+0.1\times0.97} \approx 0.23 \equiv 23\%$.
The result could seem low: the test seemed rather good. Actually, the *a priori* information is here essential: if $P_{prior}(sick) = 10\%$, the result becomes 51%. This *a priori* information is nevertheless often difficult to assess (think to the result of the exercise if, instead of testing a person randomly in the population, you test someone who volunteers....). Is the test not useful? It depends...
To answer it is worth to calculate $P(sick|-) = \frac{P(-|sick)P_{prior}(sick)}{P(-)} = \frac{0.05\times0.03}{0.05\times0.03+0.9\times0.97} \approx 0.2\%$. If you have a cheap innocuous treatment, it could be interesting to treat $P(+) = 12.5\%$ of the population, including almost all the ills. On the contrary, if you have no treatment or if the treatment has some secondary effects, the test can be only used as a first indication of illness and must be confirmed.

† We use here conditional probabilities: the vertical bar means "given". We admit also that the test gives always an answer (someone who is not tested positive is tested negative). The symbols "+" and "-" means "positive" and "negative" respectively.

3.2 Example 2

We want to measure the speed v *of a cyclist on a climb (towards the summit!) with a counter with an uncertainty* $\mathrm{c} = \mathrm{v} \pm 2\ \mathrm{km/h}$

Model of the measurement: the probability law (i.e. density) $\mathrm{p(c \mid v)}$ (Fig. 1) is Gaussian with an offset traducing a small probability assigned to outliers (failed counter) in the range of measurement ($[0\ \mathrm{c_{max}}]$).

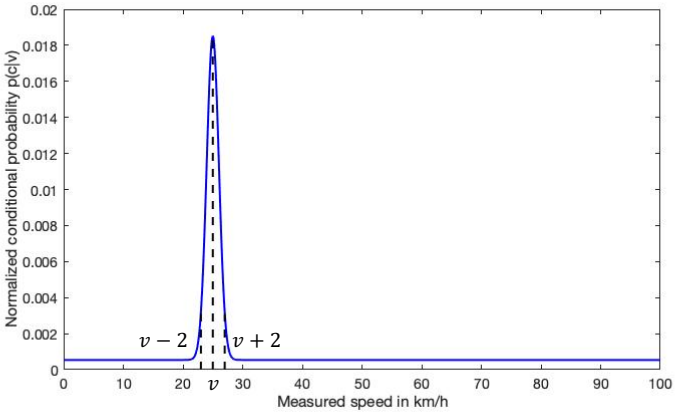

Fig. 1.1: Normalized conditional probability to measure a cyclist's speed with a counter operating in the range 0-100 km/h.

Data: $\mathrm{p(v \mid c_A)}$ if $\mathrm{c_A} = 15\ \mathrm{km/h}$ and $\mathrm{p(v \mid c_B)}$ if $\mathrm{c_B} = 80\ \mathrm{km/h}$. Of course, for $\mathrm{c_B}$, the result is not correct: the measurement process has failed. Because computers have no "common sense" (despite all the present discourses on artificial intelligence), we should define a procedure that takes into account the possibility of failure.

A priori information: because of the climb, the graph depicted in Fig. 2 reflects what we know about the cyclist's speed before its measurement (speed is in the range $0 - 40\ \mathrm{km/h}$).

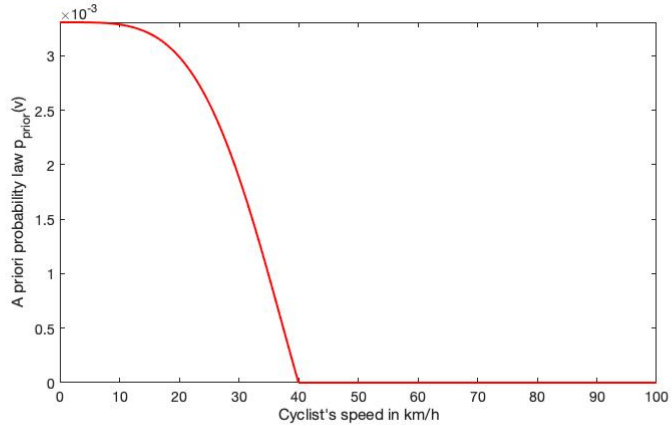

Fig. 1.2: *A priori* probability law of a cyclist's speed along the climb. The probability that speed is greater than $40\ km/h$ is null.

In this context, we use in the following a continuous version of the Bayes theorem:

$$\mathrm{p(v \mid c) \propto p(c \mid v)\ p_{prior}\ (v)} \qquad [1.2]$$

Compared to the discrete version, equality has been replaced by an operator $\propto$ which means "proportional to" and the denominator, which should be $\mathrm{p(c)}$, has been "forgotten". Indeed, from the point of view of the experimenter, c, and therefore $\mathrm{p(c)}$, are constants, which are taken into account in the form of a proportionality coefficient. If necessary, this proportionality coefficient shall be determined bearing in mind that, for any random variable x, $\int_{-\infty}^{+\infty} \mathrm{p(x)dx} = 1$.

We first formulate the direct problem by taking into account the fact that we know the measurement $\mathrm{c_0}$, which appears as a constant, and we don't know the true value $\mathrm{v_0}$, which becomes the variable.

For each value of v, one obtains a law $\mathrm{p(c|v)}$, function of c, which depends only of $\mathrm{c} - \mathrm{v}$. Hence, for the effective measurement $\mathrm{c_0}$, it is possible to draw $\mathrm{p(c_0|v)}$, function of v (Fig. 1.3).

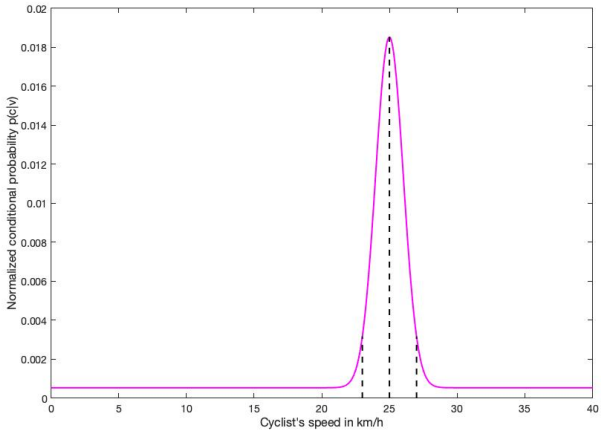

Fig. 1.3: *Conditional probability* $p(c_0|v)$ *when a speed* $c_0 = 25 \pm 2\ km/h$ is measured with the counter. This probability is defined in the range $0 - 40\ km/h$.

Then, by multiplying by $\mathrm{p_{prior}\ (v)}$ and normalizing, one obtains the *a posteriori* laws depicted in figure 1.4 when the measured speed is $\mathrm{c_A = 15\ km/h}$ and $\mathrm{c_B = 80\ km/h}$. We see that, in (A) there is no great difference between $p(c_0|v)$ and $p(v|c)$. On the other hand, in (B), the *a posteriori* law is equal to the a *priori* law: the measurement has no gained information on the cyclist speed.

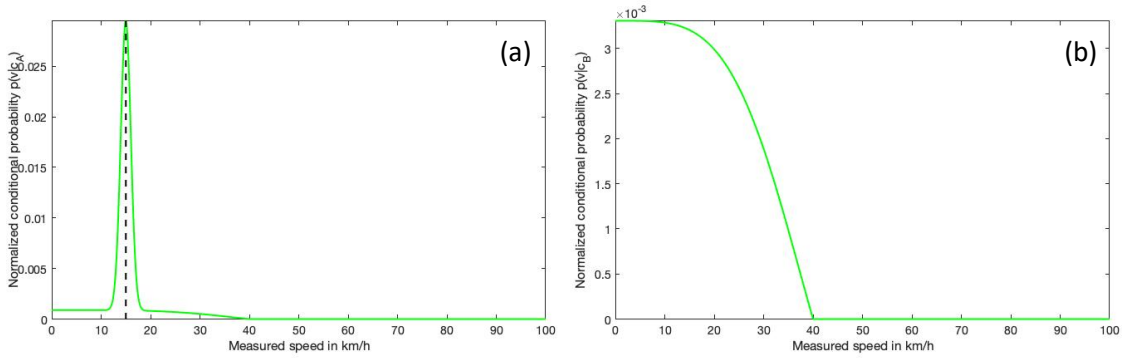

Fig. 1.4: A posteriori laws when speeds $\mathrm{c_A = 15\ km/h}$ (a) and $c_B = 80\ km/h$ (b) are measured with the counter

3.3 To summarize

We can summarize this chapter by the following table:

World	**Model**	**Experimenter**
unknown quantity θ	True (but known)	Random variable
measurements	Noised with random variables	Done (known values)
Problem to solve	Direct: the physical model of measurements determines $\mathrm{p(\theta|D)}$	Inverse: find $\mathrm{p(\theta|D)}$. The direct problem should be solved before

Actually, it could be useful to give, rather than the full $\mathrm{p(\theta|D)}$, two numbers, i.e. the mean of the law and an uncertainty range given by the standard deviation. This is the purpose of chapter 2.

Chapter 2. Estimation

1. Introduction

In the model world, θ is a true but unknown parameter of a probability law. Example: let us perform N measurements d of θ. In the model world d_n is a random variable:

$$d_i = \theta + \varepsilon_i \text{ with } i = 1, \ldots, N. \qquad [2.1]$$

The random error ε_i obeys a Gaussian distribution with zero mean, which means on the one hand that the measurement process is under control (see paragraph 5), on the other hand that the measurement process is without systematic error. Indeed, the systematic error is by definition the part of the error that is found in all measurements, therefore the mean of the error. The measurements d_n have therefore all the same mean θ.

The purpose of estimation is to construct from the measurements and *a priori* information a new random variable $\hat{\theta}$, called **estimator of** θ such as:

$$\hat{\theta} = T(d_1, \ldots, d_N, a\ priori\ information) \qquad [2.2]$$

An example of this estimator is the arithmetic average (see Annex A.1):

$$\hat{\theta} = \overline{d} = \frac{1}{N}\sum_{i=1}^{N} d_i \qquad [2.3]$$

This example shows us that an estimator is a random variable, just like the measures from which it is derived. However, it is easy to show (do it!) that, if the measurements are all independent one of each other, the variance of the arithmetic mean is $\frac{\sigma^2}{N}$. $\overline{d}$ is thus expectation of θ, just like the measurements d_i, but fluctuates less around θ, because of a standard deviation divided by $\sqrt{N}$.

Figure 2.4 illustrates the case where a true parameter $\theta = 5$ is measured 100 times with a gaussian random error with a standard deviation $\sigma = 1$. The blue curve shows the fluctuation of the measures around the true value. From these data the estimator of θ is: $\hat{\theta} = \overline{d} = \frac{1}{N}\sum_{i=1}^{N} d_i = 4.82$ with a standard deviation of 1.01 which is close to the standard deviation of the random noise. Then, the 100 measures are repeated 100 times. The red curve shows the fluctuation of the 100-arithmetic means. In that case, the arithmetic mean fluctuates with a smallest standard deviation of 0.09 with is close to $\frac{\sigma}{\sqrt{N}} = \frac{1}{\sqrt{100}} = 0.1$.

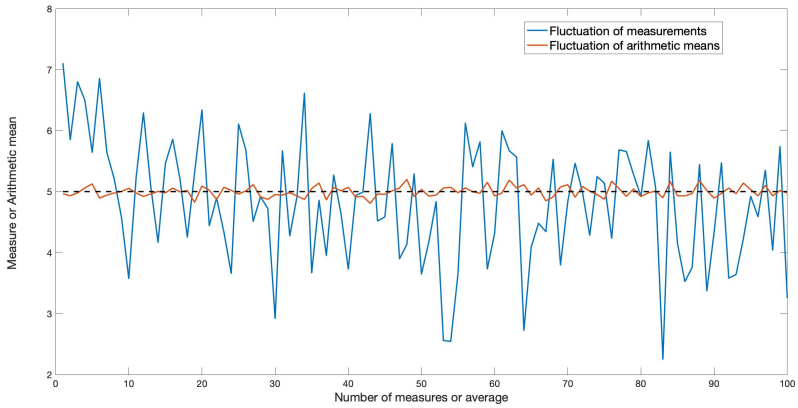

Fig. 2.1: The blue curve shows the fluctuations of 100 measures d_i of a true parameter $\theta = 5$ with a gaussian random error with a standard deviation $\sigma = 1$. The red curve corresponds to the fluctuation of the arithmetic mean performed over a set of 100 measures.

2. General properties

2.1 Consistent (asymptotically unbiased) estimator

An estimator is said consistent or asymptotically unbiased if:

$$\lim_{N\to\infty} \hat{\theta} = \theta \qquad [2.4]$$

2.2 Unbiased estimator

An estimator is unbiased if:

$$E(\hat{\theta}) = \theta \qquad [2.5]$$

Any reasonable estimator is asymptotically unbiased: if we have an infinite number of measures, we completely know the law of probability and therefore the true value. For example, it has been shown that the arithmetic mean tends towards the true mean for a very large number of measures (weak law of large numbers).

