Statistical exploitation of measurements

E. Lantz 2018

Table of contents

Chapter 1. Introduction to inverse problem

1) Direct and inverse problem

Let be "an" unknown quantity.

Some examples: a temperature, a resistance but also the geological composition in 3D of the subsoil, retrieved in order to find oil (more than one data...)

The precise value of this quantity, named θ in the following, will remain unknown, because there is always some uncertainty in the measurement process. The objective of this course can then be defined as**: How to use at best all the available knowledge to give the most precise information on θ** This available knowledge consists in:

- the measurements, N data in the following named **D**= d_1,.....,d_N

- a model of the measurement process. For example, we know in optics how an image is formed with a microscope, if we have determined the impulse response of this microscope.

- any information on θ known before the measurement process, called *a priori* information.

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 information (measurements, model of measurement, a priori information). This probability law is called *a posteriori* (after the measurements).

2) Bayes theorem

Let us resolve a small exercise to understand the above notions.

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.

A person is randomly tested in the population and the test is positive. How likely is the person tested to be sick?

***A priori* information**: $P_{prior}(sick)=3\%$

Model of the measurement: P(positive|sick)= 95%. P(negative| sick)=10%

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).

Answer to the inverse problem: P(sick|positive)?

Of course, P(sick|negative) is also a very important result!

The solution of the inverse problem uses the Bayes theorem, which is recalled in the following.

Bayes's theorem:

In the compound probability formula, A and B can be interchanged, hence:

$P(A \cap B)= P(A|B).P(B)= P(B|A).P(A)$

that you can write:

$P(A|B)= (P(B|A).P(A))/(P(B))$

or, in the context of the inverse problem, by naming **D** the data $d_1,.....,d_N$:

$P(\theta|\mathbf{D})= (P(\mathbf{D}|\theta).P_{prior}(\theta))/(P(\mathbf{D}))$

In our example:

$P(sick|positive)=(P(positive|\ sick).P_{prior}(sick))/(P(positive))$
$=(0.95\ .\ 0.03)/(0.95\ .\ 0.03 + 0.10\ .\ 0.97)=0.23$

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 0.62. 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....).

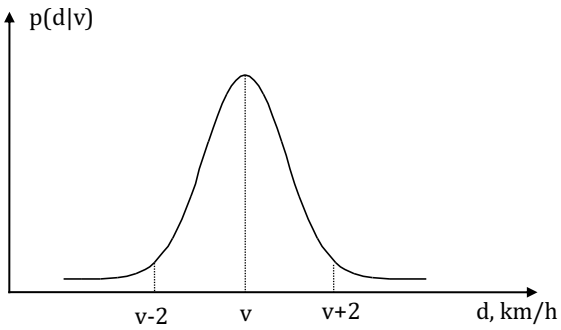
Is the test not useful? It depends... To answer it is worth to calculate P(sick|negative)=0.0017. If you have a cheap innocuous treatment, it could be interesting to treat 12% of the population, including almost all the ills. On the contrary, if you have no treatment or if the treatment has some secondary effects, the test can be only used as a first indication of illness and must be confirmed.

Second example using densities of probabilities

We want to measure the speed v of a cyclist on a climb (towards the summit!) with a counter with an uncertainty ± 2 km/h

Model of the measurement

The probability law (density) p(d|v) is Gaussian or, better, a small percentage of the probability is assigned to outliers (failed counter):



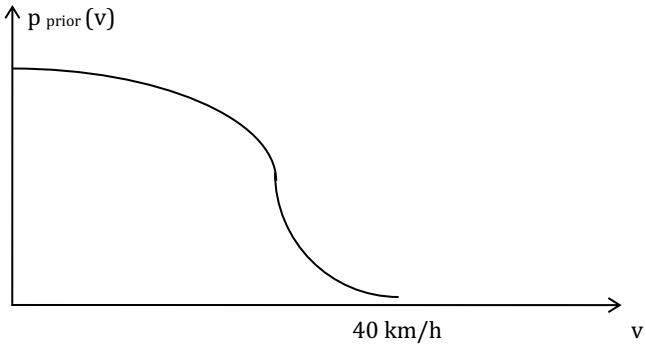
Data

p(v|d) if d_0= 15 km/h (A)?
p(v|d) if d_0= = 80 km/h (B)?

Of course, in (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 following graph reflects what we know about the cyclist's speed before its measurement.



