Lecture notes on system identification and data analysis

Revision 0.92 - April 12th, 2016

Federico Alessandro Ramponi
Dipartimento di ingegneria dell'informazione, Università degli studi di Brescia
Via Branze 38, 25123 Brescia, Italy
`federico.ramponi@unibs.it`

These notes cover the program of the undergraduate course "System identification and data analysis" (and something more), taught by me at the department of information engineering, university of Brescia, Italy, in the years 2013 and 2014. They have been inspired by a set of handwritten notes by prof. Marco C. Campi, and by many comments from the students. I wish to thank Algo Carè, Luca Ceriani, Daniele Manerba, Mattia Rizzini, and Valerio Volpe for reviewing the manuscript and providing suggestions. Some basic definitions and results in linear algebra, analysis, and probability theory, used in the text, are briefly recalled in the Appendices. I will be very grateful if you let me know whenever you spot errors or crucial omissions.

Contents

1 Least squares

1.1 Introduction and motivation

Very often it happens, in science and engineering, that one knows, suspects, or assumes a relation between a variable $x \in \mathbb{R}^m$ and a variable $y \in \mathbb{R}$ in the form of a function that further depends on an unknown parameter θ:

$$y = f(x, \theta)$$

Here, θ takes values in a set $\Theta \subseteq \mathbb{R}^p$, and as θ varies in Θ, $\{f(\cdot, \theta)\}$ describes a set of functions. For example, a linear relation between two scalars x and y takes the form

$$y = f(x, \theta) = a + bx,$$

where $\theta = (a, b)$ belongs to some subset Θ of $\mathbb{R}^2$. A typical problem, called *interpolation*, is the following: given a sequence of pairs

$$(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$$

that satisfy such a functional relation, find the particular $\theta^o \in \Theta$ that corresponds to the function; in other words, find the $\theta^o \in \Theta$ such that $y_i = f(x_i, \theta^o)$ for all $i = 1, \cdots, N$. However, if the pairs (x_i, y_i) come from N physical measures and $N > p$ (N equations in p unknowns), in general the problem does not admit an exact solution. The issue is that the functional relation is an *assumption*, but measures are always corrupted by noise. We can give for granted that there is *always* a measurement error ε_i, at least on the dependent variable[1] y_i; hence, from now on, our measurement model will be

$$y_i = f(x_i, \theta) + \varepsilon_i, \quad i = 1, \cdots, N.$$

where ε_i are disturbance terms, hopefully small, but unknown and not measurable at all. Even if the disturbances are small, it is seldom the case, if at all, that the equations $y_i = f(x_i, \theta) + \varepsilon_i$ are verified simultaneously for any θ. Other problems may occur. For example, the set of parameters Θ in which one chooses to search the solution may be wrong, i.e. too small. Even worse, the *assumption* about the functional relation may be incorrect, or only approximative; in other words, it may be false that the data are generated by a function $y = f(x, \theta)$, or that there exists a "true" θ^o. Hence, in general, the only problem that can reasonably be solved is that of searching an *approximate* function $f(\cdot, \theta)$, instead of the "true" one $f(\cdot, \theta^o)$ whose

[1] Strictly speaking, we should take into account measurement errors on the *independent* variable x_i as well. We will not do it here, both because this would complicate the model too much, and because independent variables, i.e. the ones set by the experimenter, are more likely to be known exactly.

existence may not be guaranteed. In approximating $f(\cdot, \theta^o)$ with a certain $f(\cdot, \theta)$, we commit approximation errors $\epsilon_i(\theta)$, which are called *residuals*:

$$\epsilon_i(\theta) := y_i - f(x_i, \theta), \quad i = 1, \cdots, N.$$

Please take some care in distinguishing $\epsilon_i(\theta)$ from ε_i:

- the quantities $\varepsilon_i = y_i - f(x_i, \theta^o)$ are disturbances that we assume intrinsic in the measurement model, and due to the randomness present in "nature";

- the quantities $\epsilon_i(\theta) = y_i - f(x_i, \theta)$ are errors that we commit because we select a certain function $f(\cdot, \theta)$ exploiting the only data at hand, $(x_1, y_1), \cdots, (x_N, y_N)$; the $\epsilon_i(\theta)$ are different from the ε_i except in the lucky case, which *never* happens in practice, that we succeed in finding $\theta = \theta^o$.

The goal of the least squares method is to find the particular $\hat{\theta} \in \Theta$ that provides the "best" description of the data; and what in such method is meant by "best" is that the sum of the squares of the residuals

$$Q(\hat{\theta}) := \sum_{i=1}^{N} \epsilon_i^2(\hat{\theta}) = \sum_{i=1}^{N} \left(y_i - f(x_i, \hat{\theta})\right)^2$$

is minimal. You see that the smaller that sum is, the closer the y_i are, collectively, to the quantities $f(x_i, \hat{\theta})$, so that the smaller the sum, the better the approximation. There are at least three reasons why such an approximation is useful:

- the estimate of the model can be used to investigate some internal properties of the mechanism that generates the data (for example stability, in the case of a dynamical system);

- the parameter θ^o may correspond to some physical quantity of interest: then its estimate $\hat{\theta}$ can be used to measure that quantity;

- the estimate of the model can be used to *predict* a future value y_{N+1}, when a future value x_{N+1} will be available.

Example. Historians agree that the method was devised by the *princeps mathematicorum*, Carl Friedrich Gauss. Between January and February 1801 the Italian astronomer Giuseppe Piazzi recorded 24 observations of what according to him was a new comet, and is instead Ceres, the largest asteroid or "dwarf planet" in the asteroid belt between Mars and Jupiter. He stopped recording observations due to illness; lately he was not able to find the asteroid again, but he published his observations. (By coincidence, this happened exactly while Hegel, in his dissertation, was expressing his sarcasm against the conceit of the astronomers, searching for an eighth planet.) Gauss used the least squares method to infer the trajectory of Ceres from Piazzi's measures and *predict its future position*, and published his results (Franz X. von Zach, *Monatliche Correspondenz*, September 1801), although he did *not* publish his method; the asteroid was found again, and Gauss gained a well deserved fame. The least squares method was published shortly after by the mathematician Adrien-Marie Legendre (*Sur la Méthode des moindres quarrés*, in *Nouvelles méthodes pour la détermination des orbites des comètes*, 1805). □

Summing up, given $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$ the method of least squares prescribes to find

$$\hat{\theta}_{\text{LS}} := \arg\min_{\theta \in \Theta} \; Q(\theta) = \arg\min_{\theta \in \Theta} \; \sum_{i=1}^{N} (y_i - f(x_i, \theta))^2.$$

It makes perfect sense to minimize a sum of positive functions of the errors, but why indeed the *sum of the squares* of the errors? At this stage of the discussion, the choice of the square seems, and is indeed, arbitrary. One could choose instead to minimize other functions representing the residuals collectively; for instance:

$$Q_1(\theta) := \sum_{i=1}^{N} |y_i - f(x_i, \theta)| \qquad \text{or} \qquad Q_\infty(\theta) := \max_{i=1,\cdots,N} |y_i - f(x_i, \hat{\theta})|.$$

You can see that in both these minimization criteria, the closer the objective function is to zero, the smaller are the residuals, collectively. Thus, both of them are perfectly reasonable, both of them are actually used in practice, and each one has its own pros and cons. Also, at this level of generality the only way to solve a least squares problem is by means of numerical optimization, and the minimization of a sum of *squares* has no obvious advantage over the other choices.

There is a fairly general condition, though, in which the minimization of *squares* has indeed obvious advantages; namely, when

$$f \textit{ is linear in the parameter } \theta$$

(although not necessarily in the variable x). In this chapter, we will focus on this case. When linearity holds, the theory behind the minimization is complete and elegant, it admits an intuitive explanation in terms of projections, and above all, it leads to an analytic solution. These are the reasons why the least squares method is so popular in the scientific community.

1.2 Linearity in the parameter θ

Let $x_i = \begin{bmatrix} x_i^{(1)} & \cdots & x_i^{(m)} \end{bmatrix}^\top \in \mathbb{R}^m$ and $y_i \in \mathbb{R}$ for $i = 1, \cdots, N$. In the language of least squares, if y_i is viewed as a function of x_i, we call x_i "explanatory variable", and we say that y_i are "explained" by x_i. In general the explained variable y_i may not be linked to the explanatory variable directly, but through p functions

$$\begin{aligned} \varphi_1 &: \mathbb{R}^m \to \mathbb{R} \\ \varphi_2 &: \mathbb{R}^m \to \mathbb{R} \\ &\vdots \\ \varphi_p &: \mathbb{R}^m \to \mathbb{R} \end{aligned}$$

which can be nonlinear, and which we will call *regressors*. The data generation model to which we will apply least squares is the following:

$$\begin{aligned} y_1 &= \theta_1\varphi_1\left(x_1^{(1)}, \cdots, x_1^{(m)}\right) + \theta_2\varphi_2\left(x_1^{(1)}, \cdots, x_1^{(m)}\right) + \cdots + \theta_p\varphi_p\left(x_1^{(1)}, \cdots, x_1^{(m)}\right) + \varepsilon_1; \\ y_2 &= \theta_1\varphi_1\left(x_2^{(1)}, \cdots, x_2^{(m)}\right) + \theta_2\varphi_2\left(x_2^{(1)}, \cdots, x_2^{(m)}\right) + \cdots + \theta_p\varphi_p\left(x_2^{(1)}, \cdots, x_2^{(m)}\right) + \varepsilon_2; \\ &\vdots \\ y_N &= \theta_1\varphi_1\left(x_N^{(1)}, \cdots, x_N^{(m)}\right) + \theta_2\varphi_2\left(x_N^{(1)}, \cdots, x_N^{(m)}\right) + \cdots + \theta_p\varphi_p\left(x_N^{(1)}, \cdots, x_N^{(m)}\right) + \varepsilon_N. \end{aligned}$$

Defining the vectors

$$\theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} \quad \text{and} \quad \varphi\left(x_i^{(1)}, \cdots, x_i^{(m)}\right) = \begin{bmatrix} \varphi_1\left(x_i^{(1)}, \cdots, x_i^{(m)}\right) \\ \vdots \\ \varphi_p\left(x_i^{(1)}, \cdots, x_i^{(m)}\right) \end{bmatrix}$$

and recalling that $x_i = [x_i^{(1)}, \cdots, x_i^{(m)}]^\top$, we can write the model in compact form:

$$y_i = f(x_i, \theta) + \varepsilon_i = \varphi(x_i)^\top \theta + \varepsilon_i, \qquad i = 1, \cdots, N.$$

Note that the function f is linear in the parameter θ, although not necessarily in the explanatory variable x. Let us shorten notation even more by letting $\varphi_i := \varphi(x_i)$: then the model finally reads

$$y_i = \varphi_i^\top \theta + \varepsilon_i, \qquad i = 1, \cdots, N.$$

Example. Let $x = \begin{bmatrix} x_i^{(1)} & \cdots & x_i^{(m)} \end{bmatrix}^\top \in \mathbb{R}^m$, and define (the exponent here means an index):

$$\begin{aligned}
\varphi_0\left(x^{(1)}, \cdots, x^{(m)}\right) &= 1 \\
\varphi_1\left(x^{(1)}, \cdots, x^{(m)}\right) &= x^{(1)} \\
\varphi_2\left(x^{(1)}, \cdots, x^{(m)}\right) &= x^{(2)} \\
&\vdots \\
\varphi_m\left(x^{(1)}, \cdots, x^{(m)}\right) &= x^{(m)}
\end{aligned}$$

Here, $\varphi : \mathbb{R}^m \to \mathbb{R}^{m+1}$ is an affine function of x. The application of least squares to the resulting model

$$y_i = \theta_0 + \theta_1 x_i^{(1)} + \theta_2 x_i^{(2)} + \cdots + \theta_m x_i^{(m)} + \varepsilon_i$$

is called a *linear regression*[2] of y over the m variables $x^{(1)}, \cdots, x^{(m)}$, and is very popular in applied statistics. In section 1.3 we will consider in detail a particular case with $m = 1$. □

Example. Let $x \in \mathbb{R}$, and define (the exponent here means a power):

$$\begin{aligned}
\varphi_0(x) &= 1 \\
\varphi_1(x) &= x \\
\varphi_2(x) &= x^2 \\
&\vdots \\
\varphi_m(x) &= x^n
\end{aligned}$$

Here $m = 1$, $p = n + 1$, and $\varphi : \mathbb{R} \to \mathbb{R}^{n+1}$ yields $p = n + 1$ monomials. The application of least squares to the resulting model

$$y_i = \theta_0 + \theta_1 x_i^1 + \theta_2 x_i^2 + \cdots + \theta_n x_i^n + \varepsilon_i$$

is a *polynomial interpolation*; the result is the polynomial of degree (at most) n in x_i that best approximates the y_i. □

[2] Linearity with respect to *the parameter* is a general assumption about the measurement model; here, "linear regression" means that, except for the constant θ_0, linearity holds with respect to *the explanatory data* as well, i.e. that also the regressor is linear.

Example. Let $x = \begin{bmatrix} u & v \end{bmatrix}^\top \in \mathbb{R}^2$, and define:

$$\begin{aligned}
\varphi_0(x) &= 1\\
\varphi_1(x) &= u\\
\varphi_2(x) &= v\\
\varphi_3(x) &= u^2\\
\varphi_4(x) &= uv\\
\varphi_5(x) &= v^2
\end{aligned}$$

Here, $\varphi : \mathbb{R}^2 \to \mathbb{R}^6$. The application of least squares to the resulting model

$$y_i = \theta_0 + \theta_1 u_i + \theta_2 v_i + \theta_3 u_i^2 + \theta_4 u_i v_i + \theta_5 v_i^2 + \varepsilon_i$$

is the interpolation with a polynomial of degree 2, in two variables. □

The least squares method now asks to find

$$\hat{\theta}_{\text{LS}} := \underset{\theta \in \mathbb{R}^p}{\arg\min}\ Q(\theta) = \underset{\theta \in \mathbb{R}^p}{\arg\min}\ \sum_{i=1}^{N} (y_i - \varphi_i^\top \theta)^2.$$

In order to carry on with the minimization, note the following:

1. any term $y_i - \varphi_i^\top \theta$ is an affine function of θ;

2. any term $\left(y_i - \varphi_i^\top \theta\right)^2$ is a *convex* function of θ, being the composition of the convex function $(\cdot)^2$ with an affine function;

3. the sum to be minimized is convex, being a sum of convex functions.

Thus, $Q(\theta)$ is convex *precisely due to the linearity of $f(x, \theta)$ with respect to* θ, and we are asked to minimize it; note that such sum is also obviously differentiable over the whole of $\mathbb{R}^p$. Now, in general, *nothing* guarantees that a differentiable convex function has a minimum (visualize the examples $f(t) = e^{-t}$ and $f(t) = e^{-t} - t$). Nevertheless, in view of Corollary B.3.1, *if we do* find a point such that the derivative of a convex function (that is, its gradient) is zero at that point, then such point is guaranteed to be a minimum point (the arg min). Thus, to carry on with the minimization we set the derivative with respect to θ equal to zero:

$$\begin{aligned}
\frac{\partial Q(\theta)}{\partial \theta} &= \frac{\partial}{\partial \theta} \sum_{i=1}^{N} (y_i - \varphi_i^\top \theta)^2 = \sum_{i=1}^{N} 2(y_i - \varphi_i^\top \theta)(-\varphi_i^\top)\\
&= \sum_{i=1}^{N} 2(y_i - \theta^\top \varphi_i)(-\varphi_i^\top) = 0,
\end{aligned}$$

finding

$$\theta^\top \left(\sum_{i=1}^N \varphi_i \varphi_i^\top \right) = \sum_{i=1}^N y_i \varphi_i^\top .$$

Transposing again, we find

$$\left(\sum_{i=1}^N \varphi_i \varphi_i^\top \right) \theta = \sum_{i=1}^N \varphi_i y_i. \tag{1}$$

Equation (1) is called *normal equation* (usually in the plural person: normal *equations*). It is linear in the parameter θ, namely it has the form $R\theta = b$, where $R \in \mathbb{R}^{p \times p}$ is a symmetric and positive semi-definite matrix, and $b \in \mathbb{R}^p$. Any θ that solves (1) is a minimum point for the sum of squares and a solution to the least squares problem; in particular, if $R = \sum_{i=1}^N \varphi_i \varphi_i^\top$ is invertible (and in practical applications this is generally the case), then the only solution to the normal equations reads explicitly:

$$\hat{\theta}_{\mathrm{LS}} = \left(\sum_{i=1}^N \varphi_i \varphi_i^\top \right)^{-1} \sum_{i=1}^N \varphi_i y_i. \tag{2}$$

1.3 Simple examples

Let us test the formula (2) of the previous section on two extremely simple cases, linear in both the parameter and the explanatory variable.

First, suppose that $x_i \in \mathbb{R}$, and let us test the least squares method with the measurement model $y_i = a x_i + \varepsilon_i$, hence with $\theta = a \in \mathbb{R}$. This means that, ideally, the data should be "explained" by a straight line passing through the origin; hopefully, the straight line found with the least squares method will be close to the points $\{(x_i, y_i)\}_{i=1}^N$.

Consider the problem of minimizing with respect to a the sum

$$Q(a) = \sum_{i=1}^N (y_i - a x_i)^2.$$

To do this, we note that the sum is a positive quadratic form in the variable a; to search a minimum, since Q is convex and differentiable, we set the derivative of the sum with respect to a equal to zero:

$$\frac{\partial Q(a)}{\partial a} = \sum_{i=1}^N 2(y_i - a x_i)(-x_i) = 0,$$

yielding

$$
\begin{aligned}
\sum_{i=1}^{N} x_i y_i &= a \sum_{i=1}^{N} x_i^2, \\
\hat{a}_{\mathrm{LS}} &= \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}
\end{aligned}
$$

Let us denote conventionally a sample average as $\mathsf{M}\left[h(x,y)\right] = \frac{1}{N}\sum_{i=1}^{N} h(x_i, y_i)$ for any function h of two variables. Then, dividing everywhere by N, the equations read

$$
\begin{aligned}
\mathsf{M}\left[xy\right] &= a\ \mathsf{M}\left[x^2\right], \\
\hat{a}_{\mathrm{LS}} &= \frac{\mathsf{M}\left[xy\right]}{\mathsf{M}\left[x^2\right]}.
\end{aligned}
$$

The equation

$$
y = \hat{a}_{\mathrm{LS}}\ x
$$

describes the straight line *passing through the origin* that best interpolates the data, in the sense of least squares.

The formula (2) of the previous section yields (of course) the same result. Here we have $p = 1$ (just one regressor), $\varphi(x) = \left[\ x\ \right]$, $\theta = \left[\ a\ \right]$ (don't be confused by the square brackets! A 1×1 matrix is just a scalar, i.e. a number), and

$$
\begin{aligned}
\hat{\theta}_{\mathrm{LS}} &= \arg\min_{\theta \in \mathbb{R}^2}\ \sum_{i=1}^{N} (y_i - \varphi_i^{\top}\theta)^2 \\
&= \left(\sum_{i=1}^{N} \varphi_i \varphi_i^{\top}\right)^{-1} \sum_{i=1}^{N} \varphi_i y_i \\
&= \left(\sum_{i=1}^{N} \left[\ x_i\ \right]\left[\ x_i\ \right]^{\top}\right)^{-1} \sum_{i=1}^{N} \left[\ x_i\ \right] y_i \\
&= \left[\ \sum_{i=1}^{N} x_i^2\ \right]^{-1} \left[\ \sum_{i=1}^{N} x_i y_i\ \right] \\
&\qquad \text{(now divide by } N \text{ inside brackets)} \\
&= \left[\ \mathsf{M}\left[x^2\right]\ \right]^{-1} \left[\ \mathsf{M}\left[xy\right]\ \right] \\
&= \hat{a}_{\mathrm{LS}},
\end{aligned}
$$

as we have just found.

A slightly more general case of least squares with linearity both in the parameter and the explanatory variable is when $x_i \in \mathbb{R}$, the measurement model is $y_i = a + bx_i + \varepsilon_i$, and $\theta = (a, b) \in \mathbb{R}^2$. This means that the data should be "explained" by a straight line in the plane, not necessarily passing through the origin; as before, ideally we expect the line to pass "through" the points, being close to them.

Consider the problem of minimizing with respect to a and b the sum

$$Q(a, b) = \sum_{i=1}^{N} (y_i - a - bx_i)^2.$$

We note that the sum is a positive quadratic form in the variables a, b; to search a minimum, since Q is again convex and differentiable, we set the derivatives of the sum with respect to a and b equal to zero:

$$\begin{aligned} \frac{\partial Q(a,b)}{\partial a} &= \sum_{i=1}^{N} 2(y_i - a - bx_i)(-1) = 0, \\ \frac{\partial Q(a,b)}{\partial b} &= \sum_{i=1}^{N} 2(y_i - a - bx_i)(-x_i) = 0, \end{aligned}$$

yielding

$$\begin{aligned} \sum_{i=1}^{N} y_i &= Na + b\left(\sum_{i=1}^{N} x_i\right), \\ \sum_{i=1}^{N} x_i y_i &= a\left(\sum_{i=1}^{N} x_i\right) + b\left(\sum_{i=1}^{N} x_i^2\right). \end{aligned}$$

With the same notation $\mathsf{M}[\cdot]$ of the previous example, dividing everywhere by N, the equations read

$$\begin{aligned} \mathsf{M}[y] &= a + b\ \mathsf{M}[x]; \\ \mathsf{M}[xy] &= a\ \mathsf{M}[x] + b\ \mathsf{M}\left[x^2\right]. \end{aligned}$$

The solution to this linear system is given by

$$\begin{aligned} \hat{b}_{\mathrm{LS}} &= \frac{\mathsf{M}[xy] - \mathsf{M}[x]\mathsf{M}[y]}{\mathsf{M}[x^2] - \mathsf{M}[x]^2} = \frac{\text{sample covariance of } \{x_i, y_i\}}{\text{sample variance of } \{x_i\}}; \\ \hat{a}_{\mathrm{LS}} &= \mathsf{M}[y] - \hat{b}_{\mathrm{LS}}\ \mathsf{M}[x] \\ &\quad = \text{sample average of } \{y_i\} - \hat{b}_{\mathrm{LS}} \times \text{sample average of } \{x_i\}. \end{aligned}$$

The equation

$$y = \hat{a}_{\mathrm{LS}} + \hat{b}_{\mathrm{LS}}\ x$$

describes the straight line that best interpolates the data, in the sense of least squares[3].

Let us test again the formula (2) of the previous section. We have $\varphi_1(x) = 1, \varphi_2(x) = x$, $\varphi(x) = \begin{bmatrix} 1 \\ x \end{bmatrix}$, $\theta = \begin{bmatrix} a \\ b \end{bmatrix}$; then

$$\begin{aligned}
\hat{\theta}_{\text{LS}} &= \arg\min_{\theta \in \mathbb{R}^2} \sum_{i=1}^{N} (y_i - \varphi_i^\top \theta)^2 \\
&= \left(\sum_{i=1}^{N} \varphi_i \varphi_i^\top \right)^{-1} \sum_{i=1}^{N} \varphi_i y_i \\
&= \left(\sum_{i=1}^{N} \begin{bmatrix} 1 \\ x_i \end{bmatrix} \begin{bmatrix} 1 & x_i \end{bmatrix} \right)^{-1} \sum_{i=1}^{N} \begin{bmatrix} 1 \\ x_i \end{bmatrix} y_i \\
&= \begin{bmatrix} \sum_{i=1}^N 1 & \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^N y_i \\ \sum_{i=1}^N x_i y_i \end{bmatrix} \\
&\qquad \text{(now divide by } N \text{ inside brackets)} \\
&= \begin{bmatrix} 1 & \mathsf{M}[x] \\ \mathsf{M}[x] & \mathsf{M}[x^2] \end{bmatrix}^{-1} \begin{bmatrix} \mathsf{M}[y] \\ \mathsf{M}[xy] \end{bmatrix} \\
&= \frac{1}{\mathsf{M}[x^2] - \mathsf{M}[x]^2} \begin{bmatrix} \mathsf{M}[x^2] & -\mathsf{M}[x] \\ -\mathsf{M}[x] & 1 \end{bmatrix} \begin{bmatrix} \mathsf{M}[y] \\ \mathsf{M}[xy] \end{bmatrix} \\
&= \begin{bmatrix} \frac{\mathsf{M}[y]\mathsf{M}[x^2] - \mathsf{M}[xy]\mathsf{M}[x]}{\mathsf{M}[x^2] - \mathsf{M}[x]^2} \\ \frac{\mathsf{M}[xy] - \mathsf{M}[x]\mathsf{M}[y]}{\mathsf{M}[x^2] - \mathsf{M}[x]^2} \end{bmatrix} = \begin{bmatrix} \hat{a}_{\text{LS}} \\ \hat{b}_{\text{LS}} \end{bmatrix},
\end{aligned}$$

as we have found deriving one coefficient at a time.

[3] That is to say, the data points (x_i, y_i) are hopefully close to the straight line (more on this subject in section 1.6). In most applications, this means that the $\{y_i\}$ have a strong statistical correlation with the $\{x_i\}$; indeed linear interpolation with least squares is very popular in applied statistics, where the discovery of correlations is of paramount importance. But here is an example that has nothing to do with correlation in the statistical sense: how do you compute *numerically* the derivative of a function $f : \mathbb{R} \to \mathbb{R}$ at a point x? The most naive way of doing it, that comes to mind, is by taking a small Δx and letting $f'(x) \simeq \frac{f(x+\Delta x) - f(x)}{\Delta x}$. But note that this approximation is the slope of the straight line passing through the points $(x, f(x))$ and $(x + \Delta x, f(x + \Delta x))$. Only two points! Here is another idea, just slightly less naive: why not taking many "deltas" around 0, say $\Delta x_1, \Delta x_2, \Delta x_3, \Delta x_4 = 0, \Delta x_5, \Delta x_6, \Delta x_7$, finding with least squares the straight line that interpolates the points $(x + \Delta x_i, f(x + \Delta x_i))$, and letting $f'(x) \simeq$ the slope $\hat{b}_{\text{LS}}$ of the resulting straight line? This solution is certainly not the best that numerical analysis has to offer, but it should definitely be more robust than the previous one.

1.4 Existence and uniqueness of a solution of the normal equations

As for any system of linear equations, one may ask whether a solution $\hat{\theta}_{\text{LS}}$ to Equation (1) actually exists, and if so, whether it is unique or not. The answer to the first question is positive, due to the following lemma:

Lemma 1.4.1

$$\text{range } \left(\sum_{i=1}^{N} \varphi_i \varphi_i^\top \right) = \text{span } \{\varphi_1, \cdots, \varphi_N\}.$$

Proof. Denote $R = \sum_{i=1}^{N} \varphi_i \varphi_i^\top$ and $S = \text{span } \{\varphi_1, \cdots, \varphi_N\}$. Note that R is symmetric; indeed

$$R^\top = \left(\sum_{i=1}^{N} \varphi_i \varphi_i^\top \right)^\top = \sum_{i=1}^{N} \left(\varphi_i^\top \right)^\top \varphi_i^\top = \sum_{i=1}^{N} \varphi_i \varphi_i^\top = R;$$

Therefore range $R^\top$ = range R and null R = (range $R^\top)^\perp$ = (range $R)^\perp$.
Suppose that $v \in S^\perp$. Then $\varphi_i \perp v$ for all i. Then $Rv = \sum_{i=1}^{N} \varphi_i (\varphi_i^\top v) = 0$, hence $v \in$ null R = (range $R)^\perp$.
Suppose, on the other hand, that $v \in$ (range $R)^\perp$ = null R. Then $Rv = 0$, hence

$$0 = v^\top R v = v^\top \sum_{i=1}^{N} \varphi_i \varphi_i^\top v = \sum_{i=1}^{N} (\varphi_i^\top v)^2.$$

Since the last expression is a sum of nonnegative quantities, for it to be zero it must hold $\varphi_i^\top v = 0$ for all i, that is $\varphi_i \perp v$ for all i, hence $v \in S^\perp$. We conclude that (range $R)^\perp = S^\perp$, and since the range and span under consideration are subspaces of the finite-dimensional vector space $\mathbb{R}^p$, taking the orthogonal complement on both sides we obtain range $R = S$. □

Corollary 1.4.1 *The normal equations have at least one solution.*

Proof. $\sum_{i=1}^{N} \varphi_i y_i$ belongs to span $\{\varphi_1, \cdots, \varphi_N\}$, being a linear combination of the φ_i; therefore, due to Lemma 1.4.1, it also belongs to the range of $R = \sum_{i=1}^{N} \varphi_i \varphi_i^\top$. The claim follows immediately. □

On the other hand, such a solution may not be unique. This happens precisely when R is singular, because in that case, if $\hat{\theta}$ is a solution and $n \in$ null R, then $\hat{\theta} + n$ is also a solution. Indeed, as for any linear equation, there must exist an entire affine space of solutions (infinitely many). That R is singular may happen for different reasons:

- there may not be enough data. If $N < p$, then R is singular, because for any linear mapping $R : \mathbb{R}^p \to \mathbb{R}^p$ it holds $p = \dim \operatorname{range} R + \dim \operatorname{null} R$; but $\operatorname{range} R = \operatorname{span} \{\varphi_1, \cdots, \varphi_N\}$ which has dimension $\dim \operatorname{span} \{\varphi_1, \cdots, \varphi_N\} \leq N < p$; hence $\dim \operatorname{null} R > 0$. This is, however, a trivial issue. If we did not have enough data, it would be pointless to pose the problem in the first place (e.g. find "the" parabola that passes through two points). In applications, one has $N \gg p$.

- The φ_i may be flawed by construction. Recall that $\varphi_i = \varphi(x_i) = [\varphi_1(x_i), \cdots, \varphi_p(x_i)]^\top$, where $\varphi_j(\cdot)$ are, in general, nonlinear functions from $\mathbb{R}^m$ to $\mathbb{R}$. If these functions are themselves linearly dependent, R will be singular. To see what happens, take for example $m = p = 3$, $\varphi_1(x_1, x_2, x_3) = x_1$, $\varphi_2(x_1, x_2, x_3) = x_2 + x_3$, and $\varphi_3(x_1, x_2, x_3) = x_1 + x_2 + x_3$. Then $[1, 1, -1]\varphi(x) = 0$ irrespective of x, and as you can easily see this implies that $\operatorname{rank} R \leq 2$. However, this is just a pathological condition that can be avoided with a bit of common sense in the choice of the regressors.

- The explanatory data x_i may not carry enough information. As an extreme case, if the x_i are all equal, of even if they are random but belong almost surely to a small subspace of $\mathbb{R}^m$ whose map under the function $\varphi(\cdot)$ is not the whole of $\mathbb{R}^p$, R will be singular. This is the only issue that may actually occur in applications.

Visualize what happens geometrically when $\theta \in \mathbb{R}^2$:

- If R is invertible, the solution to the normal equations is unique, hence such is the minimum point of the sum of squares. This occurs when the sum of squares happens to be *strictly* convex (its graph on the θ-plane is that of an elliptic paraboloid, going to $+\infty$ in every direction).

- If R is singular, there exists an affine subspace of solutions to the normal equations. All these solutions are minimum points for the sum of squares, hence in particular they attain the same value. Thus, the sum of squares is convex, *but not strictly* (its graph is a parabolic cylinder, i.e. a "valley", maybe going to $+\infty$ in some directions, but attaining constant height in at least one direction).

1.5 Interpretation in terms of projections

Now we will see that least squares are closely related with orthogonal projections. We start from a classical result, characterizing the point of a subspace which is closest to another point in the sense of the Euclidean distance[4]:

[4] We state it in $\mathbb{R}^N$, because this is our setting; but the result is valid in a far more general context, namely it is one of the most important properties of Hilbert spaces ($\mathbb{R}^N$,

Theorem 1.5.1 *Let V be a subspace of $\mathbb{R}^N$, and $y \in \mathbb{R}^N$. If there exists a vector $v_m \in V$ such that $\|y - v_m\| \leq \|y - v\|$ for all $v \in V$, then v_m is unique. Moreover, a necessary and sufficient condition for v_m to be the unique minimizing vector is that $y - v_m$ is orthogonal to V.*

Proof. First, we show that if v_m is a minimizing vector, then the error $y - v_m$ is orthogonal to V. Suppose, for the sake of contradiction, that $y - v_m$ is not orthogonal to V. Then there exists $v \in V$, $\|v\| = 1$, such that $\langle y - v_m, v \rangle = \alpha \neq 0$. Let then $v_{m2} = v_m + \alpha v \in V$. It holds

$$\begin{aligned}
\|y - v_{m2}\|^2 &= \|y - v_m - \alpha v\|^2 \\
&= \|y - v_m\|^2 - 2\left\langle y - v_m, \alpha v \right\rangle + \alpha^2 \|v\|^2 \\
&= \|y - v_m\|^2 - 2\alpha \left\langle y - v_m, v \right\rangle + \alpha^2 \\
&= \|y - v_m\|^2 - 2\alpha^2 + \alpha^2 \\
&= \|y - v_m\|^2 - \alpha^2 \\
&< \|y - v_m\|^2,
\end{aligned}$$

which is in contradiction with the hypothesis that v_m is a minimizing vector. Hence, $y - v_m$ must be orthogonal to V.
Now we show that, if $y - v_m$ is orthogonal to V, then v_m is a unique minimizing vector. Indeed, for any $v \in V$,

$$\begin{aligned}
\|y - v\|^2 &= \|y - v_m + v_m - v\|^2 \\
&= \|y - v_m\|^2 + \|v_m - v\|^2 \\
&> \|y - v_m\|^2 \quad \text{if and only if } v \neq v_m,
\end{aligned}$$

where the second equality is Pythagoras's theorem ($v_m - v \in V$, hence $y - v_m$ and $v_m - v$ are orthogonal by hypothesis). □

We state the next result without proof, as a logical step following and completing Theorem 1.5.1:

Theorem 1.5.2 *Let V be a subspace of $\mathbb{R}^N$, and $y \in \mathbb{R}^N$. Then there exists a vector $v_m \in V$ such that $\|y - v_m\| \leq \|y - v\|$ for all $v \in V$.*

Any such v_m is called the *orthogonal projection* of y on V. In fact, the existence proof would not be difficult (see [18, pages 50-51]); it would be based essentially on the fact that any subspace of $\mathbb{R}^N$ is a *closed* set. However, here we shall be content with the fact that the orthogonality of $y - v_m$ to V is a sufficient condition for v_m to be a minimum, because, as we shall now

endowed with the Euclidean norm, is indeed a particular Hilbert space), and a cornerstone of infinite-dimensional analysis. If you are interested in these topics, you can find a crystal-clear explanation of the subject in the excellent book [18] (the proof of theorem 1.5.1 is taken from it with minor changes).

see, such orthogonality is precisely what the normal equations are asking for, and we have already established that a solution to the normal equations exists. Consider again the model

$$y_i = \varphi_i^\top \theta + \varepsilon_i, \quad i = 1, \cdots, N.$$

Let us stack the equations of the model on each other, by defining

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \qquad \Phi = \begin{bmatrix} \varphi_1^\top \\ \varphi_2^\top \\ \vdots \\ \varphi_N^\top \end{bmatrix}, \qquad E = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix},$$

where $Y, E \in \mathbb{R}^N$ and $\Phi \in \mathbb{R}^{N \times p}$. The model then reads

$$Y = \Phi\theta + E.$$

Let $v_1, \cdots, v_p$ be the *columns* of Φ (whereas the regressors are its *rows*). Then $V = \text{span}\ \{v_1, \cdots, v_p\}$ is a subspace of $\mathbb{R}^N$, and $v = \Phi\theta$ is a vector in V. The least squares problem asks to minimize $\|Y - \Phi\theta\|^2 = \|Y - v\|^2$ with respect to v, that is, to find $v_m = \Phi\hat{\theta}$ such that $\|Y - v_m\|^2$ is minimal.
Now we apply Theorem 1.5.1: if $Y - v_m \perp V$, then v_m is a minimizing vector. Explicitly, if

$$\begin{aligned} v_i \perp Y - \Phi\hat{\theta}, &\quad \text{that is,} \\ v_i^\top (Y - \Phi\hat{\theta}) = 0 &\quad \text{for all } i = 1, \cdots, p, \end{aligned} \tag{3}$$

then $\Phi\hat{\theta}$ is a minimizing vector, and $\hat{\theta} = [\hat{\theta}_1, \cdots, \hat{\theta}_p]^\top$ is a vector of coefficients such that $\Phi\hat{\theta} = \hat{\theta}_1 v_1 + \cdots + \hat{\theta}_p v_p$ is the orthogonal projection of Y on the space spanned by the columns $v_1, \cdots, v_p$ of Φ.

Stacking the rows $v_i^\top$ on each other we get $\Phi^\top$, hence stacking the equations (3) on each other we obtain:

$$\Phi^\top (Y - \Phi\hat{\theta}) = 0,$$

which finally yields

$$\Phi^\top \Phi\hat{\theta} = \Phi^\top Y. \tag{4}$$

Equation (4) is just another way to write the normal equations (1), because

$$\Phi^\top \Phi = \sum_{i=1}^{N} \varphi_i \varphi_i^\top \qquad \text{and} \qquad \Phi^\top Y = \sum_{i=1}^{N} \varphi_i y_i.$$

Since we know that a solution $\hat{\theta}_{\text{LS}}$ to the normal equations exists, now we also know that at least one minimizer of $\|Y - \Phi\theta\|^2$ exists. Note that, according to Theorem 1.5.1, the minimizer $v_m = \Phi\hat{\theta}_{\text{LS}}$ is unique. Of course this does *not* mean that $\hat{\theta}_{\text{LS}}$ is also unique! The uniqueness of $\hat{\theta}_{\text{LS}}$ holds if and only if Φ has full rank (it has linearly independent columns), and this in turn is the case if and only if $R = \Phi^\top \Phi$ is invertible.