On the other hand, there are good estimators that are biased. Indeed, an unbiased estimator has a variance greater than or equal to a limit σ_0^2, known as the *Cramer-Rao limit*. This limit is given by a somewhat barbaric formula:

$$Var(\hat{\theta}) \geq \sigma_0^2 = \frac{1}{I(\theta)} \qquad [2.6]$$

Where $I(\theta) = E\left\{\left[\frac{\partial}{\partial\theta} ln(p(X;\theta))\right]^2\right\}$ is the *Fisher information* and $p(X;\theta)$ the likelihood function (or joint probability) of the random variable X conditioned on the value of θ. The partial derivative with respect to θ of the natural logarithm of the likelihood function is called the score. it can be shown that the expected value (the first moment) of the score, evaluated at the true parameter value θ, is 0.

The *Fisher information* is defined to be the variance of the score and it describes the probability that we observe a given outcome of X, *given* a known value of θ. If p is sharply peaked with respect to changes in θ, it is easy to indicate the "correct" value of θ from the data, or equivalently, that the data X provides a lot of information about the parameter θ. If the likelihood p is flat and spread-out, then it would take many samples of X to estimate the actual "true" value of θ that *would* be obtained using the entire population being sampled. This suggests studying some kind of variance with respect to θ.

If $ln\, p(X;\ \theta)$ is twice differentiable with respect to θ, and under certain regularity conditions, then the Fisher information may also be written as:

$$I(\theta) = -E\left\{\frac{\partial^2}{\partial\theta^2} ln(p(X;\theta))\right\} \qquad [2.7]$$

> *Example: Calculate the Cramer-Rao limit for a Bernouilli trial with two possible outcomes, "success" and "failure", with success having a probability of θ. The outcome can be thought of as determined by a coin toss, with the probability of heads being θ and the probability of tails being $1 - \theta$. The probability law of a Bernoulli trial can be defined as: $p(X;\theta) = \theta^X(1-\theta)^{1-X}$ with $X \in \{0;1\}$. Demonstrate than Cramer-Rao limit is equal to the true variance.*

An unbiased estimator of variance σ_0^2 is said to be efficient or minimum variance unbiased, and is, of course, the best unbiased estimator. On the other hand, there are sometimes biased estimators, which have therefore a mean different from the true value (this difference is called bias), whose variance is much lower than the Cramer-Rao limit. They may then be "better" than the efficient estimator, where the sense "better" will be defined in paragraph 6, devoted to Wiener filtering.

3. Estimators of mean

Two are in common use:

- Arithmetic mean (see eq. 2.3) : $\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i$. One immediately demonstrates, if the error is Gaussian and the measurements independent, that $\bar{d}$ follows a Gaussian law, of variance $\frac{\sigma^2}{N}$ and mean θ.
- Median. The measurements are ordered from the smallest (d_1) to the largest (d_N). The median is then defined as $d_{(N+1)/2}$ if N is odd, $(d_{N/2} + d_{(N+2)/2})/2$ if N is even. Median is much less sensitive than mean to outliers (N.B.: if the series of measurements include outliers, the error is no longer a Gaussian random variable, unlike in the remaining of the chapter).

We can compare the two estimators on an example: measurements of the period of a pendulum made on the chronometer by first year students: $T\ (seconds)$: $10.62\ 10.38\ 10.34\ 10.35\ 10.40\ 10.36$. Mean is $10.408\ s$ and the median value is $10.37\ s$. The measured period of $10.62\ s$ is is likely to be an outlier. If we remove this measure from data, the mean becomes $10.366\ s$.

In general, graphical representation of the data is very useful to conclude. With symmetric distribution laws, mean and median are equal.

4. Estimators of variance

With known mean: $\widehat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}(d_i - \theta)^2$

With estimated mean: $\widehat{\sigma^2} = \frac{1}{N-1}\sum_{i=1}^{N}\left(d_i - \bar{d}\right)^2$

If the multiplicative coefficient was $1/N$ and not $1/(N-1)$, $\widehat{\sigma^2}$ would be the arithmetic mean of $\left(d_i - \bar{d}\right)^2$, just as the true variance is the (true) mean of $(d_i - E\{d\})^2$. We understand the need to use $1/(N-1)$ when thinking about the case where we have only one measurement. So σ^2 is indeterminate, which seems correct since we have no idea of the dispersion of the measurements, represented by the variance. Using $1/N$ would give zero variance, which is clearly incorrect. In fact, the $d_i - \bar{d}$ are not independent, unlike the d_i.
For example, for $N = 2$ measurements: $d_1 - \bar{d} = -(d_2 - \bar{d})$.

5. Why the measurement errors are often (not always!) modeled by a Gaussian law?

The answer is a consequence of the *central-limit theorem* (CLT) or strong law of large numbers that can summarize as follows. Let $X_1, \ldots, X_i, \ldots\ldots, X_N$ N independent random variables each with the mean m and finite variance σ^2[‡]. The sum $X_1 + \ldots + X_i + \ldots\ldots + X_N$ has mean Nm and variance $N\sigma^2$. Consider now the new random variable Z_N such as:

$$Z_N = \frac{X_1 + \ldots + X_i + \ldots\ldots + X_N - Nm}{\sqrt{N\sigma^2}} = \sum_{i=1}^{N} \frac{X_i - m}{\sqrt{N\sigma^2}} = \sum_{i=1}^{N} \frac{Y_i}{\sqrt{N}} \qquad [2.8]$$

[‡] In fact, the "same variance" condition is not exactly necessary. It is sufficient that the variances have the same order of magnitude.

where $Y_i = \frac{X_i - m}{\sigma}$ is a random variable with zero mean and unit variance. We can demonstrate that: $\lim_{N\to\infty} Z_N = N(0,1)$ where $N(0,1)$ is a normal distribution (i.e. a gaussian law) of mean 0 and variance 1.

This theorem, which we will admit, applies to a measurement process of good quality, said under control, where all important causes of error have been eliminated. The residual uncertainty is due to a large number of independent causes, of various origins and of comparable weight. The measurement error is then expected to be a Gaussian random variable, whatever the probability law of each residual error.

As an example, let consider a $Bernouilli$ trial with two possible outcomes, "1" and "0" with the same probability of 0.5 (the outcome can be thought of as determined by a coin toss). This random variable has a mean $m = 0.5$ and a variance $\sigma^2 = 0.25$. We are interested in the sum of N trials. For example, with 3 trials the results that the sum can take are given in table 2.1 and the frequency of these results are given in table 2.2.

Trial 1	Trial 2	Trial 3	Sum
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	2
1	0	0	1
1	0	1	2
1	1	0	2
1	1	1	3

Table 2.1

Value of the sum	Occurrence	Frequency
0	1	12,5%
1	3	37,5%
2	3	37,5%
3	1	12,5%

Table 2.2

Figure 2.2.a illustrates the frequency distribution of the sum obtained for different values of N. When N increases, the frequency distribution $p(\sum_{i=1}^{N} X_i)$ tends towards a Gaussian centered at $Nm = \frac{N}{2}$ and a variance increasing linearly with N $Var(\sum_{i=1}^{N} X_i) = N\sigma^2 = \frac{N}{4}$ (Fig. 2.2.b). From the same data, figure 2.2.c shows the frequency distribution of Z_N. We can observe that when N increases, $p(Z_N)$ tends toward a Gaussian of mean 0 and variance 1 (Fig. 2.2.d).

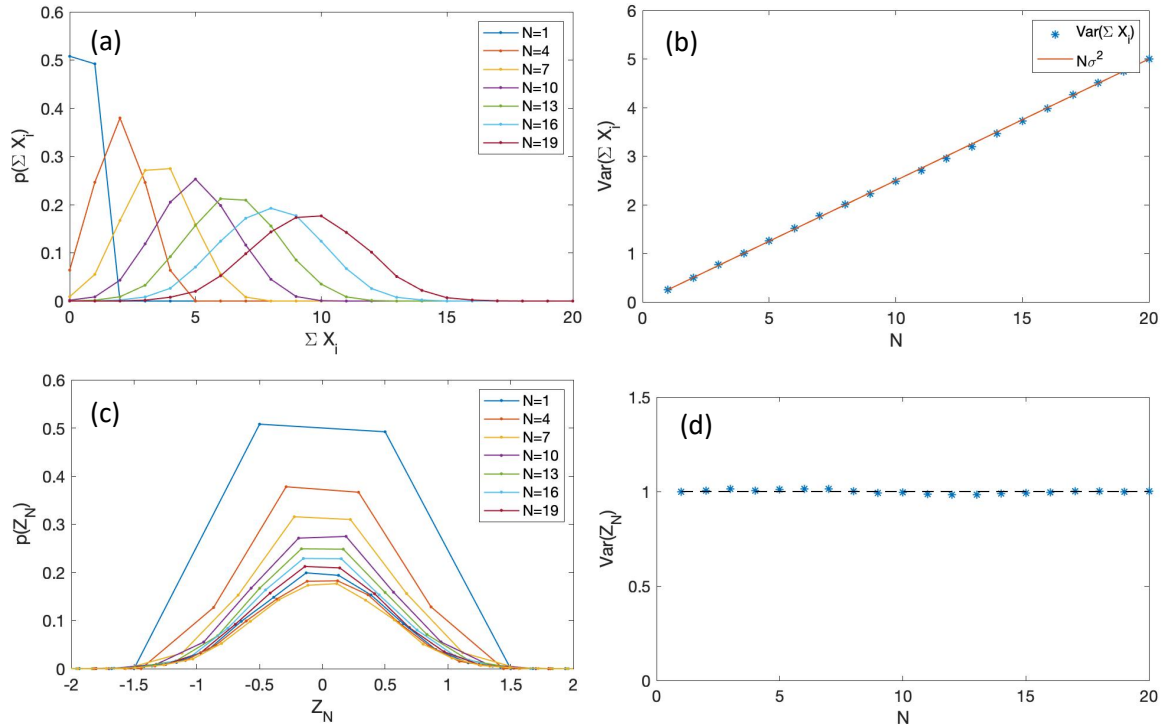

Fig. 2.2: (a) Probability distribution of the results of the sum of N Bernouilli trials. (b) Variance of the sum as a function of the number of summed trials. (c) Probability distribution of the random variable Z_N and (d) its variance as a function of the number of summed trials.

Note that a Gaussian statistic means the absence of outliers. Indeed, for a Gaussian law, $P(|d_i - \theta| > 5\,\sigma) < 2 \times 10^{-6}$. This is in agreement with the notion of process under control: an outlier comes from an important error source and happens if the process is not under control.

6. Confidence intervals

In the experimenter world, we would like to translate the *a posteriori* law $p(\theta|D)$ in a more intuitive way, by giving an interval where θ lies with a probability of 95%. In the model world, this is easy: if the process is under control, $\bar{d}$ follows a Gaussian law, of variance σ^2/N and mean θ, which allows us to write: $\theta - \frac{1.96\sigma}{\sqrt{N}} \leq \bar{d} \leq \theta + \frac{1.96\sigma}{\sqrt{N}}$ at 95% of confidence. Two reasons prevent us to simply write in the experimenter world:

$$\bar{d} - \frac{1.96\sigma}{\sqrt{N}} \leq \theta \leq \bar{d} + \frac{1.96\sigma}{\sqrt{N}} \qquad [2.9]$$

First, this expression assumes $p(\theta|D) = p(\bar{d}|\theta)$, which is true only if $p_{\text{prior}}(\theta) = Cst$. Second, σ is not known, but estimated.

As for the first assumption ($p_{\text{prior}}(\theta) = Cst$), this is a reasonable one for a controlled measurement process, where the measurement error is low and Gaussian. The uncertainty range is then low enough to consider that, within this range, the probability density of θ before measurements is a constant. Of course, if we have an explicit expression of $p_{\text{prior}}(\theta)$, we must renounce this assumption and calculate $p(\theta|\bar{d})$ with $p_{\text{prior}}(\theta)$.

As for the second reason, if $p_{\text{prior}}(\theta) = Cst$, $\frac{\bar{d}-\theta}{\sqrt{\widehat{\sigma^2}}/\sqrt{N}}$ follows a so-called $Student\ law$[§] with $N-1$ degrees of freedom, where θ is the random variable. The range around the arithmetic mean where θ has a 95% chance of being found can then be determined: this range, called the confidence interval, is given by $\bar{d} \pm \alpha \frac{\hat{\sigma}}{\sqrt{N}}$, where α depends on N:

N	**3**	**5**	**10**	**20**	**40**
$\boldsymbol{\alpha}$	4.3	2.8	2.3	2.1	2.0

Thus, from 40 measurements, Student's law merges with a Gaussian law. For a very small number of measures, however, there is a chance of underestimating the standard deviation, which gives a greater chance that the true value deviates from the arithmetic mean of more than two estimated standard deviations, as it is evident on the graph in Figure 2.3.

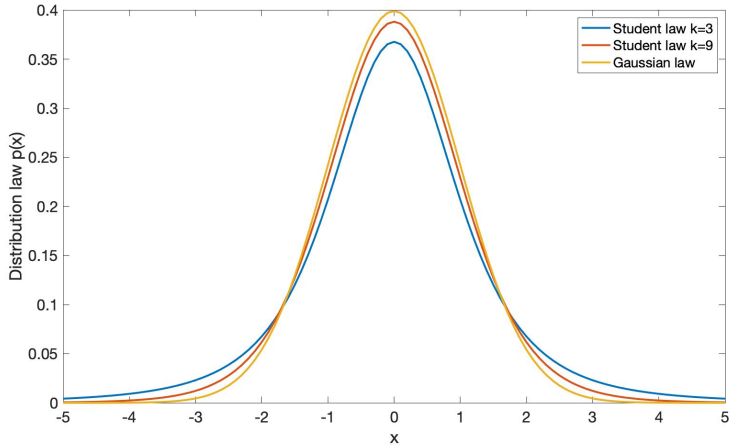

Fig. 2.3: Comparison between Student laws with 3 and 9 degrees of freedom and a Gaussian law with a standard deviation of 1. For large degree of freedom ($k \geq 40$), Student law merges with a Gaussian law.