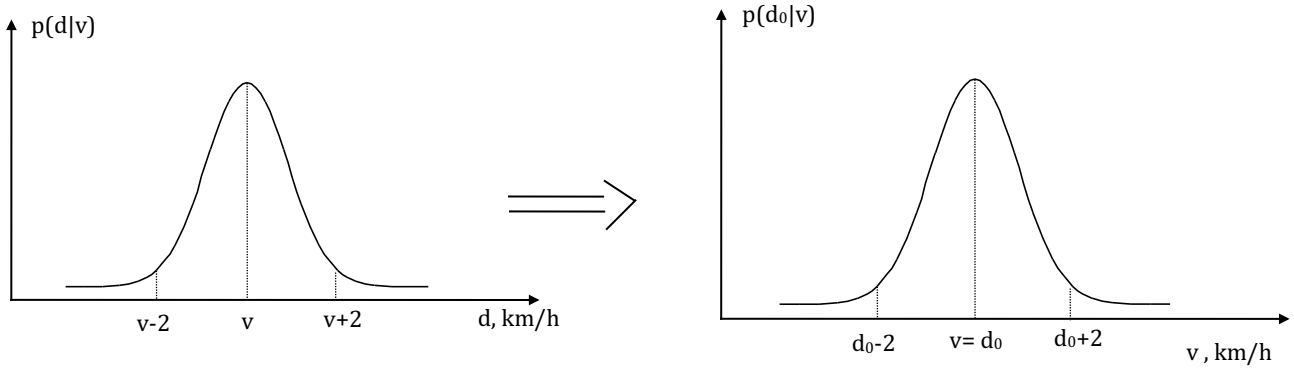
We use in the following a continuous version of the Bayes theorem:

$$p(\theta|d) \propto p(d|\theta)\, p_{prior}(\theta)$$

Compared to the discrete version, equality has been replaced by an operator $\propto$ which means "proportional to" and the denominator, which should be $p(d)$, has been "forgotten". Indeed, from the point of view of the experimenter, d, and therefore $p(d)$, 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 (r.v.) X,

$$\int_{-\infty}^{\infty} p(x)\,dx = 1.$$

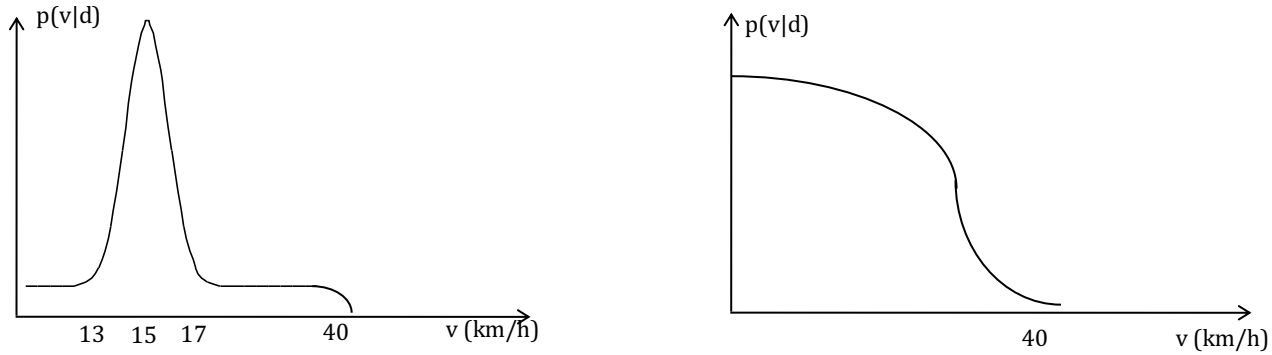
We first formulate the direct problem by taking into account the fact that we know the measurement d_0, which appears as a constant, and we don't know the true value v, which becomes the variable.
For each value of v, one obtains a law p(d|v), function of d, which depends only of $|d - v|$. Hence, for the effective measurement d_0, it is possible to draw $p(d_0|v)$, function of v:



Hence, by multiplying by $p_{prior}(v)$ and normalizing, one obtain the a posteriori laws:

(A) d_0 = 15 km/h

(B) d_0 = 80 km/h



We see that, in (A) there is no great difference between p(d|v) and p(v|d). 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.

3) Summary and introduction to chapter 2

We can summarize this chapter by the following table

World	Model	Experimenter
θ	true (but unknown)	random variable
measurements	noisy: random variables	done: known values
problem	direct: the physical model of measurements determines $p(d\|\theta)$	inverse: find $p(\theta\|D)$. The direct problem should be solved before

Actually, it could be useful to give, rather than the full $p(\theta|D)$, 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_i of θ. In the model world d_n is a random variable:

$$d_i = \theta + eral_i, i = 1, \ldots, N$$

The random error $eral_i$ 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.

$$\hat{\theta} = T(d_1, \ldots, d_N, a\ priori \text{ information})$$

with the objective of $\hat{\theta}$ as close as possible of θ.

Example, the arithmetic average:

$$\hat{\theta} = \bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i$$

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 σ^2/N. $\bar{d}$ is thus of expectation θ, just like the measurements □ᵢ, but fluctuates less around θ, because of a standard deviation divided by $\sqrt{N}$.

2) General properties

Consistent (asymptotically unbiased) estimator:

An estimator is said consistent or asymptotically unbiased if:

$$lim_{N\to\infty}\hat{\theta} = \theta$$

Unbiased estimator:

An estimator is unbiased if:

$$E(\hat{\theta}) = \theta$$

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:

$$\sigma_0^2 = \frac{1}{E\left\{\left[\frac{\partial}{\partial\theta} ln(p(\hat{\theta} and\ \theta))\right]^2\right\}}$$

The denominator of this expression is called the Fisher information I(θ). Note that the expectation involves an integral over $\hat{\theta}$ in the model world.