1.6 Goodness of fit

Suppose that we are given a collection of numbers $y_1, y_2, \cdots, y_N$, but no explanatory variables (x). Can we still apply the method of least squares to "explain" the $\{y_i\}$? Actually *yes*, because (with a slight abuse of terminology) there exists a regressor that does not depend on any data, namely the constant $\varphi = 1$. The resulting problem is to minimize

$$\sum_{i=1}^{N} (y_i - 1 \cdot \theta)^2.$$

This is of course the same as minimizing

$$\frac{1}{N} \sum_{i=1}^{N} (y_i - \theta)^2,$$

which is the average square deviation of the $\{y_i\}$ from the number θ. The solution is

$$\hat{\theta}_{\text{LS}} = \left(\sum_{i=1}^{N} [\ 1\ ] \cdot [\ 1\ ]^{\top} \right)^{-1} \sum_{i=1}^{N} [\ 1\ ] \cdot y_i = \frac{1}{N} \sum_{i=1}^{N} y_i = \mathsf{M}\,[y].$$

Hence, the *sample average* is the number from which is minimal the squared deviation of the $\{y_i\}$[5]. The attained minimum deviation is the *sample variance* of the $\{y_i\}$:

$$\mathsf{S.Var}\,[y] = \frac{1}{N} \sum_{i=1}^{N} (y_i - \mathsf{M}\,[y])^2.$$

Since the regressor $\varphi(\cdot) = 1$ is always available irrespective of the nature of the explanatory data, it is often included in least squares problems. Now note that if $\varphi_1(x_i) = 1, \varphi_2(x_i), \cdots, \varphi_p(x_i)$ are the regressors, then there always exists a parameter $\hat{\theta}_m$ such that the corresponding sum of squares is exactly the sample variance of $\{y_i\}$ as in the trivial example shown above, namely $\hat{\theta}_m = [\mathsf{M}\,[y], 0, \cdots, 0]^{\top}$. It follows as the day the night, that any self-respecting least squares solution should behave at least *better* than $\hat{\theta}_m$,

[5] Actually, this is true in a far more general context. For any random variable X having mean and variance, it holds $\arg\min_{\theta \in \mathbb{R}} \mathsf{E}\left[(X - \theta)^2\right] = \mathsf{E}\,[X]$. And recall this fact from mechanics: the rotation axis that passes through the center of mass ($\simeq$ mean) of an object is, among all the parallel axes, the one that minimizes the moment of inertia ($\simeq$ integral of squares).

that is, it should attain a lower sum of squares:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}(y_i - \varphi_i^\top \hat{\theta}_{\mathrm{LS}})^2 &< \frac{1}{N}\sum_{i=1}^{N}(y_i - \varphi_i^\top \hat{\theta}_m)^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}(y_i - \mathsf{M}\left[y\right])^2 \\ &= \mathsf{S.Var}\left[y\right],\end{aligned}$$

otherwise the "explanatory variables" x_i would not be explaining anything. We call *residual variance* the quantity

$$\mathsf{RV} := \frac{1}{N}\sum_{i=1}^{N}(y_i - \varphi_i^\top \hat{\theta}_{\mathrm{LS}})^2$$

and *explained variance* the quantity

$$\begin{aligned}\mathsf{EV} &:= \mathsf{S.Var}\left[y\right] - \mathsf{RV} \\ &= \text{total variance} - \text{residual variance.}\end{aligned}$$

The above line of reasoning means that

$$0 \leq \mathsf{EV} \leq \mathsf{S.Var}\left[y\right], \tag{5}$$

where the first inequality should in fact be strict. The inequalities (5) are in absolute terms; dividing by $\mathsf{S.Var}\left[y\right]$ we get:

$$0 \leq \rho^2 \leq 1,$$

where the quantity

$$\rho^2 := \frac{\mathsf{EV}}{\mathsf{S.Var}\left[y\right]}$$

measures the fraction of the variance of the data $\{y_i\}$ that has been explained by the least squares fit[6]. One usually expresses ρ^2 as a percentage, and says something like "the model explains 90% of the variance of the data". In general, the closer ρ^2 to 1, the better the fit[7]. On one extreme, when $\rho^2 = 0$

[6] If you carry on with the above computations for the simple example of Section 1.3, you will find that

$$\rho^2 = \left(\frac{s_{xy}}{s_x s_y}\right)^2 = \left(\frac{\text{sample covariance}(\{x_i, y_i\})}{\text{standard deviation}(\{x_i\}) \times \text{standard deviation}(\{y_i\})}\right)^2.$$

Therefore, in that case ρ^2 is the square of what in statistics is called the *correlation coefficient* between $\{x_i\}$ and $\{y_i\}$, a quantity that varies between −1 and 1.

[7] Most books on applied statistics have a chapter on "multiple regression", which is basically linear least squares (meaning linear also in the explanatory data, i.e. with the regressors $1, x^{(1)}, x^{(2)}, x^{(3)}$) with further strong assumptions, typically Gaussianity. With these assumptions it is possible to pursue *hypothesis testing* on ρ^2, i.e. to establish when ρ^2 is close to 1 enough so that there is statistical evidence to support the hypothesis "y is adequately explained by a constant term plus the variables $x^{(1)}, x^{(2)}, x^{(3)}$". However, we shall not deal with these methods here.

(equivalently $\mathsf{EV} = 0$ or $\mathsf{RV} = \mathsf{S.Var}[y]$), the explanatory variables are not adding any information, and the least squares solution is not any more useful than a sample average; on the other extreme, when $\rho^2 = 1$ (equivalently $\mathsf{EV} = \mathsf{S.Var}[y]$ or $\mathsf{RV} = 0$), the y_i are explained *perfectly* by the x_i, that is, the equations $y_i = \varphi_i^\top \hat{\theta}_{\mathrm{LS}}$ are all verified *exactly.*

Thus, in general, the closer ρ^2 is to 1, the better the explanatory variables are doing their explanatory job, and this is usually a good sign. But there should be a big *caveat* here: one possible reason why ρ^2 is very close or even equal to 1 is that *there are just too many regressors*, that is $p \simeq N$. If this is the case, for example if we try to interpolate 100 points with a polynomial of order 95 (!), what will happen is that the solution $\hat{\theta}_{\mathrm{LS}}$ will describe the noise as well as the "true" function, and the corresponding polynomial will exhibit crazy oscillations trying to pass through all the 100 points. On the sole basis of the explained variance, such a model may appear very nice for the data at hand, but it will be practically useless to predict *future* data coming from the same source. This situation is called *over-fitting.* Remember: *the spirit of the least squares method is to use many measures to average out noise, not to identify many parameters!* Here we shall be content with common sense (interpolate with a polynomial of order at most, say, 5, not 95), but there is indeed an entire theory devoted to investigate what is the most reasonable order p of a model fitting a certain data set (Akaike Information Criterion, etc.).

1.7 Statistical properties

1.7.1 L.s. solution as the estimator of a "true" parameter

So far, the functional relation $y = \varphi(x)^\top \theta^o + \varepsilon$ has been a *model* that we have used to explain the data $\{(x_i, y_i)\}_{i=1}^N$ through regressor functions, in some way, but we have nowhere really pretended that a "true" θ^o actually exists. Up to now $\{x_i\}$, and consequently $\{\varphi_i\}$, have been just vectors, and the results were "algorithmic" in the sense that they guarantee the existence of a solution to the least squares problem, and tell us how to compute it; the numbers $\{\varepsilon_i\}$ have been there just to model "disturbances", without any particular requirement other than they be preferably small. In this section, we assume that there is indeed a θ^o, and that the data conform to the model $y = \varphi(x)^\top \theta^o + \varepsilon$. We assume, moreover, that $\{\varepsilon_i\}$ and possibly $\{\varphi_i\}$ are *random variables*, subject to some hypotheses, and examine the consequences on the asymptotic behavior of $\hat{\theta}_{\mathrm{LS}}$ (its behavior "for large N"). The results that we derive justify the claim that $\hat{\theta}_{\mathrm{LS}}$ is a good estimator of θ^o.

Let the data $\{(\varphi_i, y_i)\}$ be generated by the model

$$y_i = \varphi_i^\top \theta^o + \varepsilon_i \tag{6}$$

where $\theta^o \in \mathbb{R}^p$ is understood as the "true" parameter, deterministic but unknown. To see how $\hat{\theta}_{\mathrm{LS}}$ is related to θ^o, substitute (6) in the normal equations (1) (now we know that they admit a solution $\hat{\theta}_{\mathrm{LS}}$):

$$\begin{aligned}\left(\sum_{i=1}^N \varphi_i \varphi_i^\top\right) \hat{\theta}_{\mathrm{LS}} &= \sum_{i=1}^N \varphi_i y_i \\ &= \sum_{i=1}^N \varphi_i \left(\varphi_i^\top \theta^o + \varepsilon_i\right) \\ &= \left(\sum_{i=1}^N \varphi_i \varphi_i^\top\right) \theta^o + \sum_{i=1}^N \varphi_i \varepsilon_i\end{aligned}$$

or, dividing by N,

$$\left(\frac{1}{N}\sum_{i=1}^N \varphi_i \varphi_i^\top\right) \hat{\theta}_{\mathrm{LS}} = \left(\frac{1}{N}\sum_{i=1}^N \varphi_i \varphi_i^\top\right) \theta^o + \frac{1}{N}\sum_{i=1}^N \varphi_i \varepsilon_i$$

In practical situations, we expect, or at least we wish, two things:

1. that the matrix $\frac{1}{N}\sum_{i=1}^N \varphi_i \varphi_i^\top$ becomes (and remains) invertible for $N \geq \bar{N}$, due to the fact that the φ_i carry more and more information; if so, for big N the solution $\hat{\theta}_{\mathrm{LS}}$ is unique, and we can write

$$\hat{\theta}_{\mathrm{LS}} = \theta^o + \left(\frac{1}{N}\sum_{i=1}^N \varphi_i \varphi_i^\top\right)^{-1} \frac{1}{N}\sum_{i=1}^N \varphi_i \varepsilon_i;$$

2. that the vector $\frac{1}{N}\sum_{i=1}^N \varphi_i \varepsilon_i$ tends to 0 as $N \to \infty$, thus attaining

$$\hat{\theta}_{\mathrm{LS}} \to \theta^o.$$

   For this to hold, we will require in some way that $\varphi_i \varepsilon_i$ has zero mean. And in order to satisfy the latter requirement we will assume, in turn, that ε_i has zero mean, and that φ_i is either deterministic (hence $\mathsf{E}[\varphi_i \varepsilon_i] = \varphi_i \mathsf{E}[\varepsilon_i] = 0$) or random, but independent of ε_i (hence $\mathsf{E}[\varphi_i \varepsilon_i] = \mathsf{E}[\varphi_i]\,\mathsf{E}[\varepsilon_i] = 0$) The strong law(s) of large numbers will do the rest.

These ideas conform to the principle that underlies least squares optimization:

noise tends to get averaged out and lose importance as the number of measures increases; consequently, it is usually better to take many noisy measures, than to take one single precise measure.

We can view the least squares algorithm as a function

$$\mathrm{LS} : ((\varphi_1, y_1), \cdots (\varphi_N, y_N)) \mapsto \hat{\theta}_{\mathrm{LS}}$$

used to estimate θ^o; in statistical jargon, this is called an *estimator* of θ^o. We will now show that, under two different set of hypotheses (fairly general, although with important exceptions), this estimator is unbiased and consistent. Actually, the following theorems are just two examples to show how the law of large numbers applies to least squares; more general results can be established as well.

1.7.2 Random regressors independent of the noise

The following result shows that if the noise terms have mean zero, if the regressors are *random* but independent of the noise, then the least squares estimator is unbiased and consistent.

Theorem 1.7.1 *Suppose that $\{y_i\}_{i=1}^{\infty}$ are generated by the model*

$$y_i = \varphi_i^\top \theta^o + \varepsilon_i.$$

Suppose, in addition, that

1. *$\{\varphi_i\}_{i=1}^{\infty}$ are independent and identically distributed random vectors, with correlation matrix $\Sigma := \mathsf{E}\left[\varphi_i \varphi_i^\top\right] > 0$;*

2. *$\{\varepsilon_i\}_{i=1}^{\infty}$ are independent and identically distributed random variables, with mean $\mathsf{E}\left[\varepsilon_i\right] = 0$;*

3. *ε_i is independent of φ_i for all i.*

Then

1. *if $\hat{\theta}_{\mathrm{LS}}$ exists and is unique for a certain N, then $\mathsf{E}\left[\hat{\theta}_{\mathrm{LS}}\right] = \theta^o$;*

2. *$\hat{\theta}_{\mathrm{LS}} \to \theta^o$ almost surely as $N \to \infty$.*

Proof.

1. Suppose that $\hat{\theta}_{\mathrm{LS}}$ exists and is unique:

$$\hat{\theta}_{\mathrm{LS}} = \left(\sum_{i=1}^{N} \varphi_i \varphi_i^\top \right)^{-1} \sum_{i=1}^{N} \varphi_i y_i;$$

substitute the model $y_i = \varphi_i^\top \theta^o + \varepsilon_i$ in the solution:

$$\begin{aligned}
\hat{\theta}_{\text{LS}} &= \left(\sum_{i=1}^N \varphi_i \varphi_i^\top \right)^{-1} \sum_{i=1}^N \varphi_i (\varphi_i^\top \theta^o + \varepsilon_i) \\
&= \theta^o + \left(\sum_{i=1}^N \varphi_i \varphi_i^\top \right)^{-1} \sum_{i=1}^N \varphi_i \varepsilon_i \\
&= \theta^o + \sum_{i=1}^N \left(\sum_{j=1}^N \varphi_j \varphi_j^\top \right)^{-1} \varphi_i \varepsilon_i;
\end{aligned}$$

then

$$\begin{aligned}
\mathsf{E}\left[\hat{\theta}_{\text{LS}}\right] &= \theta^o + \mathsf{E}\left[\sum_{i=1}^N \left(\sum_{j=1}^N \varphi_j \varphi_j^\top \right)^{-1} \varphi_i \varepsilon_i \right] \\
&= \theta^o + \sum_{i=1}^N \mathsf{E}\left[\left(\sum_{j=1}^N \varphi_j \varphi_j^\top \right)^{-1} \varphi_i \right] \mathsf{E}\left[\varepsilon_i\right] = \theta^o.
\end{aligned}$$

2. From the first hypothesis, we see that the components $[\varphi_i \varphi_i^\top]_{hk}$ are independent and identically distributed, and that $\mathsf{E}\left[[\varphi_i \varphi_i^\top]_{hk}\right] = \Sigma_{hk}$. Therefore, by the strong law of large numbers (Theorem D.7.2), $\frac{1}{N}\sum_{i=1}^N [\varphi_i \varphi_i^\top]_{hk} \to \Sigma_{hk}$ almost surely. Hence, the same fact holds for the entire covariance matrix:

$$\frac{1}{N} \sum_{i=1}^N \varphi_i \varphi_i^\top \to \Sigma > 0 \quad \text{almost surely.}$$

This implies that almost surely, from a certain $\bar{N}$ onwards, the matrix $\frac{1}{N}\sum_{i=1}^N \varphi_i \varphi_i^\top$ is positive definite; indeed matrices sufficiently close to a positive definite matrix are themselves positive definite. Hence, almost surely from a certain $\bar{N}$ onwards the sum is invertible, so that we can write

$$\hat{\theta}_{\text{LS}} = \theta^o + \left(\frac{1}{N} \sum_{i=1}^N \varphi_i \varphi_i^\top \right)^{-1} \frac{1}{N} \sum_{i=1}^N \varphi_i \varepsilon_i \quad \text{for all } N \geq \bar{N}$$

Now by the second and third hypotheses, the vectors $\varphi_i \varepsilon_i$ are independent and identically distributed, with mean

$$\mathsf{E}\left[\varphi_i \varepsilon_i\right] = \mathsf{E}\left[\varphi_i\right] \mathsf{E}\left[\varepsilon_i\right] = 0.$$

Consequently, by the strong law of large numbers (Theorem D.7.2),

$$\frac{1}{N} \sum_{i=1}^N \varphi_i \varepsilon_i \to 0 \quad \text{almost surely.}$$

Therefore,

$$\begin{aligned}\lim_{N\to\infty} \hat{\theta}_{\mathrm{LS}} &= \theta^o + \left(\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \varphi_i \varphi_i^\top\right)^{-1} \left(\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \varphi_i \varepsilon_i\right) \\ &= \theta^o + \Sigma^{-1} \cdot 0 = \theta^o\end{aligned}$$

almost surely; this establishes the claim.

□

1.7.3 Deterministic regressors

The following result shows that, if the regressors are deterministic and the noise terms have mean zero, then the least squares estimator is unbiased and consistent.

Theorem 1.7.2 *Suppose that $\{y_i\}_{i=1}^\infty$ are generated by the model*

$$y_i = \varphi_i^\top \theta^o + \varepsilon_i.$$

Suppose, in addition, that

1. *$\{\varphi_i\}_{i=1}^\infty$ are deterministic vectors, satisfying*

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} \varphi_i \varphi_i^\top \geq aI \quad &\textit{for all } N \geq \bar{N} \\ \|\varphi_i\|^2 \leq A \quad &\textit{for all } i\end{aligned}$$

   where a, A are real numbers, $0 < a \leq A < \infty$;

2. *$\{\varepsilon_i\}_{i=1}^\infty$ are independent random variables, not necessarily identically distributed, but such that, for all i, $\mathsf{E}[\varepsilon_i] = 0$ and $\mathsf{E}\left[\varepsilon_i^2\right] \leq c$ for a certain constant $c \in \mathbb{R}$.*

Then

1. *if $\hat{\theta}_{\mathrm{LS}}$ exists and is unique for a certain N, then $\mathsf{E}\left[\hat{\theta}_{\mathrm{LS}}\right] = \theta^o$;*

2. *$\hat{\theta}_{\mathrm{LS}} \to \theta^o$ almost surely as $N \to \infty$.*

Proof.

1. Suppose that $\hat{\theta}_{\mathrm{LS}}$ exists and is unique:

$$\hat{\theta}_{\mathrm{LS}} = \left(\sum_{i=1}^{N} \varphi_i \varphi_i^\top\right)^{-1} \sum_{i=1}^{N} \varphi_i y_i;$$

substitute the model $y_i = \varphi_i^\top \theta^o + \varepsilon_i$ in the solution:

$$\hat{\theta}_{\text{LS}} = \theta^o + \left(\sum_{i=1}^{N} \varphi_i \varphi_i^\top \right)^{-1} \sum_{i=1}^{N} \varphi_i \varepsilon_i;$$

then

$$\begin{aligned} \mathsf{E}\left[\hat{\theta}_{\text{LS}}\right] &= \theta^o + \mathsf{E}\left[\left(\sum_{i=1}^{N} \varphi_i \varphi_i^\top \right)^{-1} \sum_{i=1}^{N} \varphi_i \varepsilon_i \right] \\ &= \theta^o + \left(\sum_{i=1}^{N} \varphi_i \varphi_i^\top \right)^{-1} \sum_{i=1}^{N} \varphi_i \mathsf{E}\left[\varepsilon_i\right] \end{aligned}$$

$$= \theta^o.$$

2. From the first hypothesis, for all $N \geq \bar{N}$

$$\frac{1}{N} \sum_{i=1}^{N} \varphi_i \varphi_i^\top \geq aI > 0$$

which is invertible; hence the matrix

$$\left(\frac{1}{N} \sum_{i=1}^{N} \varphi_i \varphi_i^\top \right)^{-1} \leq (aI)^{-1} = \frac{1}{a} I$$

is deterministic *and bounded*; for $N \geq \bar{N}$ we can write

$$\hat{\theta}_{\text{LS}} = \theta^o + \left(\frac{1}{N} \sum_{i=1}^{N} \varphi_i \varphi_i^\top \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} \varphi_i \varepsilon_i$$

Now let φ_i^k be the k-th component of the vector φ_i; from the first hypothesis it holds $(\varphi_i^k)^2 \leq \|\varphi_i\|^2 \leq A$. Hence, exploiting also the second hypothesis and recalling that φ_i^k is deterministic,

$$\begin{gathered} \mathsf{E}\left[\varphi_i^k \varepsilon_i\right] = \varphi_i^k \mathsf{E}\left[\varepsilon_i\right] = 0 \\ \mathsf{E}\left[\left(\varphi_i^k \varepsilon_i\right)^2\right] = \left(\varphi_i^k\right)^2 \mathsf{E}\left[\varepsilon_i^2\right] \leq Ac \end{gathered}$$

for all i, and applying the strong law of large numbers (Theorem D.7.3):

$$\frac{1}{N} \sum_{i=1}^{N} \varphi_i^k \varepsilon_i \to 0 \quad \text{almost surely (component-wise);}$$

turning back to vectors,

$$\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varepsilon_i \to 0 \quad \text{almost surely.}$$

Therefore, recalling that $\left(\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varphi_i^\top\right)^{-1}$ is bounded, we conclude $\hat{\theta}_{\mathrm{LS}} \to \theta^o$ almost surely.

□

To complete the picture we show that, if φ_i's are deterministic and the errors are independent with the same variance, convergence holds also in the mean-square sense. This result is somewhat weaker than Theorem 1.7.2, but its proof is simple and instructive.

Theorem 1.7.3 *Suppose that $\{y_i\}_{i=1}^{\infty}$ are generated by the model*

$$y_i = \varphi_i^\top\theta^o + \varepsilon_i.$$

Suppose, in addition, that

1. *$\{\varphi_i\}_{i=1}^{\infty}$ are deterministic vectors, satisfying*

$$\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varphi_i^\top \geq aI \quad \textit{for all } N \geq \bar{N}$$

where $a > 0$;

2. *$\{\varepsilon_i\}_{i=1}^{\infty}$ are independent random variables, with mean 0 and variance σ^2.*

Then $\hat{\theta}_{\mathrm{LS}} \to \theta^o$ in the mean square as $N \to \infty$.

Proof. It is convenient, here, to use the compact notation of Section 1.5:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \qquad \Phi = \begin{bmatrix} \varphi_1^\top \\ \varphi_2^\top \\ \vdots \\ \varphi_N^\top \end{bmatrix}, \qquad E = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix},$$

First, recall that the model then reads

$$Y = \Phi\theta^o + E,$$

that

$$\Phi^\top\Phi = \sum_{i=1}^{N}\varphi_i\varphi_i^\top \qquad \text{and} \qquad \Phi^\top Y = \sum_{i=1}^{N}\varphi_i y_i,$$

and that the least squares solution, assuming that $\Phi^\top\Phi$ is invertible, is

$$\hat{\theta}_{\text{LS}} = \left(\Phi^\top\Phi\right)^{-1}\Phi^\top Y.$$

Second, note that for $N \geq \bar{N}$ the least squares estimator is *unbiased.* Indeed, by the first hypothesis $\Phi^\top\Phi$ is indeed invertible for $N \geq \bar{N}$, hence for big N

$$\begin{aligned}\hat{\theta}_{\text{LS}} &= \left(\Phi^\top\Phi\right)^{-1}\Phi^\top Y = \left(\Phi^\top\Phi\right)^{-1}\Phi^\top\left(\Phi\theta^o + E\right)\\ &= \theta^o + \left(\Phi^\top\Phi\right)^{-1}\Phi^\top E.\end{aligned}$$

By the first hypothesis Φ is deterministic, and by the second the mean of E is zero, hence

$$\mathsf{E}\left[\hat{\theta}_{\text{LS}}\right] = \theta^o + \left(\Phi^\top\Phi\right)^{-1}\Phi^\top\mathsf{E}\left[E\right] = \theta^o.$$

Third, note that the covariance matrix of E is, by the second hypothesis, $\sigma^2 I$; hence

$$\begin{aligned}\mathsf{Var}\left[\hat{\theta}_{\text{LS}}\right] &= \mathsf{E}\left[(\hat{\theta}_{\text{LS}} - \theta^o)(\hat{\theta}_{\text{LS}} - \theta^o)^\top\right]\\ &= \mathsf{E}\left[\left(\left(\Phi^\top\Phi\right)^{-1}\Phi^\top E\right)\left(\left(\Phi^\top\Phi\right)^{-1}\Phi^\top E\right)^\top\right]\\ &= \left(\Phi^\top\Phi\right)^{-1}\Phi^\top\mathsf{E}\left[EE^\top\right]\Phi\left(\Phi^\top\Phi\right)^{-1}\\ &= \left(\Phi^\top\Phi\right)^{-1}\Phi^\top\ \sigma^2 I\ \Phi\left(\Phi^\top\Phi\right)^{-1}\\ &= \sigma^2\left(\Phi^\top\Phi\right)^{-1}.\end{aligned}$$

Finally, by the first hypothesis, $\left(\Phi^\top\Phi\right) \geq aNI$, hence $\left(\Phi^\top\Phi\right)^{-1} \leq \frac{1}{aN}I$. Therefore $\lim_{N\to\infty}\mathsf{Var}\left[\hat{\theta}_{\text{LS}}\right] = 0$, and this is enough to establish the claim. □

1.7.4 Instrumental variables

If the regressor and the disturbance are both random, *and correlated*, none of the theorems about almost sure convergence of $\hat{\theta}_{\text{LS}}$ apply, and in general there is no way out of this issue within the standard theory of least squares. However, here we sketch a remedy, called the method of *instrumental variables.* The trick is this: consider the equation from which the least squares solution $\hat{\theta}_{\text{LS}}$ has been derived:

$$\sum_{i=1}^{N}\varphi_i\left(y_i - \varphi_i^\top\hat{\theta}\right) = 0;$$

divide by N and substitute the "true" model (20) into it:

$$\frac{1}{N}\sum_{i=1}^{N}\varphi_i\left(\varphi_i^\top(\theta^o-\hat{\theta})+\varepsilon_i\right)=0,$$

that is

$$\left(\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varphi_i^\top\right)\hat{\theta}=\left(\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varphi_i^\top\right)\theta^o+\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varepsilon_i.$$

The idea is that, in the limit, this equation becomes

$$\mathsf{E}\left[\varphi_i\left(\varphi_i^\top(\theta^o-\hat{\theta})+\varepsilon_i\right)\right]=0,$$

that is

$$\mathsf{E}\left[\varphi_i\varphi_i^\top\right]\hat{\theta}=\mathsf{E}\left[\varphi_i\varphi_i^\top\right]\theta^o+\mathsf{E}\left[\varphi_i\varepsilon_i\right],$$

and if two conditions hold:

- $\mathsf{E}\left[\varphi_i\varphi_i^\top\right]$ is invertible, and
- $\mathsf{E}\left[\varphi_i\varepsilon_i\right]=0$,

then its only solution is $\hat{\theta}=\theta^o$. If the regressors and the disturbance are correlated this is not the case, because the second condition fails. But if we replace (somehow heuristically) the first occurrence of the regressor φ_i with *another* variable ψ_i such that

- $\mathsf{E}\left[\psi_i\varphi_i^\top\right]$ is invertible, and
- $\mathsf{E}\left[\psi_i\varepsilon_i\right]=0$,

then the equation

$$\mathsf{E}\left[\psi_i\left(\varphi_i^\top(\theta^o-\hat{\theta})+\varepsilon_i\right)\right]=0$$

has again the only solution $\hat{\theta}=\theta^o$; therefore it makes sense to try solving

$$\left(\sum_{i=1}^{N}\psi_i\varphi_i^\top\right)\hat{\theta}=\sum_{i=1}^{N}\psi_i y_i$$

instead of the normal equations. The variable ψ_i is called *instrumental*; namely, an instrumental variable is any variable correlated with the data but *not* with the noise. Note that the square matrix in the left-hand side tends to be invertible, but it is not symmetric anymore. Note, above all, that this method is not anymore pursuing the minimization of a sum of squares. For a full treatment of the subject, refer to [17, chap. 7].

1.8 Exercises for Chapter 1

Problem 1 (price of train tickets).
The following table contains the lengths of the railway connections between the Milano Centrale train station and the central station of other cities on the way from Milan to Venice[8], and the corresponding prices of a 'regional' train ticket[9]:

Connection	**Length (km)**	**Ticket price (€)**
Milano C.le → Brescia	82.842	7.00
Milano C.le → Verona P.N.	147.480	11.55
Milano C.le → Padova	229.408	15.65
Milano C.le → Venezia S.L.	266.341	18.35

Suppose that the prices are explained by a linear model comprising a fixed price due to administrative costs plus a price proportional to the length of the connection. The prices are "noisy" because they are quantized to multiples of 5 €-cents.

1. Write down the linear model for the ticket price and the normal equations of the least squares method, and find an estimate of the fixed price and of the proportionality coefficient.

2. Estimate the price of a ticket from Milano to Vicenza, knowing that the railway between these cities is 199.138 km long.

Problem 2 (amplitude and phase of a sinusoid).
The following table contains 10 noisy measures of the values of a sinusoidal signal $y(t) = A\sin(2\pi Ft + \phi)$, taken at random times t_i:

t_i	2.188	3.043	4.207	4.937	5.675	6.104	6.260	7.150	8.600	9.655
y_i	-1.112	2.358	-1.807	1.202	-0.814	1.298	-2.520	-0.132	1.421	-0.302

We know the frequency $F = 2$ Hz of the signal, but we do not know its amplitude A and its phase ϕ.

1. Show how to apply the method of *linear* least squares in order to find an estimate of A and ϕ.
   Hint: recall trigonometry (!).

2. Compute the estimates $\hat{A}$ and $\hat{\phi}$.

[8] Retrieved from `http://it.wikipedia.org/wiki/Ferrovia_Milano-Venezia`

[9] Retrieved from `http://www.trenitalia.com` on February 25, 2013.

Problem 3 (weighted least squares).
Given some measures $(\varphi_1, y_1), \cdots, (\varphi_N, y_N)$, the canonical Least Squares estimate is the vector in $\mathbb{R}^p$ minimizing the sum of the squared residuals:

$$\hat{\theta}_{\text{LS}} := \arg\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^{N} (y_i - \varphi_i^\top \theta)^2.$$

Now suppose that we want to give more importance to some of the errors, and less to others, minimizing a *weighted* sum of squared residuals instead:

$$\hat{\theta}_{\text{WLS}} := \arg\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^{N} w_i \, (y_i - \varphi_i^\top \theta)^2,$$

where $w_i \geq 0$ for $i = 1, \cdots, N$.

1. Find the corresponding version of the normal equations.

2. Find the new Weighted Least Squares estimate in terms of the matrices

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \qquad \Phi = \begin{bmatrix} \varphi_1^\top \\ \vdots \\ \varphi_N^\top \end{bmatrix},$$

   assuming full rank whenever necessary.
   Hint: bring into the picture a new matrix containing the weights w_i.

Problem 4 (ranges).
Let Φ be a real matrix. Show that

$$\begin{aligned} \text{range } \Phi^\top \Phi &= \text{range } \Phi^\top \\ \text{rank } \Phi^\top \Phi &= \text{rank } \Phi^\top \end{aligned}$$

Hint: rewrite the proof of a Lemma about the existence of a solution to the least squares problem, but keeping it in terms of Φ and $\Phi^\top$.

Problem 5 (systematic errors).
Let the measures $\{y_i\}_{i=1}^N$ be generated according to the model $y_i = \varphi_i^\top \theta^o + \varepsilon_i$, and suppose that:

1. ε_i are independent Gaussian variables with mean μ and variance σ^2;

2. φ_i are independent and identically distributed vectors with mean $\bar{\varphi}$ and second-order moment $\Sigma = \mathsf{E}\left[\varphi_i \varphi_i^\top\right] > 0$;

3. ε_i is independent of φ_i for all i.

Does the least square estimate converge almost surely as $N \to \infty$? If so, to what does it converge?

2 Linear algebraic aspects of the least squares method

This chapter was originally conceived for a brief and fast-paced introduction to singular value decompositions, pseudo-inverses, and their applications to least squares approximation, to be covered in two or three lectures at the end of the course. It is meant as a "practical" reference, not as a rigorous one; therefore the proofs of the fundamental theorems, although crucial for a deep understanding, are omitted. The reader interested in these, and in a more detailed exposition, is referred to the Italian textbook [29], or to the standard (and quite more sophisticated) references [14], [12].
Moreover, the exposition is geared towards real matrices; however, everything can be proved in full generality for arbitrary complex matrices, with minimal conceptual differences.

2.1 The singular value decomposition

2.1.1 Motivation and definition

Real symmetric matrices, and in particular positive definite or semi-definite matrices, are those that allow for the most useful and intuitive decomposition: Any such matrix is indeed the product of three *real* matrices, one diagonal and the other two orthogonal and transposes of each other: $A = M\Lambda M^\top$. This decomposition has a very rich structure, and readily provides the eigenvalues of the matrix, and a basis of $\mathbb{R}^n$ made of corresponding eigenvectors. Moreover, in this case there are robust algorithms readily available, to compute such decomposition numerically.
Ideally, one would like to provide such a decomposition, or a similar one, for all real matrices. Unfortunately,

- There exist real normal matrices which are not symmetric, and whose eigenvalues are not real (they are, however, pairwise conjugate). The prototypical example of such a matrix is

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}.$$

  Here $A^\top = -A$ ("anti-symmetric"), so A is indeed normal, but its eigenvalues are $\pm j$.

- There exist square matrices which are not normal, but are still diagonalizable with a non-orthogonal M. For example,

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} = M\Lambda M^{-1}.$$

  (As a more general example, any matrix with *distinct* eigenvalues admits at least this decomposition.)

- There exist square matrices which are not diagonalizable at all. For example,
$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$
is not diagonalizable. There is a general decomposition that encompasses even these cases, namely the Jordan decomposition, but we shall not deal with it here.

- Finally, there exist non-square matrices (!). For these, it does not even make sense to talk about "diagonalization"; at least, it does not at first sight.

The most useful decomposition in linear algebra "resembles" the spectral decomposition of a positive semidefinite matrix: namely, it is a product of three matrices, where the second is again "diagonal", in a certain sense, and the first and third ones are again orthogonal; the main difference is that the latter are not necessarily related to each other.
Let $A \in \mathbb{R}^{m\times n}$. A *Singular Value Decomposition*, or *SVD*, of A, is a decomposition
$$A = U\Sigma V^\top,$$
where

- U is an orthogonal $m \times m$ matrix;
- V is an orthogonal $n \times n$ matrix;
- Σ is an $m \times n$ matrix (same dimensions as A; compatible dimensions with U and $V^\top$) with the following structure:
$$\Sigma = \begin{bmatrix} \sigma_1 & & & & & & \\ & \sigma_2 & & & & & \\ & & \ddots & & & & \\ & & & \sigma_k & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \\ & & & & & & \end{bmatrix}$$
where all the blank entries are meant to be zeros. In fact, all the elements of Σ are zero except for the first k (where $k \leq m, k \leq n$) on the main diagonal. The entries $\sigma_1, \cdots, \sigma_k$ are supposed to be strictly positive real numbers, and it is usually assumed that they come in decreasing order: $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k > 0$.

The numbers $\sigma_1, \cdots, \sigma_k$ are called the *singular values* of A. The main result about SVD's is the following

Theorem 2.1.1 *Every matrix $A \in \mathbb{R}^{m \times n}$ admits a singular value decomposition. Moreover, the matrix Σ appearing in such decomposition is uniquely determined by A (hence, of course, so are the singular values of A).*

We stress the fact that Σ is uniquely determined (in particular, due to the fact that the singular values appear in decreasing order), but U and V are not. As a trivial counterexample, for any orthogonal matrix U the decomposition $I = UIU^\top$ is a perfectly valid SVD of the identity matrix having $V = U$. Note, also, that any orthogonal diagonalization of a positive semi-definite real matrix,

$$A = M\Lambda M^\top$$

is also a valid SVD of A having $U = V = M$, provided that the columns of M (orthonormal eigenvectors of A) are sorted in such a way that the corresponding eigenvalues (≥ 0) appear in decreasing order on the diagonal of Λ.

2.1.2 Interpretation

What is the meaning of an SVD? Multiplying the decomposition $A = U\Sigma V^\top$ by V on the right-hand side we obtain

$$AV = U\Sigma$$

For example, suppose that A is a "tall" matrix ($m > n$), that $\{v_1, \cdots, v_n\}$ are the columns of V (and an orthonormal basis of $\mathbb{R}^n$), that $\{u_1, \cdots, u_m\}$ are the columns of U (and an orthonormal basis of $\mathbb{R}^m$), and visualize what is going on:

$$\begin{bmatrix} & & \\ & A & \\ & & \end{bmatrix}\begin{bmatrix} \vdots & \vdots & \\ v_1 & v_2 & \\ \vdots & \vdots & \end{bmatrix} = \begin{bmatrix} \vdots & \vdots & \\ u_1 & u_2 & \\ \vdots & \vdots & \\ & & \end{bmatrix}\begin{bmatrix} \sigma_1 & & & & \\ & \sigma_2 & & & \\ & & \ddots & & \\ & & & \sigma_k & \\ & & & & 0 \end{bmatrix}$$

Comparing both sides column by column,

$$
\begin{aligned}
Av_1 &= U \begin{bmatrix} \sigma_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \sigma_1 u_1 \\
Av_2 &= U \begin{bmatrix} 0 \\ \sigma_2 \\ \vdots \\ 0 \end{bmatrix} = \sigma_2 u_2 \\
Av_3 &= \sigma_3 u_3 \\
&\vdots \\
Av_k &= \sigma_k u_k \\
Av_{k+1} &= 0 \\
&\vdots \\
Av_n &= 0
\end{aligned}
$$

The first k vectors $\{v_1, \cdots, v_k\}$ are mapped by A onto multiples of $\{u_1, \cdots, u_k\}$; the others are mapped to zero. In other terms,

- $\{u_1, \cdots, u_k\}$ is an orthonormal basis of range A; its dimension is k, hence k is the rank of A;
- $\{v_{k+1}, \cdots, v_n\}$ is an orthonormal basis of null A, whose dimension is $n-k$.