[§] The Student law with k degrees of freedom is given by $f(x) = \frac{1}{\sqrt{k\pi}} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{t^2}{k}\right)^{-\left(\frac{k+1}{2}\right)}$, for $k > 0$ and where Γ is the $Gamma\ Euler$ function.

Warning: *Using 2 or α has hardly any consequences as soon as you make at least ten measurements. However, it should not be forgotten that the estimated standard deviation of θ from the arithmetic mean is not the estimated standard deviation of the measures $\hat{\sigma}$, but $\hat{\sigma}$! That's the point of repeating the measurements! ...and we will not forget either that this division by $\sqrt{N}$ is intimately linked to the assumption of independence of the measurements. If this assumption is not fully verified, the size of the confidence interval may be underestimated.*

7. Recursive estimators

To avoid memory consumption, the estimator is updated after each measurement. The simplest example is the recursive form of the mathematical average:

$$\overline{d_n} = \frac{d_n + (n-1)\overline{d_{n-1}}}{n} \qquad [2.10]$$

Clearly, only the three values that form this equation must be kept in memory, instead of N with the non-recursive definition of the arithmetic average. While we have in this example a strict equivalence between the recursive and the non-recursive definition, some new features can be gained from recursivity. For example, a numerical low-pass filter is obtained by computing:

$$\hat{\theta} = (1-b)y_n \qquad [2.11]$$

with $y_n = d_n + by_{n-1}$ and where $b = 1 - \varepsilon$ $(0 < \varepsilon \ll 1)$. It can be shown that this filter is the numerical equivalent of an analogic low-pass RC filter, with $b = exp(-\Delta t/\tau)$, with $\tau = RC$ the time constant and Δt the sampling step.

8. Wiener filtering

Wiener filtering is an important example where a biased estimator works better than its unbiased counterpart. Let be, in the Fourier domain, a signal (object) $O(\nu)$, depending of the frequency ν, on which is applied a low-pass filter, of transfer function $H(\nu)$. The output of the filter (or image) can be written as:

$$I(\nu) = H(\nu)O(\nu) + N(\nu) \qquad [2.12]$$

where $N(\nu)$ represents a Gaussian additive noise of spectral density $\langle N(\nu)^2 \rangle$ with $\langle N(\nu) \rangle = 0$.
In general, $H(\nu) \approx 1$ for low frequencies and $H(\nu) \approx 0$ at high frequencies. The problem we consider is "How to retrieve at best $O(\nu)$ from the output (the image)?". In the absence of noise, a simple multiplication of the measured $I(\nu)$ by $1/H(\nu)$ would work. However, the noise is multiplied by the same coefficient and will have catastrophic effects on the retrieval if $1/H(\nu)$ is too large, as it will be at high frequencies. Hence, we are looking for a function $G(\nu) \leq 1/H(\nu)$, such that:

$$\widehat{O(\nu)} = G(\nu)I(\nu) = G(\nu)H(\nu)O(\nu) + G(\nu)N(\nu) \quad [2.13]$$

would be as close as possible of $O(\nu)$. More precisely, we want to minimize the expectation of the quadratic error (we will omit from now the explicit mention of the frequency dependence):

$$\langle \left(\hat{O} - O\right)^2 \rangle = \langle O^2(1 + G^2H^2 - 2GH) + G^2N^2 + 2G^2HNO - 2GNO \rangle \quad [2.14]$$

We further assume that the output noise is independent of the object ($\langle NO\rangle = 0$). Hence Eq. 2.14 becomes:

$$\langle\left(\hat{O} - O\right)^2\rangle = \langle O^2(1 + G^2H^2 - 2GH)\rangle + \langle G^2N^2\rangle \quad [2.15]$$

The expectation of the quadratic error is then minimized if:

$$\frac{\partial\,\langle\left(\hat{O} - O\right)^2\rangle}{\partial G} = 0 = \langle O^2\rangle(2GH^2 - 2H) + 2G\langle N^2\rangle \quad [2.16]$$

It gives:

$$G = \frac{H\langle O^2\rangle}{H^2\langle O^2\rangle + \langle N^2\rangle} = \frac{1}{H}\frac{1}{1 + \frac{\langle N^2\rangle}{H^2\langle O^2\rangle}} = \frac{1}{H}\frac{1}{1 + \frac{1}{SNR}} \quad [2.17]$$

where SNR represents the signal-to-noise ratio in the output signal.
Two remarks:

1. Note that we use in this calculation the spectral density of the object $\langle O^2\rangle$, meaning that we have some a priori information on the statistical properties of this object, even if we don't know his exact structure (retrieving this structure is the goal of this calculation).
2. The estimator is biased: because $\langle N(\nu)\rangle = 0$, $\langle\frac{I(\nu)}{H(\nu)}\rangle = O(\nu) \Rightarrow \left|\widehat{O(\nu)}\right| = |G(\nu)I(\nu)| \leq |O(\nu)|$, with $bias = \langle|O(\nu)| - \left|\widehat{O(\nu)}\right|\rangle$.

To diminish the quadratic error, we have introduced some bias: $\langle\widehat{O(\nu)}\rangle \leq \langle O(\nu)\rangle$. In other words, not only the mean value of the noise matters. A noise of zero mean can nevertheless dominate the retrieved object.

Here we present the application of Wiener filtering to a sound $x(t)$ recorded with a microphone with a frequency transfer function $H(\nu)$ (Fig. 2.4) and an added Gaussian noise $n(t)$ with zero mean and known variance. In the time domain, the recorded sound $y(t)$ is given by:

$$y(t) = h(t) * x(t) + n(t) \quad [2.18]$$

where $h(t) = TF^{-1}(H)$ is the inverse Fourier transform of the frequency response of the microphone. In the frequency domain Eq. 2.18 becomes :

$$Y(\nu) = H(\nu)X(\nu) + N(\nu) \quad [2.19]$$

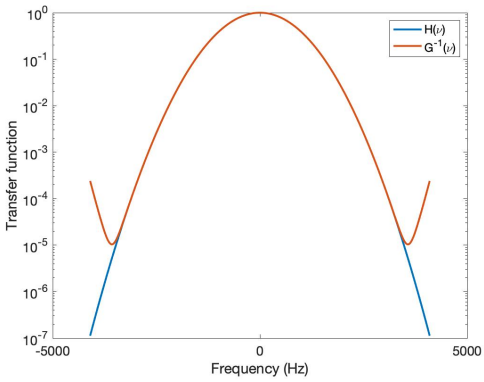

Fig. 2.4: Transfer function $H(\nu)$ of the microphone and function $G(\nu)$ minimizing the quadratic error.

Using the definition given by Eq. 2.17 for the function $G(\nu)$ (Fig. 2.4) we can compare the estimated original sound when the estimated sound is deduced from the recorded sound by:

$$\begin{cases}\hat{x}_1(t) = TF^{-1}\left(\frac{Y}{H}\right)\\ \hat{x}_2(t) = TF^{-1}(GY)\end{cases} \quad [2.20]$$

Figure 2.5a shows the graphs of the original sound and the retrieved sounds according to Eq. 2.20, in the time and frequency domains. It shows how an optimized function $G(\nu)$ allows to retrieve the

original signal with a good SNR. Figure 2.5b shows the fluctuation of the quadratic errors for both retrieved sounds as a function of time.

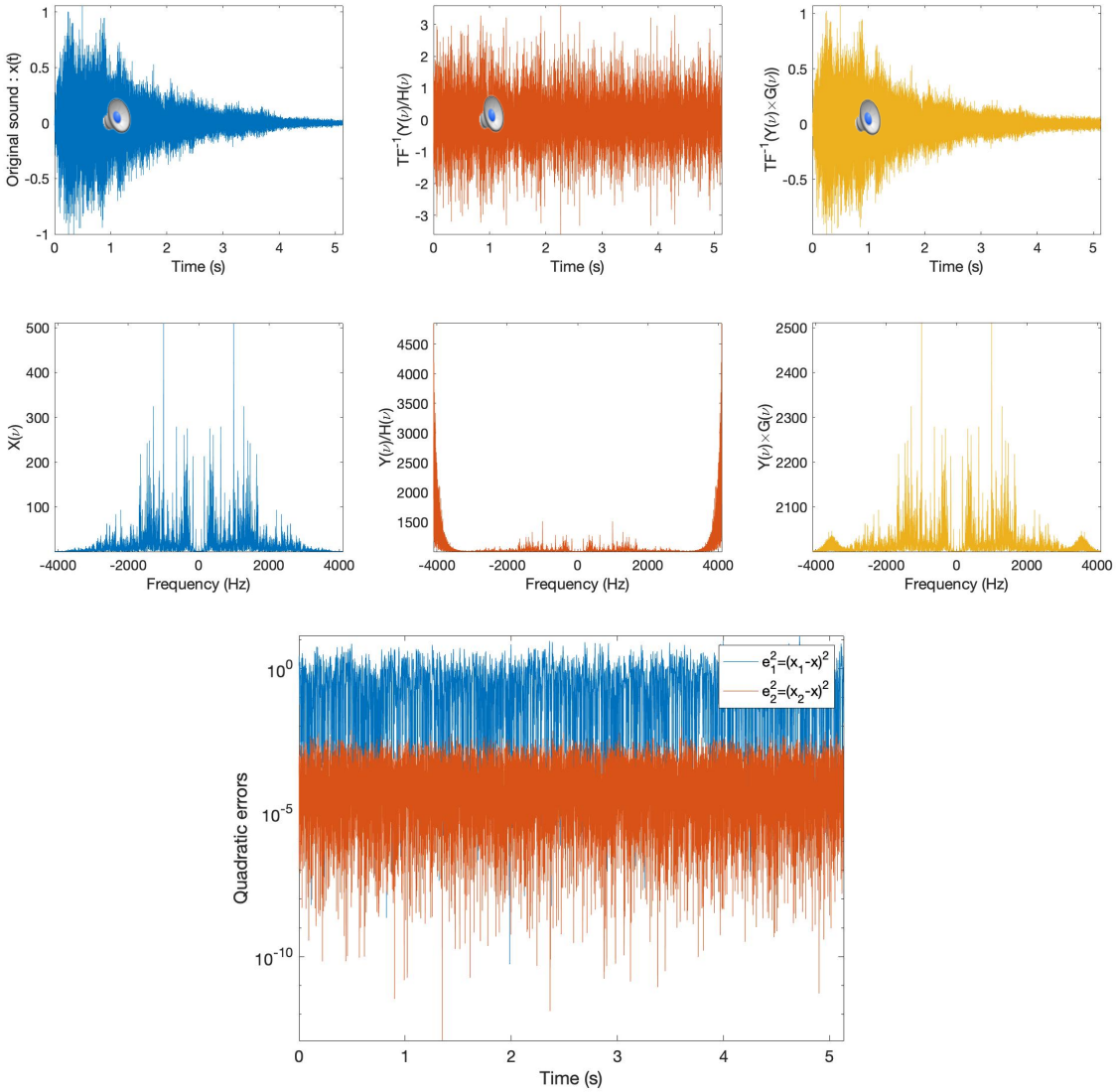

Fig. 2.5: (a) original and retrieved sounds in the time and frequency domains. Sounds can be played by clicking on small icons. (b) Quadratic errors for both retrieved sounds.

Chapter 3. Karhunen-Loève Transform or Principal Component Analysis

1. Purpose and general description

The same mathematical methods are employed on a set of data in which redundant information is coded, with two goals, linked but different, which give at least two different names to the methods.

- Karhunen-Loève Transform (KLT): compression of information by eliminating redundancy.
- Principal Component Analysis (PCA): studying the redundancy.

Another name of these methods is "factorial analysis of correspondences". It is a bit less employed, at least in the physical and technical domain, and will be left aside in the following. Let us give two generic examples of use of KLT and PCA.

1.1 Example 1 (KLT): multispectral image compression

We consider a set of K aerial images of the same landscape, obtained by using different color filters. The images are different but share some common information. If these images come from a satellite, it would be important to minimize the transmitted information and we consider the following problem: is it possible to transmit a reduced set of M images, $\mathrm{M} < \mathrm{K}$, plus some coefficients, which allow the original images to be retrieved with a negligible information loss.

1.2 Example 2 (KLT): Encoding bits reduction of an image

The goal of image compression is to store an image in a more compact form, i.e., a representation that requires fewer bits for encoding than the original image. This is possible for images because, in their "raw" form, they contain a high degree of redundant data. Most images are not random collections of arbitrary intensity transitions. Every image we see contains some forms and structures. As a result, there is some correlation between neighboring pixels. If one can find a reversible transformation that removes the redundancy by decorrelating the data, then an image can be stored more efficiently. The Karhunen-Loève Transform (KLT) is the linear transformation that accomplishes this.

1.3 Example 3 (PCA): Correlated student marks

Some students receive marks in different matters (to be specific, between 0 and 20, as in France). Is it possible to retrieve their marks from fewer "super marks", meaning for example that their physics and mathematics marks are more correlated than, say, the physics and music marks.
In both examples, the same mathematical method will be employed. Only the goal is different, information compression (KLT) or study of correlations (PCA).

2. Mathematical description of the Karhunen-Loève Transform

Let us consider a set of K measures of N joint variables. the data are arranged in a matrix $\boldsymbol{M}$ of K rows and N columns as depicted in Fig. 3.1. Each of the K measures is formed by N variables, meaning that a measure of one variable M_{kn} can be represented by a point in a N dimensional space.