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 average $\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i$

One immediately demonstrates, if the error is Gaussian and the measurements independent, that $\bar{d}$ follows a Gaussian law, of variance σ^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, $(d_{N/2} + d_{(N+1)/2})/2$ if N is even.

Median is much less sensitive than mean to outliers (N.B.: if the series of measurements includes outliers, the error is no longer a Gaussian r.v., 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

A graphical representation of the data is very useful to conclude.....

4) Estimators of variance

A) Known mean:

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

B) 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 $\sigma^2$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, $d_1 - \bar{d}$ =-$\left(d_2 - \bar{d}\right)$

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

The answer is a consequence of the:

Central-limit theorem or strong law of large numbers

Let X_1,..., X_i, ..., X_N N independent random variables, of respective expectation m_i and with the same variance σ^2, but not necessarily with the same probability distribution. We have:

$$\lim_{N \longrightarrow \infty}\left(\frac{\sum_{i=1}^{n}(X_i - m_i)}{\sqrt{N\sigma^2}}\right) \sim LG(0,1)$$

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

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 r.v., whatever the probability law of each residual error.

Note that a Gaussian statistics means the absence of outliers. Indeed, for a Gaussian law, $P(|d_n - \theta| > 5\,\sigma) < 10^{-6}$. 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: $\theta - \frac{1.96\,\sigma}{\sqrt{N}} < \bar{d} < \theta + \frac{1.96\,\sigma}{\sqrt{N}}$ at 95% of confidence. Two reasons prevent us to simply write in the experimenter world:

$$\bar{d} - \frac{1.96\,\sigma}{\sqrt{N}} < \theta < \bar{d} + \frac{1.96\,\sigma}{\sqrt{N}}.$$

First, this expression assumes $p(\theta|D) = p(\bar{d}|\theta)$, which is true only if $p_{prior}(\theta) = Cste$.

Second, σ is not known, but estimated.

As for the first assumption (constant $p_{prior}(\theta)$), 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 $p_{prior}(\theta)$, we must renounce this assumption and calculate $p(\theta|\bar{d})$ with $p_{prior}(\theta)$.

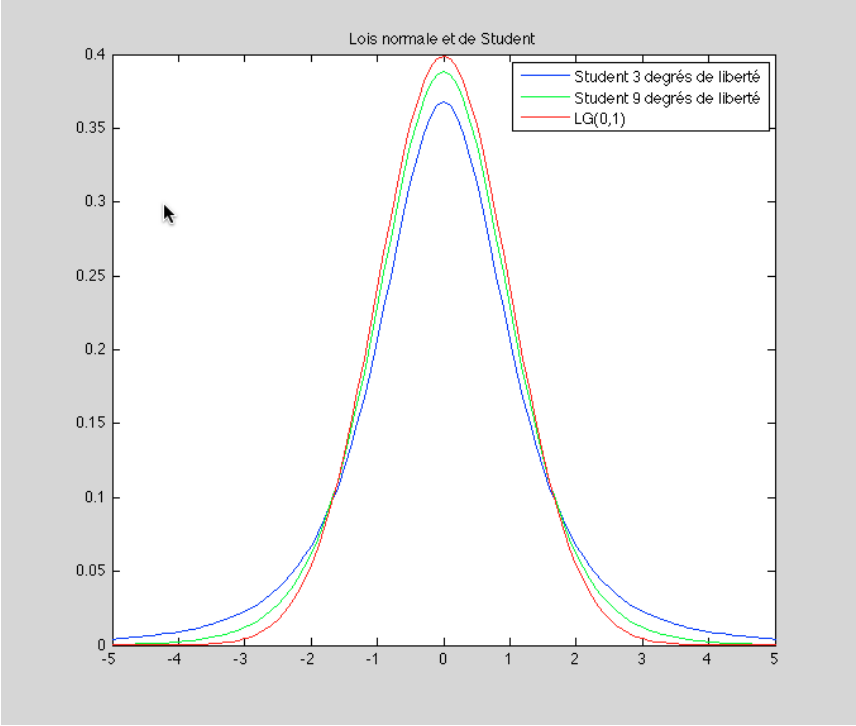
As for the second reason, if $p_{prior}(\theta)$= Cste, $\frac{\bar{d}-\theta}{\sqrt{\widehat{\sigma^2}}/\sqrt{N}}$follows a so-called Student law with N-1 degrees of freedom, where θ is the random variable. The range around the arithmetic mean where θ has a 95% chance of being found can then be determined: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

N:	3	5	10	20	40
α :	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 on the next page.