You can readily recognize a property that you already know from linear algebra courses:

$n = \dim$ (domain of A) $= \dim$ (range of A)$+\dim$ (null space of A) $= k+(n-k)$

Lemma 2.1.1 *Given two arbitrary matrices A and B with compatible dimensions, the matrices AB and BA have the same* non-zero *eigenvalues, and with the same multiplicities. (The multiplicity of any zero eigenvalue is different if A and B are not square.)*

One can show that an SVD of $A \in \mathbb{R}^{m\times n}$ can be constructed from a suitable choice of eigenvectors of the symmetric, positive semi-definite matrices $AA^\top \in \mathbb{R}^{m\times m}$ and $A^\top A \in \mathbb{R}^{n\times n}$, which have the same positive eigenvalues by Lemma 2.1.1, and that the singular values of A are precisely the square roots of those eigenvalues. We shall not deal with the details of such construction, but we can easily check the correspondences. Let $A = U\Sigma V^\top$ be an SVD of $A \in \mathbb{R}^{m \times n}$. Then

$$\begin{aligned} AA^\top &= U\Sigma V^\top V \Sigma^\top U^\top \\ &= U\Sigma\Sigma^\top U^\top \\ &= U\Lambda_m U^\top \end{aligned}$$

where $\Lambda_m \in \mathbb{R}^{m \times m}$, namely

$$\begin{aligned} \Lambda_m = \Sigma\Sigma^\top &= \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_k & \\ & & & 0 \\ 0 & \cdots & 0 & 0 \end{bmatrix} \begin{bmatrix} \sigma_1 & & & & 0 \\ & \ddots & & & \vdots \\ & & \sigma_k & & 0 \\ & & & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1^2 & & & & \\ & \ddots & & & \\ & & \sigma_k^2 & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix} \end{aligned}$$

Similarly, $A^\top A = V\Sigma^\top U^\top U \Sigma V^\top = V\Lambda_n V^\top$ where $\Lambda_n \in \mathbb{R}^{n \times n}$, namely

$$\begin{aligned} \Lambda_n = \Sigma^\top\Sigma &= \begin{bmatrix} \sigma_1 & & & & 0 \\ & \ddots & & & \vdots \\ & & \sigma_k & & 0 \\ & & & 0 & 0 \end{bmatrix} \begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_k & \\ & & & 0 \\ 0 & \cdots & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1^2 & & & \\ & \ddots & & \\ & & \sigma_k^2 & \\ & & & 0 \end{bmatrix} \end{aligned}$$

What we have written down are orthogonal diagonalizations of $AA^\top$ and $A^\top A$ respectively, hence we can conclude:

- The columns of U form an orthonormal basis of $\mathbb{R}^m$ made of eigenvectors of $AA^\top$;

- The columns of V form an orthonormal basis of $\mathbb{R}^n$ made of eigenvectors of $A^\top A$;

- The squares of the singular values $\sigma_1, \cdots, \sigma_k$ are the non-zero eigenvalues of both $AA^\top$ and $A^\top A$.

Note: the set of singular values of A is strictly related to the spectrum of $AA^\top$, but in general *not* to the spectrum of A itself. Unless A is symmetric or has some other nice structure, you can draw few conclusions on its singular values from the sole knowledge of its eigenvalues. For example, the only eigenvalue of the matrix

$$A = \begin{bmatrix} 0 & M \\ 0 & 0 \end{bmatrix},$$

is $\lambda = 0$ irrespective of the arbitrary number M; nevertheless, the largest singular value is precisely $|M|$.

2.1.3 Matrix norms

A *norm* on a vector space V is, in general, a function $\|\cdot\| : V \to \mathbb{R}$ with the following properties:

1. $\|A\| \geq 0$ for all $A \in V$, and $\|A\| = 0$ if and only if $A = 0$;
2. $\|aA\| = |a|\|A\|$ for all $A \in V$ and $a \in \mathbb{C}$;
3. $\|A + B\| \leq \|A\| + \|B\|$ for all $A, B \in V$ (triangular inequality).

Since any matrix $A \in \mathbb{R}^{m\times n}$ can be thought of as an element of a real vector space of dimension mn (it makes no essential difference whether we stack mn numbers in a column or we arrange them in a rectangle), it makes sense to apply standard norms to matrices as well. But there is more. A norm which has the further property

4. $\|AB\| \leq \|A\|\|B\|$,

for all matrices A, B of dimensions such that the product makes sense, is called a *matrix norm.*
The following is a matrix norm:

$$\|A\|_2 := \sup_{x\in\mathbb{R}^n} \frac{\|Ax\|_2}{\|x\|_2} = \sup_{x\in\mathbb{R}^n, \|x\|_2=1} \|Ax\|_2$$

It is called the “sup norm” on $\mathbb{R}^{m\times n}$ induced by the Euclidean norm in $\mathbb{R}^n$. Intuitively, $\|A\|_2$ is the “maximum amplification” that A can operate on the Euclidean norm of a vector. If $A = U\Sigma V^\top$ is an SVD of $A \in \mathbb{R}^{m\times n}$, then

$$\begin{aligned}
\|A\|_2 &= \sup_{x\in\mathbb{R}^n, \|x\|_2=1} \|Ax\|_2 = \sup_{x\in\mathbb{R}^n, \|x\|_2=1} \|U\Sigma V^\top x\|_2 \\
&= \sup_{x\in\mathbb{R}^n, \|V^\top x\|_2=1} \|\Sigma V^\top x\|_2 = \sup_{z\in\mathbb{R}^n, \|z\|_2=1} \|\Sigma z\|_2 \\
&= \sup_{z_1^2+\cdots+z_n^2=1} \sqrt{\sigma_1^2 z_1^2 + \cdots + \sigma_k^2 z_k^2 + 0 z_{k+1}^2 + \cdots + 0 z_n^2}
\end{aligned}$$

The third inequality is due to the fact that the orthogonal matrices U and $V^\top$ do not change the Euclidean norm, and the fourth to the fact that if x varies over the unit sphere, so does $z = V^\top x$. Now, since $\sigma_1 \geq \cdots \geq \sigma_k$, you can easily see that the supremum is attained when $z_1 = 1$, and all the other $z_i = 0$. Hence,

$$\|A\|_2 = \sigma_1.$$

In words: The sup norm of A induced by the Euclidean norm coincides with the greatest singular value of A.
The following quantity, defined for $A \in \mathbb{R}^{m\times n}$:

$$\|A\|_F := \sqrt{\sum_{i,j} A_{i,j}^2} = \sqrt{\text{tr } AA^\top}$$

is called the *Frobenius norm* of A. It is indeed a matrix norm, but one can easily see that it coincides with the Euclidean norm on $\mathbb{R}^{m\times n}$. Substituting an SVD of A,

$$\begin{aligned}\|A\|_F &= \sqrt{\text{tr } AA^\top} = \sqrt{\text{tr } U\Sigma V^\top V\Sigma^\top U^\top}\\ &= \sqrt{\text{tr } U\Sigma\Sigma^\top U^\top} = \sqrt{\text{tr } \Sigma^\top U^\top U\Sigma}\\ &= \sqrt{\text{tr } \Sigma^\top\Sigma} = \sqrt{\sigma_1^2 + \cdots + \sigma_k^2}\end{aligned}$$

where the fourth equality is due to the "rotation property" of the trace: tr AB = tr BA for any two matrices A, B with compatible dimensions. One of the facts that make singular value decompositions so useful is the following

Theorem 2.1.2 *Let $A = U\Sigma V^\top$ be an SVD of $A \in \mathbb{R}^{m\times n}$. Suppose that k is the rank of A, and let r be an integer less than k. The minimum*

$$\min_{\bar{A}\in\mathbb{R}^{m\times n},\ \text{rank } \bar{A}=r} \|A - \bar{A}\|_F$$

is attained by the matrix

$$\bar{A} = U\bar{\Sigma}V^\top$$

where $\bar{\Sigma}$ is the matrix obtained from Σ replacing $\sigma_{r+1}, \cdots, \sigma_k$ with zeros.

In words, the best approximation of A among the matrices of rank $r < k$, in the sense of the Frobenius norm, is the matrix obtained by the SVD of A "suppressing" the smallest $k - r$ singular values.

2.2 Example: character recognition

SVD's can be exploited for character recognition, i.e. to build a very simple OCR. (We show the basic principle, avoiding intricacies like the need of recognizing the *position* of a letter within the scanned version of a page.)

Suppose that we are given a "training set" of N images of 16×16 pixels in gray-scale, each one containing a handwritten letter 'A'. Each picture is thus a 16×16 matrix of numbers between, say, 0 and 1, representing the luminance of the pixels. Let us rearrange each of the matrices in a vector $a_i \in \mathbb{R}^{256}$, stacking column after column. Then, stacking these "big" columns row-wise, we obtain a big matrix $A \in \mathbb{R}^{256 \times N}$. Of this matrix, take an SVD: $A = U\Sigma V^\top$. In general, the matrix will have a lot of (nonzero) singular values; nevertheless, we can approximate A with a matrix of rank, say, 3, "suppressing" the singular values from σ_4 onwards, and this is the best low-rank approximation of A in the sense of the Frobenius norm. The resulting matrix $\bar{A}$ has the following range:

$$S_A := \text{range } \bar{A} = \text{span } \{u_1, u_2, u_3\} \subset \mathbb{R}^{256},$$

where u_1, u_2, u_3 are the first three columns of U. You are invited to interpret it as follows: $S_A = \text{span } \{u_1, u_2, u_3\}$ *is the* 3*-dimensional subspace of* $\mathbb{R}^{256}$ *where the "action" of A is concentrated. The image under A of the unit ball of* $\mathbb{R}^N$ *is an ellipsoid with mass concentrated "close" to this subspace; and heuristically, S_A is the* 3*-dimensional subspace of* $\mathbb{R}^{256}$ *representing the "typical" letter 'A'.*

For further reference, let us stack the three columns into a matrix $U_A = \begin{bmatrix} u_1 & u_2 & u_3 \end{bmatrix} \in \mathbb{R}^{256 \times 3}$.

Now, of course, the very same story can be repeated for a training set of N images, of 16×16 pixels, containing the letter 'B'; thus, we can construct a 3-dimensional subspace $S_B \subset \mathbb{R}^{256}$ and a matrix $U_B \in \mathbb{R}^{256 \times 3}$ representing the letter 'B'. And we go on with S_D, U_C representing 'C', $S_D, U_D, \cdots$ up to, say S_Z, U_Z.

Now a new letter, namely a 16×16 bitmap comes. Which letter is it, 'A', 'B', 'C', ... or 'Z'? Let us translate the question in algebraic language. Let y be a vector in $\mathbb{R}^{256}$ representing the new letter;

to which subspace S_i among those representing the letters 'A' ... 'Z' is y closest?

"Closest", in this chapter (and in this course), means "closest in the sense of least squares"; here it means precisely that y *is closest to the subspace* S_i *if the approximation of y obtained by taking its orthogonal projection on* S_i

is the best, i.e., if the difference between y and its orthogonal projection on S_i has the smallest norm. Hence, the algorithm we propose for recognizing a letter is the following:

1. compute the orthogonal projection y_i of y on each $S_i = S_A, \cdots, S_Z$;
2. find the $i = A, \cdots, Z$ that minimizes $\|y_i - y\|_2$;
3. that i is (probably) your letter.

If the new letter y belonged exactly to one of the 3-dimensional subspaces, then the question would translate as follows:

which one of the linear equations $U_A x = y$, $U_B x = y$, ..., $U_Z x = y$ is solvable exactly?

However, since $S_A, \cdots, S_Z$ are only *small* subspaces of range $A, \cdots,$ range Z, and since y is affected by noise anyway, in a real situation actually *none* of the linear equations will be solvable. The canonical way to circumvent this issue is the following:

If the linear system $Ux = y$ was solvable, it would attain $\|Ux - y\|_2 = 0$. If it is not solvable, try to minimize $\|Ux - y\|_2$ instead.

The algorithm translates consequently (but the substance is exactly the same):

1. for each $i = A, \cdots, Z$, compute $m_i = \min_{x \in \mathbb{R}^3} \|U_i x - y\|_2$;
2. find the i such that m_i is minimum (i.e. the linear system which is closest to exact solvability);
3. that i is (probably) your letter.

The approximate solution of a linear system $Ax = b$, that is the minimization of $\|Ax - b\|_2$, is the subject of the rest of this chapter.

2.3 The Moore-Penrose pseudo-inverse

2.3.1 Definition

Given any matrix $A \in \mathbb{R}^{m \times n}$, a matrix $A^+ \in \mathbb{R}^{n \times m}$ is called a *Moore/Penrose pseudo-inverse* of A, or a pseudo-inverse of A for short, if it satisfies the following properties:

1. The matrix $AA^+ \in \mathbb{R}^{m \times m}$ is symmetric;
2. The matrix $A^+ A \in \mathbb{R}^{n \times n}$ is symmetric;
3. $AA^+ A = A$;

4. $A^+AA^+ = A^+$.

We have the following

Theorem 2.3.1 *For any matrix $A \in \mathbb{R}^{m\times n}$, a pseudo-inverse of A exists, it is unique, and it is uniquely determined by the above four properties.*

Consequently, to show that a matrix B is the pseudo-inverse of A it is sufficient to show that B satisfies the four properties. Some examples follow:

- If A is square and invertible, then $A^+ = A^{-1}$. Indeed,

  1. $AA^{-1} = I$, which is symmetric;
  2. $A^{-1}A = I$, which is symmetric;
  3. $AA^{-1}A = IA = A$;
  4. $A^{-1}AA^{-1} = IA^{-1} = A^{-1}$.

  Since A^{-1} satisfies the four properties, it is the pseudo-inverse of A. Of course, this is where the name *pseudo*-inverse comes from: it is a generalization of the inverse of a square, full rank matrix; as we shall now see, the generalization extends to singular and even to non-square matrices.

- If A is a "tall" matrix, meaning that $m > n$, and if its columns are linearly independent, then A^+ is a *left inverse of* A, that is a matrix such that $A^+A = I$. More precisely, we have $A^+ = (A^\top A)^{-1}A^\top$. As before, one just needs to check that $(A^\top A)^{-1}A^\top$ satisfies the four properties. For example, the last one reads:

$$\begin{aligned}((A^\top A)^{-1}A^\top)A((A^\top A)^{-1}A^\top) &= (A^\top A)^{-1}(A^\top A)(A^\top A)^{-1}A^\top \\ &= (A^\top A)^{-1}A^\top\end{aligned}$$

- If A is a "flat" matrix, meaning that $m < n$, and if its rows are linearly independent, then A^+ is a *right inverse of* A, that is a matrix such that $AA^+ = I$. Namely, $A^+ = A^\top(AA^\top)^{-1}$.

2.3.2 General case

In general, $A \in \mathbb{R}^{m\times n}$ is not square and may have low rank. Let

$$A = U\Sigma V^\top$$

be a singular value decomposition of A. Given

$$\Sigma = \begin{bmatrix} \sigma_1 & & & & & & \\ & \sigma_2 & & & & & \\ & & \ddots & & & & \\ & & & \sigma_k & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \\ & & & & & & \end{bmatrix} \in \mathbb{R}^{m \times n}$$

we *define*

$$\Sigma^+ := \begin{bmatrix} 1/\sigma_1 & & & & & & & \\ & 1/\sigma_2 & & & & & & \\ & & \ddots & & & & & \\ & & & 1/\sigma_k & & & & \\ & & & & 0 & & & \\ & & & & & \ddots & & \\ & & & & & & 0 & \end{bmatrix} \in \mathbb{R}^{n \times m}$$

(you can verify that Σ^+ is indeed the pseudo-inverse of Σ). Then,

$$A^+ = V\Sigma^+U^\top.$$

Again, we only have to verify the four properties:

1.
$$\begin{aligned} AA^+ &= U\Sigma V^\top V\Sigma^+U^\top \\ &= U\Sigma\Sigma^+U^\top \\ &= U \begin{bmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & \ddots & & & & \\ & & & 1 & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{bmatrix} U^\top, \end{aligned}$$

   which is indeed symmetric;

2. similar proof;

3.

$$\begin{aligned}
AA^+A &= U\Sigma V^\top V\Sigma^+ U^\top U\Sigma V^\top \\
&= U\Sigma\Sigma^+\Sigma V^\top \\
&= U\begin{bmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & \ddots & & & & \\ & & & 1 & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{bmatrix}\Sigma V^\top \\
&= U\Sigma V^\top = A
\end{aligned}$$

4. similar proof.

Hence, $V\Sigma^+U^\top$ is the pseudo-inverse of A as claimed. (Note that, strictly speaking, $V\Sigma^+U^\top$ is *not* an SVD of A^+, because the positive numbers on the diagonal of Σ^+ come in increasing order.)
We see that the SVD allows readily to compute the pseudo-inverse of a matrix; robust and fast algorithms are available to compute SVD's, hence robust and fast algorithms are also available to compute pseudo-inverses. And as we shall now see, pseudo-inverses play a key role in the numerical solution of general linear systems.

The following simple corollary will be needed in the following section:

Corollary 2.3.1 *For any real matrix A, it holds $(A^+)^\top = (A^\top)^+$.*

Proof. Let $A = U\Sigma V^\top$. Then

$$(A^+)^\top = (V\Sigma^+U^\top)^\top = U(\Sigma^+)^\top V^\top = U(\Sigma^\top)^+V^\top = (V\Sigma^\top U^\top)^+ = (A^\top)^+.$$

□

2.4 Least squares solution of a linear system

Given a matrix $A \in \mathbb{R}^{m\times n}$ and a vector $b \in \mathbb{R}^m$, the equation

$$Ax = b$$

in the unknown x is called a linear equation (a compact form for "a system of linear equations"), and any vector $x \in \mathbb{R}^n$ such that $Ax = b$ is called a *solution* of such equation. As everybody knows, there may be no solution at all, and if a solution exists, it may not be unique; indeed, if more than one solution exists, then there exist infinitely many, and they form the translation of a subspace of $\mathbb{R}^n$, or an *affine* subspace of $\mathbb{R}^n$. Note that, if x is a solution, then $Ax - b = 0$, hence $\|Ax - b\|_2 = 0$.

When A and b come from physical measures it is seldom the case, if at all, that a solution exists, due to the presence of noise. Moreover, if the solution is supposed to be found by a computer and stored into memory, then a solution, even if it exists, may not be representable *exactly* as a floating point number[10]. At most, we can aim at an approximate solution.

In practical situations it happens, more often than not, that something resembling a solution has to be found anyway, even if strictly speaking it does not exist; and if more solution are available, one has to be chosen ("Dear Matlab, I want *the* solution to this system, and I want it now").

In the spirit of least squares, we shall stipulate that x is a good approximate solution if it minimizes the norm $\|Ax - b\|_2$ or, which is the same, the quantity $\|Ax - b\|_2^2$, which is indeed a sum of squares (the ideal goal would be to attain $\|Ax - b\|_2^2 = 0$). If both x_1 and x_2 attain such minimum, we shall prefer the one (say, x_1) which has the least squared norm $\|x_1\|_2^2$ (which is, again, a sum of squares). Formally, we pose the following

Definition 2.4.1 *Given $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$, consider the set*

$$S(A, b) = \{x \in \mathbb{R}^n \mid x \text{ minimizes } \|Ax - b\|_2\}$$

(A vector $x \in S(A, b)$ is "almost" a solution, in the sense that it attains the closest possible result to $\|Ax - b\|_2 = 0$.) Then any vector $x^* \in S(A, b)$ of minimum norm, that is

$$x^* = \arg\min_{x\in S(A,b)} \|x\|_2$$

is called a least squares solution *of the linear system $Ax = b$.*

Lemma 2.4.1 *A least squares solution of $Ax = b$ always exists and is unique.*

Proof. First, let b_c be the orthogonal projection of b on the subspace of $\mathbb{R}^m$ generated by the columns of A. Since b_c belongs to span $\{$columns of $A\}$, $b_c = A\bar{x}$ for some $\bar{x} \in \mathbb{R}^n$. Then, since the projection makes the quantity $\|b_c - b\|_2$ minimum, $\bar{x} \in S(A, b)$, so that $S(A, b)$ is not empty.

Second, notice that any x belonging to $S(A, b)$ must attain $Ax = b_c$, because b_c is the *unique* vector, among those in span $\{$columns of $A\}$, minimizing the distance from b. Hence $S(A, b)$ is an affine space, namely the affine space comprising the solutions of $Ax = b_c$ (this system *always* admits a solution).

[10] Think at this trivial case: $3x = 1$. Is the solution $\frac{1}{3}$ representable *exactly* as a vector of floating-point numbers, if the underlying system is a binary machine?

Now, a finite-dimensional affine space is always a closed convex set, and this would be enough to establish the claim, because in *any* closed convex set there is always a unique element of minimum Euclidean norm; but let us see what happens in more detail.

Any solution of $Ax = b_c$ can be decomposed *in an unique way* as $x = x_r + x^\perp$, where $x_r \in$ span {rows of A} and $x^\perp \in$ span $\{\text{rows of } A\}^\perp$ or, in other terms, $x^\perp \in$ null A. Moreover, for any two solutions x_1, x_2, it holds $Ax_1 = Ax_2 = b_c$, hence $A(x_1 - x_2) = 0$, that is $x_1 - x_2 \in$ null A. This implies that $x_1 = x_2 + x^\perp$ for some $x^\perp \in$ null A, and this in turn implies that x_r is the same for both x_1 and x_2. Hence, x_r is the same for *all* the solutions x, and it is of course a solution itself. Specifically, x_r is the orthogonal projection of any solution of $Ax = b_c$ on span {rows of A}. We claim that the least squares solution is precisely x_r. Indeed for any solution $x = x_r + x^\perp$ we have, by definition, $x_r \perp x^\perp$; hence, by Pythagoras's theorem, $\|x\|_2^2 = \|x_r\|_2^2 + \|x^\perp\|_2^2$, and we can immediately see that, x_r being unique for all x, min $\|x\|_2^2$ is attained for $x^\perp = 0$, that is $x = x_r$. □

Lemma 2.4.2 *A vector $x \in \mathbb{R}^n$ belongs to $S(A, b)$ if and only if*

$$A^\top Ax = A^\top b. \tag{7}$$

(Here we recognize an old friend: Equation (7) is none other than the *normal equation[s].*)

Proof. Let $b = b_c + b_\perp$, where

$$b_c \in \text{span } \{\text{columns of } A\}$$
$$b_\perp \in \text{span } \{\text{columns of } A\}^\perp$$

(The decomposition is unique.) It holds

$$\begin{aligned} x \in S(A, b) &\Leftrightarrow Ax = b_c \\ &\Leftrightarrow \text{for all } y \in \text{span } \{\text{columns of } A\}, \\ &\qquad y^\top(Ax - b) = y^\top(Ax - b_c - b_\perp) = y^\top(Ax - b_c) = 0 \\ &\Leftrightarrow (Ax)^\top(Ax - b) = 0 \quad \text{for all } x \in \mathbb{R}^n \\ &\Leftrightarrow x^\top(A^\top Ax - A^\top b) = 0 \quad \text{for all } x \in \mathbb{R}^n \\ &\Leftrightarrow A^\top Ax - A^\top b = 0. \end{aligned}$$

□

Now the fact that $S(A, b)$ is nonempty, as in the proof of Lemma 2.4.1, reads: the normal equations have at least one solution. And here follows the take-home message of this chapter:

Theorem 2.4.1 *The least squares solution of the system $Ax = b$ is*

$$x^* = A^+ b.$$

Proof. From Lemma 2.4.2 we have that $x \in S(A, b)$ if and only if it solves the normal equations, and from the proof of Lemma 2.4.1 we have that, among these solutions, x has minimum norm if and only if it belongs to span {rows of A} (namely, it is the projection of any $x \in S(A, b)$ on that subspace). We will now show that A^+b satisfies both these properties, and this will be enough to establish the claim.

Let us start from the second property. It holds $A^+b \in$ span {rows of A} if and only if $A^+b \perp x$ for all $x \perp$ span {rows of A} or, which is the same, for all $x \in$ null A. Now, by the properties of the pseudo-inverse, for any such x it holds

$$\begin{aligned} x^\top A^+ b &= x^\top A^+ A A^+ b \\ &= x^\top (A^+ A)^\top A^+ b \\ &= x^\top A^\top (A^+)^\top A^+ b \\ &= (Ax)^\top (A^+)^\top A^+ b \\ &= 0, \end{aligned}$$

so that indeed $A^+b \perp x$, and the second property is established.

Regarding the first property, we must show that

$$A^\top A(A^+ b) = A^\top b;$$

but this is now easy, because, by other properties of the pseudo-inverse,

$$A^\top A A^+ = A^\top (A A^+)^\top = A^\top (A^+)^\top A^\top = A^\top (A^\top)^+ A^\top = A^\top,$$

and post-multiplying by b is enough to establish the claim. □

Note that in the proof we have exploited all the four properties of the Moore-Penrose pseudo-inverse exactly once (find where!) and nothing else, to prove two properties that are essentially geometric facts related to orthogonality. This tells us that the very concept of pseudo-inverse is tightly linked to that of least squares solution.

In view of the latter theorem, the analytical solution of a least squares estimation problem, computed in section 1.4 can be restated in terms of a pseudo-inverse. In order to compute

$$\hat{\theta}_{\mathrm{LS}} = \arg\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^{N} (y_i - \varphi_i^\top \theta)^2,$$

we define

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \qquad \Phi = \begin{bmatrix} \varphi_1^\top \\ \varphi_2^\top \\ \vdots \\ \varphi_N^\top \end{bmatrix}$$

and we ask to find

$$\hat{\theta}_{\text{LS}} = \arg\min_{\theta \in \mathbb{R}^p} \ \|\Phi\theta - Y\|_2 .$$

This is the search for the least squares solution of a linear system $Y = \Phi\theta$ which strictly speaking is not solvable (indeed, recall that there is a "hidden" noise term: $Y = \Phi\theta + E$). The solution reads

$$\hat{\theta}_{\text{LS}} = \Phi^{+}Y.$$

The only news with respect to section 1.4 is that, if $\Phi^\top\Phi$ is not invertible (many "candidate solutions" exist in $S(\Phi, Y)$) then the pseudo-inverse yields the one with minimum norm. If, on the other hand, Φ is "tall" *and has full rank*, as is usually the case if many data are available, then we also know that it admits a left inverse coinciding with Φ^+, namely $\Phi^+ = (\Phi^\top\Phi)^{-1}\Phi^\top$, so that the *unique* solution reads

$$\hat{\theta}_{\text{LS}} = (\Phi^\top\Phi)^{-1}\Phi^\top Y.$$

2.5 Matlab code

When you wish to carry on with a quick and dirty least squares approximation in Matlab, you don't have to mess with sums, transposes and inverses that may not even exist: you can just use the built-in function `pinv()`, which computes pseudo-inverses and does all the hard job. So, the next time you are given some pairs of numbers $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$ and asked to provide an interpolation with a third-degree polynomial $y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$, just do the following:

1. make sure that the numbers x_i and y_i are stacked in two separate columns X and Y of the same dimension;
2. build the "data matrix" Φ having $\varphi_j(x_i) = (x_i)^j$:

$$\Phi = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ 1 & x_3 & x_3^2 & x_3^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 & x_N^3 \end{bmatrix}$$

3. compute $\hat{\theta}_{\text{LS}} = \Phi^+ Y$.

The solution takes at most two lines of code:

```
Phi = [ones(length(X),1), X, X.^2, X.^3];
thetaLS = pinv(Phi)*Y;
```

Actually, the second line can be simplified, because least squares approximation is what Matlab does by default to solve linear systems when they do not admit an unique solution in the classical sense:

```
Phi = [ones(length(X),1), X, X.^2, X.^3];
thetaLS = Phi\Y;
```

2.6 Exercises for Chapter 2

Problem 1 (SVD and pseudo-inverse).

1. Verify that
$$A = U\Sigma V^\top = \begin{bmatrix} 1/\sqrt{5} & -2/\sqrt{5} \\ 2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix} \begin{bmatrix} \sqrt{10} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}.$$
is a singular value decomposition of $A = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}$.

2. Compute the pseudo-inverse of $A = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}$.

Problem 2 (orthogonal projector).

Consider $A \in \mathbb{R}^{m\times n}$, and let A^+ be its pseudo-inverse. Show that

$$\Pi_A := AA^+ \in \mathbb{R}^{m\times m}$$

is the *orthogonal projector* onto the subspace of $\mathbb{R}^m$ generated by the columns of A; more explicitly, that for any $v \in \mathbb{R}^m$, $\Pi_A v$ is the orthogonal projection of v on span {columns of A}.

3 Identification of dynamical systems

In this chapter we apply the method of least squares to the identification of dynamical systems; the results are cited without proofs. The subject of system identification is vast; if you are interested in how the following simple examples extend to more general cases, you can refer to the standard book [17], to [30], or to the lecture notes (in Italian) [23]. Before you proceed with this chapter, check that you are familiar with the concepts about discrete-time systems reviewed in Appendix C.

3.1 Wide-sense stationary processes

In this chapter we will consider discrete-time stochastic processes like $\{y(t)\}_{-\infty}^{+\infty}$, i.e. sequences of random variables $\cdots, y(-2), y(-1), y(0), y(1), y(2), \cdots$ defined on the same probability space $(\Omega, \mathcal{F}, \mathsf{P})$, *infinite in both directions.* The process $\{y(t)\}_{-\infty}^{+\infty}$ is called *wide-sense stationary*, or *stationary of the second order*, if

1. its mean $m_y := \mathsf{E}[y(t)]$ exists finite and does not depend on t;

2. its *correlation signal*

$$R_y(\tau) := \mathsf{E}[y(t)y(t+\tau)],$$

for $\tau = \cdots, -1, 0, 1, 2, \cdots$, exists finite and does not depend on t.

It follows at once that also the *covariance signal* of the process

$$C_y(\tau) := \mathsf{E}\left[(y(t) - m_y)\left(y(t+\tau) - m_y\right)\right],$$

does not depend on t; indeed it holds $C_y(\tau) = R_y(\tau) - m_y^2$. In what follows we will focus primarily on wide-sense stationary processes; we will refer to $R_y(0) = r^2$ and $C_y(0) = \sigma^2$ as the *power* and the *variance* of the process, respectively.

Vaguely speaking, the *power spectral density*, or *spectrum* for short, of a stationary process, is the Fourier transform of its correlation signal[11]:

$$S_y(\omega) = \mathcal{F}\left[R_y\right](\omega) = \sum_{\tau=-\infty}^{+\infty} R_y(\tau)\ e^{-j\omega\tau},$$

for $\omega \in [-\pi, \pi]$. If $m_y = 0$, the spectrum coincides with the Fourier transform of the *covariance* signal. It can be shown that, due to the intrinsic

[11] The true definition of power spectral density is somewhat subtler; the fact that under rather general hypotheses it coincides with the Fourier transform of the correlation signal is the so-called Wiener-Khinchine theorem. See [27] for more details.

symmetry in the correlation signal of which it is the transform, a power spectrum is always an even function taking nonnegative values ($S_y(\omega) = S_y(-\omega)$ and $S_y(\omega) \geq 0$ for all $\omega \in [-\pi, \pi]$).

Even more vaguely speaking, the spectrum is a function of the *frequency* ω, representing how much power the process carries, *on average*, at that frequency. “Flat” spectra characterize loosely correlated variables, and in general a realization of a loosely correlated process exhibits a chaotic behavior. Conversely, a spectrum that exhibits a “peak” at a certain frequency, say ω_1, characterizes a process that has a distinct oscillatory behavior, more or less at the frequency ω_1; the more pronounced the peak, the more visible the oscillatory behavior.

On one extreme, the only processes with *constant* spectrum are those zero-mean processes such that

$$C_y(\tau) = \begin{cases} \sigma^2, & \tau = 0; \\ 0, & \text{otherwise}, \end{cases}$$

for a certain variance σ^2. Such a process is called *white noise*, in the wide sense; in words, a white noise is a sequence of pair-wise uncorrelated variables with mean zero and the same variance. To obtain stronger results, often one assumes a much stronger property: a white noise, *in the strict sense*, is a sequence of *independent and identically distributed* variables, with mean zero and a certain variance σ^2.

On the other extreme, an example of stationary process with perfect oscillatory behavior is the sampled version of a sinusoid, $y(t) = A \sin(\omega_1 t + \phi)$, where ω_1 is a constant, the amplitude A is any nonzero random variable, and the phase ϕ is a random variable with uniform density in $[-\pi, \pi]$. Strictly speaking, the correlation signal of this process does *not* possess a Fourier transform; however, you can visualize the picture as if the power spectrum had a pair of “Dirac deltas” at the frequencies $\{-\omega_1, \omega_1\}$.[12]

The processes of the first kind (white noises and *filtered* white noises) are called *completely nondeterministic*, and concern us most in this chapter; the processes of the second kind (for example finite sums of sinusoids with random amplitude and phase) are called *predictable*, because a finite number of samples — 2, for only one sinusoid — are sufficient to reconstruct without errors all the remaining samples of the process. A famous theorem by Wold says that any stationary process $\{y(t)\}$ can be decomposed as the sum of a

[12] Physical, continuous-time counterparts of the same phenomena are the light emitted by an incandescent light bulb, which is due to thermal phenomena, has more or less the same power at all the frequencies, and is distinctly “white” (this is where the name “*white* noise” comes from), and the light emitted by the state transitions of the external electrons of certain atoms, used e.g. in mercury lamps, whose power is concentrated only at some frequencies (“line spectrum”).

completely nondeterministic process $\{\tilde{y}(t)\}$ and a predictable process $\{\bar{y}(t)\}$ uncorrelated to each other.

More precisely, we will call *completely nondeterministic* a process, if it can be obtained as the output of a causal LTI system whose input is a white noise:

$$y(t) = \sum_{\tau=-\infty}^{t} w(t-\tau)e(\tau), \qquad t = \cdots, -2, -1, 0, 1, 2, \cdots \tag{8}$$

where $w(t)$, $t = 0, 1, 2, \cdots$ is the impulse response of the system, and $\{e(t)\}$ is a white noise. It can be proved that $\{y(t)\}$ is stationary *if and only if* the impulse response is summable,

$$\sum_{t=0}^{\infty} |w(t)| < \infty,$$

or, equivalently, if the transfer function of the system

$$W(z) = \sum_{t=0}^{\infty} w(t)z^{-t}$$

converges on the unit circle.

Example. The prototypical example of complete nondeterminism is a process conforming *causally* to the following equations:

$$y(t) = ay(t-1) + e(t), \qquad t = \cdots, -2, -1, 0, 1, 2, \cdots \tag{9}$$

where $a \neq 0$ is a constant; this is the simplest example of a so-called *autoregressive* process. It can be thought as the response of the causal LTI system

$$y(t) = ay(t-1) + u(t) \tag{10}$$

to a white noise fed at the input.
Recall that we suppose that $\{e(t)\}$ has mean zero. The mean of $y(t)$ goes as follows:

$$\mathsf{E}[y(t)] = a\mathsf{E}[y(t-1)] + 0 \tag{11}$$

Note that, according to (8), $y(t-1)$ depends on the noise $e(\cdot)$ only up to time $t-1$; therefore, it is uncorrelated from $e(t)$. Since the variance of the sum of two uncorrelated variables is the sum of the respective variances, we have

$$\mathsf{Var}[y(t)] = a^2\ \mathsf{Var}[y(t-1)] + \mathsf{Var}[e(t)] = a^2\ \mathsf{Var}[y(t-1)] + \sigma^2$$

It is now immediate to realize that, if $a^2 \geq 1$, the variance of $\{y(t)\}$ "explodes" as time goes on even assuming that it is finite at a certain time; in particular, this implies that $\{y(t)\}$ cannot be stationary. In fact, $\{y(t)\}$ is stationary if, and only if, $-1 < a < 1$. This implies, in view of (11), that $\mathsf{E}[y(t)] = 0$; moreover, giving for granted that $\sigma_y^2 = \mathsf{Var}[y(t)]$ is constant with respect to t, we have

$$\begin{aligned}\sigma_y^2 &= a^2\ \sigma_y^2 + \sigma^2;\\ \sigma_y^2 &= \frac{\sigma^2}{1-a^2}.\end{aligned}$$

The covariance (or correlation) signal of $\{y(t)\}$ is now easy to compute using a simple trick. By Equation (9),

$$\begin{aligned}y(t+\tau) &= ay(t+\tau-1) + e(t+\tau)\\ y(t)y(t+\tau) &= ay(t)y(t+\tau-1) + y(t)e(t+\tau)\\ \mathsf{E}[y(t)y(t+\tau)] &= a\mathsf{E}[y(t)y(t+\tau-1)] + \mathsf{E}[y(t)e(t+\tau)]\end{aligned}$$

If $\tau > 0$ the last term is zero, $e(t+\tau)$ being uncorrelated from the past samples of $y(\cdot)$; hence

$$R_y(\tau) = aR_y(\tau-1), \qquad \tau = 1,2,3,\cdots,$$

and solving the recursion,

$$R_y(\tau) = a^\tau R_y(0), \qquad \tau = 1,2,3,\cdots$$

Note that we have already computed

$$R_y(0) = C_y(0) = \sigma_y^2 = \frac{\sigma^2}{1-a^2}.$$

The case for $\tau < 0$ is, instead, just a matter of symmetry, since by definition the cross-correlation must not depend on t:

$$R_y(-\tau) = \mathsf{E}[y(t)y(t-\tau)] = \mathsf{E}[y(t+\tau)y((t+\tau)-\tau)] = R_y(\tau).$$

Resuming, $R_y(\tau) = \frac{\sigma^2}{1-a^2}a^{|\tau|}$; now we can go for the spectrum. The Fourier transform of the function $f(\tau) = a^{|\tau|}$ is

$$\begin{aligned}
F(\omega) &= \sum_{\tau=-\infty}^{+\infty} a^{|\tau|} e^{-j\omega\tau} \\
&= \sum_{\tau=-\infty}^{0} a^{|\tau|} e^{-j\omega\tau} + \sum_{\tau=0}^{\infty} a^{|\tau|} e^{-j\omega\tau} - a^{|0|} e^{-j\omega\cdot 0} \\
&= \sum_{\tau=0}^{+\infty} a^{\tau} e^{j\omega\tau} + \sum_{\tau=0}^{+\infty} a^{\tau} e^{-j\omega\tau} - 1 = \sum_{\tau=0}^{+\infty} (ae^{j\omega})^{\tau} + \sum_{\tau=0}^{+\infty} (ae^{-j\omega})^{\tau} - 1 \\
&= \frac{1}{1 - ae^{j\omega}} + \frac{1}{1 - ae^{-j\omega}} - 1 = \frac{2 - a(e^{j\omega} + e^{-j\omega})}{1 + a^2 - a(e^{j\omega} + e^{-j\omega})} - 1 \\
&= \frac{2 - 2a\cos\omega}{1 + a^2 - 2a\cos\omega} - 1 = \frac{1 - a^2}{1 + a^2 - 2a\cos\omega}
\end{aligned}$$

Therefore,

$$S_y(\omega) = \frac{\sigma^2}{1 - a^2} F(\omega) = \frac{\sigma^2}{1 + a^2 - 2a\cos\omega}.$$

You can immediately verify that S_y is an even function and that it is everywhere positive; in particular, notice that it depends on ω only through a cosine; none of these facts is a coincidence. □

It seems that, since computing the spectrum of the *simplest* possible autoregressive process involves tedious computations, the same computation for an even slightly more sophisticated model is going to be a nasty job. In many cases, though, this is not true. The following fundamental result relates directly the spectrum of a filtered process to the spectrum of the input through the transfer function of the LTI system, letting us avoid the computation of the correlation signal and its Fourier transform:

Theorem 3.1.1 *Suppose that a wide-sense stationary process $\{u(t)\}_{-\infty}^{+\infty}$ is fed as an input to a BIBO-stable LTI system with transfer function $W(z)$; then the output $\{y(t)\}_{-\infty}^{+\infty}$ is also stationary, and*

$$S_y(\omega) = \left| W\left(e^{j\omega}\right) \right|^2 S_u(\omega).$$

Proof. See e.g. [22, Section 8.4]. □

With respect to the above example, the input $\{e(t)\}$ is a white noise with variance σ^2; this means that its correlation signal is $R_e(\tau) = \sigma^2$ if $\tau = 0$, and 0 otherwise, namely an impulse multiplied by σ^2; the Fourier transform of an impulse is the constant 1, hence $S_e(\omega) = \sigma^2$ for all ω. The transfer function of the LTI system under consideration is $W(z) = \frac{1}{1-az^{-1}} = \frac{z}{z-a}$, and we suppose that its only pole, which is a, lies in the interior of the unit disc (since a is real, it means $-1 < a < 1$). We have[13]

$$
\begin{aligned}
\left|W\left(e^{j\omega}\right)\right|^2 &= W(e^{j\omega})\overline{W(e^{j\omega})} = W(e^{j\omega})W\left(\overline{e^{j\omega}}\right) = W(e^{j\omega})W(e^{-j\omega}) \\
&= \frac{1}{1-ae^{-j\omega}}\frac{1}{1-ae^{j\omega}} = \frac{1}{1+a^2-a(e^{j\omega}+e^{-j\omega})} \\
&= \frac{1}{1+a^2-2a\cos\omega},
\end{aligned}
$$

and finally, by Theorem 3.1.1,

$$
S_y(\omega) = \left|W\left(e^{j\omega}\right)\right|^2 S_e(\omega) = \frac{\sigma^2}{1+a^2-2a\cos\omega},
$$

as we have found previously.