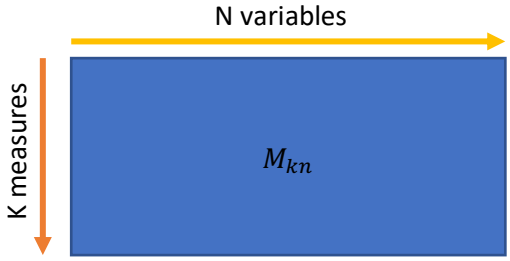

Fig. 3.1: arrangement K measures of N joint variables in a K rows and N columns matrix

Reducing the dimensionality to L will be possible if the cloud of points is contained in a L dimensional space, with some small random noise in the remaining dimensions. Let us define the average of the variable n, with $1 \leq n \leq N$:

$$\bar{M}_n = \frac{1}{K}\sum_{k=1}^{K} M_{kn} \qquad [3.1]$$

We define also the covariance between the variable p and the variable q:

$$C_{pq} = \frac{1}{K-1}\sum_{k=1}^{K}\left(M_{kp} - \bar{M}_p\right)\left(M_{kq} - \bar{M}_q\right) \qquad [3.2]$$

Hence, the $N \times N$ covariance matrix $\boldsymbol{C}$ of $\boldsymbol{M}$ can be define as:

$$\boldsymbol{C} = cov(\boldsymbol{M}) = \frac{\boldsymbol{M_c^T M_c}}{K-1} \qquad [3.3],$$

Where $\boldsymbol{M_c} = \boldsymbol{M} - \bar{\boldsymbol{M}}$ is the centered data matrix. The correlation coefficient r_{pq} is obtained by normalizing the covariances using the variances C_{pp} and C_{qq}:

$$r_{pq} = \frac{C_{pq}}{\sqrt{C_{pp}}\sqrt{C_{qq}}} \; ; \quad -1 \leq r_{pq} \leq 1 \quad [3.4]$$

All the following calculations can be made by using either the covariances or the correlations. This second choice is equivalent to first normalizing each variable by its standard deviation. This is possible but not compulsory: do we decide that each variable conveys the same quantity of information, whatever its standard deviation?
In the example 3, do we want that all matters have the same importance, even if a teacher gives marks on all the scale between 0 and 20, while a second teacher gives as lowest mark 7 and as highest 14? In some cases, there is no choice: if you want to correlate the number of children of a family and, say, the size of their home, you must use the correlation coefficients, since the units of the random variables are different.

In the following, we will use covariances. Let us define a rotation of the coordinates axes in the N-dimensional space:

$$\boldsymbol{M_c} = \boldsymbol{Y\Phi} \qquad [3.5],$$

where $\boldsymbol{\Phi}$ is a $N \times N$ rotation matrix between orthonormal bases in the N-dimensional space. For such a rotation matrix, we have $\boldsymbol{\Phi^{-1}} = \boldsymbol{\Phi^T}$, which allows us to write the transformed $K \times N$ matrix of data:

$$\boldsymbol{Y} = \boldsymbol{M_c\Phi^T} \qquad [3.6]$$

Defining respectively $\boldsymbol{V}$ and $\boldsymbol{D}$ as the $N \times N$ eigenvectors matrix[**] and the $N \times N$ diagonal eigenvalues matrix of the covariance matrix $\boldsymbol{C}$ (Eq. 3.3), we can prove that $\boldsymbol{\Phi} = \boldsymbol{V^T}$ and $\frac{\boldsymbol{Y^TY}}{K-1} = \boldsymbol{D}$. It implies that the K.L.T. data $\boldsymbol{Y}$ are uncorrelated (which does not mean independent if the random parts are not Gaussian).

[**] With Matlab, the eigenvectors matrix $\boldsymbol{V}$ and the eigenvalues matrix $\boldsymbol{D}$ are obtained with the eig function: $[\boldsymbol{V}, \boldsymbol{D}] = eig(\boldsymbol{C})$, and verify $\boldsymbol{CV} = \boldsymbol{VD}$.

Now we are looking for the rotation allowing the reconstruction of a $K \times N$ truncated data matrix $\boldsymbol{M_{trunc}}$ as close as possible of $\boldsymbol{M}$ with $L < N$ variables such that the quadratic error as a function of L: $\varepsilon_L^2 = \sum_{k=1}^{K} \sum_{n=1}^{N} (M_{kn} - M_{kn}^{trunc})^2$ is minimum. This is obtained calculating:

$$\boldsymbol{M_{trunc}} = \bar{\boldsymbol{M}} + \boldsymbol{Z\Psi} \quad [3.7]$$

where $\boldsymbol{Z}$ is a $K \times L$ matrix filled with the L last columns[††] of the matrix $\boldsymbol{Y}$ and $\boldsymbol{\Psi}$ is a $L \times N$ matrix filled with the L last rows of the matrix $\boldsymbol{\Phi}$

In matrix $\boldsymbol{D}$ the eigenvalues λ_k are ranked in ascending order ($\lambda_1 < \lambda_2 < \cdots < \lambda_N$). Consequently, the ratio $\varrho = \frac{\sum_{n=L+1}^{N} \lambda_n}{\sum_{n=1}^{N} \lambda_n}$ is the percentage of lost information. In many practical cases, there is a value of L for which this percentage drops quite abruptly to almost 0, meaning that the original data lie in a L dimensional space. Because of that, to retrieve the original data it is sufficient to conserve L K.L.T. variables, *i.e.* $K \times L$ numbers, plus the N numbers forming $\bar{\boldsymbol{M}}$ and the $L \times N$ coefficients forming $\boldsymbol{\Psi}$.

[††] In $\mathbf{V}$ and $\mathbf{D}$ matrices calculated with eig function, data are sorted in ascending order. Then, In Eq. 3.7 the calculation is made with the L last columms of the matrix $\mathbf{Y}$ and the L last rows of the matrix $\boldsymbol{\Phi}$ in order to take in account only the L highest component in the N-dimensional space.

3. Application of KLT

3.1 *Encoding bits reduction of an image*

We start with an image in grey scale of 800×600 pixels (Fig. 3.2a) which is divided in K blocks of N pixels. Then the image is rearranged into an image $\boldsymbol{I}$ of K rows and N columns. In figure 3.2b, with blocks of 40×40 pixels, $N = 40^2 = 1600$ and $K = \frac{800 \times 600}{40^2} = 300$.

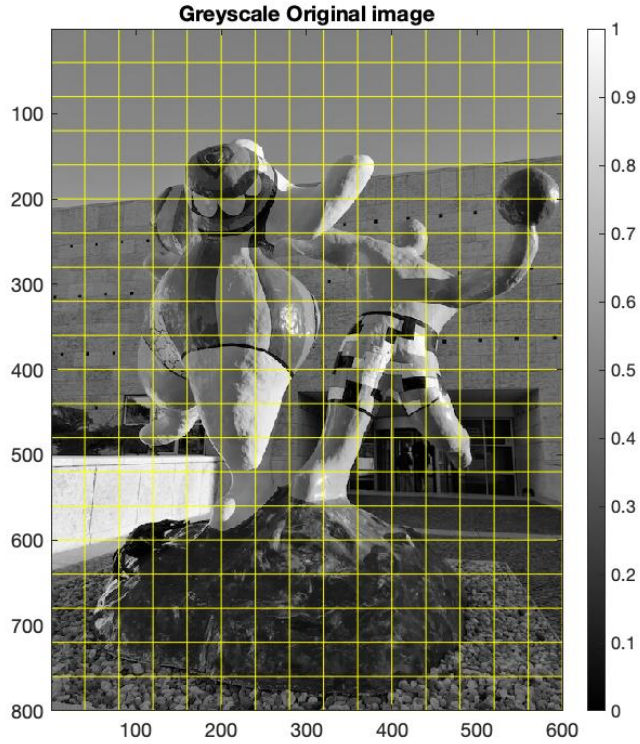

Fig. 3.2a: Normalized greyscale image of 800×600 pixels and subdivided in K blocks of N pixels. Here blocks of 40×40 pixels are illustrated.

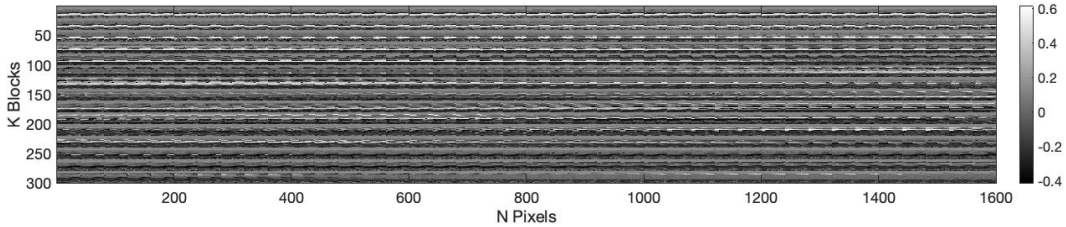

Fig. 3.2b: Rearrangement of the original image into an image of K rows and N columns.

In Fig. 3.2b a row (block) corresponds to a measure and a column (pixel) corresponds to a variable. Consequently, an image can be considered as K measures of N variables.
From the $K \times N$ rearranged image $\boldsymbol{I}$, I_{kn} is the value of the pixel n of the blocks k. The N average values of the variables are calculated over the K measures: $\bar{I}_n = \frac{1}{K}\sum_{k=1}^{K} I_{kn}$.
Then, the covariance matrix $\boldsymbol{C} = cov(\boldsymbol{I_c})$ of the centered image $\boldsymbol{I_c}$ is calculated. Figure 3.3 illustrates the $N \times N$ covariance matrix $\boldsymbol{C}$ when the image is subdivided in blocks of 8×8 pixels ($N = 64$).

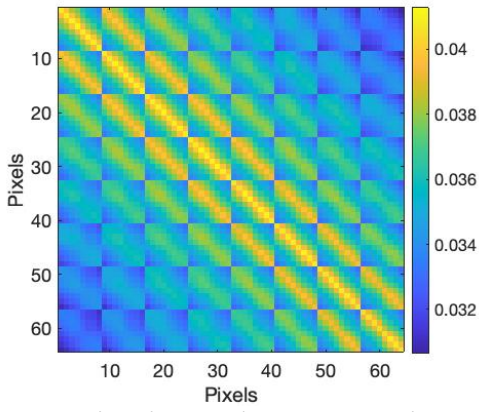

Fig. 3.3: Unbiased estimated covariance matrix between the $N = 64$ pixels of $K = 7500$ blocks when the original image is subdivided in blocks of 8×8 pixels.

The next step is to calculate the $N \times N$ diagonal matrix $\boldsymbol{D}$ of eigenvalues and a $N \times N$ full matrix $\boldsymbol{V}$ whose columns are the corresponding eigenvectors of the covariance matrix $\boldsymbol{C}$ (Fig. 3.4).

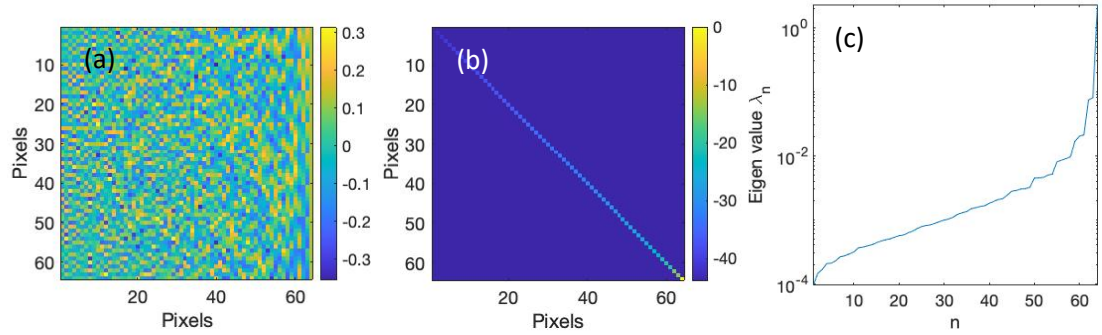

Fig. 3.4: (a) $N \times N$ Matrix where columns are the eigenvectors of the covariance matrix. (b) $N \times N$ diagonal matrix of eigenvalues given in dB. (c) Plot of the increasing ordered eigenvalues λ_n with $1 \leq n \leq N$.

From $\boldsymbol{V}$ matrix a new $K \times M$ pixels full image $\boldsymbol{Y}$ is calculated such as: $\boldsymbol{Y} = \boldsymbol{I_c V^T}$, where pixels of blocks are decorrelated (covariance matrix of $\boldsymbol{Y}$ is equal to the diagonal matrix $\boldsymbol{D}$ and $\boldsymbol{V^T} = \boldsymbol{V}^{-1}$). Figure 3.5 illustrates the reconstructed 800×600 pixels full image from $\boldsymbol{Y}$ with a zoom on of 100×100 pixels region of interest (ROI). On the ROI we clearly observed the delimitation of the blocks of 8×8 pixels where levels of pixels are decorrelated.