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 $\frac{\hat{\sigma}}{\sqrt{N}}$! 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:

$$\overline{d_n} = \frac{d_n + (n-1)\overline{d_{n-1}}}{n}$$

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:

$$\hat{\theta} = (1-b)\, y_n, with\, y_n = d_n + b\, y_{n-1}$$

Where $b = 1 - \epsilon, \quad 0 < \epsilon \ll 1$

It can be shown (exercise using the Z transform), 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 $\tau(\nu)$. The output of the filter (or image) can be written as: $I(\nu) = \tau(\nu)O(\nu) + N(\nu)$, where $N(\nu)$ represents a Gaussian additive noise of spectral density $\langle N(\nu)^2 \rangle$ with $\langle N(\nu) \rangle = 0$

In general, $\tau(\nu) \approx 1$ for low frequencies and $\tau(\nu) \approx 0$ 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 $S(\nu)$ by $1/\tau(\nu)$ would work. However, the noise is multiplied by the same coefficient and will have catastrophic effects on the retrieval if $1/\tau(\nu)$ is too large, as it will be at high frequencies.

Hence, we are looking for a coefficient $\mathrm{a}(\nu) \leq 1/\tau(\nu)$, such that $\widehat{O(\nu)} = \mathrm{a}(\nu)S(\nu)$ would be as close as possible of $\mathrm{O}(\nu)$. More precisely, we want to minimize the expectation of the quadratic error (we will omit from now the explicit mention of the frequency dependence):$\langle (\hat{O} - O)^2 \rangle$ minimum i.e.:

$$\langle (\hat{O} - O)^2 \rangle = \langle O^2(1 + a^2\tau^2) + a^2N^2 - 2a\tau O^2 - 2aNO - 2a^2\tau^2 NO \rangle \; minimum$$

We further assume that the output noise is independent of the object, $\langle NO \rangle = 0$, hence:

$$\langle (\hat{O} - O)^2 \rangle = \langle O^2(1 + a^2\tau^2 - 2a\tau) \rangle + \langle \, a^2N^2 \rangle$$

is minimized if:

$$\frac{\partial \langle (\hat{O} - O)^2 \rangle}{\partial a} = 0 = \langle O^2 \rangle (2\tau^2 a - 2\tau) + 2a \langle \, N^2 \rangle$$

giving:

$$a = \frac{\tau \langle O^2 \rangle}{\tau^2 \langle O^2 \rangle + \langle \, N^2 \rangle} = \frac{1}{\tau} \frac{1}{1 + \frac{\langle \, N^2 \rangle}{\tau^2 \langle O^2 \rangle}} = \frac{1}{\tau} \; \frac{1}{1 + \frac{1}{SNR}}$$

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 stucture is the goal of this calculation).

2) The estimator is biased: because $\langle N(\nu)\rangle = 0$, $\langle\frac{I(\nu)}{\tau(\nu}\rangle = O(\nu)$. To diminish the quadratic error, we have introduced some bias: $\langle\left|\widehat{O(\nu)}\right|\rangle < \langle|O(\nu)|\rangle$. In other words, not only the mean value of the noise matters. A noise of zero mean can nevertheless dominate the retrieved object.

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.

Example 1 (KLT): 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, M<K, plus some coefficients, which allow the original images to be retrieved with a negligible information loss.

Example 2 (PCA): 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

Each of the K images is formed by N pixels, meaning that a pixel j can be represented by a point in a K dimensional space. 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 image i, with i between 1 and K:

$$\overline{I_i} = \frac{1}{N}\sum_{j=1}^{N} I_{ij}$$

We define also the covariance between the image k and the image l:

$$C_{kl} = \frac{1}{N-1}\sum_{j=1}^{N}(I_{kj} - \bar{I_k})(I_{lj} - \bar{I_l})$$

The correlation coefficient r_{kl} is obtained by normalizing the covariances using the variances C_{kk} and C_{ll}:

$$r_{kl} = \frac{C_{kl}}{\sqrt{C_{kk}}\sqrt{C_{ll}}}, \qquad -1 \leq r_{kl} \leq 1$$

All the following calculations can be made by using either the covariances or the correlations. This second choice is equivalent to first normalizing each image by its standard deviation. This is possible but not compulsory: do we decide that each image conveys the same quantity of information, whatever its contrast (standard deviation)? In the second example, 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 K-dimensional space:

$$\boldsymbol{I} - \bar{\boldsymbol{I}} = \boldsymbol{\Phi Y}$$

Where $\boldsymbol{I}$ is the matrix of the images (N lines, K columns) and $\boldsymbol{\Phi}$ a K x K rotation matrix between orthonormal bases.

For such a rotation matrix, we have $\boldsymbol{\Phi^{-1} = \Phi^T}$, which allows us to write the transformed image matrix as: $\boldsymbol{Y = \Phi^T (I - \bar{I})}$.