3.2 Model classes

The purpose of system identification is to find a suitable finite-dimensional model for a discrete-time random process. The most popular models in control engineering describe such process as the output of a causal LTI system whose inputs are a white noise, which cannot be observed by the experimenter, and possibly another "exogenous" input, either deterministic or random, which is instead known to the experimenter (or even *set by* the experimenter).

An *autoregressive* (AR) process $\{y(t)\}_{-\infty}^{+\infty}$ is the output of a *causal* LTI system described by the following model:

$$
y(t) - a_1 y(t-1) - \cdots - a_n y(t-n) = e(t)
$$

where $a_1, \cdots, a_n \in \mathbb{R}$, and $\{e(t)\}_{-\infty}^{+\infty}$ is a sequence of independent and identically distributed random variables ("process noise"), with mean zero and unknown variance (to simplify the picture, assume it is Gaussian) that brings randomness into the picture. Note that $\{e(t)\}$ is not necessarily a disturbance: it is just another input to the system, invisible to the experimenter. Since $\{e(t)\}$ is a white noise, and since the system is supposed to be causal, $e(t)$ is independent of $y(\tau)$ for all $\tau < t$.

Let us interpret z and z^{-1} as the anticipation operator and the delay operator respectively, that is, those operators that act on a sequence by translating it by one time step to the left or to the right ($zy(t) = y(t+1)$ and

[13] Note that the second equality holds because W is analytic in the region of the plane that includes the unit circle.

$z^{-1}y(t) = y(t-1)$). Then the model reads

$$\begin{aligned} y(t) - a_1 z^{-1} y(t) - \cdots - a_n z^{-n} y(t) &= e(t); \\ (1 - a_1 z^{-1} - \cdots - a_n z^{-n}) y(t) &= e(t); \end{aligned}$$

$$\begin{aligned} y(t) &= \frac{1}{1 - a_1 z^{-1} - \cdots - a_n z^{-n}}\ e(t) \\ &= \frac{z^n}{z^n - a_1 z^{n-1} - \cdots - a_n}\ e(t). \end{aligned}$$

The operator

$$W(z) = \frac{z^n}{z^n - a_1 z^{n-1} - \cdots - a_n}$$

is the transfer function of a causal LTI system whose input is $e(t)$. If, with a slight abuse of terminology, we interpret z as a complex variable, and the poles of $W(z)$, that is the roots of the polynomial $z^n - a_1 z^{n-1} - \cdots - a_n$, lie inside the interior of the unit disc ($\{z \in \mathbb{C} \mid |z| < 1\}$), then such system is BIBO-stable, and $y(t)$ is wide-sense stationary. In particular, since the coefficients $a_1, \cdots, a_n$ are real, either the roots are real, or they come in conjugate pairs; any two conjugate poles induce in $y(t)$ an oscillatory behavior (the frequency depending on their phase), and the closer they are to the unit circle $\{z \in \mathbb{C} \mid |z| = 1\}$, the more pronounced is such behavior.

The process $\{y(t)\}_{-\infty}^{+\infty}$ is *autoregressive with exogenous input* (ARX) if it conforms to the following model:

$$y(t) - a_1 y(t-1) - \cdots - a_n y(t-n) = b_1 u(t-1) + b_2 u(t-2) + \cdots + b_m u(t-m) + e(t)$$

where $m \leq n$, $\{e(t)\}$ is a white noise as above, and $\{u(t)\}_{-\infty}^{+\infty}$ is another signal, either random or deterministic, but known to the experimenter. We suppose that $e(t)$ is independent of $y(s)$ for all $s < t$, and independent of $u(s)$ for all s. In terms of transfer functions, the model reads

$$\begin{aligned} &y(t) - a_1 z^{-1} y(t) - \cdots - a_n z^{-n} y(t) = b_1 z^{-1} u(t) + b_2 z^{-2} u(t) + \cdots + b_m z^{-m} u(t) + e(t); \\ &y(t) = \frac{b_1 z^{-1} + \cdots + b_m z^{-m}}{1 - a_1 z^{-1} - \cdots - a_n z^{-n}}\ u(t) + \frac{1}{1 - a_1 z^{-1} - \cdots - a_n z^{-n}}\ e(t) \\ &\quad = \frac{b_1 z^{n-1} + \cdots + b_m z^{n-m}}{z^n - a_1 z^{n-1} - \cdots - a_n}\ u(t) + \frac{z^n}{z^n - a_1 z^{n-1} - \cdots - a_n}\ e(t) \\ &\quad = G(z)\ u(t) + W(z)\ e(t) \end{aligned}$$

As before, if the roots of $z^n - a_1 z^{n-1} - \cdots - a_n$ lie in the open unit disc, then $y(t)$ is stationary, provided $u(t)$ is stationary as well.

With respect to both AR and ARX models, the system identification problem is the task of reconstructing $W(z)$ or, which is the same, finding the parameters $a_1, \cdots, a_n$, and possibly $b_1, \cdots, b_m$, from a sequence of measures, respectively $(y(0), y(1), \cdots, y(N))$ or $((u(0), y(0)), (u(1), y(1)), \cdots, (u(N-1), y(N-1)), y(N))$. As we shall soon see, AR and ARX models are of particular interest to us, because their identification can be accomplished in a natural way with the least squares method.

Example. The identification of AR models is used in telecommunications to save bandwidth in the transmission of speech through digital channels. Indeed, the frequencies of interest in a voice signal span a range of about 4 kHz, thus the transmission of such a signal requires at least a sampling rate of 8 kHz[14] (after a suitable prefiltering), and a channel capacity of 8k samples per second[15]. However, the human voice is generally accepted as stationary in time "windows" of 20 ms, say of 160 samples, if we sample at 8 kHz. Any such window can therefore be modeled as a short realization of an AR process, whose parameters can be identified accordingly. Modern voice transmission techniques prescribe to split the signal in 20 ms windows, extract some features (e.g. sinusoids, pulse sequences) from each window and subtract them from the signal, identify an AR model for the remaining process, and to *transmit the features and the model*, say $20 \sim 30$ parameters along with an estimate of the variance of the model noise $e(t)$ inferred from the residuals, instead of the 160 samples. In this way, a lot of channel capacity is saved. See [24] for further details.
Particularly with respect to this example, computing the analytical solution to the normal equations is not a fast algorithm and not the one used in practice in telecommunications, because the blind computation of an inverse or of a pseudo-inverse does not exploit the symmetry inherent in such equations; however, in the rest of this chapter we shall pursue this approach, it being the most intuitive, mathematically speaking. □

Other model classes of which you should be aware are the following:

- *Moving average* (MA) models:

$$\begin{aligned} y(t) &= e(t) + c_1 e(t-1) + \cdots + c_k e(t-k) \\ &= (1 + cz^{-1} + \cdots + c_k z^{-k})\ e(t). \end{aligned}$$

  Here, $y(t)$ is modelled as the output of a finite-impulse-response (FIR) filter, whose input is a white noise.

[14] Recall Shannon's sampling theorem...

[15] The *de facto* standard is 8 bits per sample, implying a rate of 64 kbps.

- *Autoregressive/moving average* (ARMA) models:

$$\begin{aligned}
&y(t) - a_1 y(t-1) - \cdots - a_n y(t-n) = e(t) + c_1 e(t-1) + \cdots + c_k e(t-k) \\
&\qquad (1 - a_1 z^{-1} - \cdots - a_n z^{-n}) y(t) = (1 + c_1 z^{-1} + \cdots + c_k z^{-k}) e(t) \\
&y(t) = \frac{1 + c_1 z^{-1} + \cdots + c_k z^{-k}}{1 - a_1 z^{-1} - \cdots - a_n z^{-n}}\ e(t).
\end{aligned}$$

- Autoregressive/moving average models with exogenous input (ARMAX):

$$\begin{aligned}
&y(t) - a_1 y(t-1) - \cdots - a_n y(t-n) = b_1 u(t-1) + \cdots + b_m u(t-m) \\
&\qquad + e(t) + c_1 e(t-1) + \cdots + c_k e(t-k) \\
&y(t) = \frac{b_1 z^{-1} + \cdots + b_m z^{-m}}{1 - a_1 z^{-1} - \cdots - a_n z^{-n}}\ u(t) + \frac{1 + c_1 z^{-1} + \cdots + c_k z^{-k}}{1 - a_1 z^{-1} - \cdots - a_n z^{-n}}\ e(t).
\end{aligned}$$

- Other more sophisticated linear models are in widespread use (e.g. Box-Jenkins, state-space).

3.3 Prediction

3.3.1 Linear prediction of stationary processes

Suppose that $\{y(t)\}_{-\infty}^{+\infty}$ is a stationary, nondeterministic process modelled as the output of a causal LTI system:

$$y(t) = \sum_{\tau=-\infty}^{t} w(t-\tau) e(\tau),$$

where $w(t)$, $t = 0, 1, 2, \cdots$ is the impulse response of the system, and $\{e(t)\}$ is a white noise with variance σ_e^2. The impulse response of the system is summable,

$$\sum_{t=0}^{\infty} |w(t)| < \infty,$$

and its transfer function is the $\mathscr{Z}$-transform of $\{w(t)\}$,

$$W(z) = \sum_{t=0}^{+\infty} w(t) z^{-t}.$$

Then, in symbolic notation,

$$y(t) = W(z)\ e(t),$$

and the power spectrum of $\{y(t)\}$ is

$$S_y(\omega) = S_e(\omega) \left| W\left(e^{j\omega}\right) \right|^2 = \sigma_e^2\ \left| W\left(e^{j\omega}\right) \right|^2 = \sigma_e^2\ \left[W\left(z\right) W\left(z^{-1}\right) \right]_{z=e^{j\omega}}.$$

Without loss of generality, we can assume $w(0) = W(\infty) = 1$.

A *linear predictor* of $y(t)$, given the "past" values $e(t-1), e(t-2), \cdots$ is a linear function $\hat{y}(e(t-1), e(t-2), \cdots)$ of "the past" used as an estimator of the "present" value $y(t)$. The meaning being clear given the context, we will denote any such function $\hat{y}(t|t-1)$.
Since $y(t)$ can be expressed as

$$y(t) = w(0)e(t) + \sum_{\tau=-\infty}^{t-1} w(t-\tau)e(\tau) = e(t) + \sum_{\tau=-\infty}^{t-1} w(t-\tau)e(\tau),$$

it turns out that the predictor

$$\hat{y}(t|t-1) = \sum_{\tau=-\infty}^{t-1} w(t-\tau)e(\tau) = w(1)e(t-1) + w(2)e(t-2) + \cdots$$

is the "best" possible one, in the sense that it attains the least possible variance (that of $e(t)$, that is σ_e^2) among all the possible linear functions of $e(t-1), e(t-2), \cdots$; this is fairly intuitive. In can be written in symbolic form:

$$\begin{aligned} \hat{y}(t|t-1) &= \sum_{\tau=-\infty}^{t} w(t-\tau)e(\tau) - e(t) \\ &= (W(z) - 1)\ e(t). \end{aligned}$$

Such expression could be computed, in principle, if one knew the variables $e(t-1), e(t-2), \cdots$; but this is not the case, since the noise is not observable. Thus, it would be far more useful to express $\hat{y}(t|t-1)$ as a function of the past values of $\{y(t)\}$, that are indeed known.
Suppose that there exists a BIBO-stable *causal inverse* of the LTI system generating $\{y(t)\}$, that is, there exists a sequence $\bar{w}(0), \bar{w}(1), \bar{w}(2), \cdots$ such that

$$\sum_{t=0}^{\infty} |\bar{w}(t)| < \infty$$

$$\sum_{t=0}^{+\infty} \bar{w}(t)z^{-t} = \bar{W}(z) = \frac{1}{W(z)}.$$

When $W(z)$ is a rational transfer function, this happens precisely when all its *zeros* belong to the interior of the unit disc; the zeros of $W(z)$ are indeed the poles of $\frac{1}{W(z)}$.
If this is the case, $\{\bar{w}(t)\}$ is the impulse response of a causal LTI system that takes $\{y(t)\}$ as the input and yields $\{e(t)\}$ as the output:

$$e(t) = \bar{W}(z)\ y(t) = \frac{1}{W(z)}\ y(t).$$

But then,

$$\begin{aligned}\hat{y}(t|t-1) &= (W(z)-1)\ e(t) = (W(z)-1)\ \frac{1}{W(z)}\ y(t) \\ &= \left(1-\frac{1}{W(z)}\right)\ y(t).\end{aligned} \tag{12}$$

This expression *is* computable, at least in principle, because the right-hand member depends only on the past of $\{y(t)\}$.

Example. Let

$$\begin{aligned}y(t) &= \frac{1}{2}y(t-1)+e(t)+\frac{1}{3}e(t-1) \\ y(t) &= W(z)\ e(t) = \frac{1+\frac{1}{3}z^{-1}}{1-\frac{1}{2}z^{-1}}\ e(t) = \frac{z+\frac{1}{3}}{z-\frac{1}{2}}\ e(t).\end{aligned}$$

This is an ARMA model having a pole at $\frac{1}{2}$ and a zero at $-\frac{1}{3}$. Since the pole is in the interior of the unit disc, the system represented by $W(z)$ is BIBO-stable. And since the *zero* is in the interior of the unit disc too, then

$$\frac{1}{W(z)} = \bar{W}(z) = \frac{z-\frac{1}{2}}{z+\frac{1}{3}} = \frac{1-\frac{1}{2}z^{-1}}{1+\frac{1}{3}z^{-1}}$$

is also the representative of a causal BIBO-stable system; filtering $\{y(t)\}$ with such system we recover $\{e(t)\}$:

$$\begin{aligned}e(t) &= \frac{1-\frac{1}{2}z^{-1}}{1+\frac{1}{3}z^{-1}}\ y(t) \\ e(t) &= -\frac{1}{3}e(t-1)+y(t)-\frac{1}{2}y(t-1).\end{aligned}$$

The best linear predictor $\hat{y}(t|t-1)$ is the following:

$$\begin{aligned}\hat{y}(t|t-1) &= \left(1-\frac{1}{W(z)}\right)\ y(t) = \left(1-\frac{z-\frac{1}{2}}{z+\frac{1}{3}}\right)\ y(t) \\ &= \frac{\frac{5}{6}}{z+\frac{1}{3}}\ y(t) = \frac{\frac{5}{6}z^{-1}}{1+\frac{1}{3}z^{-1}}\ y(t); \\ \hat{y}(t|t-1) &= -\frac{1}{3}\hat{y}(t-1|t-2)+\frac{5}{6}y(t-1).\end{aligned}$$

This is a recursive algorithm that updates the current prediction $\hat{y}(t|t-1)$ given the past prediction $\hat{y}(t-1|t-2)$ and the past value $y(t-1)$ of the process. □

Any BIBO-stable transfer function ($W(z)$ with poles in the open unit disc) of a causal LTI system, that admits a causal BIBO-stable inverse ($W(z)$ with also *zeros* in the open unit disc), is called a *minimum phase* transfer function; the LTI system is said to be *causally invertible.* This is the case of the above example. The inverse $\bar{W}(z) = \frac{1}{W(z)}$ is called the *whitening filter*, and the one-step predictor $\hat{y}(t|t-1) = (1 - 1/W(z))y(t)$ is a particular case of what in literature is known as the Wiener/Kolmogorov filter.

What is actually going on would be better understood in terms of Euclidean geometry. We are working under the assumption that all the variables involved, $y(t)$, $e(t)$, $y(t-1)$, $e(t-1)$, and so on, have zero mean and finite variance. The vector space of *all* such variables can be endowed with a scalar product:

$$\langle y, e \rangle := \mathsf{E}\,[ye] = \mathsf{Cov}\,[y, e]$$

(see Appendix D.3.3). Now

$$y(t) = w(0)e(t) + w(1)e(t-1) + w(2)e(t-2) + \cdots$$

is a "linear combination" of $e(t), e(t-1), e(t-2), \cdots$, i.e. it belongs to the space of the "past and present", $V = \text{span}\,\{e(t), e(t-1), e(t-2), \cdots\}$. Since all the $\{e(t)\}$ are uncorrelated, that is *orthogonal* in geometric language, the set $\{e(t), e(t-1), e(t-2), \cdots\}$ forms an "orthogonal basis" for V. (We did not provide a proper interpretation for either an infinite "linear combination" nor for an infinite "basis", hence our discussion remains at the intuitive level.) But indeed, since such basis is orthogonal, the expression for the predictor

$$\hat{y}(t|t-1) = w(1)e(t-1) + w(2)e(t-2) + \cdots$$

is none other than the *orthogonal projection* of $y(t)$ onto span $\{e(t-1), e(t-2), \cdots\}$, which is the "past" subspace of V. Finally, the whitening filter operates, time after time, a "change of basis" between $\{e(t-1), e(t-2), \cdots\}$ and $\{y(t-1), y(t-2), \cdots\}$, allowing us to express $\hat{y}(t|t-1)$ as a linear combination of the latter.

3.3.2 Non-minimum phase and spectral factorization

When a process is expressed in terms of a transfer function which is BIBO-stable, but *not* minimum phase, one has to resort to the so-called problem of *spectral factorization.* We will provide just a short example, because in the rest of the chapter we will deal only with AR and ARX models, for which this issue does not exist.

Example. Let

$$y(t) = W(z)\ e(t) = \frac{1 + 3z^{-1}}{1 - \frac{1}{2}z^{-1}}\ e(t) = \frac{z+3}{z - \frac{1}{2}}\ e(t).$$

This is again an ARMA model having a pole at $\frac{1}{2}$ (BIBO-stable transfer function) and a zero at -3, well outside the unit disc. The "naïve" inverse

$$\frac{1}{W(z)} = \frac{z-\frac{1}{2}}{z+3}$$

would *not* be BIBO-stable, hence it would be useless for prediction purposes. In order to carry on with the prediction procedure, one starts not from $W(z)$, but from the *spectrum of* $\{y(t)\}$:

$$\begin{aligned} S_y(\omega) &= \sigma_e^2 \ \left[W\left(z\right) W\left(z^{-1}\right)\right]_{z=e^{j\omega}} \\ &= \sigma_e^2 \ \left[\frac{z+3}{z-\frac{1}{2}} \cdot \frac{z^{-1}+3}{z^{-1}-\frac{1}{2}}\right]_{z=e^{j\omega}} \end{aligned}$$

Then he or she builds a new *equivalent representation of the process* $\{y(t)\}$, in the sense that the same process is expressed in terms of a different transfer function and of a different white noise. We want to get rid of the zero at -3; here is the trick:

$$\begin{aligned} S_y(\omega) &= \sigma_e^2 \ \left[\frac{(z+3)(z^{-1}+3)}{(z-\frac{1}{2})(z^{-1}-\frac{1}{2})}\right]_{z=e^{j\omega}} = \sigma_e^2 \ \left[\frac{(z+3)z^{-1}z(z^{-1}+3)}{(z-\frac{1}{2})(z^{-1}-\frac{1}{2})}\right]_{z=e^{j\omega}} \\ &= \sigma_e^2 \ \left[\frac{(1+3z^{-1})(1+3z)}{(z-\frac{1}{2})(z^{-1}-\frac{1}{2})}\right]_{z=e^{j\omega}} = \sigma_e^2 \ \left[\frac{1+3z}{z-\frac{1}{2}} \cdot \frac{1+3z^{-1}}{z^{-1}-\frac{1}{2}}\right]_{z=e^{j\omega}} \\ &= (3\sigma_e)^2 \ \left[\frac{1}{3}\ \frac{1+3z}{z-\frac{1}{2}} \cdot \frac{1}{3}\ \frac{1+3z^{-1}}{z^{-1}-\frac{1}{2}}\right]_{z=e^{j\omega}} \\ &= \sigma_{\bar{e}}^2 \ \left[W_{\text{mp}}\left(z\right) W_{\text{mp}}\left(z^{-1}\right)\right]_{z=e^{j\omega}}. \end{aligned}$$

The process $\{y(t)\}$ can now be expressed in terms of an equivalent filtering:

$$y(t) = W_{\text{mp}}(z)\ \bar{e}(t),$$

where

$$W_{\text{mp}}(z) = \frac{1}{3}\ \frac{1+3z}{z-\frac{1}{2}} = \frac{z+\frac{1}{3}}{z-\frac{1}{2}}$$

is a minimum phase transfer function (the zero is now at $-\frac{1}{3}$), hence it can be seen as the $\mathcal{Z}$-transform of the impulse response $\{w_{\text{mp}}(t)\}$ of a causal LTI system, in particular such that $w_{\text{mp}}(0) = W_{\text{mp}}(\infty) = 1$, and $\{\bar{e}(t)\}$ is a white noise different from $\{e(t)\}$.

Indeed, defining the function

$$H(z) = \frac{1+3z}{z+3},$$

we find

$$\begin{aligned} y(t) = W(z)\ e(t) &= \frac{z+3}{z-\frac{1}{2}}\ e(t) = \frac{1}{3}\ \frac{z+3}{z-\frac{1}{2}}\ H(z)H^{-1}(z)\ 3e(t) \\ &= \left(\frac{1}{3}\ \frac{z+3}{z-\frac{1}{2}}\ \frac{1+3z}{z+3}\right) \cdot \left(\frac{z+3}{1+3z}\ 3e(t)\right) = W_{\text{mp}}(z)\ \bar{e}(t) \end{aligned}$$

that is, $W_{\text{mp}}(z) = \frac{1}{3}W(z)H(z)$ and $\bar{e}(t) = H^{-1}(z)3e(t)$. It is easy to check that $H(z)$ and its inverse $H^{-1}(z)$, and in general any function of the form

$$\bar{H}(z) = \frac{z+a}{1+\bar{a}z},$$

where $a \in \mathbb{C}$, attain $\left|\bar{H}(z)\left(e^{j\omega}\right)\right| \equiv 1$ on the unit circle[16]; they are called *all-pass filters*[17].

In particular, $H^{-1}(z) = \frac{z+3}{1+3z}$ is a *BIBO-stable* all-pass filter (its only pole is at $-\frac{1}{3}$). Thus, $\{\bar{e}(t)\}$ can be seen as the output of a causal, BIBO-stable all-pass filter whose input is the white noise $\{3e(t)\}$; consequently the spectrum of $\{\bar{e}(t)\}$ is a constant ($= 9\sigma_e^2$), and $\{\bar{e}(t)\}$ is a white noise. Summing up, $y(t) = W_{\text{mp}}(z)\bar{e}(t)$ is another representation of the process $\{y(t)\}$, in terms of a white noise filtered by a causal, minimum phase filter. One then carries out analogous computations as those of the previous example obtaining, by coincidence, the same minimum variance predictor:

$$\begin{aligned} \hat{y}(t|t-1) &= \left(1 - \frac{1}{W_{\text{mp}}(z)}\right)\ y(t); \\ \hat{y}(t|t-1) &= -\frac{1}{3}\hat{y}(t-1|t-2) + \frac{5}{6}y(t-1). \end{aligned}$$

□

3.4 Linear predictors for AR and ARX models

A linear predictor for an AR process

$$\begin{aligned} y(t) - a_1 z^{-1}y(t) - \cdots - a_n z^{-n}y(t) &= e(t); \\ (1 - a_1 z^{-1} - \cdots - a_n z^{-n})y(t) &= e(t); \end{aligned}$$

is *always* computable as in Equation (12), because in symbolic notation

$$y(t) = \frac{1}{1 - a_1 z^{-1} - \cdots - a_n z^{-n}}\ e(t) = \frac{z^n}{z^n - a_1 z^{n-1} - \cdots - a_n}\ e(t),$$

[16] Indeed $\left|\bar{H}\left(e^{j\omega}\right)\right|^2 = \bar{H}\left(e^{j\omega}\right)\overline{\bar{H}\left(e^{j\omega}\right)} = \frac{e^{j\omega}+a}{1+\bar{a}e^{j\omega}}\frac{e^{-j\omega}+\bar{a}}{1+ae^{-j\omega}}$. Multiplying by $e^{j\omega}$ above and below in the second fraction we find $\left|\bar{H}\left(e^{j\omega}\right)\right|^2 = 1$.

[17] Any *product* of such functions is also an all-pass filter.

and the transfer function

$$W(z) = \frac{z^n}{z^n - a_1 z^{n-1} - \cdots - a_n}$$

is always supposed to be BIBO-stable; since, moreover, all its zeros are at the origin, it is also minimum phase. Then one has

$$\begin{aligned} \hat{y}(t|t-1) &= \left(1 - \frac{1}{W(z)}\right)\ y(t) \\ &= \left(1 - (1 - a_1 z^{-1} - \cdots - a_n z^{-n})\right)\ y(t) \\ &= a_1 y(t-1) + \cdots + a_n y(t-n). \end{aligned} \tag{13}$$

Consider now the ARX model:

$$y(t) - a_1 y(t-1) - \cdots - a_n y(t-n) = b_1 u(t-1) + \cdots + b_m u(t-m) + e(t);$$

in symbolic notation, the model reads

$$\begin{aligned} y(t) &= \frac{b_1 z^{-1} + \cdots + b_m z^{-m}}{1 - a_1 z^{-1} - \cdots - a_n z^{-n}}\ u(t) + \frac{1}{1 - a_1 z^{-1} - \cdots - a_n z^{-n}}\ e(t) \\ &= G(z)\ u(t) + W(z)\ e(t). \end{aligned}$$

We define

$$v(t) := y(t) - G(z)\ u(t)$$

and note that, since the variables $\{u(\tau)\}$ are known to the experimenter up to time $t-1$, and since $G(z)$ has by definition at least one delay at the numerator, it holds

$$\hat{v}(t|t-1) = \hat{y}(t|t-1) - G(z)\ u(t).$$

With the above definition the model reads $v(t) = W(z)e(t)$, so that $\{v(t)\}$ now has the form of an AR process, for which we know the best predictor:

$$\begin{aligned} \hat{v}(t|t-1) &= \left(1 - \frac{1}{W(z)}\right)\ v(t) = \left(1 - \frac{1}{W(z)}\right)(y(t) - G(z)\ u(t)) \\ &= \left(1 - \frac{1}{W(z)}\right) y(t) - \left(1 - \frac{1}{W(z)}\right) G(z)\ u(t). \end{aligned}$$

Substituting back $\hat{v}(t|t-1)$ we obtain, finally,

$$\begin{aligned} \hat{y}(t|t-1) &= \hat{v}(t|t-1) + G(z)\ u(t) \\ &= \left(1 - \frac{1}{W(z)}\right) y(t) + \frac{1}{W(z)} G(z)\ u(t) \\ &= (a_1 z^{-1} + \cdots + a_n z^{-n})\ y(t) + (b_1 z^{-1} + \cdots + b_m z^{-m})\ u(t) \\ &= a_1 y(t-1) + \cdots + a_n y(t-n) + b_1 u(t-1) + \cdots + b_m u(t-m). \end{aligned} \tag{14}$$

3.5 Identification of AR and ARX models

3.5.1 The Prediction-Error-Minimization method

As we have seen, given a process conforming to an AR, ARX, ARMA or other model, whose parameters $a_1, \cdots, a_n$, $b_1, \cdots, b_n$, $c_1, \cdots, c_k$ are known, a linear predictor $\hat{y}(t|t-1)$ yields a "best" linear estimate of $y(t)$ given the past values of $\{u(\tau)\}$ and $\{y(\tau)\}$ itself up to time $t-1$. Any such estimate is subject to a random error; for example, referring to (12), this error is $e(t)$. This is the picture, *if the model is known*. Vice versa, given a time series $(u(0), y(0)), (u(1), y(1)), \cdots, (u(N-1), y(N-1)), y(N)$ of measures coming from a system, one can exploit the *empirical* prediction errors yielded by an arbitrary predictor (not necessarily the best one) running through the time series, in order to test whether it is good, and if it happens to be, to estimate the model. Finding the dynamical model, given the input/output measures, is the goal of the branch of systems engineering called *system identification*.

We have seen some examples in which the predictor is just a recursive or instantaneous function of the data, depending on the parameters $a_1, \cdots, a_n$, $b_1, \cdots, b_n$, $c_1, \cdots, c_k$; the ideal goal of system identification would be to find the "true" parameters. Suppose that $(u(0), y(0)), (u(1), y(1)), \cdots, (u(N-1), y(N-1)), y(N)$ are known, and suppose that they come from a realization of two stationary processes $\{u(t)\}$ and $\{y(t)\}$.
The so-called *prediction-error-minimization* (PEM) method prescribes to estimate the parameters by finding finding those $\hat{a}_1, \cdots, \hat{a}_n$, $\hat{b}_1, \cdots, \hat{b}_n$, $\hat{c}_1, \cdots, \hat{c}_k$ that minimize the sum of the squares of the prediction errors (the *residuals*):

$$Q(\hat{a}_1, \cdots, \hat{a}_n, \hat{b}_1, \cdots, \hat{b}_n, \hat{c}_1, \cdots, \hat{c}_k) = \sum_{t=1}^{N} (y(t) - \hat{y}(t|t-1))^2 .$$

In general, this is a nonlinear problem. But in this respect, AR and ARX model are exceptional. The predictors of AR and ARX models have particularly simple expressions: indeed the *crucial* facts about the predictors (13) and (14) are that

- the predictor depends *only* on the measurements of $y(t-1), \cdots, y(t-n)$ and possibly $u(t-1), \cdots, u(t-m)$, but not on past values of the predictor itself (this is not the case for ARMA predictors, see e.g. the examples in Section 3.3);
- except for particular cases that we will consider later, the parameters of the model are in one-to-one correspondence with the expression of the predictor;
- the expressions of both the predictors are *linear* with respect to the parameters of the model.

Consider, for example, the simple ARX model

$$y(t) = a^o y(t-1) + b^o u(t-1) + e(t). \tag{15}$$

As we know, the best linear predictor of $y(t)$ given the past has the form

$$\hat{y}^o(t|t-1) = a^o y(t-1) + b^o u(t-1).$$

The ideal goal of the PEM method is to compute $\hat{y}^o$, but we do not know a^o and b^o, therefore the best that we can do is to go for an approximation based on data. In order to find a suitable estimate of the parameters, we build a predictor with the same structure:

$$\hat{y}(t|t-1) = ay(t-1) + bu(t-1);$$

running $\hat{y}(t|t-1)$ through the data, we collect the empirical errors (i.e. *residuals*) $\epsilon(t) := y(t) - \hat{y}(t|t-1)$, and PEM prescribes to find the predictor that minimizes $Q(a,b) := \sum_{t=1}^N \epsilon(t)^2$, that is to find

$$\begin{aligned}(\hat{a}, \hat{b}) &= \arg\min_{a,b} \sum_{t=1}^{N} \left(y(t) - \hat{y}(t|t-1)\right)^2 \\ &= \arg\min_{a,b} \sum_{t=1}^{N} \left(y(t) - ay(t-1) - bu(t-1)\right)^2 .\end{aligned}$$

Needless to say, this is a job for the method of least squares. Indeed, let

$$\begin{aligned} y_t &:= y(t) \\ \varphi_t &:= \begin{bmatrix} y(t-1) \\ u(t-1) \end{bmatrix} \\ \theta^o &:= \begin{bmatrix} a^o \\ b^o \end{bmatrix} \\ \varepsilon_t &:= e(t) \end{aligned}$$

Then (15) can be rewritten

$$y_t = \varphi_t^\top \theta^o + \varepsilon_t, \quad t = 1, \cdots, N$$

which is the typical model of the least squares theory. Now, the method of least squares finds

$$\hat{\theta}_{\mathrm{LS}} = \arg\min_{\theta \in \mathbb{R}^2} \sum_{t=1}^{N} \left(y_t - \varphi_t^\top \theta\right)^2$$

and $\hat{\theta}_{\mathrm{LS}} = (\hat{a}, \hat{b})$ is the PEM estimate of the model parameters. Note that there is nothing strange in having, among the regressors that explain the sample $y(t)$, a sample of the same process, namely $y(t-1)$. Indeed this is precisely the reason for the name *autoregressive*: the process *regresses on itself.*

With respect to the general AR model

$$y(t) - \sum_{i=1}^{n} a_i^o y(t-i) = e(t) \tag{16}$$

a linear predictor has the form

$$\begin{aligned}\hat{y}(t|t-1) &= \sum_{i=1}^{n} a_i y(t-i) = \left[\begin{array}{ccc} y(t-1) & \cdots & y(t-n) \end{array}\right] \left[\begin{array}{c} a_1 \\ \vdots \\ a_n \end{array}\right] \\ &= \varphi_t^\top \theta;\end{aligned}$$

and with respect to the ARX model

$$y(t) - \sum_{i=1}^{n} a_i^o y(t-i) = \sum_{i=1}^{m} b_i^o u(t-i) + e(t) \tag{17}$$

a linear predictor has the form

$$\begin{aligned}\hat{y}(t|t-1) &= \sum_{i=1}^{n} a_i y(t-i) + \sum_{i=1}^{m} b_i u(t-i) \\ &= \left[\begin{array}{cccccc} y(t-1) & \cdots & y(t-n) & u(t-1) & \cdots & u(t-m) \end{array}\right] \left[\begin{array}{c} a_1 \\ \vdots \\ a_n \\ b_1 \\ \vdots \\ b_m \end{array}\right] \\ &= \varphi_t^\top \theta.\end{aligned}$$

The method of least squares applies without relevant changes. The only difference with the example shown above is that, since now we have

$$\begin{aligned}\varphi_t &= \left[\begin{array}{ccc} y(t-1) & \cdots & y(t-n) \end{array}\right]^\top \quad \text{or} \\ \varphi_t &= \left[\begin{array}{cccccc} y(t-1) & \cdots & y(t-n) & u(t-1) & \cdots & u(t-m) \end{array}\right]^\top\end{aligned}$$

(recall that $m \leq n$ by assumption), and since the first measures are $y(0)$ and possibly $u(0)$, the regressor makes sense only for $t \geq n$. Hence, the normal equations become

$$\left(\sum_{t=n}^{N} \varphi_t \varphi_t^\top\right) \theta = \sum_{t=n}^{N} \varphi_t y_t.$$

3.5.2 Convergence

We pose now the question as whether or not, if (16) or (17) is the *true* model that generates $\{y(t)\}$, $\hat{\theta}_{\text{LS}}$ converges to $\theta^o := \begin{bmatrix} a_1^o & \cdots & a_n^o & b_1^o & \cdots & b_m^o \end{bmatrix}^\top$. Since $e(t)$ is independent of $\{u(t)\}$ and from the past of $\{y(t)\}$, ε_t is independent from φ_t, and you could be tempted to apply Theorem 1.7.1. This cannot be done directly, because here the regressors $\{\varphi_t\}$ are not independent. However, this is not really an issue; the *crucial* fact, here, is indeed that they are independent of ε_t. We state the following results without proof; the interested reader can refer to [17, chapter 8], which treats the subject in much greater detail and in a far more general setting.

Theorem 3.5.1 *Suppose that the AR process $\{y(t)\}_{-\infty}^{+\infty}$ is generated by*

$$y(t) - \sum_{i=1}^{n} a_i^o y(t-i) = e(t),$$

and let $\theta^o = \begin{bmatrix} a_1^o & \cdots & a_n^o \end{bmatrix}^\top$. Suppose, moreover, that

1. *$e(t)$ is independent of $y(t-1), y(t-2), \cdots$;*
2. *the roots of the polynomial $z^n - a_1^o z^{n-1} - \cdots - a_n^o$ lie in the open unit disc ($\{z \in \mathbb{C} \mid |z| < 1\}$);*

then the least squares-estimate

$$\begin{aligned} \hat{\theta}_{\text{LS}} &= \arg\min_{\theta \in \mathbb{R}^n} \sum_{t=n}^{N} (y_t - \varphi_t^\top \theta)^2 \\ &= \textit{solution of the normal equations} \end{aligned}$$

converges almost surely to θ^o as $N \to \infty$.