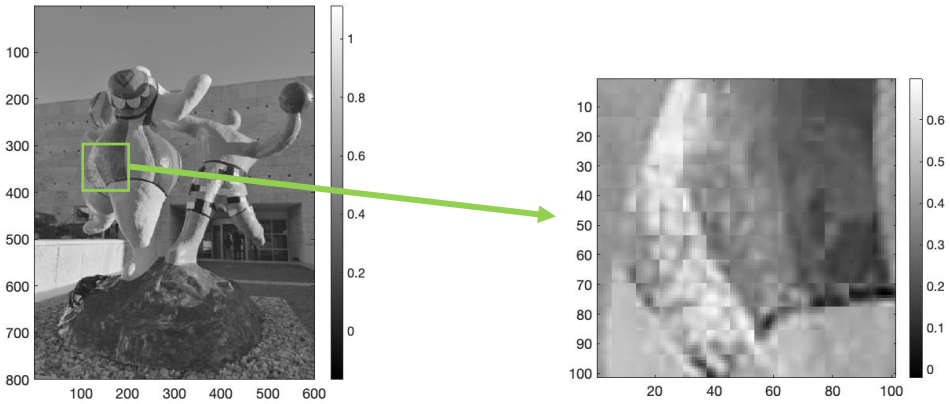

Fig. 3.5: Reconstructed image from the eigenvector's matrix and a zoom of a 100×100 pixels ROI

Using the L eigenvectors corresponding to the L last highest eigenvalues of the covariance matrix $\boldsymbol{C}$, a truncated matrix $\boldsymbol{I_{trunc}}$ of $K \times N$ pixels is obtained using Eq. 3.7.

Finally, the compressed image is obtained by rearranging the $K \times N$ pixels truncated matrix into an image of 800×600 pixels. Figure 3.6 shows the compression results for $L = 1,\ 5, 9, 13$ that can be compared to the original picture (1st row). To emphasize the effect of the compression, zooms of the same small ROIs are given (right column). For $L = 1$, the highest eigenvalue is used and its corresponding eigenvector. In the ROI we can observe that level in a block of 8×8 pixels is constant leading to a strong pixelization (i.e. low resolution) in the compressed image. When L increases, resolution increases too and we can observe in Fig. 3.6 than less of 10 eigenvalues are necessary to obtain a compressed image with a good resolution.

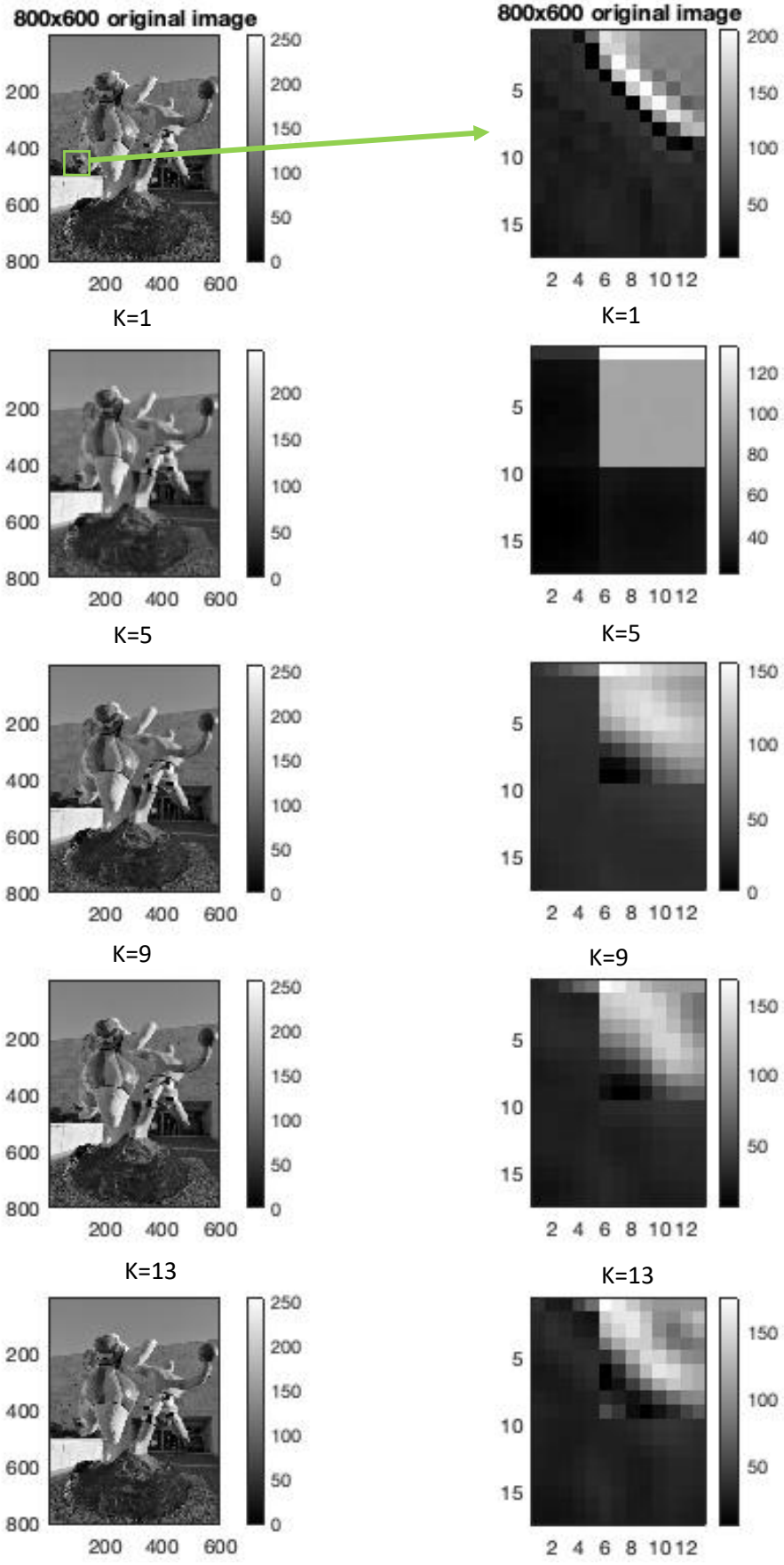

Fig. 3.6: Comparison between compressed images and the original image as a function of the number of eigenvalues used to calculated the compressed image. Images of the right column are the same ROI of 16 × 12 pixels from images depicted in the left column.

In conclusion, to retrieve a compressed image, we only need to transmit the average image $\bar{\boldsymbol{I}}$ ($1 \times N$ matrix), the L eigenvalues and the $L \times N$ eigenvectors matrix and the $K \times L$ matrix corresponding to the L last columns of the decorrelated matrix $\boldsymbol{Y}$. From the compressed data, lost information and compression ratios can be calculated as a function of the number of eigenvalues kept for image compression. In Fig. 3.7, we can observe that with $L = 1$, only 12% of the information is lost while the compression rate is about 97%. We can also notice, that the compression rate decreases linearly as a function of L.

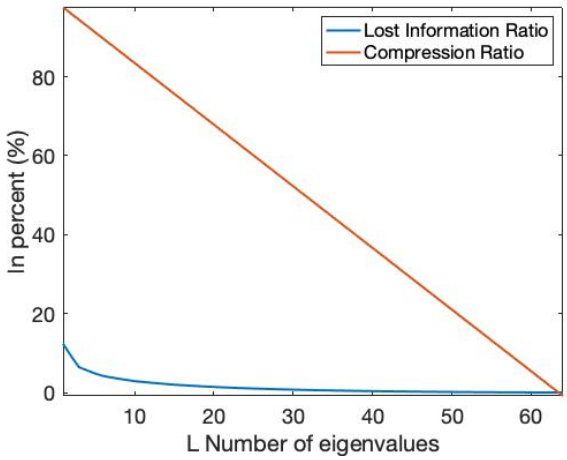

Fig. 3.7: lost information and compression ratio (in percent) as a function of the number of highest eigenvalues used to compress the image.

3.2 PCA of students' marks

Table 3.1 gives the marks of $K = 9$ students in $N = 5$ subjects. This is the matrix $\boldsymbol{M}$ of $K \times N$ joint data. Marks are the variables and the students correspond to the measures.

Students	**Math**	**Sciences**	**French**	**Latin**	**Music**
Jean	6	6	5	5,5	8
Aline	8	8	8	8	9
Annie	6	7	11	9,5	11
Monique	14,5	14,5	15,5	15	8
Didier	14	14	12	12	10
André	11	10	5,5	7	13
Pierre	5,5	7	14	11,5	10
Brigitte	13	12,5	8,5	9,5	12
Evelyne	9	9,5	12,5	12	18

Table 3.1: Students' marks in 5 different subjects

Subject	**Math**	**Sciences**	**French**	**Latin**	**Music**
Math	1,00	0,98	0,23	0,49	0,01
Sciences	0,98	1,00	0,40	0,63	0,01
French	0,23	0,40	1,00	0,96	0,04
Latin	0,49	0,63	0,96	1,00	0,09
Music	0,01	0,01	0,04	0,09	1,00

Table 3.2: Correlation coefficients between the marks of different subjects

From the data correlation coefficients between marks in the different subjects are calculated. In table 3.2 we can notice that marks in $Math$ and $Science$ are strongly correlated as well as $French$ and $Latin$. Marks in $Latin$ and $Science$ are partially correlated while marks in $Music$ are not correlated with any other subject. Eigenvalues of the covariance matrix of data deduced with eig Matlab function are respectively in ascending order:

Students	**Math**	**Sciences**	**French**	**Latin**	**Music**
Jean	6,02	6,02	5,10	5,36	8,01
Aline	7,95	8,08	8,03	7,94	9,00
Annie	6,04	6,91	10,92	9,64	10,99
Monique	14,40	14,67	15,62	14,80	8,01
Didier	14,11	13,82	11,90	12,19	9,99
André	10,91	10,10	5,47	7,02	13,00
Pierre	5,51	6,96	13,93	11,61	9,99
Brigitte	13,07	12,39	8,45	9,59	11,99
Evelyne	8,99	9,56	12,59	11,86	18,01

Table 3.-3: Reconstructed Students' marks from the 3 principal components associated to the 3 highest eigenvalues

$$\{\lambda_n\} = \{0.0048; 0.0400; 9.6901; 13.6756; 31.4089\}.$$

We can observe that only the 3 last eigenvalues are significative. It means that the 5-dimensional space of the original data can be contracted to a 3-dimensional space defined by the 3 principal components associated to the eigenvalues. Fist one will be associated to the highest eigenvalue (λ_N), the second to λ_{N-1} and the third to λ_{N-2}. Consequently, marks can be reconstructed with small uncertaincies from the reduced eigenvectors basis still using Eq. 3.7.

It is also interesting to plot a 3D-diagram where each subject is represented in the 3D space defined by the principal components. Coordinates of each subject are the 3 last coefficients of the eigenvectors matrix $\boldsymbol{V}$.
Figure 3.8 shows this 3D diagram. We can observe that $Math$ and $Science$ are very close as well as $French$ and $Latin$ and $Music$ is almost orthogonal with the other subjects. These results are consistent with the values of the correlation coefficients given in table 3.2.

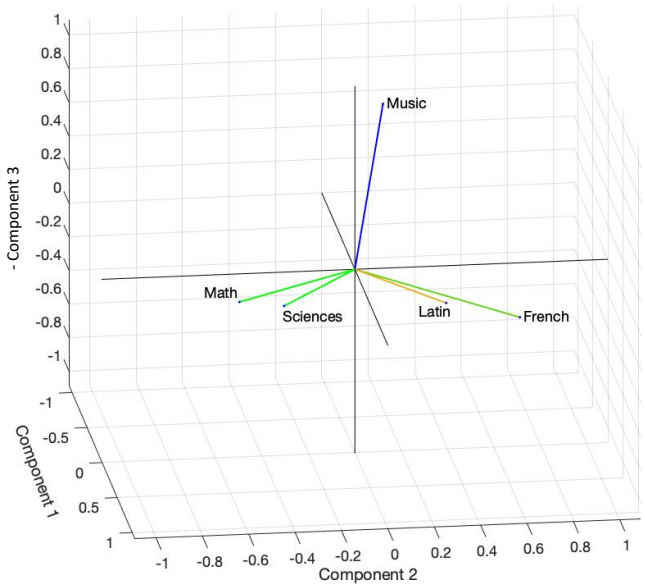

Fig. 3.8: Representation of the location of the different subjects in the 3D space corresponding to the 3 principal components of the data.

The decomposition of student's marks in this basis gives the following information:

- Positive value of component 1 means that student is good (or bad for negative value of the component 1). However, component 1 does not give information about the music skills of student.
- Positive value of component 2 means students are better at literary than scientific subjects (the contrary for negative value of component 2).
- Positive value of component 3 only means that student has good music skills.

To represent the data in the same basis, we have to consider the 3 last columms of the 9×5 matrix $\boldsymbol{Y}$ that contains the coordinates of each student. To simplify the analysis of this plot we select a 2D view along the components 1 and 2 plane (Fig. 3.9). This diagram is in good agreement with the marks given in table 3.1 : we can notice that $Monique$ has a strong component 1 and a small component 2. It means that she is a good student in all subject except in $Music$ where she has a negative value along the component 3 (not shown here).
$Jean$ is a weak student in all subjects because its component 1 has a large negative value.
For all other students try to estimate the students' skills and weaknesses and compare your analysis with table 3.1.