We are looking for the rotation allowing the reconstruction $\boldsymbol{I^{truncated}}$ as close as possible of $\boldsymbol{I}$ with M<K images. For the image i, we want to obtain:

$$I_{ij}^{truncated} = \bar{I_\iota} + \sum_{k=1}^{M}\Phi_{ik}Y_{kj}$$

such that:

$$e^2(M) = \sum_{i=1}^{K}\sum_{j=1}^{N}\left(I_{ij} - I_{ij}^{truncated}\right)^2 \quad \text{minimum.}$$

Some simple calculations lead to:

$$e^2(M) = \sum_{k=M+1}^{K}\Phi_k^T\, Cov(\boldsymbol{I})\Phi_k$$

where Φ_k is the k[th] vector of $\boldsymbol{\Phi}$. This quadratic error is minimum if $\boldsymbol{\Phi^T}$ transforms $\boldsymbol{I}$ in $\boldsymbol{Y}$ with a diagonal covariance matrix:

$$Cov(\boldsymbol{Y}) = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_K \end{bmatrix}$$

which implies that the K.L. images Y are uncorrelated (which does not mean independent if the random parts are not Gaussian). The eigenvalues are ranked in descending order.

$\frac{\sum_{k=M+1}^{K} \lambda_k}{\sum_{k=1}^{K} \lambda_k}$ is the percentage of lost information.

In many practical cases, there is a value of M for which this percentage drops quite abruptly to almost 0, meaning that the original data lie in a M dimensional space. Because of that, to retrieve the original images it is sufficient to conserve M K.L. images, i.e. M.N numbers, plus the K numbers forming $\bar{\boldsymbol{I}}$ and the K^2 coefficients forming $\boldsymbol{\Phi}$. If N is large, as usual for an image, the gain in compression (i.e. not transmitted information) is almost equal to (K-M) N numbers.

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: $if\ \theta = 0,\ P(-2 < \bar{d} < 2) \approx 95\%$. 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. 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 implicitely 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:

- Are data compatible with a value (mean, standard deviation)?
- Are data compatible with a probability distribution (Gaussian, for example)?
- 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

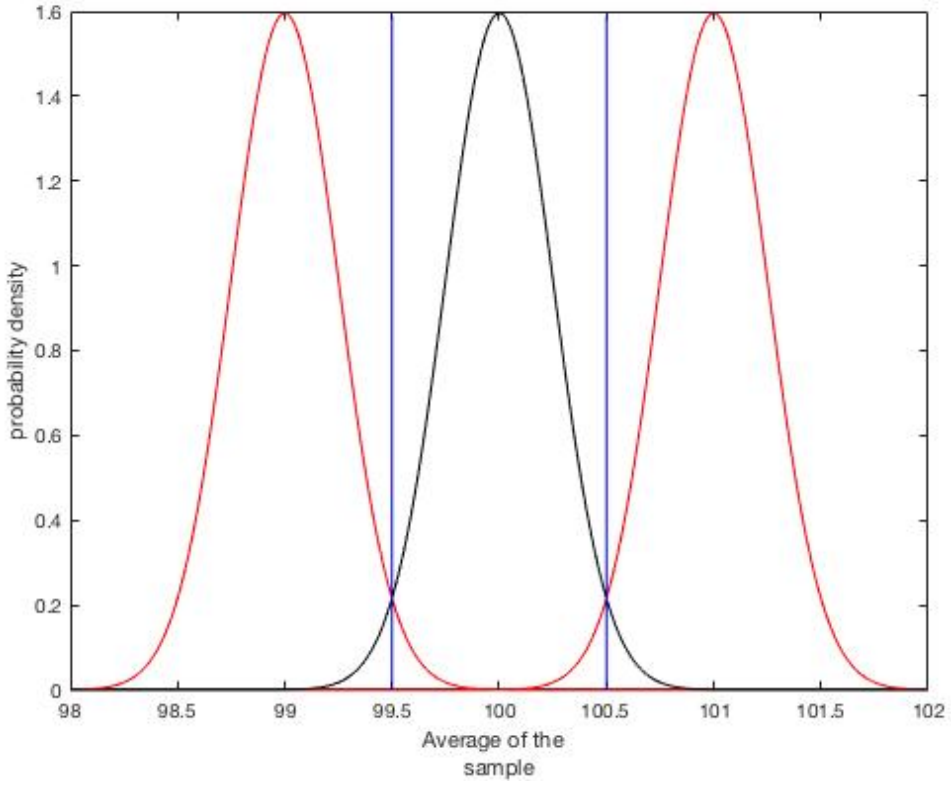
2.1) Reception quality control

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.



The black curve corresponds to the p.d.f. of the dimension of the pieces for H_0, θ=100 (supplier hypothesis). The red curves correspond to the maximum risk of acceptation with θ outside the tolerances for the customer i.e. θ=9□□□□ θ=101. We want to define a decision rule: the lot is accepted between two limits. 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 $\frac{\hat{\sigma}}{\sqrt{N}} = 0.25$.