Theorem 3.5.2 *Suppose that the ARX process $\{y(t)\}_{-\infty}^{+\infty}$ is generated by*

$$y(t) - \sum_{i=1}^{n} a_i^o y(t-i) = \sum_{i=1}^{m} b_i^o u(t-i) + e(t),$$

and let $\theta^o = \begin{bmatrix} a_1^o & \cdots & a_n^o & b_1^o & \cdots & b_m^o \end{bmatrix}^\top$. Suppose, moreover, that

1. *the process $\{u(t)\}_{-\infty}^{+\infty}$ is wide sense stationary, with correlation sequence $R_u(\tau) = \mathsf{E}[u(t)u(t+\tau)]$;*
2. *the Toeplitz matrix*

$$M = \begin{bmatrix} R_u(0) & R_u(1) & R_u(2) & \cdots & R_u(m-1) \\ R_u(1) & R_u(0) & R_u(1) & \cdots & R_u(m-2) \\ R_u(2) & R_u(1) & R_u(0) & \cdots & R_u(m-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R_u(m-1) & R_u(m-2) & R_u(m-3) & \cdots & R_u(0) \end{bmatrix}$$

is positive definite;

3. *$e(t)$ is independent of $y(t-1), y(t-2), \cdots$, and of $u(s)$ for all s;*

4. *the roots of the polynomial $z^n - a_1^o z^{n-1} - \cdots - a_n^o$ lie in the open unit disc;*

then the least squares estimate

$$\hat{\theta}_{\mathrm{LS}} = \arg \min_{\theta \in \mathbb{R}^{n+m}} \sum_{t=n}^{N} (y_t - \varphi_t^\top \theta)^2$$

converges almost surely to θ^o as $N \to \infty$.

Some remarks about Theorem 3.5.2 are in order.

The second assumption of Theorem 3.5.2 is called *persistent excitation* of the input signal, and ensures that the latter carries enough information to make the identifiability possible. From the point of view of least squares, persistent excitation ensures that the estimate $\hat{\theta}_{\mathrm{LS}}$, that is the solution to the normal equations, is unique, at least for big N; recall, indeed, that the only real-world issue that could prevent uniqueness was "the regressors φ_t do not carry enough information".

Example. To see what happens when this does not hold, consider the extreme case $u(t) \equiv 0$ (this is a perfectly legitimate stationary process having correlation signal $R_u(\tau) \equiv 0$ and $M = 0$): it should be obvious that the values, or even the presence, of $b_1, \cdots, b_m$ cannot be seen from either $\{y(t)\}$ or $\{u(t)\}$. Indeed, what happens to the normal equations if $u(t) \equiv 0$? It holds

$$\varphi_t = \begin{bmatrix} y(t-1) \\ u(t-1) \end{bmatrix} = \begin{bmatrix} y(t-1) \\ 0 \end{bmatrix},$$

hence the matrix R at the left-hand side of the normal equations reads

$$\begin{aligned} R = \sum_{t=1}^{N} \varphi_t \varphi_t^\top &= \sum_{t=1}^{N} \begin{bmatrix} y(t-1) \\ 0 \end{bmatrix} \begin{bmatrix} y(t-1) & 0 \end{bmatrix} \\ &= \sum_{t=1}^{N} \begin{bmatrix} y(t-1)^2 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^{N} y(t-1)^2 & 0 \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

Since R is singular, the solution $\begin{bmatrix} \hat{a} & \hat{b} \end{bmatrix}^\top$ to the normal equations cannot be unique, and there is no hope for identifiability. □

The hypothesis that $e(t)$ is independent of $y(t-1), y(t-2), \cdots$ is quite natural, and tells us that the relation between $\{e(t)\}$ and $\{y(t)\}$ is causal.

However, the hypothesis that $e(t)$ is independent of $u(s)$ for all s "hides" an implicit assumption, which may or may not happen in reality, but should anyway be taken seriously: namely that *there is no feedback between* $y(t)$ *and* $u(t)$. Identification in closed loop is indeed a delicate subject.

Example. To see why feedback could lead to trouble, consider for example the simple ARX model

$$y(t) = a^o y(t-1) + b^o u(t-1) + e(t),$$

and assume that $u(t) = ky(t)$ for all t (closed loop). Then

$$y(t) = (a^o + b^o k)y(t-1) + e(t) = \alpha y(t-1) + e(t),$$

and you can already see that there is no way to decouple a^o from b^o, not even if k is known. But what happens to the normal equations? We have

$$\varphi_t = \left[\begin{array}{c} y(t-1) \\ u(t-1) \end{array} \right] = \left[\begin{array}{c} y(t-1) \\ ky(t-1) \end{array} \right],$$

hence the matrix R at the left-hand side of the normal equations reads

$$\begin{aligned} R = \sum_{t=1}^{N} \varphi_t \varphi_t^\top &= \sum_{t=1}^{N} \left[\begin{array}{c} y(t-1) \\ ky(t-1) \end{array} \right] \left[\begin{array}{cc} y(t-1) & ky(t-1) \end{array} \right] \\ &= \sum_{t=1}^{N} \left[\begin{array}{cc} y(t-1)^2 & ky(t-1)^2 \\ ky(t-1)^2 & k^2 y(t-1)^2 \end{array} \right] = \left(\sum_{t=1}^{N} y(t-1)^2 \right) \left[\begin{array}{cc} 1 & k \\ k & k^2 \end{array} \right]. \end{aligned}$$

Again, R is singular, hence the solution to the normal equations cannot be unique, and identifiability is lost. □

We insist on the fact that Theorems 3.5.1 and 3.5.2 are rather particular cases in a specific setting, reported here for the sole purpose of illustrating how the least squares method applies to dynamical systems. But system identification is a branch of information engineering in its own right, coping with much broader issues: identification of ARMAX models, choice of a suitable class of models (ARX? ARMAX?), choice of the *order* of the model (what are some suitable m and n?), deterministic or non-stationary inputs, identification in closed loop, identification of state-space models, multi-variable models, time-varying models, nonlinear models... and so on.

A simulation of 100 least-squares estimations, each based on $N = 1000$ samples of the process (15) with $a = 0.8$, $b = 0.2$, $u(t) \sim \mathcal{N}(0, 1)$ and $e(t) \sim \mathcal{N}(0, 0.09)$, yielded on average $\hat{a}_{\text{LS}} = 0.79872$ (with variance 0.00020 over the 100 runs) and $\hat{b}_{\text{LS}} = 0.19991$ (with variance 0.00008 over the 100 runs), which is rather good. The code is available in Appendix E.1.

3.6 A model with noise in the measures

Consider now a slightly different situation, in which the process $\{y(t)\}_{-\infty}^{+\infty}$ is generated by the linear system

$$y(t) = a^o y(t-1) + b^o u(t-1) \tag{18}$$

without process noise. Now the output $\{y(t)\}$ is not accessible anymore; instead, the experimenter has access to measures of the output corrupted by noise:

$$y_m(t) = y(t) + e(t) \tag{19}$$

where $\{e(t)\}$ are independent and identically distributed random variables with mean zero and a certain variance σ^2. We assume that each $e(t)$ is independent of $u(\tau)$ and $y(\tau)$ for all $\tau \in \mathbb{Z}$. Is it still possible to apply the method of least squares to estimate a^o and b^o? It would seem so, because substituting (19) into (18) we obtain

$$\begin{aligned} y_m(t) - e(t) &= a^o(y_m(t-1) - e(t-1)) + b^o u(t-1) \\ y_m(t) &= a^o y_m(t-1) + b^o u(t-1) + (e(t) - a^o e(t-1)) \end{aligned}$$

and letting

$$\begin{aligned} y_t &:= y_m(t) \\ \varphi_t &:= \begin{bmatrix} y_m(t-1) \\ u(t-1) \end{bmatrix} \\ \theta^o &:= \begin{bmatrix} a^o \\ b^o \end{bmatrix} \\ \varepsilon_t &:= e(t) - a^o e(t-1) \end{aligned}$$

(note that both y_t and φ_t are available to the experimenter), the model becomes, as before,

$$y_t = \varphi_t^\top \theta^o + \varepsilon_t. \tag{20}$$

Nevertheless, a simulation of 100 least-squares estimations, each based on $N = 1000$ samples of the process (18) with $a^o = 0.8$, $b^o = 0.2$, $u(t) \sim \mathcal{N}(0,1)$, and $e(t) \sim \mathcal{N}(0, 0.09)$, yielded on average $\hat{a}_{\mathrm{LS}} = 0.50657$ (with variance 0.00051 over the 100 runs) and $\hat{b}_{\mathrm{LS}} = 0.20026$ (with variance 0.00010 over the 100 runs; the code is available in Appendix E.2).

Bad.

The estimate of $\hat{a}_{\mathrm{LS}}$ is completely wrong. And the fact is, if you try to simulate with more and more data, the situation will not improve at all.

Why so? The reason is that

$$\begin{aligned}
\mathsf{E}\left[\varphi_t\varepsilon_t\right] &= \left[\begin{array}{c} \mathsf{E}\left[y_m(t-1)(e(t)-a^o e(t-1))\right] \\ \mathsf{E}\left[u(t-1)(e(t)-a^o e(t-1))\right] \end{array}\right] \\
&= \left[\begin{array}{c} \mathsf{E}\left[(y(t-1)+e(t-1))(e(t)-a^o e(t-1))\right] \\ \mathsf{E}\left[u(t-1)(e(t)-a^o e(t-1))\right] \end{array}\right] \\
&= \left[\begin{array}{c} -a^o\sigma^2 \\ 0 \end{array}\right].
\end{aligned}$$

The regressor and the disturbance are correlated (in the first component, and indeed it is the first component of $\hat{\theta}_{\mathrm{LS}}$ that yields a bad estimate). Therefore, none of the theorems about almost sure convergence of $\hat{\theta}_{\mathrm{LS}}$ apply.

However, for a broad class of input signals $\{u(t)\}$ here we can apply the method of *instrumental variables* (see Section 1.7.4), and it so happens that the vector $\psi_t := \left[\begin{array}{c} u(t-2) \\ u(t-1) \end{array}\right]$ is often a good instrumental variable.
To keep the example simple enough, suppose that $\{u(t)\}$ is *a white noise* with variance σ_u^2. Note that, if this is the case, then $u(t)$ is independent (hence uncorrelated) from $y(\tau)$ for all $\tau \leq t$, because the model is always supposed to be *causal.* Suppose, moreover, that $b^o \neq 0$.
Then we have

$$\begin{aligned}
\mathsf{E}\left[\psi_t\varphi_t^\top\right] &= \mathsf{E}\left[\left[\begin{array}{c} u(t-2) \\ u(t-1) \end{array}\right]\left[\begin{array}{cc} y_m(t-1) & u(t-1) \end{array}\right]\right] \\
&= \left[\begin{array}{cc} \mathsf{E}\left[u(t-2)(a^o y(t-2)+b^o u(t-2)+e(t-1))\right] & \mathsf{E}\left[u(t-2)u(t-1)\right] \\ \mathsf{E}\left[u(t-1)(a^o y(t-2)+b^o u(t-2)+e(t-1))\right] & \mathsf{E}\left[u(t-1)u(t-1)\right] \end{array}\right] \\
&= \left[\begin{array}{cc} \mathsf{E}\left[b^o u(t-2)^2\right] & 0 \\ 0 & \mathsf{E}\left[u(t-1)^2\right] \end{array}\right] = \sigma_u^2\left[\begin{array}{cc} b^o & 0 \\ 0 & 1 \end{array}\right],
\end{aligned}$$

which is invertible since $b^o \neq 0$; moreover,

$$\mathsf{E}\left[\psi_t\varepsilon_t\right] = \mathsf{E}\left[\left[\begin{array}{c} u(t-2)(e(t)-a^o e(t-1)) \\ u(t-1)(e(t)-a^o e(t-1)) \end{array}\right]\right] = 0.$$

Solving

$$\left(\sum_{t=2}^{N}\psi_t\varphi_t^\top\right)\hat{\theta} = \sum_{t=2}^{N}\psi_t y_t$$

we find a good estimate of a^o and b^o (the sums start from $t = 2$ because the instrumental variable contains $u(t-2)$, and the first available measure is $u(0)$).

A simulation of 100 instrumental-variable estimations, each based on $N = 1000$ samples of the process (18) with $a^o = 0.8$, $b^o = 0.2$, $u(t) \sim \mathcal{N}(0,1)$, and

$e(t) \sim \mathcal{N}(0, 0.09)$, yielded on average $\hat{a}_{\text{LS}} = 0.80565$ (with variance 0.00364 over the 100 runs) and $\hat{b}_{\text{LS}} = 0.19820$ (with variance 0.00013 over the 100 runs), which is now fairly good (the code is available in Appendix E.3).

Finding *good* instrumental variables is a delicate problem (in multivariate statistics, not just in system identification). For the dynamical case at hand, you may refer to [30, Chapter 8], which is dedicated to the subject.

3.7 Example: the periodicity of solar activity

Sunspots are small regions that appear periodically on the surface of the Sun, due to the magnetic activity of its photosphere, that are visible as 'dark spots' since they emit less radiation than the surrounding environment (although they *do* emit a lot of radiation). The number of sunspots present in each year has been collected for about three centuries, and aggregated in an index, called *Wolf's* (or *Wolfer's*) number, which takes into account the fact that they usually come in pairs and other details. Here is a table of Wolf's numbers from 1749 to 1924:

year	num.	year	num.	year	num.	year	num.	year	num.	year	num.
1749	80.9	1779	125.9	1809	2.5	1839	85.7	1869	74.0	1899	12.1
1750	83.4	1780	84.8	1810	0.0	1840	64.6	1870	139.0	1900	9.5
1751	47.7	1781	68.1	1811	1.4	1841	36.7	1871	111.2	1901	2.7
1752	47.8	1782	38.5	1812	5.0	1842	24.2	1872	101.6	1902	5.0
1753	30.7	1783	22.8	1813	12.2	1843	10.7	1873	66.2	1903	24.4
1754	12.2	1784	10.2	1814	13.9	1844	15.0	1874	44.7	1904	42.0
1755	9.6	1785	24.1	1815	35.4	1845	40.1	1875	17.0	1905	63.5
1756	10.2	1786	82.9	1816	45.8	1846	61.5	1876	11.3	1906	53.8
1757	32.4	1787	132.0	1817	41.1	1847	98.5	1877	12.4	1907	62.0
1758	47.6	1788	130.9	1818	30.1	1848	124.7	1878	3.4	1908	48.5
1759	54.0	1789	118.1	1819	23.9	1849	96.3	1879	6.0	1909	43.9
1760	62.9	1790	89.9	1820	15.6	1850	66.6	1880	32.3	1910	18.6
1761	85.9	1791	66.6	1821	6.6	1851	64.5	1881	54.3	1911	5.7
1762	61.2	1792	60.0	1822	4.0	1852	54.1	1882	59.7	1912	3.6
1763	45.1	1793	46.9	1823	1.8	1853	39.0	1883	63.7	1913	1.4
1764	36.4	1794	41.0	1824	8.5	1854	20.6	1884	63.5	1914	9.6
1765	20.9	1795	21.3	1825	16.6	1855	6.7	1885	52.2	1915	47.4
1766	11.4	1796	16.0	1826	36.3	1856	4.3	1886	25.4	1916	57.1
1767	37.8	1797	6.4	1827	49.6	1857	22.7	1887	13.1	1917	103.9
1768	69.8	1798	4.1	1828	64.2	1858	54.8	1888	6.8	1918	80.6
1769	106.1	1799	6.8	1829	67.0	1859	93.8	1889	6.3	1919	63.6
1770	100.8	1800	14.5	1830	70.9	1860	95.8	1890	7.1	1920	37.6
1771	81.6	1801	34.0	1831	47.8	1861	77.2	1891	35.6	1921	26.1
1772	66.5	1802	45.0	1832	27.5	1862	59.1	1892	73.0	1922	14.2
1773	34.8	1803	43.1	1833	8.5	1863	44.0	1893	85.1	1923	5.8
1774	30.6	1804	47.5	1834	13.2	1864	47.0	1894	78.0	1924	16.7
1775	7.0	1805	42.2	1835	56.9	1865	30.5	1895	64.0		
1776	19.8	1806	28.1	1836	121.5	1866	16.3	1896	41.8		
1777	92.5	1807	10.1	1837	138.3	1867	7.3	1897	26.2		
1778	154.4	1808	8.1	1838	103.2	1868	37.6	1898	26.7		

Figure 1 shows a plot of the numbers tabulated above. You will of course notice that the numbers exhibit some kind of periodicity; for a lot of time this somewhat unexpected regularity has been of great interest to astronomers. Can we design a systematic procedure to infer something about it? The first answer that comes to mind is: let's model the phenomenon as a constant plus a sinusoid plus some noise,

$$n(t) = K + A\sin(\omega t + \phi) + e(t),$$

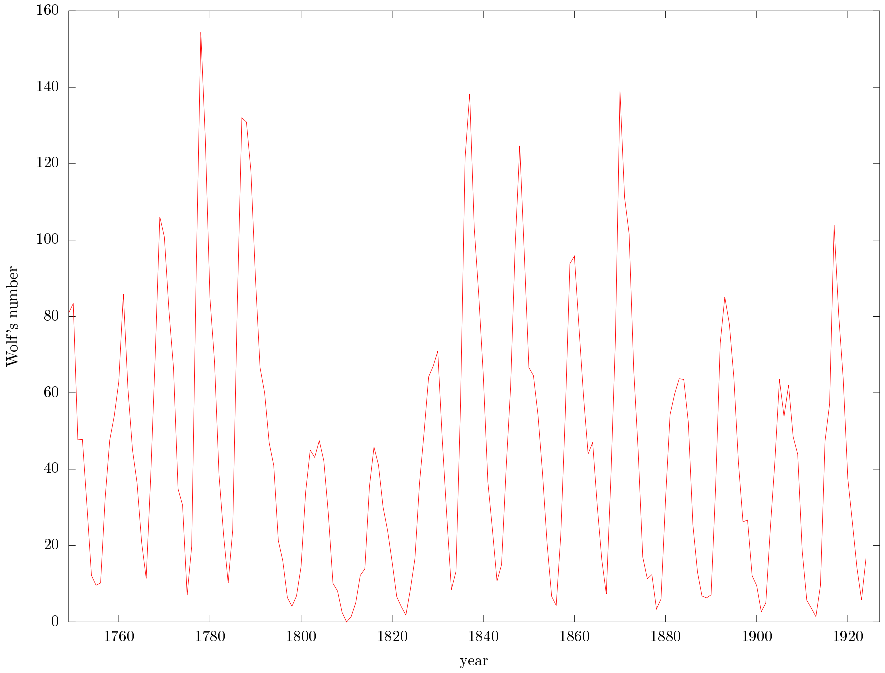

Figure 1: Wolf's sunspot numbers from 1749 to 1924.

and extract the interesting parameters, above all ω, with some numerical algorithm (for example *nonlinear* least squares, see Chapter 4). The philosophy behind the this method is that the phenomenon under investigation is perfectly deterministic and regular (a sinusoid!), and that the randomness in the numbers is only due to errors in the measurements.

To the author's knowledge, George Udny Yule, in his 1927 paper [31], was the first statistician to model the numbers themselves as the realization of a stationary stochastic process, and to use autoregressive models in time series analysis for the first time. He subtracted from Wolf's number their mean $(y(t) := n(t) - \frac{1}{176}\sum_{t=1749}^{1924} n(t))$ to 'de-trend' their time series, and supposed that the result was a realization of an AR model of order 2,

$$y(t) = ay(t-1) + by(t-2) + e(t), \tag{21}$$

where $\{e(t)\}$ were independent and identically distributed random variables, with mean zero and a certain variance σ^2. The order 2 is indeed the minimum order such that the transfer function of the system can have a pair of complex conjugate poles; this is a necessary condition for a *resonance*, that is a a "peak" in the power spectrum, and a more or less pronounced oscillatory behavior in realizations.

Yule found an estimate of a and b with the method of least squares, exactly as we have done in Section 3.5. One could carry on with the computations in the usual way, by solving the normal equations (see e.g. the code in Appendix E.4), but of course ad-hoc software is available and in widespread use. The most common software package for system identification is Matlab's System Identification Toolbox, originally written by Lennart Ljung (see [20]). Here follows, instead, a short script in the statistically-oriented programming language R[18], which has both the advantage of providing Wolf's numbers as a standard dataset, and that of being probably the most powerful free-software alternative to Matlab for time series analysis:

```
require(datasets)            # This package contains many standard datasets,
                             # among them Wolf's numbers from the 18th to the 20th century
x <- sunspot.year[50:225]    # Wolf's numbers from 1749 to 1924, as in Yule's paper
x <- x - mean(x)             # "Detrend", that is center around 0

# Fit an AR(2) model. aic=FALSE and order.max=2 instruct the routine not to choose
# the model order automatically, but to let it be 2.
# method="ols" means "use ordinary least squares"
model <- ar(x, aic=FALSE, order.max=2, method="ols")

print( round(as.vector(model$ar), digits=2) )  # Print the coefficients
resid_var <- var(model$resid, na.rm=TRUE)      # The first two residuals are not available;
                                               # na.rm=TRUE tells var() to ignore them
print( round(resid_var, digits=2) )            # Print the variance of residuals
```

We find $\hat{a} = 1.34$ and $\hat{b} = -0.65$, exactly as in Yule's paper, and we recover an estimate of the variance σ^2 as the sample variance of the residuals $\{\epsilon(t) = y(t) - \hat{a}y(t-1) - \hat{b}y(t-2)\}$, which turns out to be $\hat{\sigma}^2 = 239.31$.

Since the transfer function of (21) is $W(z) = \frac{1}{1-az^{-1}-bz^{-2}}$, in view of Theorem 3.1.1, the spectrum of y is

$$\begin{aligned}
S_y(\omega) &= W\left(e^{j\omega}\right) W\left(e^{-j\omega}\right) \sigma^2 \\
&= \frac{1}{1 - ae^{-j\omega} - be^{-2j\omega}} \cdot \frac{1}{1 - ae^{j\omega} - be^{2j\omega}} \, \sigma^2 \\
&= \frac{\sigma^2}{1 + a^2 + b^2 - a(e^{-j\omega} + e^{-j\omega}) - b(e^{2j\omega} + e^{-2j\omega}) + ab(e^{j\omega} + e^{-j\omega})} \\
&= \frac{\sigma^2}{1 + a^2 + b^2 - 2a\cos\omega - 2b\cos 2\omega + 2ab\cos\omega} \\
&= \frac{\sigma^2}{1 + a^2 + b^2 + 2a(b-1)\cos\omega - 2b\cos 2\omega}.
\end{aligned}$$

Hence, our *estimate* of the spectrum, based on the *assumption* that Wolf's numbers conform to an AR model of order 2, is

$$\hat{S}_y(\omega) = \frac{\hat{\sigma}^2}{1 + \hat{a}^2 + \hat{b}^2 + 2\hat{a}(\hat{b}-1)\cos\omega - 2\hat{b}\cos 2\omega}. \qquad (22)$$

[18] See e.g. `http://www.r-project.org/` .

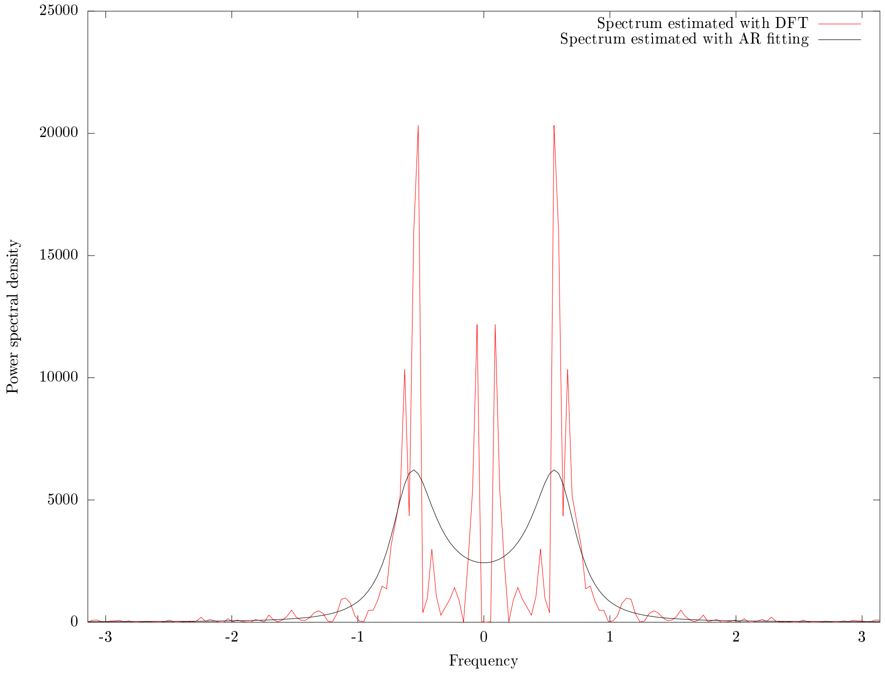

Figure 2: Spectrum of the de-trended Wolf's numbers: estimate with AR(2) fitting and with periodogram.

Another popular estimate of the spectrum of a sequence $\{y(t)\}$, given N of its samples, is the so-called *periodogram*, that is a rescaled version of the square modulus of its discrete Fourier transform (DFT, colloquially called also "FFT" due to the extreme convenience of the mainstream algorithm used to compute it):

$$\tilde{S}_y(\omega) = \frac{1}{N} \mid \mathrm{DFT}[y](\omega) \mid^2 = \frac{1}{N} \left| \sum_{t=0}^{N-1} y(t) e^{j\omega t} \right|^2 .$$

There are techniques to smoothen the periodogram and extract salient information from it, but despite these, the periodogram is widely regarded as a *bad* estimator of the true spectrum (in particular, it is biased and not at all consistent). It is, nevertheless, a frequently used estimator. In Figure 2 you can find a comparison of the estimates of the spectrum obtained by the above two methods.

You can see that a pair of pronounced peaks, symmetric around 0, is present in both estimates, the positive one being at some frequency between 0 and 1 rad; it is precisely those peaks, in the frequency domain, that represent the oscillatory behavior so evident in the time domain. We are now interested in their position: from the periodogram, it can only be guessed by inspection,

but from the expression of Yule's estimate it can be computed analytically, being nothing else than the arg max of $\hat{S}_y(\omega)$.
So, what is the frequency $\bar{\omega}$ that maximizes (22)? It has to be the same $\bar{\omega}$ that minimizes the denominator, since the numerator is a constant. We let

$$\begin{aligned}
0 &= \frac{\partial}{\partial\omega}\left[1 + \hat{a}^2 + \hat{b}^2 + 2\hat{a}(\hat{b}-1)\cos\omega - 2\hat{b}\cos 2\omega\right] \\
&= 2\hat{a}(1-\hat{b})\sin\omega + 4\hat{b}\sin 2\omega \\
&= 2\hat{a}(1-\hat{b})\sin\omega + 8\hat{b}\sin\omega\cos\omega;
\end{aligned}$$

we see by inspection that the interesting frequency is neither 0 nor $\pm\pi$; hence we can divide on both sides by $2\sin\omega \neq 0$ and obtain[19]

$$\begin{aligned}
0 &= \hat{a}(1-\hat{b}) + 4\hat{b}\cos\omega; \\
\bar{\omega} &= \arccos\frac{\hat{a}(\hat{b}-1)}{4\hat{b}} = \arccos\frac{1.34\cdot(0.65+1)}{4\cdot 0.65} \simeq 0.554.
\end{aligned}$$

A simpler method to estimate the position of the peaks, which gives a similar (but not the same) result if the peaks are "high", is simply to compute the phase of the corresponding poles of $W(z)$. We have

$$\hat{W}(z) = \frac{z^2}{z^2 - \hat{a}z - \hat{b}} = \frac{z^2}{z^2 - 1.34z + 0.65},$$

the roots of $z^2 - 1.34z + 0.65$ are $0.67 \pm 0.45j \simeq 0.8e^{\pm j0.59}$, and we recover the second estimate $\hat{\omega} \simeq 0.590$.

Finally, the *periods* corresponding to $\bar{\omega} = 0.554$ and $\hat{\omega} \simeq 0.590$ are

$$\bar{T} = \frac{2\pi}{\bar{\omega}} \simeq 11.3, \qquad \hat{T} = \frac{2\pi}{\hat{\omega}} \simeq 10.6.$$

Indeed, it is now a well-known fact that solar activity has a periodicity of about 11 *years.* The periodicity of Wolf's sunspot numbers is but one of many experimental evidences of this fact. Yule's method has allowed us to recover a sufficiently accurate estimate *of a physical quantity* from the estimate of a model: this is a typical example in which system identification is employed to investigate a "hidden" property of a complex system, as opposed to its other main objective, which is the prediction of future samples.

[19] We recover $\bar{\omega} = 0.554$ by the standard definition of the arc cosine, and 0.554 is the position of the *right* peak in the spectrum; of course -0.554, the position of the *left* peak, is also a solution.

3.8 Recursive least squares

Suppose that we are given input and output measures from the ARX system of order p, m:

$$y(t) = \sum_{t=1}^{p} a_i^o y(t-i) + \sum_{i=1}^{m} b_i^o u(t-i) + e(t) \tag{23}$$

where $m \leq p$, and let

$$y_t := y(t) \qquad \varphi_t := \begin{bmatrix} y(t-1) \\ \vdots \\ y(t-p) \\ u(t-1) \\ \vdots \\ u(t-m) \end{bmatrix} \qquad \theta^o := \begin{bmatrix} a_1^o \\ \vdots \\ a_p^o \\ b_1^o \\ \vdots \\ b_m^o \end{bmatrix} \qquad \varepsilon_i \quad := e(i)$$

Then (23) reads

$$y_t = \varphi_t^\top \theta^o + \varepsilon_t \quad \text{for all } t \geq 1.$$

But suppose, now, that the measures form a potentially infinite sequence:

$$t = 1, \cdots, n_0, \cdots, n, \cdots$$

that is, the system is "running" and it will continue to do so potentially forever. We pose the problem of identifying θ^o; since in this case the method of least squares is consistent, the least squares estimate will improve as more and more measures (φ_t, y_t) come, and it will (almost surely) converge to θ^o. Of course, the same method of section 3.5 can be applied without changes (compute $\hat{\theta}_{\text{LS}}$ at each time), but you can easily guess that in this way computations become more and more expensive as time passes and measures become available, because sums and inverses have to be computed each time, and at each time the old solution is discarded. This seems, and is indeed, a waste of resources.

There is a smarter way to proceed. Suppose that at a certain time n_0 the least squares solution is available *and unique*, in other words that the matrix

$$R(n_0) := \sum_{t=1}^{n_0} \varphi_t \varphi_t^\top$$

has full rank. This implies, of course, that all the sums that will follow

$$R(n) := \sum_{t=1}^{n} \varphi_t \varphi_t^\top = R(n_0) + \sum_{t=n_0+1}^{n} \varphi_t \varphi_t^\top$$

will also have full rank for all $n \geq n_0$. For $n \geq n_0$, denote the least squares solution at time n

$$\hat{\theta}_{\text{LS}}(n) := \left(\sum_{t=1}^{n} \varphi_t \varphi_t^\top\right)^{-1} \left(\sum_{t=1}^{n} \varphi_t y_t\right).$$

The fact is that, if at time n we store the two sums

$$R(n) = \sum_{t=1}^{n} \varphi_t \varphi_t^\top, \qquad S(n) = \sum_{t=1}^{n} \varphi_t y_t,$$

(hence $\hat{\theta}_{\text{LS}}(n) = R(n)^{-1}S(n)$), then they can be updated to yield the corresponding sums at time $n+1$ quite trivially:

$$\begin{aligned} R(n+1) &= \sum_{t=1}^{n+1} \varphi_t \varphi_t^\top = R(n) + \varphi_{n+1}\varphi_{n+1}^\top, \\ S(n+1) &= \sum_{t=1}^{n+1} \varphi_t y_t = S(n) + \varphi_{n+1} y_{n+1}, \end{aligned}$$

and of course $\hat{\theta}_{\text{LS}}(n{+}1) = R(n{+}1)^{-1}S(n{+}1)$. If now we store $R(n{+}1), S(n{+}1)$ in place of $R(n), S(n)$, then we can use them to compute $R(n{+}2), S(n{+}2)$, obtain $\hat{\theta}_{\text{LS}}(n+2)$... and so on. The solution will require, at each time, a matrix inversion and some additions to compute the new solution. And this is already something noteworthy: storing a *finite* amount of information, namely the two matrices, we can update the least square estimate without recomputing the same sums over and over again[20].

But there is more. Since $\hat{\theta}_{\text{LS}}(n) = R(n)^{-1}S(n)$, of course $S(n) = R(n)\hat{\theta}_{\text{LS}}(n)$, and

$$\begin{aligned} \hat{\theta}_{\text{LS}}(n+1) &= R(n+1)^{-1}S(n+1) \\ &= R(n+1)^{-1}\left(S(n) + \varphi_{n+1}y_{n+1}\right) \\ &= R(n+1)^{-1}\Big(R(n)\hat{\theta}_{\text{LS}}(n) + \varphi_{n+1}y_{n+1}\Big) \\ &= R(n+1)^{-1}\Big(R(n+1)\hat{\theta}_{\text{LS}}(n) - \varphi_{n+1}\varphi_{n+1}^\top\hat{\theta}_{\text{LS}}(n) + \varphi_{n+1}y_{n+1}\Big) \\ &= \hat{\theta}_{\text{LS}}(n) + R(n+1)^{-1}\varphi_{n+1}\Big(y_{n+1} - \varphi_{n+1}^\top\hat{\theta}_{\text{LS}}(n)\Big). \end{aligned}$$

[20] As an one-shot exercise, you should now write down a recursive algorithm to update an *average*: given the average $M(n)$ of the numbers $x_1, \cdots, x_n$, and a new incoming number x_{n+1}, find the average $M(n+1)$ of all the numbers. Note that the algorithm storing $R(n)$ and $S(n)$ is fragile, because both the sum tend to "explode" as more and more terms are added. The issue is solved storing the averages instead: $\bar{R}(n) = \frac{1}{n}\sum_{i=1}^{n}\varphi_i\varphi_i^\top$, $\bar{S}(n) = \frac{1}{n}\sum_{i=1}^{n}\varphi_i y_i$.

On the other hand,

$$R(n+1)^{-1} = \left(R(n) + \varphi_{n+1}\varphi_{n+1}^\top\right)^{-1}.$$

Apply the *matrix inversion lemma* (Lemma A.6.1 in the Appendix) with $A = R(n)$, $B = \varphi_{n+1}$, $C = 1$, and $D = \varphi_{n+1}^\top$. Then

$$\begin{aligned}
R(n+1)^{-1} &= \left(R(n) + \varphi_{n+1} 1 \varphi_{n+1}^\top\right)^{-1} \\
&= R(n)^{-1} - R(n)^{-1}\varphi_{n+1}\left(1 + \varphi_{n+1}^\top R(n)^{-1}\varphi_{n+1}\right)^{-1}\varphi_{n+1}^\top R(n)^{-1} \\
&= R(n)^{-1} - \frac{R(n)^{-1}\varphi_{n+1}\varphi_{n+1}^\top R(n)^{-1}}{1 + \varphi_{n+1}^\top R(n)^{-1}\varphi_{n+1}}; \\
R(n+1)^{-1}\varphi_{n+1} &= R(n)^{-1}\varphi_{n+1}\left(1 - \frac{\varphi_{n+1}^\top R(n)^{-1}\varphi_{n+1}}{1 + \varphi_{n+1}^\top R(n)^{-1}\varphi_{n+1}}\right) \\
&= \frac{R(n)^{-1}\varphi_{n+1}}{1 + \varphi_{n+1}^\top R(n)^{-1}\varphi_{n+1}}; \\
\hat{\theta}_{\mathrm{LS}}(n+1) &= \hat{\theta}_{\mathrm{LS}}(n) + R(n+1)^{-1}\varphi_{n+1}\left(y_{n+1} - \varphi_{n+1}^\top\hat{\theta}_{\mathrm{LS}}(n)\right) \\
&= \hat{\theta}_{\mathrm{LS}}(n) + \frac{R(n)^{-1}\varphi_{n+1}}{1 + \varphi_{n+1}^\top R(n)^{-1}\varphi_{n+1}}\left(y_{n+1} - \varphi_{n+1}^\top\hat{\theta}_{\mathrm{LS}}(n)\right).
\end{aligned}$$

Finally, let $P(n) := R(n)^{-1}$ and

$$L(n+1) := \frac{P(n)\varphi_{n+1}}{1 + \varphi_{n+1}^\top P(n)\varphi_{n+1}},$$

so that, recursively,

$$\begin{aligned}
\hat{\theta}_{\mathrm{LS}}(n+1) &= \hat{\theta}_{\mathrm{LS}}(n) + L(n+1)\left(y_{n+1} - \varphi_{n+1}^\top\hat{\theta}_{\mathrm{LS}}(n)\right); \\
P(n+1) &= \left(I - L(n+1)\varphi_{n+1}^\top\right)P(n).
\end{aligned}$$

This is a *new* recursive algorithm; to implement it, we must store $\hat{\theta}_{\mathrm{LS}}(n)$ and $P(n)$ instead of $R(n)$ and $S(n)$, and to compute $L(n+1)$ on-the-fly before updating. Its job is the same as before: given an old estimate $\hat{\theta}_{\mathrm{LS}}(n)$ and a new measure (φ_{n+1}, y_{n+1}), it provides the new estimate $\hat{\theta}_{\mathrm{LS}}(n+1)$. Why all these nasty computations, then, if it is just another recursive algorithm that does the same thing?

Because matrix inversions at each step are not there anymore.

The tricky application of the matrix inversion lemma has indeed turned a matrix inversion into a division by a scalar $(1 + \varphi_{n+1}^\top P(n)\varphi_{n+1})$. This algorithm is *faster*, and it is your algorithm of choice for on-line least squares estimation. In the literature, it is known as *Recursive Least Squares (RLS)*. Some remarks are in order:

1. In the present form, the algorithm cannot be started at time $n = 1$ (that is, at the first sample); it must be started at the first n such that $R(n)$ is invertible, so that $P(n)$ exists. In the above derivation, we have denoted such time n_0. There are workarounds to this (minor) issue that allow to start at $n = 1$, if one accepts a sub-optimal estimate until the algorithm approaches steady state (see [17]).

2. In the above formulas, both $R(n)$ and $P(n)$ are supposed to be symmetric and positive definite. This is vital for RLS to work properly. However, although the update equation

   $$P(n+1) = \left(I - L(n+1)\varphi_{n+1}^\top\right) P(n)$$

   guarantees such assumption mathematically by construction, it does not do so *numerically*; indeed numerical errors due to computations may lead this assumption to fail rather quickly. Thus, in order to implement the algorithm correctly, one should adopt a different, "symmetrized" version of the same equation in order to guarantee symmetry by construction (again, refer to [17] for details).