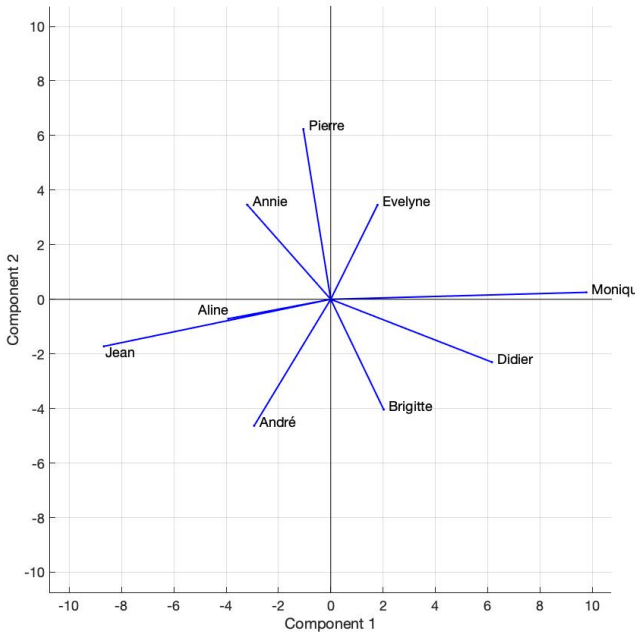

Fig. 3.9 : 2D view in the components 1 and 2 of the students' coordinates in the 3D contracted space.

Chapter 4. Hypothesis testing

1. Introduction

1.1 Definition

Hypothesis testing means taking a decision with respect to a given hypothesis. Example: are data **compatible** with $\theta = \theta_0$? If yes, we **cannot reject** the hypothesis $\theta = \theta_0$. However, we have **not proved** $\theta = \theta_0$, in any case.

A small sub-example to understand: are the measurements $\bar{d} = 0.5$, $\frac{\hat{\sigma}}{\sqrt{N}} = 1$ compatible with $\theta = 0$?

This is a question in the experimenter world and the answer is given by the following reasoning in the model world: $if\ \theta = 0,\ P(-2 \leq \bar{d} \leq 2) \sim 95\%$. Hence, since $\bar{d} = 0.5$ is comprised in the confidence interval, we cannot reject $\theta = 0$. Two important remarks should be immediately added:

- First, many other values of θ are compatible with the data, which is an evident demonstration that we have not proved our hypothesis. But we can reject safely $\theta = 4$. See however the second remark just below.
- Second, the result is given in the experimenter world, while the reasoning is done in the model world. This is not logically consistent, and actually there is always an implicit assumption on the a priori distribution of θ. Most often, this implicit assumption is a constant prior, which is more or less correct for a small Gaussian error. In other cases where the $a\ priori$ distribution is not constant, hypothesis testing leads to absurdities. For example, let us come back to our illness test where 3% of the population is ill (see Chapter 1). A positive test is **not** compatible, at a confidence of 90%, with a health person.
  This is troubling because $P(health|positive) = 0.77$. This example shows clearly that, when choosing between two alternative hypotheses, hypothesis testing implicitly assumes an equal $a\ priori$ probability between these hypotheses. **Do not** use hypothesis testing if such assumption is false.

1.2 Types of tests

The most common tests can be divided in three categories:

1. Are data compatible with a value (mean, standard deviation)?
2. Are data compatible with a probability distribution (Gaussian, for example)?
3. Could two samples come from the same population?

In the third category, we have used two definitions.

- *Population*: ensemble of objects statistically equivalent with respect to a quantitative criterion; only chance gives different values to different objects.
- *Sample:* randomly chosen subset of a population.

The next paragraphs describe two important examples of tests in the two first categories. The third type of test will not be treated in this course.

2. Statistical control of quality

2.1 Reception quality control

A customer receives a bundle of N pieces and has defined tolerance limits on the mean (not on the individual pieces!). He wants to be sure, at a given risk of $2{,}5\%$, that the mean is actually within the tolerance limits. A lot of pieces fabricated with a mean outside the tolerance limits must be rejected with a probability $P > 97.5\%$.

To be specific, let us consider that θ must be greater than 99 or smaller than 101, (customer or H_1 hypothesis) in some unity. If $\theta \leq 99$ or $\theta \geq 101$, the lot must be rejected with $P > 97.5\%$.

On the other hand, the supplier pretends $\theta = 100$ If this (null) hypothesis is true, the lot must be accepted with $P > 95\%$. $1 - P$ is the first kind risk.

Note that the customer is much less demanding than the supplier.
Because of the statistical fluctuations, this is compulsory. We are looking now for a rule allowing the customer's and supplier's demands to be both fulfilled. We assume a Gaussian fabrication and draw the following graph (Fig. 4.1). The black curve corresponds to the pdf of the dimension of the pieces for H_0, $\theta = 1$ (supplier hypothesis). The red curves correspond to the maximum risk of acceptation with θ outside the tolerances for the customer (i.e., $\theta = 99$ or $\theta = 101$). We want to define a decision rule: the lot is accepted between two limits.

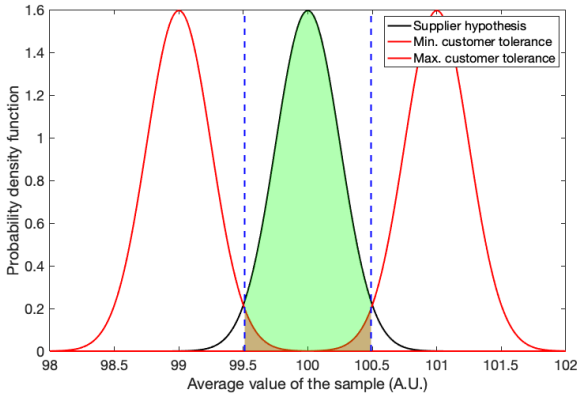

Fig. 4.1 : Black curve corresponds to the p.d.f. of the dimension of the pieces for H_0 : θ=100 (supplier hypothesis). Red curves correspond to the maximum risk of acceptation with θ outside the tolerances for the customer (i.e. θ=99 or θ=101).

We immediately see that accepting the lot between the blue limits allows both the supplier and the customer risks to lie at the above given limits. All the probability densities curves are Gaussian with a standard deviation of $0.25 = (101 - 100)/4$. This only value ensures the right risks for both the supplier and the customer. The customer's risk is, for example, the integral of the left red curve in the acceptation region i.e., at the right of the low acceptation threshold, 99.5. The supplier's risk is the integral of the black curve outside the acceptation region, on both sides.

In practice, it means that we have to take a lot of N pieces such that $\frac{\hat{\sigma}}{\sqrt{N}} = 0.25$.

2.2 Statistical mastering of production

While of great historical importance, the method developed in the preceding paragraph has some important drawbacks. In particular, a drift of the true mean could lead to the rejection of the pieces, with painful consequences for both the supplier and the customer. In this paragraph, we are looking for rules that allow the machine to be adjusted before any risk of production rejection. However, another important requirement is the necessity of avoiding no necessary adjustments: an adjustment has a cost.

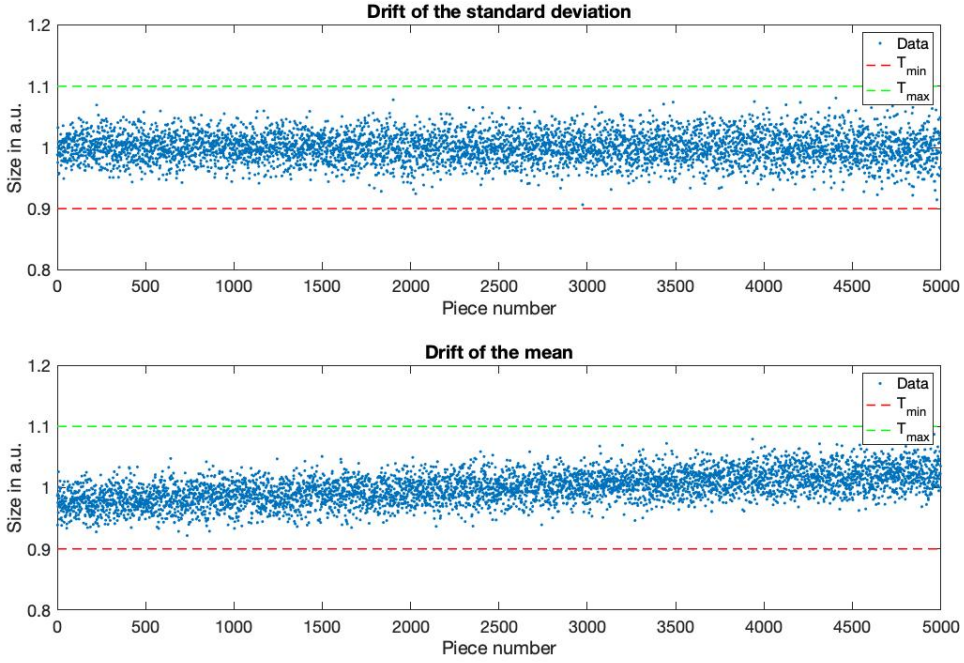

Fig. 4.2: Example of mastering production when drifts of the mean or of the standard deviation of the size of a piece is observed

The proposed approach consists in two separate steps:

- Step 1: is the production capable, i.e., can fulfill the customer tolerances on the individual pieces?
- Step 2: is the fabrication stable, without drift of either the mean or the standard deviation?

This second step **does not use tolerances**: even if largely compatible with a production within the tolerances, a drift must be evidenced. The graphs in figure 4.2 show two examples of drift, concerning the mean and the standard deviation. In both cases, no size value is outside the tolerances. However, the drift is significant and must be evidenced in order to decide corrective actions before any risk.

In more details, the step 1 is fulfilled if the capability verifies:

$$C_{pk} = Inf\left(\frac{\bar{d} - T_{min}}{3\sigma}, \frac{T_{max} - \bar{d}}{3\sigma}\right) \geq 1.33 \qquad [4.1],$$

where T_{min}, T_{max} are the low and high tolerance limit (horizontal dotted lines in Fig. 4.1). Note that σ is the standard deviation of the pieces' size (not of the arithmetic average as above).
Once we know that the fabrication is capable, we pass to step 2 by computing the arithmetic average and estimated standard deviation on successive samples. The fabrication is stable if the variation of these values can be attributed to randomness: tables allow the hypothesis of stability to be rejected with a given confidence level (of course, the stability can never be proved: see the introduction of the chapter). The use of these tables is left to exercises.

3. Fit test: χ^2 test

3.1 Definition of a χ^2_k distribution law of k degrees of freedom

In probability theory and statistics, the **chi-squared distribution** (or χ^2-distribution) with *k* degrees of freedom is the distribution of a sum of the squares of *k* independent standard normal random variables. The probability density of the distribution law is:

$$p_k(x) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} \qquad [4.2],$$

where k is the degree of freedom and Γ is the $gamma$ function. With Matlab such law is obtain using pdf function. Figure 4.3 shows the χ^2_k distribution law for different degree of freedom.

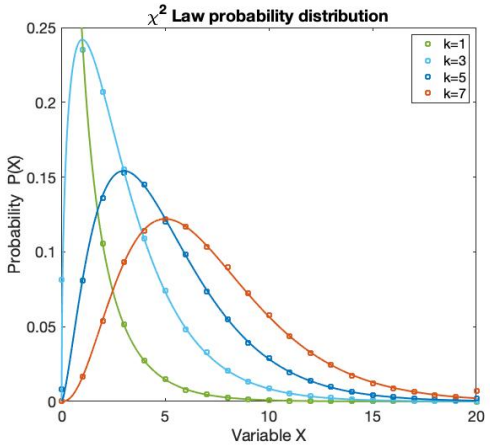

Fig. 4.3: χ^2_k distribution law for different degree of freedom k. Squares correspond to data given by the $chi2rnd$ function of Matlab and curves corresponds to the plots of Eq. 4.2.

3.2 χ^2 test

In this course, the χ^2 test will be introduced as an important example of the category of hypothesis test:

Is the data behavior compatible with a specific, probability law, here the Gaussian law?

The idea is to compare the measured frequency to the expected probability. More precisely, we define M classes C_j with bounds $[k_j, k_{j+1}]$. A measurement $d_i \in C_j \iff k_j \leq d_i < k_{j+1}$.
We experimentally find n_j elements in the class C_j. The limits of classes are defined such that $n_j > 6$. The total number of data is $N = \sum_{j=1}^{M} n_j$. Moreover, we calculate the arithmetic average $\bar{d}$ and the

estimated standard deviation $\hat{\sigma}$ (see chapter 2). We now compare to a Gaussian distribution. If the data are Gaussian, we can find the probability of inclusion in the class C_j:

$$P(C_j) = P(k_j \leq d_i < k_{j+1}) = P(d_i < k_{j+1}) - P(k_j < d_i) \quad [4.2]$$

The probabilities in the last equality are directly given by the cumulative density function (cdf function with Matlab) of a Gaussian of mean $\bar{d}$ and standard deviation $\hat{\sigma}$.
We then compute the distance D^2 between the measured frequencies and the expected probabilities as follows:

$$D^2 = \sum_{j=1}^{M} \frac{(n_j - NP_j)^2}{NP_j} \quad [4.3]$$

If the data follow a Gaussian distribution, D^2 follows a χ^2 distribution of $M-1$ degrees of freedom. The corresponding confidence interval can be determined from tables or with a computer routine ($icdf$ function with Matlab). We can either conclude that the data are compatible with the Gaussian hypothesis, if D^2 is included in the interval, or that the data are not compatible with this Gaussian hypothesis.