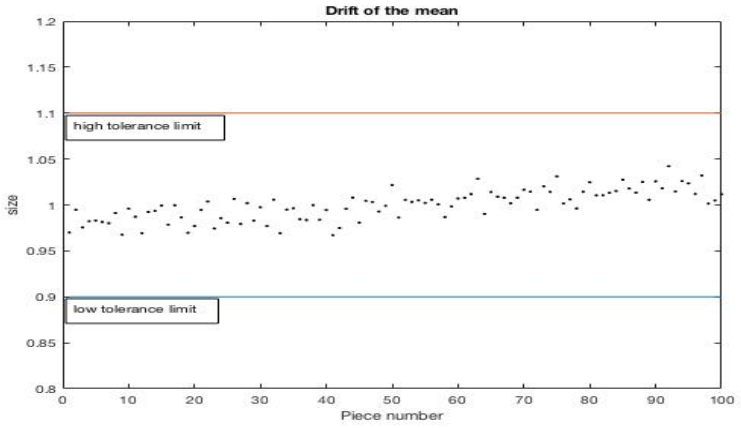
2.2) Statistical mastering of production

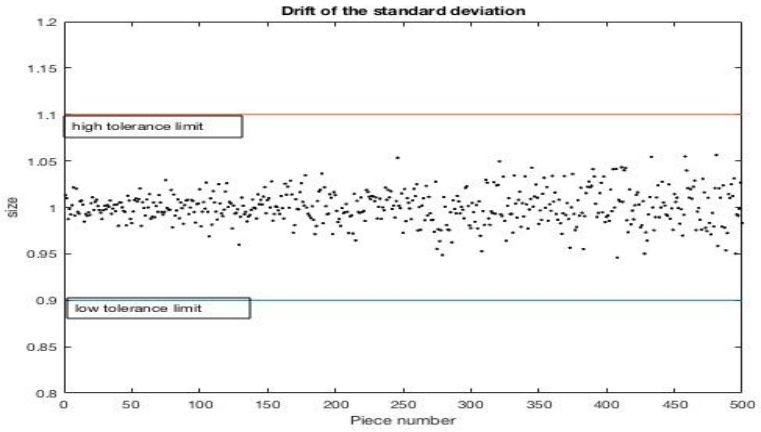
While of great historical importance, the method developed in the preceding paragraph has some important drawbacks. In particular, a drift of the true mean could lead to the rejection of the pieces, with painful consequences for both the supplier and the customer. In this paragraph, we are looking for 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.

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 below 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 by defining the capability:

Capability: $C_{pk} = Inf\left(\frac{\bar{d}-T_i, T_s-\bar{d}}{3\sigma}\right)$ T_i: low tolerance limit T_s : high tolerance limit.

C_{pk} must be greater than 1,33. Note that σ is the standard deviation of the pieces (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 stables is left to exercises.

3) Fit test: $\chi 2$ test

In this course, the $\chi 2$ 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 $d_i \in C_j \Leftrightarrow\ k_j \leq d_i < k_{j+1}$

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 $N = \sum_{k=1}^{M} n_j$

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:

$$P_j(C_j) = P(k_j \leq d_i < k_{j+1}) = P(d_i < k_{j+1}) - P(d_i < k_j)$$

The probabilities in the last equality are directly given by the cumulative density function 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:

$$D^2 = \sum_{k=1}^{M} \frac{(n_j - NP_j)^2}{NP_j}$$

If the data follow a Gaussian distribution, D^2 follows a $\boldsymbol{\chi 2}$ distribution**.** The corresponding confidence interval can be determined from tables or with a computer routine. 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. 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) regression (fitting)

1) General frame

Data (measurements) $\boldsymbol{D} = \begin{Bmatrix} \vdots \\ d_i \\ \vdots \end{Bmatrix}$ at abscissae $\begin{Bmatrix} \vdots \\ x_i \\ \vdots \end{Bmatrix}$ are modeled by a model M using K parameters $\boldsymbol{P} = (p_1, \dots, p_K)$:

$$d_i = M(x_i; p_1, \dots, p_K)$$

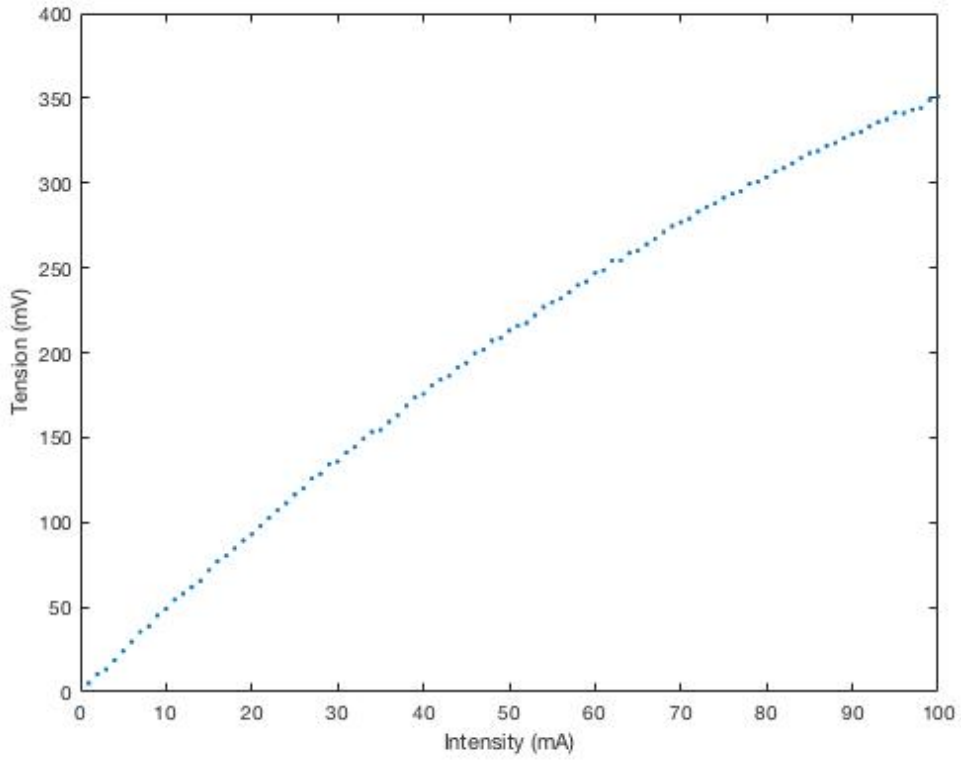
Let us define, as in chapter 1:

- the direct problem: model the data **D** (find M) for a given value of **P**

-the inverse problem: retrieve **P** from **D**, 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 R with a model $U_i = R\ I_i$. The graph of the data is the following one:



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!**

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 statistics, 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_i are performed at abscissae x_i. Let be a model $M_K(x_i)$ depending on $(p_1, \dots, p_K)$. We assume:

$$d_i = M_K(x_i) + n_i$$

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

Hence, we have in the model world:

$$p\big(d_i \big| M_k(x_i)\big) = \frac{1}{\sigma_i\sqrt{2\pi}} exp\left(-\frac{\big(d_i - M_k(x_i)\big)^2}{2\sigma_i^2}\right)$$

$$\Rightarrow p(\boldsymbol{D}|M_k) = \prod_i p\big(d_i \big| M_k(x_i)\big) \propto \quad exp\left(-\sum_i \frac{\big(d_i - M_k(x_i)\big)^2}{2\sigma_i^2}\right)$$

which means that $p(\boldsymbol{D}|M_k)$ is maximum if $\sum_i \frac{\big(d_i - M_k(x_i)\big)^2}{\sigma_i^2}$ is minimum.

If $p_{prior}(M_k)$ is constant around the estimated parameters, this is also the maximum of $p(M_k|\boldsymbol{D})$. 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: $a = \frac{\sum_i x_i(\bar{y} - y_i)}{\sum_i x_i(\bar{x} - x_i)} = \frac{\sum_i y_i(x_i - \bar{x})}{\sum_i (\bar{x} - x_i)^2}$, b=$\bar{y} - a\,\bar{x}$

The uncertainty on the coefficients is given by:

$$\sigma_a^2 = \frac{\widehat{\sigma^2}}{\sum_i (\bar{x} - x_i)^2} = a^2\left(\frac{1/\rho^2}{N-2}\right), \qquad \sigma_b^2 = \widehat{\sigma^2}\left(\frac{1}{N} + \frac{\bar{x}^2}{\sum_i (\bar{x} - x_i)^2}\right)$$

where: - $\widehat{\sigma^2}$ is estimated by: $\widehat{\sigma^2} = \frac{1}{N-2}\sum_i (y_i - (ax_i + b))^2$

- 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.

Confidence interval on the fitted points:

Let be $y'_i = ax_i + b = \bar{y} + a(x - \bar{x})$, we have : $\sigma^2_{y'_i} = \widehat{\sigma^2}\left(\frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum_i (\bar{x} - x_i)^2}\right)$

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, normalized by their standard deviation:

$$\begin{matrix}\vdots \\ \frac{y_i}{\sigma_i} \\ \vdots\end{matrix} = \frac{1}{\sigma_i}\left(p_1 f_1(x_i) + \cdots + p_K f_K(x_i)\right)$$

This system of N equations with K unknowns can be written in a matrix form as:

Y=AP, with:

K columns

$$\boldsymbol{Y} = \begin{bmatrix}\vdots \\ \frac{y_i}{\sigma_i} \\ \vdots\end{bmatrix}, \text{ N lines,} \qquad \boldsymbol{A} = \begin{bmatrix}\cdots & \begin{matrix}\vdots \\ \frac{f_j(x_i)}{\sigma_i} \\ \vdots\end{matrix} & \cdots\end{bmatrix} \text{Nlines ,} \qquad \boldsymbol{P} = \begin{bmatrix}p_1 \\ \vdots \\ p_K\end{bmatrix}, \text{ K lines}$$

It can be shown that the least squares solution ***P*** is the solution of the system of K equations with K unknowns obtained after a left multiplication by $\boldsymbol{A^T}$:
$\boldsymbol{A^T Y = A^T A P}$

Exercise: apply to $y_i = a\,x_i + b$.

A particular case is polynomial least squares $f_1(x_i)=1$, $f_2(x_i)= x_i$, $f_3(x_i)= (x_i)^2$, 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 ***P.*** 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 exp(-t/τ). This non linear model can be linearized as ln(y)= ln(a) -t/τ, with parameters ln(a) and 1/τ.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.