3. Note the similarity between the update equation

   $$\hat{\theta}_{\text{LS}}(n+1) = \hat{\theta}_{\text{LS}}(n) + L(n+1)\left(y_{n+1} - \varphi_{n+1}^\top \hat{\theta}_{\text{LS}}(n)\right)$$

   and the (Luenberger) *asymptotic observer* of linear system theory. The structure is the same, because $\hat{y}_{n+1} := \varphi_{n+1}^\top \hat{\theta}_{\text{LS}}(n)$ is the best prediction, in the sense of least squares, of the incoming observation y_{n+1}, given the "old" estimate $\hat{\theta}_{\text{LS}}(n)$ and the "new" explanatory data φ_{n+1}; hence $y_{n+1} - \varphi_{n+1}^\top \hat{\theta}_{\text{LS}}(n)$ is a *prediction error*, and the update equation reads

   $$\text{new estimate} \;=\; \text{old estimate} \;+\; \text{gain} \times \text{prediction error.}$$

   The main differences with the Luenberger observer are that $\hat{\theta}_{\text{LS}}(n)$ is the estimate of a fixed quantity θ^o, not of a time-varying state (but this is not really a difference, just a particular case), and that the gain $L(n+1)$ is time-varying, whereas in the Luenberger observer it is a constant. The second difference is due to the fact that $L(n+1)$ comes from a precise optimization criterion (minimize the overall sum of squares) instead of a choice of eigenvalues that drive the estimation error to zero with a prescribed rate.

3.9 Tracking time-varying models

The RLS can be adapted to estimate the parameters of a *time-varying* model. For simplicity, consider an AR model of order p:

$$y(t) = \sum_{i=1}^{p} a_i y(t-i) + e(t) \tag{24}$$

This model describes a stochastic process "generated" by the noise $\{e(t)\}$. As we know, if the model is BIBO-stable, then it describes a wide-sense stationary process. Think at the sampled version of a sustained note played by a flute. We can estimate its parameters, as usual, letting

$$y_t := y(t) \qquad \varphi_t := \begin{bmatrix} y(t-1) \\ \vdots \\ y(t-p) \end{bmatrix} \qquad \theta^o := \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix} \qquad \varepsilon_t := e(t)$$

and applying the RLS algorithm.

In many applications, though, the process at hand is inherently *non*-stationary. Think now at a flute playing a melody. We can try to describe the situation by means of a *slowly time-varying* model, that is, an auto-regressive model whose parameters are not anymore constant with respect to time, but change *slowly* in comparison with the dynamics of the model that they represent:

$$y(t) = \sum_{i=1}^{p} a_i(t) y(t-i) + e(t) \tag{25}$$

We pose the problem of estimating the parameters $a_1(t), \cdots, a_p(t)$ of the process, which we collect in the vector

$$\theta^t := \begin{bmatrix} a_1(t) \\ \vdots \\ a_p(t) \end{bmatrix}$$

(the other quantities being defined as before, the model then reads $y_t = \varphi_t^\top \theta^t + \varepsilon_t$). The objective that we have had until now, that of *converging* to the one and only, "true" parameter θ^o, is gone. Still, we may try to *track* θ^t, time after time.

In this case, the least squares solution computed over *all* the data $\{(\varphi_t, y_t)\}_{t=1}^n$ is meaningless and useless (whether or not it is computed recursively with RLS, it does not matter), because it gives equal importance to regressors and measures belonging to the past (hence following a "past" model, which should be "forgotten" instead) and more recent regressors and measures. Instead, in estimating the "recent" model we should care only about "recent" information. An obvious way to pursue this goal is to compute a least squares estimate on the last N samples $\{(\varphi_{n-N}, y_{n-N}), \cdots, (\varphi_{n-1}, y_{n-1}), y_n\}$, where $N < n$ is a fixed time lag small enough to ensure that not too many measurements are "remembered", but big enough so that the solution exists and is unique. This is a feasible way, because the amount of information to store is always the same (the last N samples, in a buffer), and it involves a matrix inversion at each time.

However, there is a nicer algorithm that does a similar job, and resembles RLS strongly. Let us start, indeed, from the same objective of ordinary least squares:

$$\hat{\theta}_{\text{LS}} := \arg\min_{\theta \in \mathbb{R}^p} \sum_{t=1}^{n} (y_t - \varphi_t^\top \theta)^2,$$

We modify the least squares criterion *weighting* the sum (compare with exercise 3 in Section 1.8), in such a way that the past errors matter less than recent ones:

$$\hat{\theta}_{\text{WLS}}(n) := \arg\min_{\theta \in \mathbb{R}^p} \sum_{t=1}^{n} \lambda^{n-t} (y_t - \varphi_t^\top \theta)^2,$$

where λ is a constant, $0 < \lambda < 1$. We call $\hat{\theta}_{\text{WLS}}(n)$ the "exponentially weighted" least squares estimate of θ^t at time $t = n$. The fact that remote errors matter less and less, with exponential rate of decay, means in other terms that past information gets "forgotten" with exponential rate of decay. The constant λ is indeed called *forgetting factor*; the higher is λ, the more $\hat{\theta}_{\text{WLS}}(n)$ will "remember" past information. The sum

$$M = \sum_{t=-\infty}^{n} \lambda^{n-t} = \sum_{t=0}^{+\infty} \lambda^t = \frac{1}{1-\lambda}$$

is called the *asymptotic memory* of the estimator, that is the "effective" number of measurements that will be taken into consideration, in the limit. The higher the forgetting factor, the higher the asymptotic memory. The solution to the weighted least squares problem is of course $\hat{\theta}_{\text{WLS}}(n) = R(n)^{-1} S(n)$, where

$$R(n) = \sum_{t=1}^{n} \lambda^{n-t} \varphi_t \varphi_t^\top, \qquad S(n) = \sum_{t=1}^{n} \lambda^{n-t} \varphi_t y_t.$$

The trick is now to recognize that

$$\begin{aligned}
R(n+1) &= \sum_{t=1}^{n+1} \lambda^{n+1-t} \varphi_t \varphi_t^\top \\
&= \sum_{t=1}^{n} \lambda^{n+1-t} \varphi_t \varphi_t^\top + \varphi_{n+1}\varphi_{n+1}^\top \\
&= \lambda \sum_{t=1}^{n} \lambda^{n-t} \varphi_t \varphi_t^\top + \varphi_{n+1}\varphi_{n+1}^\top \\
&= \lambda R(n) + \varphi_{n+1}\varphi_{n+1}^\top
\end{aligned}$$

and similarly

$$S(n+1) = \lambda S(n) + \varphi_{n+1} y_{n+1},$$

and then proceed in the same way as in ordinary RLS:

$$\begin{aligned}
\hat{\theta}_{\mathrm{WLS}}(n+1) &= R(n+1)^{-1} S(n+1) \\
&= R(n+1)^{-1} \left(\lambda S(n) + \varphi_{n+1} y_{n+1}\right) \\
&= R(n+1)^{-1} \left(\lambda R(n) \hat{\theta}_{\mathrm{WLS}}(n) + \varphi_{n+1} y_{n+1}\right) \\
&= R(n+1)^{-1} \left(R(n+1) \hat{\theta}_{\mathrm{WLS}}(n) - \varphi_{n+1}\varphi_{n+1}^\top \hat{\theta}_{\mathrm{WLS}}(n) + \varphi_{n+1} y_{n+1}\right) \\
&= \hat{\theta}_{\mathrm{WLS}}(n) + R(n+1)^{-1} \varphi_{n+1} \left(y_{n+1} - \varphi_{n+1}^\top \hat{\theta}_{\mathrm{WLS}}(n)\right) \\
R(n+1)^{-1} &= \left(\lambda R(n) + \varphi_{n+1} 1 \varphi_{n+1}^\top\right)^{-1} \\
&\quad \text{(apply again the matrix inversion lemma...)} \\
&= \frac{1}{\lambda} \left(R(n)^{-1} - \frac{R(n)^{-1}\varphi_{n+1}\varphi_{n+1}^\top R(n)^{-1}}{\lambda + \varphi_{n+1}^\top R(n)^{-1} \varphi_{n+1}}\right) \\
R(n+1)^{-1}\varphi_{n+1} &= \frac{R(n)^{-1}\varphi_{n+1}}{\lambda + \varphi_{n+1}^\top R(n)^{-1} \varphi_{n+1}}
\end{aligned}$$

And letting again $P(n) := R(n)^{-1}$, we obtain the recursive algorithm

$$\begin{aligned}
L(n+1) &= \frac{P(n)\varphi_{n+1}}{\lambda + \varphi_{n+1}^\top P(n) \varphi_{n+1}} \\
\hat{\theta}_{\mathrm{WLS}}(n+1) &= \hat{\theta}_{\mathrm{WLS}}(n) + L(n+1) \left(y_{n+1} - \varphi_{n+1}^\top \hat{\theta}_{\mathrm{WLS}}(n)\right) \\
P(n+1) &= \frac{1}{\lambda} \left(I - L(n+1)\varphi_{n+1}^\top\right) P(n),
\end{aligned}$$

which is called *Exponentially Weighted Least Squares (EWLS).*

4 Nonlinear least squares

4.1 Structure and issues of the problem

Let us return to the first measurement model of Chapter 1:

$$y_i = f(x_i, \theta) + \varepsilon_i, \quad i = 1, \cdots, N.$$

where $y_i \in \mathbb{R}$ are *measures*, $x_i \in \mathbb{R}^m$ are *explanatory data*, $\theta \in \mathbb{R}^p$ is a vector parameterizing the function family "f", and $\varepsilon_i \in \mathbb{R}$ are *disturbance terms.* The general least squares problem is to minimize

$$Q(\theta) = \sum_{i=1}^{N} (y_i - f(x_i, \theta))^2 \tag{26}$$

with respect to the parameter θ. Recall that a great advantage of the least squares method is that if the function f is linear *with respect to* θ, i.e. $f(x_i, \theta) = \varphi(x_i)^\top \theta = \varphi_i^\top \theta$,

$$Q(\theta) = \sum_{i=1}^{N} \left(y_i - \varphi_i^\top \theta \right)^2, \tag{27}$$

then $Q(\theta)$ is convex and differentiable, hence its minimum can be found by equating to zero its derivative (i.e. gradient) with respect to θ; this leads to the normal equations

$$\left(\sum_{i=1}^{N} \varphi_i \varphi_i^\top \right) \theta = \sum_{i=1}^{N} \varphi_i y_i$$

or, with compact notation,

$$\left(\Phi^\top \Phi \right) \theta = \Phi^\top Y,$$

which can be solved by algebraic methods. To distinguish this solution from the general case studied in this chapter, minimizing (27) is often called the *ordinary* least squares problem (OLS).

We ask now the question: what if f is *not* linear with respect to θ? Can we still devise a method to find

$$\arg\min_{\theta \in \mathbb{R}^p} Q(\theta) = \arg\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^{N} (y_i - f(x_i, \theta))^2 \tag{28}$$

in order to fit the data $\{x_i, y_i\}$ with a suitable $f(\cdot, \theta)$? This is called the *nonlinear least squares* problem.

If we do not require from $f(x,\theta)$ any regularity with respect to θ, $Q(\theta)$ can be as nasty as a function can be; hence, the general answer being "no", we assume at least that, for all $x \in \mathbb{R}^m$, $f(x,\theta)$ is sufficiently smooth, that is, *twice continuously differentiable with respect to* θ, so that $Q(\theta)$ is also smooth. Even with this assumption, we immediately stumble across two issues.

For notational simplicity, let us denote $f_i(\theta) := f(x_i,\theta)$. If we blindly set $\frac{\partial Q(\theta)}{\partial \theta} = 0$ to try finding a minimum, we obtain

$$\frac{\partial Q(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta}\sum_{i=1}^{N}(y_i - f_i(\theta))^2 = 2\sum_{i=1}^{N}(y_i - f_i(\theta))\left(-\frac{\partial f_i(\theta)}{\partial \theta}\right) = 0;$$

$$V(\theta) := \sum_{i=1}^{N}\frac{\partial f_i(\theta)}{\partial \theta} f_i(\theta) - \sum_{i=1}^{N}\frac{\partial f_i(\theta)}{\partial \theta} y_i = 0,$$

which in general is a highly nonlinear equation. Our first issue is that finding an analytical solution with an explicit formula, like the one for $\hat{\theta}_{\mathrm{LS}}$ that we found in ordinary least squares, is in general impossible.

However, finding *one* solution to the above equation is not really troublesome, because the *numerical* solution of a nonlinear equation, in our case the problem of finding a root of the function $V(\theta)$, is a well-known and thoroughly studied problem of numerical analysis: namely, there are a lot of general-purpose iterative algorithms around, designed to solve it. An iterative "root-finding" algorithm works more or less like this: *an initial guess point* $\theta^{(0)}$ *is provided by the user*; the algorithm then uses the information on the derivative of V at $\theta^{(0)}$ to guess a direction that will lead to a point $\theta^{(1)}$ attaining a smaller value, i.e. such that $|V(\theta^{(1)})| < |V(\theta^{(0)})|$; then it uses the information at $\theta^{(1)}$ to move to a point $\theta^{(2)}$ such that $|V(\theta^{(2)})| < |V(\theta^{(1)})|$ ... and so on goes the algorithm, finding $\theta^{(3)}$, $\theta^{(4)}$ etc., until a certain $|V(\theta^{(n)})|$ is so small that it may be regarded as zero for all practical purposes (say, 10^{-10}). The corresponding $\theta^{(n)}$ is the numerical solution to the "root-finding" problem.

However, since we have to resort to numerical analysis and iterative algorithms anyway, it turns out that aiming directly at the minimization of $Q(\theta)$ is better, because this optimization problem has a richer structure[21]. Numerical optimization, possibly subject to constraints, is a branch of mathematics standing by its own, and many algorithms to find minima, whether constrained or not, exist and are in widespread use. In particular, if the goal

[21] For example, a zero of $V(\theta)$ could correspond to a *maximum*, to a *saddle*, or it could be a point having no intrinsic extremality property (e.g. consider $V : \mathbb{R} \to \mathbb{R}$, $V(\theta) = \theta^3$: its derivative vanishes only at $\theta = 0$, which has no extremality property at all). On the other hand, optimization algorithms always aim at least at a local minimum.

function to be minimized is *convex*, constrained and unconstrained minimization are regarded as "easy", standard problems, "practically solved". In the words of Stephen Boyd (see [2]), convex optimization is "almost a technology", i.e. it is approaching a stage of maturity in which "it can be reliably used by many people who do not know, and do not need to know, the details"[22]. Thus, if $Q(\theta)$ happens to be convex, there is no significant issue in solving the nonlinear least squares problem (28).

In fact the main, *real* issue is that whenever the functions $f_i(\theta)$ are not linear, there is *hardly* any hope for $Q(\theta)$ to be convex. And with few exceptions in very specific contexts, non-convex optimization is *hard*; the higher the dimension of the problem (that is p, since $\theta \in \mathbb{R}^p$), the harder the problem. There are heuristics, randomized methods and other techniques that try to cope with the issue, but to this date no general-purpose algorithm can ensure, with sufficiently high reliability, that it will be able to find the *global* minimum of an arbitrary non-convex function $Q : \mathbb{R}^{10} \to \mathbb{R}$ in reasonable time.

Since $Q(\theta)$ is supposed to be sufficiently smooth we will, indeed, employ algorithms similar to those commonly adopted in convex optimization. But they can only pursue their goal successfully, and eventually find a solution, in a region of the space $\mathbb{R}^p$ where $Q(\theta)$ is *locally* convex; hence, in the end our algorithm of choice will find a *local* minimum point. As reasonable as such local minimum point may be, in general there is not, and there cannot be, any guarantee that it also attains a *global* minimum. For comparison, keep in mind that the most notable property of convex functions is precisely that any local minimum is also a global one. Now, it is *the user* who selects the region where the algorithm is supposed to find a minimum, *precisely by guessing the initial point* $\theta^{(0)}$; the choice of such point may be obvious for some problems, or difficult for others: in any case it boils down to guesswork, arguably more of an art than of a science.

Unfortunately, there is no general, "off the shelf" way out of this issue. But fortunately for us, our problem is of a particular nature, that of fitting data with a function: for this purpose, the data themselves work on our side. Suppose, for example, that we have to fit some points $(t_1, y_1), \cdots, (t_N, y_N)$ with a sinusoid:

$$y_i = A \sin(2\pi F t_i + \phi) + \varepsilon_i,$$

where the parameter to find is $\theta = (A, F, \phi) \in \mathbb{R}^3$. This problem is clearly nonlinear[23]; it will be solved by means of an iterative algorithm, hence we

[22] In comparison, according to him the method of ordinary least squares is a mature technology.

[23] because the frequency F is unknown and has to be estimated along with A and ϕ. Compare with Problem 2 (amplitude and phase of a sinusoid) at the end of Chapter 1,

must provide an initial guess $\theta^{(0)} = (A^{(0)}, F^{(0)}, \phi^{(0)})$. If the sinusoidal model comes from someone else's guess and, by chance, a plot of the data reveals clearly that they would better fit by a *parabola*, or by an *exponential*, or anything different from a sinusoid, we may wonder whether the model is worthy *before* running any algorithm, and maybe reject it altogether. On the other hand, suppose that they really conform to the sinusoidal model. If, as it may happen, the sequence $(t_1, y_1), \cdots, (t_N, y_N)$ is the sampled version of an analog voltage signal, where the variables t_i represent time and are equally spaced, and if the sampling frequency is, say, 8 kHz, both common sense and Shannon's sampling theorem will tell us that it is totally pointless to choose any $F^{(0)} > 4$ kHz. Instead, a plot of the data may reveal, by inspection, that the frequency is "something between 10 Hz and 100 Hz". Choose, for instance $F^{(0)} = 50$ Hz! The same plot may show that the highest measures among the $\{y_i\}$ attain, more or less, values between 310 V and 330 V. Then choose $A^{(0)} = 320$ V! But perhaps for $\phi^{(0)}$ you have no clue? Let $\phi^{(0)} = 0$ rad and have faith: if the sinusoidal model is really worth, the method will work.

Of course engineers should always use common sense, not just *before* running a procedure, but also *after* having obtained the results. Plot the data and the fitting curve together: you will see immediately whether the fit makes sense or not. Fitting mathematical models to data is an art! Always inspect your data before applying an algorithm blindly, always check your results, please do not exaggerate with those significant digits, and please aim at least at the same kind of common sense that would lead, say, a physician to reject a negative blood pressure with disdain, an employee to regard as 'suspect' a payment of $2,000,000$ € for this month's salary, and your electrician to mistrust a watt-meter measuring 50 kW as the consumption of your TV set. You got the idea.

4.2 Iterative algorithms

4.2.1 General discussion

We proceed with a brief discussion of some standard algorithms adopted for the minimization of *convex* functions. The fundamental fact about convex problems is that if a point is *locally* a minimum, i.e. it satisfies some minimality condition in the neighborhood of a point, then it is a minimum also *globally*, hence it is a solution to the problem.

We suppose that the function $Q : \mathbb{R}^p \to \mathbb{R}$ to be minimized is everywhere *differentiable*; then, in view of Theorem B.3.1, for all $\theta, \bar{\theta} \in \mathbb{R}^p$ it holds

$$Q(\theta) \geq Q\left(\bar{\theta}\right) + \nabla Q\left(\bar{\theta}\right)^\top \left(\theta - \bar{\theta}\right)$$

where F is known; that problem can be reduced to an OLS problem by means of a simple trigonometrical trick.

where $\nabla Q\left(\bar{\theta}\right)$ denotes the *gradient* of Q computed at $\bar{\theta}$. Note that the gradient, by definition, should always be interpreted as a *row* vector; however, in numerical analysis its transpose is used instead:

$$\nabla Q(\theta) = \left[\begin{array}{cccc} \frac{\partial Q(\theta)}{\partial \theta_1} & \frac{\partial Q(\theta)}{\partial \theta_2} & \cdots & \frac{\partial Q(\theta)}{\partial \theta_p} \end{array} \right]^\top.$$

If we find a particular $\bar{\theta}$ such that $\nabla Q\left(\bar{\theta}\right) = 0$, then $Q(\theta) \geq Q\left(\bar{\theta}\right)$ for all $\theta \in \mathbb{R}^p$, so that $\bar{\theta}$ is a global minimum; thus, for convex differentiable functions a local test of optimality on the gradient is sufficient to establish global optimality.
Iterative algorithms for convex minimization proceed in the following way:

- an initial guess $\theta^{(0)}$ is given;
- at the i-th step, $\theta^{(i)}$ is the current candidate for the optimum point $\bar{\theta}$; if it is not good enough, it is improved by setting $\theta^{(i+1)} \leftarrow \theta^{(i)} + \Delta\theta^{(i)}$ in such a way that $Q\left(\theta^{(i+1)}\right) < Q\left(\theta^{(i)}\right)$;
- the iteration is stopped when $|\Delta\theta^{(i)}|$ is sufficiently small, e.g. $|\Delta\theta^{(i)}| < \varepsilon = 10^{-10}$, which means that $\theta^{(i)}$ has 'settled' close to a minimum point, or when the gradient is sufficiently small, e.g. $\|\nabla Q\left(\theta^{(i)}\right)\| < \varepsilon = 10^{-10}$, which means that we are close to $\nabla Q(\theta) = 0$ (here the norm $\|\cdot\|$ is not necessarily the Euclidean one).

At each step (i), the displacement $\Delta\theta$ must be chosen in order to improve the current guess θ. The canonical way is to choose $\Delta\theta = \alpha v$, where v is a "descent direction", and $\alpha > 0$ a constant. Recall that the *directional derivative* of Q along a certain direction v is defined as

$$\frac{\delta Q(\theta)}{\delta v} := \lim_{\alpha \to 0^+} \frac{Q(\theta + \alpha v) - Q(\theta)}{\alpha}, \tag{29}$$

and under regularity assumptions on Q, one finds that

$$\frac{\delta Q(\theta)}{\delta v} = \nabla Q(\theta)^\top v; \tag{30}$$

now, v is by definition a *descent direction* if

$$\frac{\delta Q(\theta)}{\delta v} < 0, \tag{31}$$

that is, taking a little step in the direction of the vector v, the function Q decreases; the classical minimization algorithms differ above all in the choice of a suitable descent direction.

Note that even if *locally* v is a descent direction, taking the "full step" v, that is setting $\Delta\theta = v$, may in general be a bad choice, for it can actually attain $Q(\theta + v) > Q(\theta)$. The step should not be too large, and this is the reason for introducing a multiplier α in the choice $\Delta\theta = \alpha v$. Let v be a descent direction, and $\beta \in (0,1)$, for example $\beta = \frac{1}{2}$. A simple procedure to select a suitable α is the following one, called *backtracking*:

- let $\alpha^{(0)} = 1$;
- if $Q(\theta + \alpha^{(k)} v) < Q(\theta)$, then $\alpha = \alpha^{(k)}$ is OK;
- otherwise, set $\alpha^{(k+1)} \leftarrow \beta\alpha^{(k)}$ and repeat the previous test.

That the backtracking procedure finds a suitable α is ensured by (29) and (31), because v is assumed to be a descent direction and the sequence $\{\alpha^{(k)}\}$ tends to zero from above.

4.2.2 Steepest descent

Looking at equation (30) you will immediately recognize that, unless $\nabla Q(\theta) = 0$ (sufficient condition for optimality), $v := -\nabla Q(\theta)$ is always a descent direction; indeed

$$\frac{\delta Q(\theta)}{\delta v} = \nabla Q(\theta)^\top v = \nabla Q(\theta)^\top (-\nabla Q(\theta)) = -\left\|\nabla Q(\theta)\right\|^2 < 0.$$

Actually, among the vectors v with the same modulus as $\nabla Q(\theta)$, $-\nabla Q(\theta)$ is the one that guarantees the *least* possible value of the expression $\nabla Q(\theta)^\top v$; for this reason the algorithm adopting such descent direction is named *steepest descent.*

One may think that the above one is the best possible choice; this is not true unless in rather particular cases, because the property of attaining the steepest descent at a given point is only *local*: it attains "greedily" a seemingly fast descent at each step, but in ill-conditioned problems it may require a huge lot of steps to converge to the minimizing solution $\bar{\theta}$. There are indeed better choices (the Newton step, discussed below) that do not pursue the direction with minimum slope at each step, but that attain a faster convergence rate, in term of the number of steps. However, it can be proven that the steepest descent algorithm *does* converge to the optimal $\bar{\theta}$; its main advantages are that it is the simplest to understand and to implement, and it requires small computational burden to compute the descent direction.

4.2.3 The Newton step

A *quadratic form on* $\mathbb{R}^p$ is a function $q : \mathbb{R}^p \to \mathbb{R}$ like the following one:

$$q(v) = a + b^\top v + \frac{1}{2} v^\top A v,$$

where $a \in \mathbb{R}$, $b \in \mathbb{R}^p$, and $A \in \mathbb{R}^{p \times p}$.
If A is symmetric and positive semi-definite ($A \geq 0$), then q is called a *positive semi-definite* quadratic form; if, moreover, $A > 0$, q is called *positive definite*, its graph is that of a convex elliptic paraboloid, and it admits a unique minimum point (the vertex of the paraboloid). The minimum is found equating to zero its derivative (i.e. gradient) with respect to v:

$$\begin{aligned} 0 &= \frac{\partial q(v)}{\partial v} = b^\top + v^\top A; \\ Av &= -b; \end{aligned}$$

If $A > 0$ (invertible), then the minimum point is

$$v = -A^{-1}b.$$

Basically, the Newton algorithm works in the following way:

- an initial guess $\theta^{(0)}$ is given;
- at the i-th step, $\theta^{(i)}$ is the current candidate for the optimum point $\bar{\theta}$; the algorithm *approximates the function Q locally around $\theta^{(i)}$ with a quadratic form*;
- it sets $\theta^{(i+1)}$ = the minimum point of the quadratic form;
- the iteration is stopped when $|\nabla Q\left(\theta^{(i)}\right)|$ is sufficiently small.

To approximate the function Q around a certain $\theta = \theta^{(i)}$ with a quadratic form means to expand it in a Taylor polynomial of order 2 around θ:

$$Q(\theta + v) \simeq q_\theta(v) := Q(\theta) + \nabla Q(\theta)^\top v + \frac{1}{2} v^\top H(\theta) v. \tag{32}$$

This is a quadratic form in the variable v. Here, $\nabla Q(\theta) \in \mathbb{R}^p$ is the gradient of Q at θ, and $H(\theta) \in \mathbb{R}^{p \times p}$ is the *Hessian matrix*:

$$H(\theta) = \begin{bmatrix} \frac{\partial^2 Q(\theta)}{\partial \theta_1^2} & \frac{\partial^2 Q(\theta)}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 Q(\theta)}{\partial \theta_1 \partial \theta_p} \\ \frac{\partial^2 Q(\theta)}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 Q(\theta)}{\partial \theta_2^2} & \cdots & \frac{\partial^2 Q(\theta)}{\partial \theta_2 \partial \theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 Q(\theta)}{\partial \theta_p \partial \theta_1} & \frac{\partial^2 Q(\theta)}{\partial \theta_p \partial \theta_2} & \cdots & \frac{\partial^2 Q(\theta)}{\partial \theta_p^2} \end{bmatrix}$$

Since Q is supposed to be *twice* continuously differentiable, by Schwarz's theorem it is a symmetric matrix (the order of differentiation does not matter); supposing, further, that Q is at least locally convex, $H(\theta)$ is at least positive semi-definite ($H(\theta) \geq 0$). If it happens that $H(\theta) > 0$, then the minimum point of (32) is

$$v = -H(\theta)^{-1} \nabla Q(\theta). \tag{33}$$

The v in (33) is the typical displacement of Newton's algorithm. It is immediate to check that it is a descent direction; indeed if $H(\theta) > 0$, then $H(\theta)^{-1} > 0$ as well, and

$$\frac{\delta Q(\theta)}{\delta v} = \nabla Q(\theta)^\top v = -\nabla Q(\theta)^\top H(\theta)^{-1} \nabla Q(\theta) < 0.$$

Thus, the update step in Newton's algorithm is as follows,

- set $\theta^{(i+1)} = \theta^{(i)} - \alpha H(\theta)^{-1} \nabla Q(\theta)$, where α is a suitable backtracking constant chosen in $(0, 1]$.

In most cases, especially close to the solution $\bar{\theta}$, $\alpha = 1$ is just fine.

Newton's algorithm is the workhorse of convex optimization, and a benchmark for every other unconstrained optimization algorithm. Its convergence rate is *very* high compared to other methods (meaning that the number of iterations needed to converge to an acceptable solution is much smaller in comparison).

4.3 Application to nonlinear least squares

4.3.1 The steepest descent algorithm

Now we will adapt the two iterative algorithms seen in Section 4.2 to the sum of squares of the main problem:

$$Q(\theta) = \sum_{i=1}^{N} (y_i - f_i(\theta))^2 .$$

Let us define $Y \in \mathbb{R}^N$ and $F : \mathbb{R}^p \to \mathbb{R}^N$ as follows:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \qquad F(\theta) = \begin{bmatrix} f_1(\theta) \\ f_2(\theta) \\ \vdots \\ f_N(\theta) \end{bmatrix}$$

(then $Q(\theta) = \|Y - F(\theta)\|^2$). Let $J(\theta) \in \mathbb{R}^{N \times p}$ be the Jacobian matrix of F computed at θ:

$$J(\theta) = \begin{bmatrix} \frac{\partial f_1(\theta)}{\partial \theta_1} & \frac{\partial f_1(\theta)}{\partial \theta_2} & \cdots & \frac{\partial f_1(\theta)}{\partial \theta_p} \\ \frac{\partial f_2(\theta)}{\partial \theta_1} & \frac{\partial f_2(\theta)}{\partial \theta_2} & \cdots & \frac{\partial f_2(\theta)}{\partial \theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_N(\theta)}{\partial \theta_1} & \frac{\partial f_N(\theta)}{\partial \theta_2} & \cdots & \frac{\partial f_N(\theta)}{\partial \theta_p} \end{bmatrix}$$

The goal function $Q(\theta)$ has a very rich structure; in particular,

$$\frac{\partial Q(\theta)}{\partial \theta_k} = \sum_{i=1}^{N} 2\left(y_i - f_i(\theta)\right)\left(-\frac{\partial f_i(\theta)}{\partial \theta_k}\right)$$

Then the gradient of Q at θ is

$$\begin{aligned}\nabla Q(\theta) &= \left[\begin{array}{cccc} \frac{\partial Q(\theta)}{\partial \theta_1} & \frac{\partial Q(\theta)}{\partial \theta_2} & \cdots & \frac{\partial Q(\theta)}{\partial \theta_p} \end{array}\right]^\top \\ &= -2J(\theta)^\top (Y - F(\theta))\end{aligned}$$

Thus, the descent direction in the steepest descent algorithm is $v = -\nabla Q(\theta) = 2J(\theta)^\top (Y - F(\theta))$, and its update step is as follows:

- set $\theta^{(i+1)} = \theta^{(i)} + \alpha 2J(\theta)^\top (Y - F(\theta))$, where α is a suitable backtracking constant chosen in $(0, 1]$.

4.3.2 The Newton algorithm

The Hessian of $Q(\theta)$ is the matrix $H(\theta) \in \mathbb{R}^{p\times p}$ having as its k, l-th component

$$[H(\theta)]_{kl} = \frac{\partial^2 Q(\theta)}{\partial \theta_k \partial \theta_l} = \sum_{i=1}^{N} 2\left(\frac{\partial f_i(\theta)}{\partial \theta_k}\frac{\partial f_i(\theta)}{\partial \theta_l} - (y_i - f_i(\theta))\frac{\partial^2 f_i(\theta)}{\partial \theta_k \partial \theta_l}\right).$$

Due to the burden of computing the second derivatives $\frac{\partial^2 f_i(\theta)}{\partial \theta_k \partial \theta_l}$, the Newton iteration can be computationally rather demanding. Note that, since convexity is not guaranteed anymore, it is not necessarily true that $H(\theta)$ is non-singular, nor that it is even positive semi-definite. It can be actually any symmetric matrix; hence the corresponding quadratic approximation at θ,

$$\begin{aligned}q_\theta(v) &= Q(\theta) + \nabla Q(\theta)^\top v + \frac{1}{2} v^\top H(\theta) v \\ &= Q(\theta) - 2(Y - F(\theta))^\top J(\theta) v + \frac{1}{2} v^\top H(\theta) v,\end{aligned}$$

is not necessarily positive semi-definite. Computing the Newton step would in principle require to solve for v the equation

$$H(\theta)v = -\nabla Q(\theta);$$

this may be impossible in general, or, since convexity fails to hold, it could yield a wrong direction, i.e. not a descent direction.

Note, however, that *if it so happens that θ is sufficiently close to the solution $\bar{\theta}$*, then the terms $(y_i - f_i(\theta))$ are relatively small, the terms in the summation containing second derivatives become hopefully negligible with respect to the other ones, and it may become convenient to approximate $H(\theta)$ with a matrix $\tilde{H}(\theta) \in \mathbb{R}^{p\times p}$ having as its k,l-th component

$$\left[\tilde{H}(\theta)\right]_{kl} := 2\sum_{i=1}^{N} \frac{\partial f_i(\theta)}{\partial \theta_k}\frac{\partial f_i(\theta)}{\partial \theta_l};$$

this leads to the approximate Hessian matrix

$$\tilde{H}(\theta) = 2J(\theta)^\top J(\theta).$$

The corresponding quadratic approximation at θ,

$$\tilde{q}_\theta(v) = Q(\theta) - 2(Y - F(\theta))^\top J(\theta)v + v^\top J(\theta)^\top J(\theta)v,$$

is *always* at least positive semi-definite. Correspondingly, we have an *approximate Newton step*, which is found solving for v the equation $\tilde{H}(\theta)v = -\nabla Q(\theta)$, that is:

$$J(\theta)^\top J(\theta)\ v = J(\theta)^\top (Y - F(\theta)) \tag{34}$$

Besides seeming a convenient trick to simplify formulas, due to an apparently arbitrary approximation of the Hessian, the above step has manifold advantages:

1. equation (34) requires only that $f(x,\theta)$ is *once* continuously differentiable with respect to θ, not *twice* as before;

2. equation (34) demands less computational burden; but above all

3. *equation* (34) *is an instance of the normal equations!* (Indeed, recall that its unknown is v.) Since we know how to solve normal equations, the approximate Newton descent direction is trivial to compute, once $J(\theta)$ is known.

The third point demands further inquiry. In practice, dropping the terms containing second derivatives from the Hessian, we have pretended that they are negligible; the "official" justification for this is the assumption that we are close to the optimal solution, so that their multipliers $(y_i - f_i(\theta))$ are small. But another reason for which we could want to discard those terms is that *we choose to neglect the second derivatives* $\frac{\partial^2 f_i(\theta)}{\partial\theta_k\partial\theta_l}$ *altogether*, in other words to employ a *linear* local approximation of each $f_i(\theta)$ around θ instead of a quadratic one.

There is, in particular, a case in which the linear approximation does not lose any information at all, and of course this is when the $f_i(\theta)$ themselves are already linear. In this case, all the second derivatives are zero, and the approximate Hessian of $Q(\theta)$ equals the true one. And now an obvious question arises:

what happens if we apply a Newton step when the $f_i(\theta)$ are linear?

You already know the answer in your heart. If $f_i(\theta)$ is linear in θ, it means that we are in the old case $f_i(\theta) = \varphi_i^\top \theta$; but then

$$
\begin{aligned}
F(\theta) &= \begin{bmatrix} \varphi_1^\top \theta \\ \varphi_2^\top \theta \\ \vdots \\ \varphi_N^\top \theta \end{bmatrix} = \begin{bmatrix} \varphi_1^\top \\ \varphi_2^\top \\ \vdots \\ \varphi_N^\top \end{bmatrix} \theta = \Phi\theta; \\
J(\theta) &= \frac{\partial F(\theta)}{\partial \theta} = \Phi, \quad \text{irrespective of } \theta.
\end{aligned}
$$

We are supposed to run the Newton algorithm, hence we need to provide an initial guess $\theta^{(0)}$; in this case, though, we know that Q is *convex*, hence the initial guess does not matter very much. We choose $\theta^{(0)} = 0$, we obtain $F\left(\theta^{(0)}\right) = 0$, and equation (34) reads

$$\Phi^\top \Phi \; v = \Phi^\top Y.$$

Except for the change of a symbol from θ to v, these are the normal equations of ordinary least squares. Thus,

when applied to an ordinary least squares problem, the Newton algorithm solves it in just one step by solving the standard normal equations.

Coming back to nonlinear least squares and resuming, the approximate Newton descent direction is, assuming that $J(\theta)^\top J(\theta)$ is invertible, $v = \left(J(\theta)^\top J(\theta)\right)^{-1} J(\theta)^\top (Y - F(\theta))$; the corresponding update step is as follows:

- set $\theta^{(i+1)} = \theta^{(i)} + \alpha \left(J(\theta)^\top J(\theta)\right)^{-1} J(\theta)^\top (Y - F(\theta))$, where α is a suitable backtracking constant chosen in $(0, 1]$.

So far we have introduced two iterative algorithms, the steepest descent method and the Newton method; the second of them can be employed in a simplified version when the current $\theta^{(i)}$ is close to the optimal solution. The advantages of both the algorithm are exploited in the next (and last) algorithm that we will cover, which is more or less a standard for the solution of nonlinear least squares problems.

4.4 The Levenberg-Marquardt algorithm

4.4.1 A compromise between steepest descent and Newton's algorithm

Compare the equations for the displacement, at the step (i), of the steepest descent algorithm,

$$\begin{aligned}\Delta\theta &= -\alpha\nabla Q(\theta)\\ &= 2\alpha J(\theta)^\top(Y - F(\theta))\\ &\text{that is}\\ \lambda\Delta\theta &= J(\theta)^\top(Y - F(\theta)),\end{aligned}$$

where $\lambda := \frac{1}{2\alpha}$, and that of the approximate Newton algorithm:

$$\begin{gathered}\tilde{H}(\theta)\ \Delta\theta = -\nabla Q(\theta),\\ \text{that is}\\ J(\theta)^\top J(\theta)\ \Delta\theta = J(\theta)^\top(Y - F(\theta)).\end{gathered} \tag{35}$$

Recall that $\alpha \in (0, 1]$ is the so-called backtracking constant, that reduces the full step if it does not lead to a decrease $Q(\theta + \Delta\theta) < Q(\theta)$ in the goal function. The approximate Newton displacement is valid only in the proximity of a minimum: in this case, the full displacement is often just good.