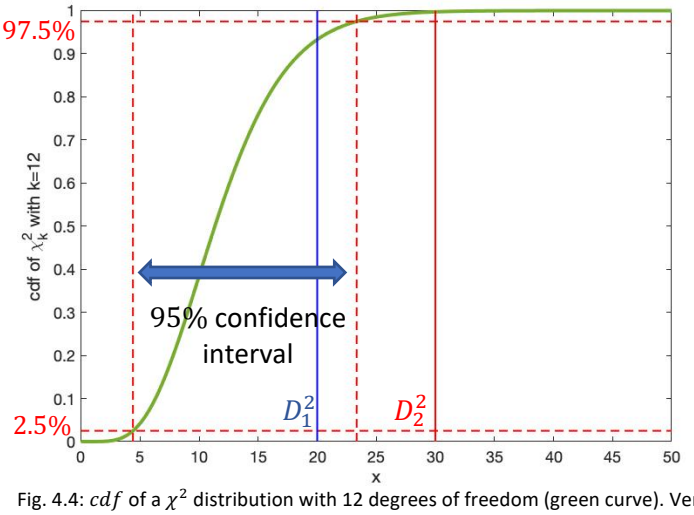

Fig. 4.4: cdf of a χ^2 distribution with 12 degrees of freedom (green curve). Vertical dotted lines give the limits of the 95% confidence interval. Using eq. 4.3 two cases are represented: D_1^2=20 is included in the confidence interval, then the Gaussian hypothesis cannot be rejected. D_2^2=30 is outside the confidence interval, then the Gaussian hypothesis is rejected.

Figure 4.4 shows an example where the green curve corresponds to the cdf of a χ^2 distribution with 12 degrees of freedom. The vertical dotted black lines give the limit of the 95% confidence interval. We assume 2 set of data for which distances D^2 are calculated using Eq. 4.3. For the first set of data, $D_1^2 = 20$ is included in the confidence interval. Consequently, it means that the Gaussian hypothesis cannot be rejected. For the second one, $D_2^2 = 30$ is outside the confidence interval meaning that Gaussian hypothesis is rejected.
Note however that a value of D^2 smaller than the inferior limit of the interval means that the data follow perfectly a Gaussian distribution, without the fluctuations we were expected. It could seem a bit strange to conclude that the data are not issued from a Gaussian distribution. It remains that such a distribution has a low probability to occur if the data are random. Are they? is in this case the good question...

Chapter 5. Inverse problem and least-squares fitting

1. General frame

Assume data (measurements) $D = \{\cdots, d_i, \cdots\}$ at abscissae $\{\cdots, x_i, \cdots\}$ modeled by a model M using K parameters $P = (p_1, \ldots, p_K)$ such as:

$$d_i = M(x_i; p_1, \ldots, p_K) \qquad [5.1]$$

Let us define, as in chapter 1:

- the direct problem: model the data D (find the model M) for a given value of parameters P.
- the inverse problem: retrieve parameters P from data D, knowing the model M and any other a priori information.

In some case, the process can be iterative, i.e. the model can be modified because of the measurements.

Example: find the resistance R of an electric dipole with a current-voltage characteristic model $U_i = R\, I_i$ from voltage and current intensity data $\{\cdots, U_i, \cdots\}$ and $\{\cdots, I_i, \cdots\}$. Figure 5.1 shows the plot $U_i = f(I_i)$ of data. Clearly, the linear model is not correct and the direct problem must be reconsidered.

Warning: *all tests of linearity will show that the data on the above graph are compatible with a linear assumption, while a look on the graph shows they are not linear.*

Make a graph first!

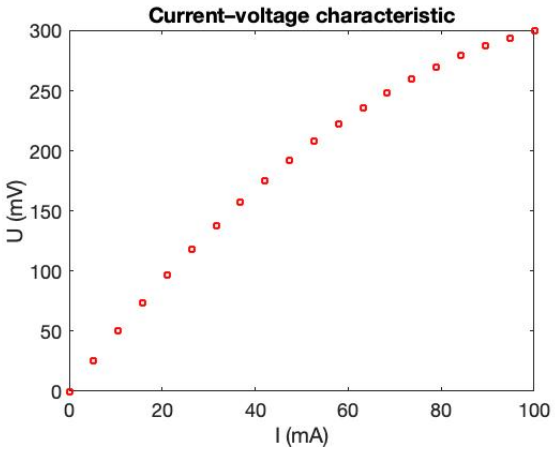

Fig. 5.1: plot of current-voltage characteristic of an electric dipole.

2. Why using least squares?

Using least squares is justified if the errors are Gaussian, for example for a measurement under control (see chapter 2 paragraph 5). If the errors do not obey a Gaussian statistic, least squares fitting can give poor results. In particular, least- squares fitting is extremely sensitive to a (some) outlier(s). If outliers are suspected, use equivalents of the median, rather than least-squares estimators. The arithmetic average is a least-squares estimator and is sensitive to outliers (see exercise in the paragraph 3 of Chapter 2).

In more details, measurements $D = \{\cdots, d_i, \cdots\}$ are performed at abscissae $\{\cdots, x_i, \cdots\}$. Let be a model $M_K(x_i)$ depending on parameters $P = (p_1, \ldots, p_K)$. We assume:

$$d_i = M_K(x_i) + n_i \quad [5.2],$$

where n_i is a centered Gaussian additive noise, independent from a measurement to another, of variance σ_i^2 which can depend on x_i.

Hence, in the model world we have an additive noise modelled by the Gaussian probability law:

$$p\left(d_i \middle| M_K(x_i)\right) = \frac{1}{\sigma_i^2\sqrt{2\pi}} e^{-\frac{\left(d_i - M_K(x_i)\right)^2}{2\sigma_i^2}}$$

$$\Rightarrow p(D|M_K) = \prod p\left(d_i \middle| M_K(x_i)\right) \propto e^{-\sum\frac{\left(d_i - M_K(x_i)\right)^2}{2\sigma_i^2}} \quad [5.3].$$

which means that $p(D|M_K)$ is maximum if $\sum \frac{\left(d_i - M_K(x_i)\right)^2}{2\sigma_i^2}$ is minimum. If $p_{prior}(M_K)$ is constant around the estimated parameters, this is also the maximum of $p(M_K|D)$. This is a reasonable hypothesis for a measurement under control.

3. Linear least squares with two parameters a and b : $y_i = a\,x_i + b$

We suppose that there is no uncertainty on x (this is a troubling hypothesis...) and that the uncertainty on the ordinates does not depend on x: $\sigma_i^2 = \sigma^2$. Least squares solution is obtained by minimizing the quadratic error $\varepsilon^2 = \sum\left(y_i - (ax_i + b)\right)^2$. Consequently, the problem is to find a and b coefficients of the linear model such as $\frac{\partial \varepsilon^2}{\partial a} = 0$ and $\frac{\partial \varepsilon^2}{\partial b} = 0$. It leads to the solutions:

$$\begin{cases} a = \dfrac{\sum_i x_i(\bar{y} - y_i)}{\sum_i x_i(\bar{x} - x_i)} \\ b = \bar{y} - a\bar{x} \end{cases} \quad [5.4].$$

The uncertainty on the coefficients is given by:

$$\begin{cases} \sigma_a^2 = \dfrac{\hat{\sigma}^2}{\sum_i (\bar{x} - x_i)^2} = a^2\left(\dfrac{\frac{1}{\rho^2} - 1}{N - 2}\right) \\ \sigma_b^2 = \hat{\sigma}^2\left(\dfrac{1}{N} + \dfrac{\bar{x}^2}{\sum_i (\bar{x} - x_i)^2}\right) \end{cases} \quad [5.5],$$

where $\hat{\sigma}^2$ is estimated by: $\hat{\sigma}^2 = \frac{1}{N-2}\sum_i\left(y_i - (ax_i + b)\right)^2$ and ρ is the correlation coefficient, given by all calculators and computer routines. In these expressions, the division by $N - 2$ can be understood

by considering the case $N = 2$: there is no deviation between the data and the fitted straight line, and no possibility of assessing the uncertainty.
The confidence interval of the fitted points is obtained by considering the new ordinates y'_i such as:

$$y'_i = ax_i + b = a(x_i - \bar{x}) + \bar{y} \qquad [5.6],$$

with variance $\sigma^2_{y'_i} = \hat{\sigma}^2 \left(\frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum_i (\bar{x} - x_i)^2}\right)$. Note that this uncertainty is lower than the uncertainty on the data, because of some average, at best near $\bar{x}$. It is easy to use y' for interpolation between data. On the other hand, extrapolation outside the data range is almost always catastrophic. **Do not** extrapolate, unless you are a meteorologist (a politician, a journalist...), with considerable means and average results.

4. Linear least squares: matrix formulation

Let be N equations with K parameters, $N > K$:

$$y_i = \theta_1 f_1(x_i) + \cdots + \theta_K f_K(x_i) \qquad [5.7].$$

This system of N equations with K unknowns can be written in a matrix form as $\boldsymbol{Y} = \boldsymbol{J\theta}$, with:

$$\boldsymbol{Y} = \begin{bmatrix} \vdots \\ y_i \\ \vdots \end{bmatrix} N\ lines; \boldsymbol{J} = \begin{bmatrix} \ddots & \vdots & \iddots \\ \cdots & f_j(x_i) & \cdots \\ \iddots & \vdots & \ddots \end{bmatrix} N \times K\ matrix; \boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_K \end{bmatrix} K\ columns \qquad [5.8].$$

It can be shown that the linear least squares method consists in minimizing the quadratic error weighted by the variance of the measures:

$$\varepsilon^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left(\sum_{k=1}^{K} \theta_k f_k(x_i) - y_i \right)^2 \qquad [5.9].$$

Minimizing Eq. 5.9 consist in finding the solution of the K equations $\frac{\partial \varepsilon^2}{\partial \theta_k} = 0$. It can be shown that solution $\boldsymbol{\theta}_{min}$ can be written with in matrix format as:

$$\boldsymbol{\theta}_{min} = \left(\boldsymbol{J^T W J}\right)^{-1} \boldsymbol{J^T W Y} \qquad [5.10],$$

where $\boldsymbol{W} = \begin{bmatrix} \frac{1}{\sigma_1^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sigma_N^2} \end{bmatrix}$ is the diagonal weighting matrix. With no ponderation $\boldsymbol{W}$ is the identity matrix.

Figure 5.2 shows an example of fitted data obeying to the linear function $f(x) = x + 2$. Red square corresponds to the "experimental" data where a Gaussian noise (with increasing standard deviation with x_i) is added.
For the least squares method, three cases are considered.

- First one considers a linear model where data a weighted by the variance. The model deduced from the linear least squares is $M_1(x) = (1.00 \pm 0.06)x + (2.1 \pm 0.1)$ and the quadratic error is $\varepsilon_1^2(\boldsymbol{\theta}) = 0.02$.
- The second one is also a linear model but the data are not weighted by the variance. In that case the model is $M_2(x) = (1.00 \pm 0.08)x + (2.0 \pm 0.3)$ and the quadratic error is $\varepsilon_2^2(\boldsymbol{\theta}) = 0.06$.

- The last one considers a quadratic model with weighted data.
  The model is: $M_3(x) = (0.00 \pm 0.02)x^2 + (1.0 \pm 0.1)x + (2.0 \pm 0.1)$ and the quadratic error is $\varepsilon_3^2(\boldsymbol{\theta}) = 0.03$.

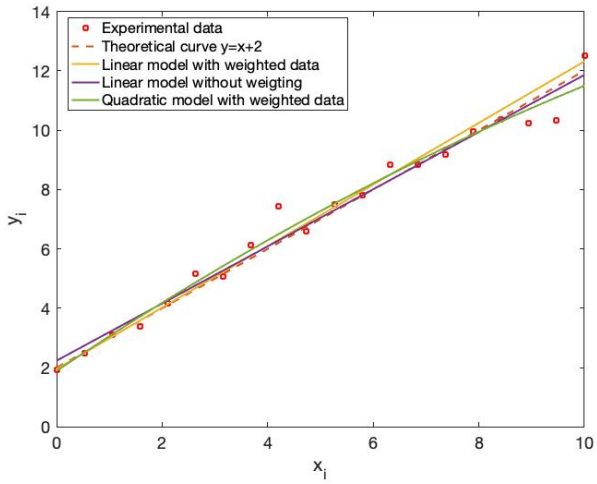

Fig. 5.2: fitting example of data obeying to a linear function

A particular case is polynomial least squares $f_1(x) = 1, f_2(x) = x\,, f_3(x) = x^2$, and so on... The resulting matrix is ill conditioned: very close curves can be issued from a very different set of coefficients. This is not a problem if we are interested by the fitted curve, but it results in great uncertainties on $\boldsymbol{\theta}_{min}$**.** Do not use polynomial least squares to retrieve the coefficients of the polynomial. Orthogonal polynomials should be rather used.

5. Linearized least squares

As an example, let us retrieve a and τ in a model $y = a\, e^{-t/\tau}$. This nonlinear model can be linearized as $ln(y) = ln(a) - t/\tau$, with parameters $ln(a)$ and $1/\tau$ given by linear regression.
In practice, the results are not so good: after linearization, the variance becomes no constant. Even weighting by the inverse of the standard deviation (proportional to the differential $1/y$) does not give good results, because the hypothesis underlying differentiation, i.e. small relative noise, is not fulfilled at low y.

6. Nonlinear least squares: the Gauss-Newton method

The nonlinear method of fitting given in the paragraph below is one of the many possibilities. Even if not used to retrieve $\boldsymbol{\theta}_{min}$, the formalism employed here allows the uncertainties on the parameters to be assessed, which is of paramount importance.

6.1 Newton method

Before to expose the least squares method, let us recall the Newton method, to retrieve a unique parameter θ for a unique abscissa x, with a nonlinear model M*:*

Find θ such that $M(x, \theta) = y$ where M is a nonlinear function of θ.