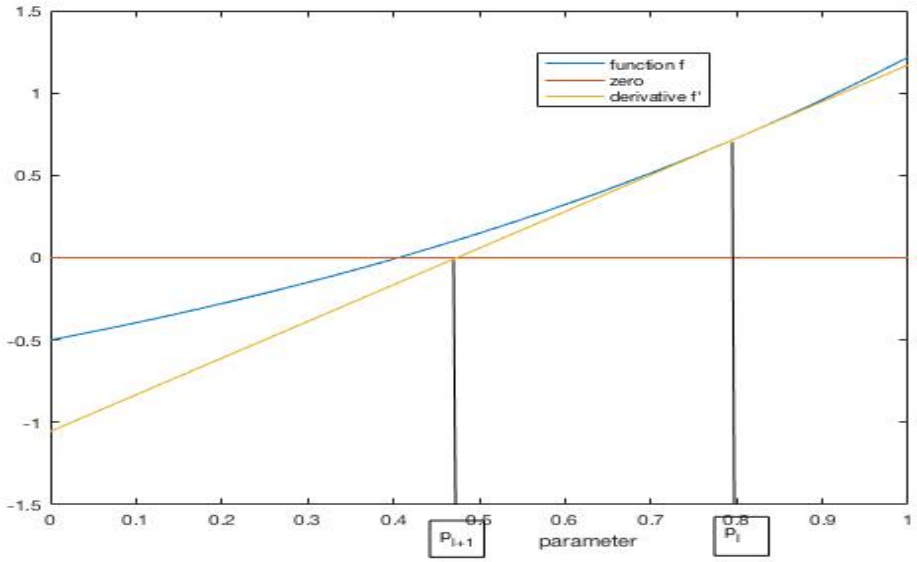
6) Nonlinear least squares: the Gauss-Newton method

The nonlinear method of fitting given in the paragraph below is one of the many possibilities. Even if not used to retrieve ***P***, 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 p for a unique abscissa x, with a nonlinear model *M:* find *p* such that *M(x,p)=y* where *M* is a nonlinear function of *p.*

If M is not "too" nonlinear, the solution of *0=f(p)= M(x,p)-y* can be found by using the Newton algorithm.



Let be p_l the value of the parameter obtained after l iterations. We obtain p_{l+1} by calculating the derivative $f'(p_l)$ of $f(p)$ at p_l and writing (see the above graph)*:* $f(p_l) = f'(p_l)(p_l - p_{l+1})$.

This method converges rapidly for small nonlinearities but is not robust. Other methods exist, like the dichotomy method: find p_1 such that $f(p_1)>0$ and p_2 such that $f(p_2)<0$. Determinate $p_3=(p_1+ p_2)/2$. If $f(p_3)>0$, calculate $p_4=(p_3+ p_2)/2$. Else, calculate $p_4=(p_3+ p_1)/2$, and so on. At each iteration, the interval of possible solutions is divided by 2: slower but safer convergence.

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:

$$y_i = M_K(p_1, \cdots, p_K; x_i), i = 1, \cdots, N$$

The quantity to be minimized is $\sum_i \frac{(y_i - M_k(x_i))^2}{\sigma_i^2}$

At each iteration, we will use a linearized model.

Let be $\boldsymbol{P_l} = \begin{bmatrix} \vdots \\ p_j \\ \vdots \end{bmatrix}$ obtained at the l[th] iteration.

The following system will be solved in the least squares sense:

$$\boldsymbol{M_K(X) - Y = M'(P_l)(P_l - P_{l+1})}$$

With $\boldsymbol{M'} = \begin{bmatrix} & \vdots & \\ \cdots & \frac{\partial M(x_i)}{\partial p_j} & \cdots \\ & \vdots & \end{bmatrix}$

As in paragraph 4, the mean squares solution is given by:

$$\boldsymbol{M'^T(M_K(X) - Y) = M'^T\, M'(P_l - P_{l+1})}$$

This is a system of K equations with K unknowns easily solved by a computer (use a specific routine and not a matrix inversion).

The convergence is obtained if the model is not too nonlinear and the starting point not too different of the solution.

<u>Uncertainties on the parameters</u>

Even if you have obtained the final set of parameters by another algorithm, you must compute $\boldsymbol{M'(P_{solution})}$ in order to evaluate the uncertainty on $\boldsymbol{P}$, given by the covariance matrix:

$$\boldsymbol{C_P} = \left[\boldsymbol{M'^T C_D^{-1} M'}\right]^{-1}$$

If the data are independent each other, the covariance of the data $\boldsymbol{C_D}$ is diagonal**:**

$$\boldsymbol{C_D} = \begin{bmatrix} \ddots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_i^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots \end{bmatrix}$$

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=a.b x. Clearly, only the product a.b, can be retrieved and not the individual parameters a and b. Consider now a model *y= a sin(b.x).* a and b are not correlated, ... except if |*b.x*|<< *1* whatever x... Fortunately, the correlation matrix of the parameters, obtained from the covariance matrix $\boldsymbol{C_P}$, (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=a.b 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.