The iterative method adopting the displacement $\Delta\theta$ that solves the following equation,

$$\begin{gathered}\left(\tilde{H}(\theta) + 2\lambda I\right)\ \Delta\theta = -\nabla Q(\theta),\\ \text{that is}\\ \left(J(\theta)^\top J(\theta) + \lambda I\right)\ \Delta\theta = J(\theta)^\top(Y - F(\theta)),\end{gathered} \tag{36}$$

where λ is a positive constant, is called the *Levenberg-Marquardt* algorithm. Note that (36) does *always* admit a *unique* solution because, since $\lambda > 0$, the matrix $\tilde{H}(\theta) + 2\lambda I = J(\theta)^\top J(\theta) + \lambda I$ is positive definite. Furthermore, $\Delta\theta$ is always a descent direction; indeed since $\tilde{H}(\theta) + 2\lambda I > 0$, then $(\tilde{H}(\theta) + 2\lambda I)^{-1} > 0$ as well, hence

$$\frac{\delta Q(\theta)}{\delta(\Delta\theta)} = \nabla Q(\theta)^\top\Delta\theta = -\nabla Q(\theta)^\top\left(\tilde{H}(\theta) + 2\lambda I\right)^{-1}\nabla Q(\theta) < 0$$

unless $\nabla Q(\theta) = 0$ (optimality condition).

Now, if the constant λ is very small, then $\left(J(\theta)^\top J(\theta) + \lambda I\right) \simeq J(\theta)^\top J(\theta)$, so that

$$J(\theta)^\top J(\theta)\ \Delta\theta \simeq J(\theta)^\top(Y - F(\theta)),$$

which resembles (35) (in the limit case $\lambda = 0$, equation (36) would *coincide* with (35)). On the other hand, if λ is large ($\lambda \gg 1$), then $\left(J(\theta)^\top J(\theta) + \lambda I\right) \simeq \lambda I$, so that

$$\Delta\theta \simeq \frac{1}{\lambda} J(\theta)^\top (Y - F(\theta)),$$

which is a steepest descent displacement, with a “kind-of-backtracking” constant already in place.

Therefore, applying the displacement $\Delta\theta$ provided by (36) is “almost” a Newton step when λ is small, and “almost” a rescaled steepest descent step when λ is large.

4.4.2 The Levenberg-Marquardt algorithm

The Levenberg-Marquardt algorithm prescribes that *the scaling factor* λ *is changed adaptively at each step* (i), being in fact a $\lambda^{(i)}$, starting from a steepest-descent-like behavior when $\theta^{(i)}$ is distant from the minimum, and decreasing it in order to attain a Newton-like behavior in its proximity (precisely when the *approximate* version of the Newton equation, which is (35), works fine).
The choice of a suitable λ for the next step depends on the so-called *gain ratio*

$$\rho := \frac{Q(\theta) - Q(\theta + \Delta\theta)}{\tilde{q}_\theta(0) - \tilde{q}_\theta(\Delta\theta)}, \tag{37}$$

which measures the ratio between the actual decrease in the goal function Q and the decrease predicted by its Newton approximation $\tilde{q}_\theta$ computed at θ, when the displacement $\Delta\theta$ is applied. Recall that

$$\tilde{q}_\theta(\Delta\theta) = Q(\theta) - 2(Y - F(\theta))^\top J(\theta)\Delta\theta + \Delta\theta^\top J(\theta)^\top J(\theta)\Delta\theta,$$

hence $\tilde{q}_\theta(0) = Q(\theta)$, and since the Levenberg-Marquardt step (36) yields $J(\theta)^\top J(\theta)\ \Delta\theta = J(\theta)^\top (Y - F(\theta)) - \lambda\Delta\theta$, we have

$$\begin{aligned}
\tilde{q}_\theta(0) - \tilde{q}_\theta(\Delta\theta) &= 2(Y - F(\theta))^\top J(\theta)\Delta\theta - \Delta\theta^\top J(\theta)^\top J(\theta)\Delta\theta \\
&= 2\Delta\theta^\top J(\theta)^\top (Y - F(\theta)) - \Delta\theta^\top J(\theta)^\top J(\theta)\Delta\theta \\
&= \Delta\theta^\top \left(2J(\theta)^\top (Y - F(\theta)) - J(\theta)^\top (Y - F(\theta)) + \lambda\Delta\theta\right) \\
&= \Delta\theta^\top \left(-\frac{1}{2}\nabla Q(\theta) + \lambda\Delta\theta\right) \\
&= -\frac{1}{2}\nabla Q(\theta)^\top \Delta\theta + \lambda \left\|\Delta\theta\right\|^2 .
\end{aligned}$$

Recalling that the Levenberg-Marquardt displacement $\Delta\theta$ is a descent direction, we have that the denominator in (37) is always a positive quantity (unless $\Delta\theta = 0$, which denotes optimality).

Hence, $\rho > 0$ if and only if the displacement $\Delta\theta$ actually attains a decrease in the goal function. At step (i), the acceptance of the displacement and the choice of the next λ should be adapted consequently:

- if $\rho < 0$, reject the displacement $\Delta\theta$ and increase λ;
- if ρ is positive but small, then the displacement $\Delta\theta$ can be accepted, but the quadratic approximation is not very good, hence λ should be increased, aiming at a steepest-descent-like behavior;
- if ρ is high, the displacement $\Delta\theta$ should be accepted, the quadratic approximation is good, and λ should be decreased, aiming at a Newton-like behavior.

In [19], the following strategy is proposed:

- at the beginning, set $\nu \leftarrow 2$;
- at each step (i),
  - if $\rho > 0$, set $\theta \leftarrow \theta + \Delta\theta$, $\lambda \leftarrow \lambda \max\{\frac{1}{3}, 1 - (2\rho - 1)^3\}$, and $\nu \leftarrow 2$;
  - otherwise, set $\lambda \leftarrow \nu\lambda$ and $\nu \leftarrow 2\nu$.

4.5 Example: localization of a transmitter

5 Machine learning

5.1 Introduction and motivation

In the theory of least squares we have supposed that the data were generated by a model $y_i = \varphi^\top(x_i)\theta^o + \varepsilon_i$ linear in the "true" parameter θ^o, and the underlying objective was to estimate or approximate θ^o, a quantity intrinsic in the model. Recall that such estimate could be useful for three objectives:

- investigate internal properties of the mechanism that generates the data;
- measure indirectly some physical quantity of interest;
- predict a future value of the output, given a future value of the input.

In many applications, though, either there is no parameter θ^o, or such parameter is so complex that its estimation is hopeless and meaningless.

Example. Think at the estimation of the pitch rate (output) of a fixed-wing airplane, given the inputs to its ailerons, flaps, and canards. In fact, the estimation of the pitch rate requires a model of the entire airplane and of the surrounding environment. Models of this kind exist and are used in the design stage for the structural analysis of aircraft, but are far too complex for the online computations carried by the on-board computer to help the pilot.

Example. Think at weather forecasts. Given the current state of a sector of the atmosphere (pressure, temperature, wind speed etc., i.e. the input), the problem is that of estimating, say, tomorrow's minimal or average temperature at a certain location (the output). What is θ^o here?

In many cases of interest, the "output variable" belongs to a finite set; in these cases, the problem of "predicting" the output is called *classification.*

Example. Think at the recognition of a hand-written character (and maybe have another look at Section 2.2). Here, the input is a bitmap containing a scan of the character, and the output is an element of the set $\{A, B, C, \cdots, Z\}$.

Example. Consider the diagnosis of a disease; say, the presence of a melanoma on the skin of a patient. The definitive diagnosis is typically based on a biopsy; however, such procedure is expensive and invasive, and for a first assessment (subject to a certain probability of error), less invasive techniques are preferred. For example, a technique for automatic recognition of a melanoma relies on the analysis of a photograph; the input, here, is a bitmap, and the output is an answer in $\{\text{Yes}, \text{No}\}$. [24] A conceptually similar problem is the diagnosis of a heart disease given the observation of an ECG over a time interval $[0, T]$ (thus the input is, ideally, a function $f : [0, T] \to \mathbb{R}^{12}$, or if you wish twelve functions, one per each "lead"; the output is again $\{\text{Yes}, \text{No}\}$). In these examples, can you imagine any θ^o linking the input to the output?

What matters, in these example, is not really any "true" parameter θ^o, belonging to a finite-dimensional vector space, but the relation between the input and the output, considered as a mapping. In this chapter we drop the assumption that we know anything about the particular mechanism that generates the data (in particular about linearity), and we study the problem of selecting an entire *function* to match how nature generates data, with the purpose of establishing a bound on the probability of producing a wrong prediction, which is intrinsic to the problem and cannot be avoided. As you can guess, estimating a function is in general way more complex than estimating a vector; hence we shall introduce other kinds of constraints on the problems that we are going to deal with, and we will reduce the complexity of the problem to the minimum possible; still, you will see that establishing such bound is difficult enough.

We assume that there is a certain random quantity U, called the *input*, which we can measure; that the input is fed through a "black box", managed by "nature", whose behavior we don't know; and that at the other side of the black box we can measure a second random quantity $Y \in \{0, 1, \cdots, M\}$, called the *output*. Finally, we assume that we have measured finitely many realizations $(U_1, Y_1), \cdots, (U_N, Y_N)$ (the "data") of the generation process. The goal is to find a function $\hat{Y} = \hat{f}(U)$ that approximates the behavior of the black box in the best possible way. In view of the hypothesis that Y can take only finitely many values, this is called a *classification* problem (we "classify" U sticking to it a label $\hat{Y}$ chosen from the set $\{0, 1, \cdots, M\}$).
Unlike what we did with respect to least squares, the main goal here is not to provide robust and reliable algorithms, but to understand what is going on, and to investigate the limits inherent in the process of matching our function to nature's data generation mechanism, a process that we call *machine learning*. Even with the hypothesis that Y can take finitely many values, the general theory of machine learning is beyond the scope of these notes; we restrict our investigation to the following particular case:

- the input U is scalar (that is, a real random variable);

[24] We stipulate to call an automatic diagnosis, or a similar problem, a "prediction" even if it really is not; indeed, there is no time variable in this example, and a $\{\text{Yes}, \text{No}\}$ answer tells something about *now* with a certain probability, not anything about the future.

- the output Y is another random variable (hopefully, highly correlated with U) that can only take the values 0 and 1;
- the goal is, given the data $(U_1, Y_1), \cdots, (U_N, Y_N)$, to find a function $\hat{f} : \mathbb{R} \to \{0, 1\}$ that will make, on average, the least possible error in guessing the future data Y_i when U_i will be given.

A problem with these assumptions is called a *binary classification* problem. Loosely speaking, $\hat{f}$ now takes the place of what before was θ^o; but note that while θ^o was supposed to be *intrinsic* in the data generation mechanism, now $\hat{f}$ has nothing to do, in general, with the internals of the "black box", and instead pertains to *our* point of view on how the behavior of the black box should be approximated. This, as we will see, will have to do above all with the class of functions in which we search a solution. In particular, the objectives of investigating internal properties of the data generation rule and measuring physical quantities (which we stated among the goals of the least squares method) are gone; the only one that remains is *prediction*.

5.2 Binary classification and the Bayesian classifier

Let us make the problem a bit more precise. We assume the following:

- a sequence of random pairs $(U_1, Y_1), \cdots, (U_N, Y_N)$ is observed, where $U_i \in \mathbb{R}$ and $Y_i \in \{0, 1\}$. *We assume that the pairs are independent and identically distributed.* We hope that in some sense Y_i is "predictable" from U_i, but we do not actually know either the distribution of U_i or how Y_i was generated from U_i.
- We wish to study how a function $\hat{f} : \mathbb{R} \to \{0, 1\}$ behaves in providing an estimate $\hat{Y} = \hat{f}(U)$, and we wish to compare such functions. We will call the functions $\hat{f} : \mathbb{R} \to \{0, 1\}$ *binary classifiers*; note that a binary classifier is the indicator function of a subset of $\mathbb{R}$. As we shall soon see with some examples, it is not necessarily the case that the data $(U_1, Y_1), \cdots, (U_N, Y_N)$ are actually generated by means of a function $Y_i = f(U_i)$, hence we will match with a function a rule which is not necessarily a function; still, this is the best that we can do.
- "All the indicator functions" is a set far too complex to choose from, and would immediately make the problem intractable. Therefore we restrict our attention to *families* of functions $\mathcal{F} = \{\hat{f}_c\}_{c \in C}$ parameterized by some set C of indices. We will see that even families that look relatively simple may be still too complex and inhibit the correct choice of a classifier.
- The objective is to find the index c that parameterizes the "best" classifier $\hat{f}_c \in \mathcal{F}$. *What does "best" mean?* Similarly to what we did

with the theory of least squares, we define as the "best" $\hat{f}_c$ the one that minimizes the expected square error:

$$\bar{J} = \mathsf{E}\left[(Y - \hat{Y})^2\right] = \mathsf{E}\left[(Y - \hat{f}(U))^2\right]$$

We shall now investigate how $\bar{J}$ behaves. Keep in mind that this is an abstract definition concerning some random variables U and Y, but what we will have available at the end of the day will always be finitely many pairs of data $(U_1, Y_1), \cdots, (U_N, Y_N)$. The point is the following: today, while we are sitting in class and studying the problem from an abstract point of view, $(U_1, Y_1), \cdots, (U_N, Y_N)$ are independent random pairs in their own respect, and we may study them as such; but tomorrow, when someone will come with "the data" and ask us to run a software to actually compute the best classifier, they will be a particular realization of the random variables, that is a bunch of *numbers*. What we investigate today is, indeed, the average behavior of the best solution that the software will be able to give tomorrow, on the sole basis of those numbers.

Example. Consider the following data generation rules:

1. U is Gaussian $N(0, 1)$, and $Y = f(U)$, where

   $$f(u) = \mathbb{1}_{[-1,0]}(u) + \mathbb{1}_{[\frac{1}{2},1]}(u).$$

   In this case, the generation is actually done by a function.

2. U is Gaussian $N(0, 1)$; if it happens that $U < 0$, then

   $$Y = \begin{cases} 1 & \text{with probability } 0.8; \\ 0 & \text{with probability } 0.2. \end{cases}$$

   If instead $U \geq 0$, then

   $$Y = \begin{cases} 1 & \text{with probability } 0.1; \\ 0 & \text{with probability } 0.9. \end{cases}$$

   In this case the mechanism is intrinsically probabilistic, i.e. it is *not* produced by a function.

3. U is Gaussian $N(0, 1)$; if it happens that $U < 0$, then $Y = 1$; if instead $U \geq 0$, then

   $$Y = \begin{cases} 1 & \text{with probability } 0.5; \\ 0 & \text{with probability } 0.5. \end{cases}$$

   In this case the mechanism is also probabilistic.

In all the three cases, there is indeed a binary classifier $\hat{f}$ that minimizes the expected error $\bar{J} = \mathsf{E}\left[(Y - \hat{f}(U))^2\right]$. Note that, since the expression $(Y - \hat{f}(U))^2$ is 0 when $Y = \hat{f}(U)$ and 1 when $Y \neq \hat{f}(U)$, we can write

$$\bar{J} = \mathsf{E}\left[\mathbb{1}_{Y \neq \hat{f}(U)}\right] = \mathsf{P}\left[Y \neq \hat{f}(U)\right]$$

Hence, in order to find the best classifier $\hat{f}$ it is sufficient to minimize such probability, point-wise with respect to u.

1. In the first example the optimal classifier is $\hat{f} = f$, the same function that generated the data, which attains $\bar{J} = 0$.
2. In the second example, it is not difficult to realize that the classifier that minimizes the probability of error at each u is

   $$\hat{f}(u) = \begin{cases} 1 & \text{if } u < 0; \\ 0 & \text{if } u \geq 0. \end{cases}$$

   It attains

   $$\bar{J} = 0.2 \cdot \mathsf{P}\left[U < 0\right] + 0.1 \cdot \mathsf{P}\left[u \geq 0\right] = 0.15.$$
3. The third example is similar, but exhibits an issue: if $u < 0$ it is of course correct to assign $\hat{f}(u) = 1$, but if $u \geq 0$, it makes no difference whether $\hat{f}(u) = 0$ or 1: the error probability will be 0.5 in any case. The choice is somewhat arbitrary; therefore we choose the constant function $\hat{f}(u) = 1$, and attain $\bar{J} = 0.25$.

□

The best function $\hat{f}$ that we have found in the three examples is called the *Bayesian classifier*. Computing it does not seem a big deal, after all. Is this the solution to the classification problem?

No.

We have neglected at least three of the assumptions made in the statement of the problem. First, to compute $\hat{f}$ we have exploited a lot of knowledge on the data generation mechanism (i.e. functions, probabilities), while we had assumed that of such mechanism we should have no knowledge at all. Second, we have disregarded the only information that we assumed to have, that is the data $(U_1, Y_1), \cdots, (U_N, Y_N)$. Third, we have ignored that we were supposed to search the classifier in a class of functions $\mathcal{F} = \{\hat{f}_c\}_{c \in C}$ and the goal was to find an optimal $c \in C$, implicitly assuming that the Bayesian classifier belongs to $\mathcal{F}$. Moreover, the Bayesian classifier could be *very* hard to compute, although this is not apparent from the above simple examples. Summing up: yes, there exists a best classifier, called the Bayesian one, which could be found if everything were known; but actually very little is known, and in general we have no hope to compute it exactly.

5.3 Minimization of the error

To sort things out, the first step to do is to re-define properly the "cost" that we use to evaluate different classifiers in our family $\mathcal{F}$. To start with, it must depend on the choice of c, hence it must be a function on the set C:

Definition 5.3.1 *The* error function *is the function $\bar{J} : C \to [0,1]$, defined as follows:*

$$\bar{J}(c) := \mathsf{E}\left[(Y - \hat{f}_c(U))^2\right] = \mathsf{E}\left[\mathbb{1}_{Y \neq \hat{f}_c(U)}\right] = \mathsf{P}\left[Y \neq \hat{f}_c(U)\right]$$

The "true" error $\bar{J}(c)$ will be key to our comprehension of what can be accomplished with machine learning. The ideal goal of our theory would be to find its minimum:

$$\bar{c} = \arg\min_{c \in C} \bar{J}(c)$$

But now recall that the only information available comes from the data $(U_1, Y_1), \cdots, (U_N, Y_N)$, hence we have no hope to compute either $\bar{J}$ or its minimum. The best that we can do is to *approximate* $\bar{J}$, and dedicate our future efforts to show that the approximation is good.

Definition 5.3.2 *The* empirical error function*, based on the data*

$$(U_1, Y_1), \cdots, (U_N, Y_N),$$

is the function $\hat{J}_N : C \to [0,1]$, defined as follows:

$$\hat{J}_N(c) := \frac{1}{N} \sum_{i=1}^{N} (Y_i - \hat{f}_c(U_i))^2 = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{Y_i \neq \hat{f}_c(U_i)}$$

The empirical error $\hat{J}_N(c)$ is our approximation of the true error $\bar{J}(c)$. In fact, you can think of $\hat{J}_N(c)$ as a replica of $\bar{J}(c)$ where the expectation has been replaced by a sample average over the available data:

$$\mathsf{E}\left[\cdot\right] \quad \rightsquigarrow \quad \mathsf{M}\left[\cdot\right] = \frac{1}{N} \sum_{i=1}^{N} (\cdot)$$

More informally, think that if (U_i, Y_i) had a joint density $p(u, y)$ (here it is not the case), the true error would be

$$\bar{J}(c) = \mathsf{E}\left[(Y - \hat{f}_c(U))^2\right] = \int_{\mathcal{U} \times \mathcal{Y}} (y - \hat{f}_c(u))^2 \; p(u, y) \; du \; dy$$

In the same spirit, the empirical error would be

$$\hat{J}_N(c) = \frac{1}{N} \sum_{i=1}^{N} (Y_i - \hat{f}_c(U_i))^2 = \int_{\mathcal{U} \times \mathcal{Y}} (y - \hat{f}_c(u))^2 \; p_N(u, y) \; du \; dy$$

with the true density replaced by an "empirical density", which would now be a weighted sum of "deltas", one "delta" per each pair of data:

$$p_N(u,y) = \frac{1}{N}\sum_{i=1}^{N}\delta_{U_i,Y_i}(u,y)$$

(There is nothing rigorous in this reasoning. It is just to support intuition.)

Whatever the interpretation of the empirical error, with the sole knowledge of the data the best that we can do is to pursue its minimization:

$$\hat{c}_N = \arg\min_{c\in C} \hat{J}_N(c)$$

The main point of this chapter is the following:

- *we expect that, for big N, $\hat{J}_N(c)$ is close to $\bar{J}(c)$ with high probability;*
- *we expect that, as $N \to \infty$, $\hat{J}_N$ tends to $\bar{J}$ (in which sense?);*
- *consequently we expect that for big N the minimum of $\hat{J}_N$ is close to the minimum of $\bar{J}$ with high probability, and that as $N \to \infty$ the former converges to the latter.*

Whether these facts hold or not is a deep subject regarding the limits of learning; in essence, it depends on both the data-generation mechanism and the family of classifier functions in which we pursue the empirical minimization. In general, our expectations are going to be frustrated, as we shall immediately see with a pathological example.

Example. Consider the following data generation rule: U_i is Gaussian $N(0,1)$ and $Y_i = f(U_i)$, where

$$f(u) = \mathbb{1}_{[-2,0]}(u) + \mathbb{1}_{[\frac{1}{3},2]}(u).$$

First, imagine that we search the best possible classifier among the indicator functions of closed intervals on the real line. That is to say, the family of classifiers is

$$\mathcal{F}_1 = \{\mathbb{1}_{[a,b]}(\cdot)\}.$$

parameterized by the index $c = (a,b)$, where $C = \{(a,b) \mid a \leq b\}$. A little reasoning reveals that the optimal classifier in $\mathcal{F}_1$ is attained for $\bar{c} = (-2,2)$, and $\hat{f}_{\bar{c}}(u) = \mathbb{1}_{[-2,2]}(u)$. The corresponding true error is

$$\bar{J}(\bar{c}) = \frac{1}{\sqrt{2\pi}}\int_0^{1/3} e^{-t^2/2}\,dt \simeq 0.13$$

Having a lot of data available, say $N = 10000$, with very high probability we will find a lot of ones in the interval $[-2, 0]$, a few zeros in the interval $[0, 1/3]$, again a lot of ones in $[1/3, 2]$, and some other zeros elsewhere. Following common sense, or heuristics, or actually minimizing numerically the empirical cost, we will choose to forget about the zeros in $(0, 1/3)$ and select an interval resembling $[-2, 2]$, say $\hat{c}_N = (-2 - \varepsilon_1, 2 + \varepsilon_2)$, yielding $\hat{f}_{\hat{c}_N}(u) = \mathbb{1}_{[-2-\varepsilon_1, 2+\varepsilon_2]}(u)$. The corresponding empirical error will be comparable to the optimal one, say

$$\hat{J}_N(\hat{c}_N) = 0.125.$$

The real cost will be something more than the optimal one, say:

$$\bar{J}(\hat{c}_N) = \frac{1}{\sqrt{2\pi}} \left(\int_{-2-\varepsilon_1}^{-2} e^{-t^2/2}\, dt + \int_0^{1/3} e^{-t^2/2}\, dt + \int_2^{2+\varepsilon_2} e^{-t^2/2}\, dt \right) = 0.14$$

As you can see, this situation is fairly good.
Consider now the same data generation rule, but this time suppose that our family of classifiers is

$$\mathcal{F}_2 = \mathcal{F}_1 \cup \{\mathbb{1}_{(\text{finite set})}(\cdot)\}.$$

This new family comprises the interval classifiers of before, plus all the functions that take the value 1 on finitely many points, and 0 elsewhere ("needle functions"). (C is a rather messy set which has no real importance here.) Given the data $(U_1, Y_1), \cdots, (U_N, Y_N)$, the choice of a classifier in $\mathcal{F}_2$ minimizing the empirical error is obvious: indeed choosing $\hat{c}_N =$ the index of the needle function

$$\hat{f}_{\hat{c}_N} = \begin{cases} 1 & \text{at all those } U_i \text{ for which the corresponding } Y_i = 1, \\ 0 & \text{elsewhere,} \end{cases}$$

the corresponding empirical cost is

$$\hat{J}_N(\hat{c}_N) = 0;$$

nevertheless, for the same choice it holds

$$\bar{J}(\hat{c}_N) = \frac{1}{\sqrt{2\pi}} \left(\int_{-2}^{0} e^{-t^2/2}\, dt + \int_{1/3}^{2} e^{-t^2/2}\, dt \right) \simeq 0.82.$$

Actually, the choice of *any* needle function attains the same true error 0.82, which is rather bad; therefore the choice of a needle function, obtained by minimizing the empirical error, is completely wrong, and the correct choice would have been the interval classifier of before, which still belongs to $\mathcal{F}_2$. □

Something got wrong. What? The general, common sense answer is

The family of classifiers $\mathcal{F}_2$ is too complex, and
in general it is advisable to stay away from complexity.

But let us examine more closely what is going on in the second part of the example. We have seen three quantities of interest:

- $\bar{J}(\bar{c})$, the minimum true error achievable by choosing among all classifiers in the family;

- $\hat{J}_N(\hat{c}_N)$, the minimum empirical error; and

- $\bar{J}(\hat{c}_N)$, the true error that corresponds to the choice that is optimal for the empirical error.

The issue is of course that the situation $\bar{J}(\hat{c}_N) \gg \bar{J}(\bar{c})$ is unavoidable: no matter how many data are available, at least one needle function will always attain empirical error 0 and true error 0.82. More in detail, it so happens that at all c that select interval classifiers, the empirical error is always more or less *close* to the true one, and this is the reason why in the first part of the example there was no problem; but among the indices c corresponding to needle functions there always exist one that attains a difference of 0.82. As N increases, we expect that $\hat{J}_N$ gets closer and closer to $\bar{J}$ in the subset of C containing interval classifiers; not so for the subset of needle classifiers. This is precisely the situation that we wish to avoid; only if $\hat{J}_N(c)$ and $\bar{J}(c)$ are everywhere close, in general, we can hope that their respective minima will be close, and maybe equal in the limit as $N \to \infty$. Guided by these few considerations, here we formulate the principle, to establish which the rest of this chapter is dedicated:

as $N \to \infty$, the function $\hat{J}_N$ should tend to $\bar{J}$ uniformly with respect to c.

This will indeed be the case in some settings of interest; among them, the choice among interval classifiers. This is the program that we will follow:

1. we will show that, if $\hat{J}_N \to \bar{J}$ uniformly, the minimum of $\hat{J}_N$ indeed tends to the minimum of $\bar{J}$, therefore the minimization of the empirical error is "good";

2. we will study in detail the behavior of the so-called empirical distribution of a real random variable, and establish probabilistic guarantees of "uniform closeness" between empirical distribution and true distribution;

3. we will show that for certain simple families of classifiers the results on empirical distributions transport naturally to the empirical cost $\hat{J}_N$, therefore establishing that in these families it actually happens that $\hat{J}_N \to \bar{J}$ uniformly.

5.4 A lemma on uniform convergence

Forget for a moment that $\hat{J}_N$ is a random quantity. Suppose that a realization of the data has been drawn, so that $\hat{J}_N$ is just a *function*. The first point of the program is precisely the following result:

Lemma 5.4.1 *(Uniform convergence). Let C be a set in a normed space, $\{\hat{J}_N : C \to \mathbb{R}\}_{N=1}^{\infty}$ a sequence of functions, and $\bar{J} : C \to \mathbb{R}$ another function. Suppose that*

1. $\hat{c}_N := \arg\min_{c \in C} \hat{J}_N(c)$ *exists for all N;*

2. $\bar{c} := \arg\min_{c \in C} \bar{J}(c)$ *exists;*

3. $\hat{J}_N \to \bar{J}$ *uniformly, that is,*

$$\lim_{N \to \infty} \sup_{c \in C} |\hat{J}_N(c) - \bar{J}(c)| = 0.$$

Then

$$\lim_{N \to \infty} \bar{J}(\hat{c}_N) = \bar{J}(\bar{c}).$$

If, in addition,

4. *C is compact;*

5. *$\bar{J}$ is continuous on C;*

6. *$\bar{c}$ is unique,*

then

$$\lim_{N \to \infty} \hat{c}_N = \bar{c}.$$

Proof. Suppose that hypotheses 1, 2, and 3 hold. Then

$$\begin{aligned}
\bar{J}(\hat{c}_N) &= \hat{J}_N(\hat{c}_N) + \Big(\bar{J}(\hat{c}_N) - \hat{J}_N(\hat{c}_N)\Big) \\
&\le \hat{J}_N(\hat{c}_N) + \sup_{c \in C} |\hat{J}_N(c) - \bar{J}(c)| \\
&= \min_{c \in C} \hat{J}_N(c) + \sup_{c \in C} |\hat{J}_N(c) - \bar{J}(c)| \\
&= \min_{c \in C} \Big(\bar{J}(c) + \Big(\hat{J}_N(c) - \bar{J}(c)\Big)\Big) + \sup_{c \in C} |\hat{J}_N(c) - \bar{J}(c)| \\
&\le \min_{c \in C} \left(\bar{J}(c) + \sup_{c' \in C} |\hat{J}_N(c') - \bar{J}(c')|\right) + \sup_{c \in C} |\hat{J}_N(c) - \bar{J}(c)| \\
&= \min_{c \in C} \bar{J}(c) + 2 \sup_{c \in C} |\hat{J}_N(c) - \bar{J}(c)| \\
&= \bar{J}(\bar{c}) + 2 \sup_{c \in C} |\hat{J}_N(c) - \bar{J}(c)|.
\end{aligned}$$

Due to uniform convergence, the latter quantity converges to $\bar{J}(\bar{c})$ from above. On the other hand, $\bar{J}(\hat{c}_N) \geq \bar{J}(\bar{c})$ by definition. Hence, $\bar{J}(\hat{c}_N)$ converges to $\bar{J}(\bar{c})$ as claimed.
Suppose now that hypotheses 4, 5, and 6 hold as well, and suppose for the sake of contradiction that $\hat{c}_N$ does not converge to $\bar{c}$. Then there exists an ε for which it is possible to extract a sub-sequence $\{\hat{c}_{N_i}\}$ in such a way that $\|\hat{c}_{N_i} - \bar{c}\| \geq \varepsilon$ for all i. Since C is compact, from the latter sequence it is possible to extract a sub-sub-sequence $\{\hat{c}_{N_k}\}$ which converges in C. Let

$$\hat{c}_\infty = \lim_{k\to\infty} \hat{c}_{N_k}$$

By continuity of the norm it holds $\|\hat{c}_\infty - \bar{c}\| \geq \varepsilon$, hence $\hat{c}_\infty \neq \bar{c}$. Finally, due to the continuity of $\bar{J}$,

$$\lim_{k\to\infty} \bar{J}(\hat{c}_{N_k}) = \bar{J}(\hat{c}_\infty) > \bar{J}(\bar{c}),$$

because $\bar{c}$ is the *unique* minimum point of $\bar{J}$. But this contradicts the claim of the first part of the lemma, which must follow from hypotheses 1, 2, and 3. Therefore, $\{\hat{c}_N\}$ converges to $\bar{c}$ as claimed. □

Example. We provide two counterexamples where one hypothesis of the second part of Lemma 5.4.1 fails, and its conclusion is false (of course the counterexamples that follow do *not* prove that the hypotheses are *necessary*; they are meant only to provide insight).
Consider the case in which $C = [0, \infty)$ (unbounded, hence not compact), $\bar{J}(c) = ce^{-c}$ (continuous), and

$$\hat{J}_N(c) = \begin{cases} \frac{1}{N}, & \text{if } c = 0; \\ 0, & \text{if } c = N; \\ \bar{J}(c) & \text{otherwise.} \end{cases}$$

It is easy to see that the unique minimum point of $\bar{J}$ is $\bar{c} = 0$, the unique minimum point of $\hat{J}_N$ is $\hat{c}_N = N$, and $\hat{J}_N$ converges to $\bar{J}$ uniformly as $N \to \infty$; nevertheless hypothesis 4 of the Lemma does not hold, and indeed $\lim_{N\to\infty} \hat{c}_N = \infty \neq \bar{c}$.
Consider now the case in which $C = [0, 1]$ (compact), $\bar{J}$ is defined as follows:

$$\bar{J}(c) = \begin{cases} 1, & \text{if } c = 0; \\ 0, & \text{if } c = 1; \\ c & \text{otherwise,} \end{cases}$$

(not continuous), and

$$\hat{J}_N(c) = \begin{cases} 1, & \text{if } c = 0; \\ \frac{1}{N}, & \text{if } c = 1; \\ 0, & \text{if } c = \frac{1}{N}; \\ c & \text{otherwise.} \end{cases}$$

Here, again, hypotheses 1 to 3 of the Lemma are satisfied, since the unique minimum point of $\bar{J}$ is $\bar{c} = 1$, the unique minimum point of $\hat{J}_N$ is $\hat{c}_N = \frac{1}{N}$, and $\hat{J}_N$ converges to $\bar{J}$ uniformly as $N \to \infty$; but hypothesis 5 of the Lemma does not hold, and indeed $\lim_{N\to\infty} \hat{c}_N = 0 \neq \bar{c}$. □

Lemma 5.4.1 is an asymptotic result. However, provided that an uniform bound is available, that is, if we know that for a certain N $\hat{J}_N$ is "uniformly ε-close" to $\bar{J}$, then given the empirical error $\hat{J}_N(\hat{c}_N)$ we can still say something on the quantities that really matter, $\bar{J}(\hat{c}_N)$ and $\bar{J}(\bar{c})$. Fix N, ε. If

$$\sup_{c\in C} |\hat{J}_N(c) - \bar{J}(c)| \leq \varepsilon,$$

then

$$\begin{aligned}
\bar{J}(\hat{c}_N) &= \Big(\bar{J}(\hat{c}_N) - \hat{J}_N(\hat{c}_N)\Big) + \Big(\hat{J}_N(\hat{c}_N) - \hat{J}_N(\bar{c})\Big) + \Big(\hat{J}_N(\bar{c}) - \bar{J}(\bar{c})\Big) + \bar{J}(\bar{c}) \\
&= (\leq \varepsilon) + (\leq 0) + (\leq \varepsilon) + \bar{J}(\bar{c}) \\
&\leq \bar{J}(\bar{c}) + 2\varepsilon.
\end{aligned}$$

On the other hand, under the same hypothesis,

$$\bar{J}(\hat{c}_N) \leq \hat{J}_N(\hat{c}_N) + \varepsilon.$$

5.5 Convergence of the empirical distribution

To establish whether it happens that $\hat{J}_N \to \bar{J}$ uniformly, we need some results on the convergence of the empirical distribution of the data to the actual distribution. We will focus on real, mono-dimensional data, although the results can be extended to multivariate data (the so-called Vapnik/Chervonenkis theory).

Let $X_1, \cdots X_N$ be independent real random variables, identically distributed and with distribution function

$$F(x) := \mathsf{P}\left[X \leq x\right] = \mathsf{E}\left[\mathbb{1}_{X\leq x}\right]$$

Here, for any fixed $x \in \mathbb{R}$, $\mathbb{1}_{X\leq x}$ is a random variable that takes value 1 if $X \leq x$, and 0 if $X > x$. On the other hand, since X is a random variable, $\mathbb{1}_{X\leq x}$ can be understood also as a *random* function of x which takes value 0 if $x < X$, and 1 if $x \geq X$ (in other words, it is the indicator function of the random interval $[X, \infty)$).

Suppose that the distribution F is unknown; we estimate it with the following function, called the *empirical distribution* of the X_i:

$$\hat{F}_N(x) := \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{X_i\leq x}$$

Think of it as the true distribution with the expectation replaced by a sample average. You can easily verify that $\hat{F}_N$ is actually a distribution function in its own right: it is monotone non-decreasing, continuous from the right, and its limits at $-\infty$ and $+\infty$ are 0 and 1 respectively. Note that, since the X_i are identically distributed,

$$\mathsf{E}\left[\hat{F}_N(x)\right] = \frac{1}{N}\sum_{i=1}^{N}\mathsf{E}\left[\mathbb{1}_{X_i \le x}\right] = \frac{1}{N}\sum_{i=1}^{N}F(x) = F(x)$$

so that, for a fixed x, $\hat{F}_N(x)$ is an unbiased estimator of $F(x)$. Does this estimator enjoy consistency as well? The answer is *yes!*

Lemma 5.5.1 *For any fixed $x \in \mathbb{R}$, $\hat{F}_N(x)$ converges to $F(x)$ almost surely.*

Proof. Indeed, for a fixed x, since X_i are i.i.d. random variables, $\mathbb{1}_{X_i \le x}(x)$ are i.i.d. random variables as well (but taking only the values 0 and 1). Therefore, applying the strong law of large numbers (Theorem D.7.2),

$$\lim_{N\to\infty}\hat{F}_N(x) = \lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}_{X_i \le x} = \mathsf{E}\left[\mathbb{1}_{X_i \le x}\right] = F(x),$$

almost surely. □

This is an asymptotic result. We wish to say something more on the *rate of convergence* to $F(x)$, since we only have N data available. For example, given a fixed x, define the *distribution error*:

$$\begin{aligned} E_N(x) &:= F(x) - \hat{F}_N(x) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\mathsf{P}\left[X_i \le x\right] - \mathbb{1}_{X_i \le x}\right) \\ &= \frac{1}{N}\sum_{i=1}^{N}\nu_i(x), \end{aligned}$$

where $\nu_i(x) := \mathsf{P}\left[X_i \le x\right] - \mathbb{1}_{X_i \le x}(x)$ are now i.i.d. random variables taking values in $[-1, 1]$, with mean $\mathsf{E}\left[\nu_i(x)\right] = \mathsf{P}\left[X_i \le x\right] - \mathsf{E}\left[\mathbb{1}_{X_i \le x}(x)\right] = 0$ and a certain unknown variance σ_ν^2. Of course, then,

$$\begin{aligned} \mathsf{E}\left[E_N(x)\right] &= 0 \\ \mathsf{Var}\left[E_N(x)\right] &= \mathsf{Var}\left[\frac{1}{N}\sum_{i=1}^{N}\nu_i(x)\right] = \frac{1}{N^2}\mathsf{Var}\left[\sum_{i=1}^{N}\nu_i(x)\right] \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\sigma_\nu^2 = \frac{\sigma_\nu^2}{N} \end{aligned}$$

and applying Čebyšev's inequality for a certain $\varepsilon > 0$

$$\mathsf{P}\left[|E_N(x)| \geq \varepsilon\right] \leq \frac{\mathsf{Var}\left[E_N(x)\right]}{\varepsilon^2} = \frac{\sigma_\nu^2}{N\varepsilon^2} \to 0$$

as $N \to \infty$. Hence, $E_N(x) \to 0$ or, which is the same, $\hat{F}_N(x) \to F(x)$, *in probability*, with rate of convergence $1/N$.
This result is nice, but not yet satisfactory. Indeed, due to the central limit theorem, we expect that for big N, the quantity

$$\sqrt{N} E_N(x) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \nu_i(x)$$

is approximately Gaussian with mean 0 and variance σ_ν^2. Hence

$$\begin{aligned} \mathsf{P}\left[|E_N(x)| \geq \varepsilon\right] &= \mathsf{P}\left[\left|\frac{1}{N}\sum_{i=1}^{N} \nu_i(x)\right| \geq \varepsilon\right] \\ &= \mathsf{P}\left[\left|\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \nu_i(x)\right| \geq \sqrt{N}\varepsilon\right] \\ &\simeq \mathsf{P}\left[\left|\mathcal{N}(0, \sigma_\nu^2)\right| \geq \sqrt{N}\varepsilon\right], \end{aligned}$$

and the latter quantity, that is the area of the tails of a Gaussian distribution beyond $\pm\sqrt{N}\varepsilon$, tends to zero *exponentially* as $N \to \infty$. Therefore, we expect that the empirical distribution converges with exponential rate. We will establish it rigorously with the following result.