If M is not "too" nonlinear, the solution of $f(\theta) = M(x, \theta) - y = 0$ can be found by using the Newton or dichotomy algorithms (methods to find the zero of a function, $fzero$ function with Matlab).

6.2 The Gauss-Newton method

It is the combination of the Newton and the least squares (Gauss) method. Let be N nonlinear equations with K parameters, $N > K$: $y_i = M_K(x_i; \boldsymbol{\theta})$. The objective is to determine the K parameters of the vector $\boldsymbol{\theta}$ by minimizing in an iterative manner the quadratic error given by Eq. 5.9.

Let be $\boldsymbol{\theta}_l = \begin{bmatrix} \vdots \\ \theta_k \\ \vdots \end{bmatrix}$ the parameter vector obtained after l^{th} iteration. The following matrix system will be solved in the least squares sense:

$$\boldsymbol{M_K}(\boldsymbol{X};\boldsymbol{\theta}_l) - \boldsymbol{Y} = \boldsymbol{M'_K}(\boldsymbol{X};\boldsymbol{\theta}_l)[\boldsymbol{\theta}_l - \boldsymbol{\theta}_{l+1}] \qquad [5.11],$$

with $\boldsymbol{M'_K}(\boldsymbol{X};\boldsymbol{\theta}_l) = \begin{bmatrix} \ddots & \vdots & \iddots \\ \cdots & \frac{\partial M_K(x_i;\boldsymbol{\theta}_l)}{\partial \theta_k} & \cdots \\ \iddots & \vdots & \ddots \end{bmatrix}$ is a $N \times K$ matrix. Starting with a parameter vector $\boldsymbol{\theta}_0$ deduced from data and with Eq. 5.11, parameter vector $\boldsymbol{\theta}_{l+1}$ at $l+1$ iteration is obtained calculating:

$$\boldsymbol{\theta}_{l+1} = \boldsymbol{\theta}_l - \left(\boldsymbol{M'_K}(\boldsymbol{X};\boldsymbol{\theta}_l)^{\boldsymbol{T}}\boldsymbol{M'_K}(\boldsymbol{X};\boldsymbol{\theta}_l)\right)^{-1}\boldsymbol{M'_K}(\boldsymbol{X};\boldsymbol{\theta}_l)^{\boldsymbol{T}}(\boldsymbol{M_K}(\boldsymbol{X};\boldsymbol{\theta}_l) - \boldsymbol{Y}) \qquad [5.12].$$

Iterative process is stopped when a given tolerance is reached. The convergence is obtained if the model is not too nonlinear and the starting parameters not too different of the solution.

6.3 Uncertainties on the parameters

Whatever the numerical method used to calculate the optimized parameters of the model $\boldsymbol{\theta}_{opt}$, you must compute $\boldsymbol{M'_K}(\boldsymbol{X};\boldsymbol{\theta}_{opt})$ in order to evaluate the uncertainty on $\boldsymbol{\theta}_{num}$, given by the $K \times K$ covariance matrix:

$$\boldsymbol{C_\theta} = \left(\boldsymbol{M'_K}(\boldsymbol{X};\boldsymbol{\theta}_{opt})^{\boldsymbol{T}}\boldsymbol{C_Y}\boldsymbol{M'_K}(\boldsymbol{X};\boldsymbol{\theta}_{opt})\right)^{-1} \qquad [5.13],$$

where $\boldsymbol{C_Y}$ is the $N \times N$ covariance matrix of data. If the data are independent each other, $\boldsymbol{C_Y}$ is a diagonal matrix whose elements are the variances of the measurements y_i. Even small uncertainties on the data and a good agreement between the data and the model do not guarantee small uncertainties on the parameters. Indeed, very different sets of parameters can lead to (almost) the same data if:

1. A parameter has little influence on the data.
2. Less evidently, it exists a strong correlation between parameters. An extreme example is given by the model $y = ab\,x$. Clearly, only the product ab, can be retrieved and not the individual parameters a and b. Consider now a model $y = a\,sin(bx)$. a and b are not correlated, ... except if $|bx| << 1$ whatever x... Fortunately, the correlation matrix of the parameters, obtained from the covariance matrix $\boldsymbol{C_\theta}$, (see chapter 3 to pass from covariances to correlations), indicates such links. If two parameters have a strong (anti) correlation, it is compulsory to remove one of the two. In the model $y = ab\,x$, the correlation coefficient between a and b is equal to -1: an increase of a can be exactly compensated by a decrease of b.

6.3 Example: fitting of Gaussian distribution of data

We consider a set of $N = 20$ data $Y = \{\cdots, y_i, \cdots\}$ measured at abscissa equidistant $X = \{\cdots, x_i, \cdots\}$ with Δx as sampling step.

Hypothesis : data obeys to a Gaussian law: $y = f(x) = Ae^{-\frac{(x-x_0)^2}{2\sigma^2}}$ where the amplitude A, the origin offset x_0 and the standard deviation σ are the parameters to determine : $\theta = \begin{bmatrix} A \\ x_0 \\ \sigma \end{bmatrix}$.

First, plot of data allows to verify if the initial hypothesis is valid or not. Then, from data the initial values of the parameters have to be estimated. For a Gaussian function the starting values can be estimated as follows[‡‡]:

$$\boldsymbol{\theta_0} = \begin{bmatrix} \max(Y) \\ X(Y == \max(Y)) \\ length(\frac{Y}{\max(Y)} > 0.5) \times \Delta x/2 \end{bmatrix} \quad [5.14].$$

From the model, the matrices $\boldsymbol{M_K}(\boldsymbol{X};\boldsymbol{\theta_0})$ and $\boldsymbol{M'_K}(\boldsymbol{X};\boldsymbol{\theta_0})$ are calculated and the iterative process is done using Eq. 5.12.

Figure 5.3 shows a plot of the data and the curve obtained with the initial value of the parameters and the curve obtained with the optimized parameters. Here, the convergence is obtained after 4 iterations.

The theoretical parameters are $\boldsymbol{\theta_{th}} = \begin{bmatrix} 1 \\ 0.1 \\ 0.2 \end{bmatrix}$.

The initial parameters are $\boldsymbol{\theta_0} = \begin{bmatrix} 0.88 \\ 0.16 \\ 0.26 \end{bmatrix}$.

The optimized parameters are $\boldsymbol{\theta_{opt}} = \begin{bmatrix} 0.89 \\ 0.09 \\ 0.23 \end{bmatrix}$.

The covariance matrix is:

$$\boldsymbol{C_\theta} = \begin{bmatrix} 0.0040 & 0.0000 & -0.0007 \\ 0.0000 & 0.0004 & -0.0000 \\ -0.0007 & -0.0000 & 0.0004 \end{bmatrix}.$$

From the diagonal terms of the covariance matrix we can estimate the 95% confidence interval for optimized parameters:

$$\begin{cases} \hat{A} = 0.9 \pm 0.1 \\ \hat{x}_0 = 0.09 \pm 0.04 \\ \hat{\sigma} = 0.23 \pm 0.04 \end{cases} \quad [5.15].$$

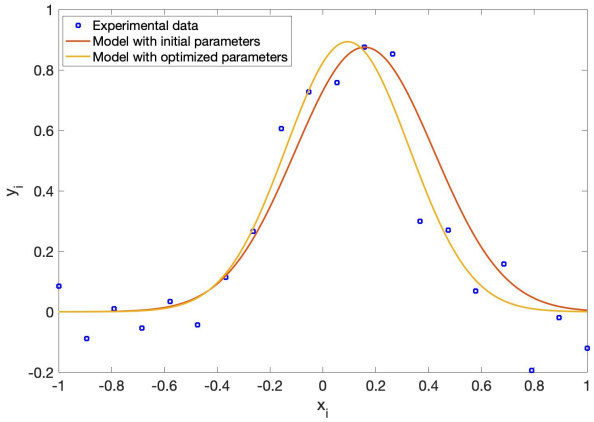

Fig. 5.3: Fitting of data with Gaussian law model with 3 parameters to optimize (amplitude, offset, standard deviation).

From covariance matrix we can also observe that other coefficients of the covariance matrix are very small proving that parameters of the model are uncorrelated.

[‡‡] The syntax of Matlab software is used.

Annex

A.1 Arithmetic mean (or arithmetic average)

The arithmetic mean is the sum of a collection of numbers divided by the count of numbers in the collection. The collection is often a set of results of an experiment or an observational study, or frequently a set of results from a survey. The term "arithmetic mean" is preferred in some contexts in mathematics and statistics, because it helps distinguish it from other means, such as the geometric mean and the harmonic mean.

Given a data set $X = \{x_1, x_2, x_3, \cdots x_n\}$, the arithmetic mean (or mean or average), denoted $\bar{x}$, is the mean of the n values $x_1, x_2, x_3, \cdots x_n$. $\bar{x}$ is defined by the formula:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad [1]$$

A.2 Expected value

In probability theory, the expected value of a random variable X, often denoted $E(X)$, is a generalization of the weighted average, and is intuitively the arithmetic mean of a large number of independent realizations of X. In the finite case, let X be a (discrete) random variable with a finite number of finite outcomes $x_1, x_2, x_3, \cdots x_n$ occurring with probabilities $p_1, p_2, p_3, \cdots p_n$ respectively. The expectation of X is defined as:

$$E(X) = p_1.x_1 + p_2.x_2 + p_3.x_3 + \cdots + p_n.x_n = \sum_{i=1}^{n} p_i.x_i \qquad [2]$$

Since $p_1 + p_2 + p_3 + \cdots + p_n = 1$, the expected value is the weighted sum of the x_i values, with the probabilities p_i as the weights.

If all outcomes x_i are equiprobable ($p_1 = p_2 = \cdots p_n$), then the weighted average turns into the simple average. On the other hand, if the outcomes x_i are not equiprobable, then the simple average must be replaced with the weighted average, which takes into account the fact that some outcomes are more likely than others.

If X is a random variable with a probability density function of $f(x)$, then the expected value is defined as the Lebesgue integral:

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx \qquad [3]$$

where the values on both sides are well defined or not well defined simultaneously.

A.3 Standard deviation

In statistics, the standard deviation is a measure of the amount of variation or dispersion of a set of values X. A low standard deviation indicates that the values tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the values are spread out over a wider range.

Let μ be the expected value (the average) of random variable X with density $f(x)$:

$$\mu = E(X) = \int_{-\infty}^{+\infty} x f(x) dx$$

The standard deviation σ of X is defined as:

$$\sigma = \sqrt{E[(X-\mu)^2]} = \sqrt{E[X^2] - (E[X])^2} \qquad [4]$$

In the case where X takes random values from a finite data set $x_1, x_2, x_3, \cdots x_n$ with each value having the same probability, the standard deviation is:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \qquad [5]$$

where $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$.

If, instead of having equal probabilities, the values $x_1, x_2, x_3, \cdots x_n$ have different probabilities $p_1, p_2, p_3, \cdots p_n$. In this case, the standard deviation will be:

$$\sigma = \sqrt{\sum_{i=1}^{n} p_i (x_i - \mu)^2} \qquad [6]$$

where $\mu = \sum_{i=1}^{n} p_i x_i$.

The standard deviation of a continuous real-valued random variable X with probability density function $f(x)$ is:

$$\sigma = \sqrt{\int_{\mathbb{X}} (x-\mu)^2 f(x) dx} \qquad [7]$$

where $\mu = \int_{\mathbb{X}} xf(x)dx$.

A.4 sample standard deviation

One can find the standard deviation of an entire population in cases (such as standardized testing) where every member of a population is sampled. In cases where that cannot be done, the standard deviation σ is estimated by examining a random sample taken from the population and computing a statistic of the sample, which is used as an estimate of the population standard deviation. Such a statistic is called an estimator, and the estimator (or the value of the estimator, namely the estimate) is called a sample standard deviation, and is denoted by s (possibly with modifiers).

Unlike in the case of estimating the population mean, for which the sample mean is a simple estimator with many desirable properties (unbiased, efficient, maximum likelihood), there is no single estimator for the standard deviation with all these properties, and unbiased estimation of standard deviation is a very technically involved problem. Most often, the standard deviation is estimated using the corrected sample standard deviation (using $N - 1$), defined below, and this is often referred to as the "sample standard deviation", without qualifiers. However, other estimators are better in other respects: the uncorrected estimator (using N) yields lower mean squared error.

Uncorrected sample standard deviation: $s_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Corrected sample standard deviation: $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

A.5 Bayes theorem

In probability theory and statistics, Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event. Bayes' theorem is stated mathematically as the following equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \ , \qquad [8]$$

where A, B are different events with probabilities to be observed with any given conditions $P(A)$ and $P(B)$ respectively. These probabilities are known as the marginal probability or prior probability. Here $P(B) \neq 0$ is assumed.
$P(A|B)$ is a conditional probability: the probability of event A occurring given that B is true. It is also called the posterior probability of A given B.
$P(B|A)$ is a conditional probability: the probability of event B occurring given that A is true. It is also called the posterior probability of B given A.

Bayes' theorem may be derived from the definition of conditional probability if $P(B) \neq 0$:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \ , \qquad [9]$$

where $P(A \cap B)$ is the probability of both A and B being true. Similarly, $P(B|A) = \frac{P(A \cap B)}{P(A)}$ if $P(A) \neq 0$. Solving for $P(A \cap B)$ and substituting into the above expression for $P(A|B)$ yields Bayes' theorem (Eq. 8)