5.6 Hoeffding's inequality

Hoeffding's inequality is a bound on the probability that a sample average is distant from the mean. It has a similar purpose to Čebyšev's inequality, but it requires more (boundedness) and yields a far tighter bound. We start with a preliminary result, also by Hoeffding:

Lemma 5.6.1 *(Hoeffding). Let Z be a bounded random variable with $\mathsf{E}[Z] = 0$, namely let $Z \in [a, b]$, where $a \leq 0 \leq b$. Then for any $s \geq 0$*

$$\mathsf{E}\left[e^{sZ}\right] \leq e^{\frac{s^2(b-a)^2}{8}}.$$

Proof.
If $s = 0$ or $a = b$ (which implies $a = b = 0$ and hence $Z = 0$ almost surely), the statement reads $1 \leq 1$ and is trivially true. Hence, let us suppose in the following that $s > 0$ and $a < b$.

Note that e^{sZ} is a convex function in Z for any $s > 0$. Let $\lambda = \frac{Z-a}{b-a}$; then $1 - \lambda = \frac{b-Z}{b-a}$, and since $Z \in [a, b]$ we have $\lambda \geq 0$, $1 - \lambda \geq 0$ and moreover

$$\lambda b + (1 - \lambda)a = Z.$$

Therefore

$$\begin{aligned} e^{sZ} &= e^{s(\lambda b + (1-\lambda)a)} \\ &\leq \lambda e^{sb} + (1 - \lambda)e^{sa} \qquad \text{(by convexity)} \\ &= \frac{Z-a}{b-a}e^{sb} + \frac{b-Z}{b-a}e^{sa}; \end{aligned}$$

taking expectations,

$$\mathsf{E}\left[e^{sZ}\right] \leq \frac{be^{sa} - ae^{sb}}{b-a},$$

and taking logarithms,

$$\log \mathsf{E}\left[e^{sZ}\right] \leq \log \frac{be^{sa} - ae^{sb}}{b-a} = sa + \log \frac{b - ae^{s(b-a)}}{b-a}$$

(for small s the argument of the logarithm is not zero, because $a < b$). Let $c = \frac{-a}{b-a}$, and apply the change of variable $t = s(b - a)$:

$$\log \mathsf{E}\left[e^{sZ}\right] \leq -ct + \log(1 - c + ce^t) := f(t).$$

(Again, for small t the expression $1 - c + ce^t$ is not zero since $a < b$.)
Now we expand f in a Taylor polynomial around $t = 0$ (which corresponds to $s = 0$). By Taylor's theorem

$$f(t) = f(0) + f'(0)t + \frac{f''(u)}{2}t^2$$

for a certain u between 0 and t. Consider that

$$\begin{aligned} f(0) &= 0 \\ f'(t) &= -c + \frac{ce^t}{1 - c + ce^t} \\ f'(0) &= 0 \\ f''(t) &= \frac{ce^t(1 - c + ce^t) - (ce^t)^2}{(1 - c + ce^t)^2} = \frac{(1-c)ce^t}{(1 - c + ce^t)^2} \end{aligned}$$

The last expression is of the form $\frac{\alpha\beta}{(\alpha+\beta)^2}$, and since for any numbers α, β such that $\alpha + \beta \neq 0$ it holds

$$\begin{aligned} 0 &\leq \alpha^2 - 2\alpha\beta + \beta^2 = (\alpha - \beta)^2 \\ 4\alpha\beta &\leq \alpha^2 + 2\alpha\beta + \beta^2 = (\alpha + \beta)^2 \\ \frac{\alpha\beta}{(\alpha+\beta)^2} &\leq \frac{1}{4}, \end{aligned}$$

we have $f''(t) \leq \frac{1}{4}$ for all t. But then

$$\log \mathsf{E}\left[e^{sZ}\right] \leq f(t) = \frac{f''(u)}{2} t^2 \leq \frac{t^2}{8} = \frac{s^2(b-a)^2}{8}.$$

Finally, taking exponentials establishes the claim. □

Theorem 5.6.1 *(Hoeffding's inequality). Let $Z_i \in [a_i, b_i]$ be independent, bounded random variables for $i = 1, \cdots, N$, and let $S_N = \frac{1}{N}\sum_{i=1}^N Z_i$. Then*

$$\mathsf{P}\left[S_N - \mathsf{E}\left[S_N\right] \geq \varepsilon\right] \leq \exp\left(\frac{-2N^2\varepsilon^2}{\sum_{i=1}^N (b_i - a_i)^2}\right);$$
$$\mathsf{P}\left[S_N - \mathsf{E}\left[S_N\right] \leq -\varepsilon\right] \leq \exp\left(\frac{-2N^2\varepsilon^2}{\sum_{i=1}^N (b_i - a_i)^2}\right).$$

Note that the variables $\{Z_i\}$ are not supposed to be identically distributed, and $\mathsf{E}\left[S_N\right]$ is not necessarily a constant.

Proof. For all $s \geq 0$,

$$\begin{aligned}
\mathsf{P}\left[S_N - \mathsf{E}\left[S_N\right] \geq \varepsilon\right] &= \mathsf{P}\left[e^{s(S_N - \mathsf{E}[S_N])} \geq e^{s\varepsilon}\right] \\
&\leq \frac{\mathsf{E}\left[e^{s(S_N - \mathsf{E}[S_N])}\right]}{e^{s\varepsilon}} \qquad \text{(Chernoff's bound)} \\
&= \frac{\mathsf{E}\left[e^{\frac{s}{N}\sum_{i=1}^N (Z_i - \mathsf{E}[Z_i])}\right]}{e^{s\varepsilon}} \\
&= \frac{\prod_{i=1}^N \mathsf{E}\left[e^{\frac{s}{N}(Z_i - \mathsf{E}[Z_i])}\right]}{e^{s\varepsilon}} \qquad \text{(by independence).}
\end{aligned}$$

Now, $Z_i - \mathsf{E}\left[Z_i\right]$ is a random variable with mean 0 taking values in $[a_i - \mathsf{E}\left[Z_i\right], b_i - \mathsf{E}\left[Z_i\right]]$. Therefore, from Lemma 5.6.1 we get

$$\begin{aligned}
\mathsf{E}\left[e^{\frac{s}{N}(Z_i - \mathsf{E}[Z_i])}\right] &\leq \exp\left(\frac{\frac{s^2}{N^2}(b_i - \mathsf{E}\left[Z_i\right] - a_i + \mathsf{E}\left[Z_i\right])^2}{8}\right) \\
&= \exp\left(\frac{s^2(b_i - a_i)^2}{8N^2}\right);
\end{aligned}$$

hence

$$\begin{aligned}
\mathsf{P}\left[S_N - \mathsf{E}\left[S_N\right] \geq \varepsilon\right] &\leq \frac{\prod_{i=1}^N \exp\left(\frac{s^2(b_i - a_i)^2}{8N^2}\right)}{e^{s\varepsilon}} \\
&= \exp\left(\frac{s^2 \sum_{i=1}^N (b_i - a_i)^2}{8N^2} - s\varepsilon\right)
\end{aligned}$$

for all $s \geq 0$. In particular, the inequality holds for the minimum over $s \geq 0$ of the right-hand side, which is attained at the minimum exponent, which in turn is attained for $s = 4N^2\varepsilon / \sum_{i=1}^{N}(b_i - a_i)^2$; namely,

$$\mathsf{P}\left[S_N - \mathsf{E}\left[S_N\right] \geq \varepsilon\right] \leq \exp\left(\frac{-2N^2\varepsilon^2}{\sum_{i=1}^{N}(b_i - a_i)^2}\right),$$

and the first part of the claim is proven. The proof of the second part of the claim is obtained in a similar way, starting from a Chernoff bound on $\mathsf{P}\left[\mathsf{E}\left[S_N\right] - S_N \geq \varepsilon\right]$. □

Corollary 5.6.1 *Under the same hypotheses of Theorem 5.6.1,*

$$\mathsf{P}\left[\left|S_N - \mathsf{E}\left[S_N\right]\right| \geq \varepsilon\right] \leq 2\exp\left(\frac{-2N^2\varepsilon^2}{\sum_{i=1}^{N}(b_i - a_i)^2}\right).$$

Proof. The statements follows trivially from the Theorem, since the events whose probabilities are bounded by Hoeffding's inequality are disjoint:

$$\mathsf{P}\left[\left|S_N - \mathsf{E}\left[S_N\right]\right| \geq \varepsilon\right] = \mathsf{P}\left[S_N - \mathsf{E}\left[S_N\right] \geq \varepsilon\right] + \mathsf{P}\left[S_N - \mathsf{E}\left[S_N\right] \leq -\varepsilon\right].$$

□

Example. Let us test our new result, and compare it with Čebyšev's inequality. We toss a fair coin $N = 100$ times, and compute bounds on the probability that the number of heads that we get is at most 10 or at least 90. For $i = 1, \cdots, 100$ define the independent Bernoulli variables

$$Z_i = \begin{cases} 1, & \text{with probability } p = \frac{1}{2}, \text{ if at the } i\text{-th outcome we get a head;} \\ 0, & \text{with probability } 1 - p = \frac{1}{2}, \text{ if at the } i\text{-th outcome we get a tail.} \end{cases}$$

The total number of heads is $\sum_{i=1}^{100} Z_i$, which is a binomial variable $B(N, p) = B(100, \frac{1}{2})$ with mean $Np = 50$ and variance $Np(1-p) = 25$. The event "the number of heads is at most 10 or at least 90" reads $\left|\sum_{i=1}^{100} Z_i - 50\right| \geq 40$. Therefore Čebyšev's inequality yields

$$\mathsf{P}\left[\left|\sum_{i=1}^{100} Z_i - 50\right| \geq 40\right] \leq \frac{25}{40^2} \simeq 1.56 \times 10^{-2}.$$

On the other hand, the Z_i are bounded random variables taking values in $[0,1]$, hence Hoeffding's inequality yields

$$\begin{aligned}
\mathsf{P}\left[\left|\sum_{i=1}^{100} Z_i - 50\right| \geq 40\right] &= \mathsf{P}\left[\left|\frac{1}{100}\sum_{i=1}^{100} Z_i - \frac{50}{100}\right| \geq \frac{40}{100}\right] \\
&= \mathsf{P}\left[\left|S_{100} - \mathsf{E}\left[S_{100}\right]\right| \geq \frac{2}{5}\right] \\
&\leq 2\exp\left(\frac{-2\cdot 100^2\cdot\left(\frac{2}{5}\right)^2}{\sum_{i=1}^{100}(1-0)^2}\right) \\
&= 2e^{-32} \simeq 2.53\times 10^{-14}.
\end{aligned}$$

You can see that Hoeffding's bound is *much* tighter than Čebyšev's one. Compare the bounds with the true probability:

$$\begin{aligned}
&\mathsf{P}\left[\left(\sum_{i=1}^{100} Z_i \leq 10\right) \vee \left(\sum_{i=1}^{100} Z_i \geq 90\right)\right] \\
&= \sum_{k=0}^{10}\binom{100}{k}\left(\frac{1}{2}\right)^k\left(\frac{1}{2}\right)^{100-k} + \sum_{k=90}^{100}\binom{100}{k}\left(\frac{1}{2}\right)^k\left(\frac{1}{2}\right)^{100-k} \\
&= 2\cdot 2^{-100}\cdot\sum_{k=0}^{10}\binom{100}{k} \simeq 3.06\times 10^{-17}.
\end{aligned}$$

□

5.7 Exponential convergence of the empirical distribution

We can now apply Hoeffding's inequality to the empirical distribution. For a fixed x, let

$$\begin{aligned}
Z_i &= \mathbb{1}_{X_i \leq x} \\
\mathsf{E}\left[Z_i\right] &= \mathsf{P}\left[X_i \leq x\right] = F(x) \\
S_N &= \frac{1}{N}\sum_{i=1}^{N} Z_i = \hat{F}_N(x) \\
\mathsf{E}\left[S_N\right] &= F(x) \\
[a_i, b_i] &= [0,1]
\end{aligned}$$

We obtain

$$\begin{aligned}
\mathsf{P}\left[\left|\hat{F}_N(x) - F(x)\right| \geq \varepsilon\right] &\leq 2\exp\left(\frac{-2N^2\varepsilon^2}{\sum_{i=1}^{N}(1-0)^2}\right) \\
&= 2e^{-2N\varepsilon^2}.
\end{aligned}$$

In words: at all x, the empirical distribution converges *in probability* to the true distribution, with exponential rate of convergence.

It so happens that the exponential rate of convergence is enough to establish *almost sure* convergence as well. Although we already know that almost sure convergence holds from the strong law of large numbers (Lemma 5.5.1), in Appendix F.1 we provide an alternative proof based on Hoeffding's inequality. The proof is somewhat subtle, but instructive and worth some effort.

Summing up, we would say that the empirical distribution behaves quite nicely, but these results are still not sufficient to say something about the empirical *error* $\hat{J}_N$, and its uniform convergence to the real error $\bar{J}$. For this, we need to establish that the empirical distribution converges to the true one *uniformly*; this is the statement of the Glivenko/Cantelli theorem.

5.8 Uniform convergence of the empirical distribution

Theorem 5.8.1 *(Glivenko/Cantelli). Let $\{X_i\}$ be a sequence of independent variables, identically distributed according to the distribution $F(x)$, and let the empirical distribution $\hat{F}_N(x)$ be defined as before. Then, almost surely, $\hat{F}_N \to F$ uniformly, that is,*

$$\lim_{N\to\infty} \sup_{x\in\mathbb{R}} |\hat{F}_N(x) - F(x)| = 0 \quad \textit{almost surely.}$$

Proof. Although the theorem is true in general, we provide a simple proof for the particular case when F is continuous. You can find a general proof in [5] or [1, Theorem 20.6, p. 269].

Since F is continuous, for any $\varepsilon > 0$ there exist finitely many points

$$x_1 < \cdots < x_j < \cdots < x_K$$

such that $F(x_1) \leq \varepsilon/2$, $1 - F(x_K) \leq \varepsilon/2$, and $F(x_{j+1}) - F(x_j) \leq \varepsilon/2$ for all $j = 1, \cdots, K-1$. Given the K points, let us denote, for the sake of clarity, "$x_0 = -\infty$" and "$x_{K+1} = +\infty$" so that $F(x_0) = 0$, $F(x_{K+1}) = 1$, and the requirement reads: $F(x_{j+1}) - F(x_j) \leq \varepsilon/2$ for all $j = 0, \cdots, K$.

Now, by the strong law of large numbers, at all points x_j it holds $\hat{F}_N(x_j) \to F(x_j)$ almost surely. Hence, almost surely for all $\varepsilon > 0$ there exists N_j such that $|\hat{F}_N(x_j) - F(x_j)| \leq \varepsilon/2$ for all $N \geq N_j$. Since the x_j are finitely many, it is well defined the index $\bar{N} := \max_{j=1\cdots K} N_j$, such that for all $N \geq \bar{N}$ the inequalities

$$|\hat{F}_N(x_j) - F(x_j)| \leq \varepsilon/2, \quad j = 1, \cdots, K$$

hold simultaneously. The other two inequalities

$$|\hat{F}_N(x_0) - F(x_0)| \leq \varepsilon/2, \quad |\hat{F}_N(x_{K+1}) - F(x_{K+1})| \leq \varepsilon/2$$

hold trivially, since $\hat{F}_N(-\infty) = F(-\infty) = 0$ and $\hat{F}_N(+\infty) = F(+\infty) = 1$. Note that $\bar{N}$ is random, but this is inessential here; the crucial point is that since the $\{x_j\}$ are finitely many, $\bar{N}$ exists finite almost surely.
Consider now any real number x. There exists a certain index $j \in \{0, \cdots, K\}$ such that x belongs to an interval delimited by x_j and x_{j+1}, namely $x \in (x_0, x_1)$ if $x < x_1$, or $x \in [x_j, x_{j+1})$ for a certain $j \in \{1, \cdots, K\}$ otherwise. Thus, for all $N \geq \bar{N}$,

$$\begin{aligned}\hat{F}_N(x) &\geq \hat{F}_N(x_j) \quad \text{(by the monotonicity of } \hat{F}_N\text{)}\\ &\geq F(x_j) - \frac{\varepsilon}{2} \quad \text{(by the inequalities that hold for } N \geq \bar{N}\text{)}\\ &\geq F(x) - \varepsilon \quad \text{(by construction of the } \{x_j\}\text{)},\end{aligned}$$

and by similar arguments

$$\begin{aligned}\hat{F}_N(x) &\leq \hat{F}_N(x_{j+1})\\ &\leq F(x_{j+1}) + \frac{\varepsilon}{2}\\ &\leq F(x) + \varepsilon.\end{aligned}$$

The latter inequalities together yield $|\hat{F}_N(x) - F(x)| \leq \varepsilon$.
Summing up, almost surely for all ε there exists $\bar{N}$ such that, for all $N \geq \bar{N}$, it holds $|\hat{F}_N(x) - F(x)| \leq \varepsilon$ for all x, and consequently

$$\sup_{x\in\mathbb{R}} |\hat{F}_N(x) - F(x)| \leq \varepsilon.$$

In other terms,

$$\sup_{x\in\mathbb{R}} |\hat{F}_N(x) - F(x)| \to 0 \quad \text{almost surely.}$$

□

As before, we wish to investigate the behavior of $\hat{F}_N(x)$ for finite N, to establish a result on “uniform closeness”. To do this, first notice that in the proof of the Glivenko/Cantelli theorem the number K of points is a simple function of ε, namely $K = \lceil 2/\varepsilon \rceil - 1$; for the sake of clarity, take $K = 2/\varepsilon$, and assume that it is an integer number.
Then, for fixed ε, the probability that $|\hat{F}_N(x_j) - F(x_j)| > \varepsilon/2$ at *any* of the K points is

$$\begin{aligned}\mathsf{P}\left[\bigcup_{j=1}^{K} \left\{|\hat{F}_N(x_j) - F(x_j)| > \varepsilon/2\right\}\right] &\leq \sum_{j=1}^{K} \mathsf{P}\left[|\hat{F}_N(x_j) - F(x_j)| > \varepsilon/2\right]\\ &\leq \sum_{j=1}^{K} 2e^{-2N(\varepsilon/2)^2} \qquad \text{(Hoeffding)}\\ &= \frac{4}{\varepsilon} e^{-N\varepsilon^2/2}.\end{aligned}$$

Hence, with probability at least $1 - \frac{4}{\varepsilon}e^{-N\varepsilon^2/2}$ the inequalities

$$|\hat{F}_N(x_j) - F(x_j)| \leq \varepsilon/2, \quad j = 1, \cdots, K$$

hold simultaneously. Repeating the other steps in the proof of Glivenko/Cantelli's theorem, we obtain the following result:

Lemma 5.8.1 *Fix $\varepsilon > 0, N$. Then*

$$\mathsf{P}\left[\sup_{x\in\mathbb{R}} |\hat{F}_N(x) - F(x)| \leq \varepsilon\right] \geq 1 - \frac{4}{\varepsilon}e^{-N\varepsilon^2/2}.$$

This probability tends to 1 with exponential rate as $N \to \infty$. We remark that this convergence rate would be enough to establish again almost sure convergence, as in Glivenko/Cantelli's theorem.

The last step that we need before returning to classification problems is an extension to two variables, one of which is binary.
Let $\{(X_i, Y_i)\}$ be a sequence of independent and identically distributed random pairs, where $X_i \in \mathbb{R}$ and $Y_i \in \{0, 1\}$. Let $\mathsf{P}[Y_i = 0] = \alpha, \mathsf{P}[Y_i = 1] = 1 - \alpha$, and define

$$\begin{aligned}
F^0(x) &:= \mathsf{P}[X_i \leq x, Y_i = 0], \\
F^1(x) &:= \mathsf{P}[X_i \leq x, Y_i = 1].
\end{aligned}$$

Then of course

$$\begin{aligned}
F(x) &= \mathsf{P}[X_i \leq x] \\
&= \mathsf{P}[(X_i \leq x \wedge Y_i = 0) \vee (X_i \leq x \wedge Y_i = 1)] \\
&= \mathsf{P}[X_i \leq x \wedge Y_i = 0] + \mathsf{P}[X_i \leq x \wedge Y_i = 1] \\
&= F^0(x) + F^1(x);
\end{aligned}$$

moreover,

$$\begin{aligned}
\lim_{x\to-\infty} F^0(x) &= \lim_{x\to-\infty} \mathsf{P}[X_i \leq x \wedge Y_i = 0] \\
&= \lim_{n\to\infty} \mathsf{P}[\{X_i \leq -n\} \cap \{Y_i = 0\}] \\
&= \mathsf{P}\left[\bigcap_{n=1}^{\infty} \{X_i \leq -n\} \cap \{Y_i = 0\}\right] \\
&= \mathsf{P}[\varnothing \cap \{Y_i = 0\}] = 0; \\
\lim_{x\to\infty} F^0(x) &= \lim_{x\to\infty} \mathsf{P}[X_i \leq x \wedge Y_i = 0] = \mathsf{P}[Y_i = 0] = \alpha; \\
\lim_{x\to-\infty} F^1(x) &= 0; \\
\lim_{x\to\infty} F^1(x) &= \mathsf{P}[Y_i = 1] = 1 - \alpha.
\end{aligned}$$

The functions F^0 and F^1 behave like distribution functions, except for their limits at $+\infty$. Define now

$$\hat{F}^0_N(x) := \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{X_i \le x, Y_i = 0} \ ,$$
$$\hat{F}^1_N(x) := \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{X_i \le x, Y_i = 1} \ .$$

These are the respective "empirical" counterparts of F^0 and F^1. Since $\mathsf{E}\left[\mathbb{1}_{X_i \le x, Y_i = 0}\right] = \mathsf{P}\left[X_i \le x, Y_i = 0\right] = F^0(x)$, and analogously for $F^1(x)$, by the strong law of large numbers we get, for all $x \in \mathbb{R}$,

$$\begin{aligned} \hat{F}^0_N(x) &\to F^0(x) \quad \text{almost surely as } N \to \infty; \\ \hat{F}^1_N(x) &\to F^1(x) \quad \text{almost surely as } N \to \infty. \end{aligned}$$

Hence we can repeat the proof of Glivenko/Cantelli theorem, with minor changes, to show that

$$\begin{aligned} \sup_{x\in\mathbb{R}} |\hat{F}^0_N(x) - F^0(x)| &\to 0 \quad \text{almost surely;} \\ \sup_{x\in\mathbb{R}} |\hat{F}^1_N(x) - F^1(x)| &\to 0 \quad \text{almost surely.} \end{aligned}$$

For example, to prove the convergence of $\hat{F}^0_N$ to F^0, the differences from the "standard" Glivenko/Cantelli proof are that

1. the limit of F^0 as $x \to \infty$ is α, not 1;
2. the limit of $\hat{F}^0_N$ as $x \to \infty$ is $\alpha_N := \frac{\text{number of } Y_i \text{ equal to } 0}{N}$, that is different from α; hence, with the same notation of Theorem 5.8.1, the inequality

   $$|\hat{F}^0_N(x_{K+1}) - F^0(x_{K+1})| \le \varepsilon/2$$

   is *not* satisfied trivially. However, $\alpha_N \to \alpha$ almost surely by the strong law of large numbers, hence this issue is easily solved for $\bar{N}$ large enough.

Finally, we can repeat our considerations for the probabilistic bound of Lemma 5.8.1. Since now the "distributions" F^0 and F^1 do not have the co-domain $[0, 1]$, but $[0, \alpha]$ and $[0, 1 - \alpha]$ respectively, the number of points to "split" the co-domains in sub-intervals of length $\varepsilon/2$ and carry on with Glivenko/Cantelli-like proofs are now $K_0 = 2\alpha/\varepsilon$ for F^0, and $K_1 = 2(1 - \alpha)/\varepsilon$ for F^1. The result reads as follows:

Lemma 5.8.2 *Let $\{(X_i, Y_i)\}$ be a sequence of independent and identically distributed random pairs, where $X_i \in \mathbb{R}$ and $Y_i \in \{0, 1\}$. Let* $\mathsf{P}\left[Y_i = 0\right] = \alpha$

and let the functions $F^0(\cdot), F^1(\cdot), \hat{F}_N^0(\cdot), \hat{F}_N^1(\cdot)$ *be defined as before. Then, for any* $\varepsilon > 0$ *and* N*,*

$$\mathsf{P}\left[\sup_{x\in\mathbb{R}} |\hat{F}_N^0(x) - F^0(x)| \leq \varepsilon\right] \geq 1 - \frac{4\alpha}{\varepsilon} e^{-N\varepsilon^2/2},$$
$$\mathsf{P}\left[\sup_{x\in\mathbb{R}} |\hat{F}_N^1(x) - F^1(x)| \leq \varepsilon\right] \geq 1 - \frac{4(1-\alpha)}{\varepsilon} e^{-N\varepsilon^2/2}.$$

5.9 Threshold classifiers

Armed with a lot of results from the previous sections, namely

- the strong law of large numbers,
- Hoeffding's inequality and its consequences,
- the Glivenko/Cantelli theorem and its generalizations, and
- the lemma on uniform convergence,

we can finally return to classifiers and the uniform convergence of the empirical error. Consider the following family of mono-dimensional classifiers:

$$\mathcal{F}_T = \{\mathbb{1}_{(-\infty,c]}(\cdot)\}.$$

These are called *threshold classifiers*, and are parameterized by a number $c \in \mathbb{R}$. (The parameter set is $C = \mathbb{R}$.) A classifier in this family has the form

$$\hat{f}_c(u) = \begin{cases} 1, & \text{if } u \leq c; \\ 0, & \text{if } u > c. \end{cases}$$

Suppose that the data $(U_1, Y_1), \cdots, (U_N, Y_N)$ are available, and recall the following definitions regarding true and empirical error:

$$\bar{J}(c) = \mathsf{P}\left[Y_i \neq \hat{f}_c(U_i)\right],$$
$$\hat{J}_N(c) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{Y_i \neq \hat{f}_c(U_i)} \ ,$$
$$\bar{c} = \arg\min_{c\in C} \bar{J}(c),$$
$$\hat{c}_N = \arg\min_{c\in C} \hat{J}_N(c).$$

To establish the convergence of the minimum empirical error, we are going to exploit Lemma 5.4.1, which requires the existence of $\hat{c}_N$ and $\bar{c}$, and possibly their uniqueness.

It is easy to convince ourselves that $\hat{c}_N$ always exists. Indeed $\hat{J}_N(c)$ takes constant values in each of the intervals $(-\infty, U_1), [U_1, U_2), [U_2, U_3), \cdots, [U_N, \infty)$;

since these intervals are finitely many, $\min_{c \in \mathbb{R}} \hat{J}_N(c)$ is the minimum of those finitely many constant values. On the other hand, $\hat{c}_N$ is never unique, because $\hat{J}_N(c)$ takes the same value over the entire interval that attains the minimum.

The same questions arise with respect to $\bar{c}$. In fact, neither its existence nor its uniqueness are automatically guaranteed for the threshold classifiers, as the following example shows:

Example. Let U_i be Gaussian with mean 0 and variance 1, and let

$$Y_i = \begin{cases} 0 & \text{with probability } \alpha = 0.1, \\ 1 & \text{with probability } 1 - \alpha = 0.9, \end{cases}$$

irrespective of U_i. Then

$$\bar{J}(c) = 0.1 \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{c} e^{-u^2/2} \, du + 0.9 \cdot \frac{1}{\sqrt{2\pi}} \int_{c}^{\infty} e^{-u^2/2} \, du.$$

Then $\inf_{c \in \mathbb{R}} \bar{J}(c) = \lim_{c \to \infty} \bar{J}(c) = 0.1$, but $\bar{J}(c) = 0.1$ is not attained for any $c \in \mathbb{R}$, hence the minimum point $\bar{c}$ does not exist. As another example, let U_i be uniform in $[0, 3]$, and

$$Y_i = \begin{cases} 1, & \text{if } U_i \in [0,1] \text{ or } U_i \in [2,3]; \\ 0, & \text{otherwise.} \end{cases}$$

Here, $\inf_{c \in \mathbb{R}} \bar{J}(c) = \frac{1}{3}$, which is attained at both $\bar{c} = 1$ and $\bar{c} = 3$. Thus, the minimum point is not unique. □

Regarding uniqueness, we do not really care, because we will use only the first part of Lemma 5.4.1, which does not require it. On the other hand, the existence of $\bar{c}$ is needed, but it depends crucially on the data-generation rule, which we suppose unknown. Therefore the best that we can do is to assume it as a hypothesis.

Our final result shows that under this and other fairly general hypotheses, convergence holds, so that the minimization of the empirical error in $\mathcal{F}_T$ actually makes sense:

Theorem 5.9.1 *Let $\mathcal{F}_T = \{\mathbb{1}_{(-\infty,c]}(\cdot)\}$, parameterized by $c \in C$, where $C = \mathbb{R}$. Suppose that $(U_1, Y_1), \cdots, (U_N, Y_N)$ are independent and identically distributed, where U_i has continuous distribution $F(u)$, $Y_i \in \{0, 1\}$, and $\mathsf{P}[Y_i = 0] = \alpha$. Define $\bar{J}(c)$, $\hat{J}_N(c)$, $\bar{c}$, and $\hat{c}_N$ as usual, and assume that $\bar{c}$ exists. Then:*

1. *almost surely, $\hat{J}_N \to \bar{J}$ uniformly;*

2. *almost surely,* $\bar{J}(\hat{c}_N) \to \bar{J}(\bar{c})$*;*

3. *for fixed* $\varepsilon > 0$ *and* N*, it holds*

$$\mathsf{P}\left[\sup_{c\in C} |\hat{J}_N(c) - \bar{J}(c)| \geq \varepsilon\right] \leq \frac{8}{\varepsilon} e^{-N\varepsilon^2/8}.$$

Proof. For each U_i define $V_i = -U_i$, and denote $G(u)$ the distribution of V_i. Define, as we did in the previous section,

$$\begin{aligned}
F^0(u) &= \mathsf{P}\left[U_i \leq u, Y_i = 0\right] \\
\hat{F}^0_N(u) &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{U_i \leq u, Y_i = 0} \\
G^1(u) &= \mathsf{P}\left[V_i \leq u, Y_i = 1\right] \quad \text{(it tends to } \mathsf{P}\left[Y_i = 1\right] = 1 - \alpha \text{ as } u \to \infty) \\
\hat{G}^1_N(u) &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{V_i \leq u, Y_i = 1}
\end{aligned}$$

Note that:

$$\begin{aligned}
\bar{J}(c) &= \mathsf{P}\left[\left(\hat{f}_c(U_i) = 1 \wedge Y_i = 0\right) \vee \left(\hat{f}_c(U_i) = 0 \wedge Y_i = 1\right)\right] \\
&= \mathsf{P}\left[U_i \leq c, Y_i = 0\right] + \mathsf{P}\left[U_i > c, Y_i = 1\right] \\
&= \mathsf{P}\left[U_i \leq c, Y_i = 0\right] + \mathsf{P}\left[V_i < -c, Y_i = 1\right] \\
&= F^0(c) + G^1(-c)
\end{aligned}$$

where the last equality holds in particular because, since U_i and consequently V_i have continuous distributions, $\mathsf{P}\left[V_i = -c\right] = \mathsf{P}\left[U_i = c\right] = 0$. Moreover,

$$\begin{aligned}
\hat{J}_N(c) &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\left\{\left(\hat{f}_c(U_i)=1 \wedge Y_i=0\right) \vee \left(\hat{f}_c(U_i)=0 \wedge Y_i=1\right)\right\}} \\
&= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{U_i \leq c, Y_i = 0} + \mathbb{1}_{U_i > c, Y_i = 1} \\
&= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{U_i \leq c, Y_i = 0} + \mathbb{1}_{V_i < -c, Y_i = 1} \\
&= \hat{F}^0_N(c) + \hat{G}^1_N(-c) - \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{V_i = -c, Y_i = 1}
\end{aligned}$$

Exploiting these expressions for $\bar{J}(c)$ and $\hat{J}_N(c)$,

$$\begin{aligned}
\left|\hat{J}_N(c) - \bar{J}(c)\right| \leq \left|\hat{F}^0_N(c) - F^0(c)\right| + \left|\hat{G}^1_N(-c) - G^1(-c)\right| \\
+ \left|\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{U_i = c, Y_i = 1}\right|.
\end{aligned}$$

As $N \to \infty$, the supremum over C of the first two terms on the right-hand side tends almost surely to zero by the Glivenko/Cantelli theorem. On the other hand, the third term is almost surely equal to zero because U_i has continuous distribution (the event $\{U_i = c\}$ has probability zero). Therefore the supremum of $|\hat{J}_N(c) - \bar{J}(c)|$ tends also almost surely to zero; this establishes the first part of the claim.

In other words, $\hat{J}_N(c) \to \bar{J}(c)$ uniformly almost surely; applying Lemma 5.4.1 establishes the second part of the claim.

Now fix ε and N. By Lemma 5.8.2,

$$\mathsf{P}[\mathcal{A}] := \mathsf{P}\left[\sup_{c\in\mathbb{R}} \left|\hat{F}_N^0(c) - F^0(c)\right| \geq \varepsilon\right] \leq \frac{4\alpha}{\varepsilon} e^{-N\varepsilon^2/2};$$
$$\mathsf{P}[\mathcal{B}] := \mathsf{P}\left[\sup_{c\in\mathbb{R}} \left|\hat{G}_N^1(-c) - G^1(-c)\right| \geq \varepsilon\right] \leq \frac{4(1-\alpha)}{\varepsilon} e^{-N\varepsilon^2/2}.$$

On the other hand, the event $\left\{\left|\frac{1}{N}\sum_{i=1}^N \mathbb{1}_{U_i=c,Y_i=1}\right| \geq \varepsilon\right\}$ has probability 0 for all $\varepsilon > 0$, since $U_i = -V_i$ has continuous distribution. Putting together the inequalities,

$$\mathsf{P}\left[\sup_{c\in\mathbb{R}} |\hat{J}_N(c) - \bar{J}(c)| \geq 2\varepsilon\right] \leq \mathsf{P}[\mathcal{A} \cup \mathcal{B}] \leq \mathsf{P}[\mathcal{A}] + \mathsf{P}[\mathcal{B}] \leq \frac{4}{\varepsilon} e^{-N\varepsilon^2/2}.$$

The third part of the claim follows immediately from a change of variable $2\varepsilon = \varepsilon'$. □

Using the simple arguments shown after the proof of Lemma 5.4.1, one can also show that for fixed $\varepsilon > 0$ and N, both the events $\bar{J}(\hat{c}_N) \leq \hat{J}_N(\hat{c}_N) + \varepsilon$ and $\bar{J}(\hat{c}_N) \leq \bar{J}(\bar{c}) + 2\varepsilon$ have probability at least $1 - \frac{8}{\varepsilon}e^{-N\varepsilon^2/8}$. Unfortunately, the probabilistic bound established by Theorem 5.9.1 is not really tight *unless N is very large*, as the following example shows.

Example. Suppose we wish that the empirical error is within distance at most $\varepsilon = 0.1$ from the true one. The theorem says that this happens with probability at least $1 - \mathsf{P}$, where

$$\mathsf{P} = \frac{8}{1/10} e^{-N\frac{(1/10)^2}{8}} = 80 e^{-\frac{N}{800}}.$$

Solving for N,

$$N = 800 \cdot \log\left(\frac{80}{\mathsf{P}}\right).$$

To attain, for instance, $\mathsf{P} = 0.0001$, one needs $N \simeq 11000$ measures.

Suppose now that we ask that the empirical error is within distance at most $\varepsilon = 0.05$ from the true one. This happens with probability at least $1 - \mathsf{P}$, where

$$\mathsf{P} = \frac{8}{5/100} e^{-N\frac{(5/100)^2}{8}} = 160 e^{-\frac{N}{3200}},$$

and solving for N,

$$N = 3200 \cdot \log\left(\frac{160}{\mathsf{P}}\right).$$

Even to attain $\mathsf{P} = 0.01$, now we need $N \simeq 31000$ measures. □

Thus, the applicability and practical usefulness of Theorem 5.9.1 depends on how many data are available. In fact, the bound established by this result is not the tightest available. Moreover, it comes in the form of a probabilistic inequality for an *uniform* bound over a large set; hence, the bound is loose because we are asking a lot from it. In practice, the empirical error will be more likely to be close to the true one than what the bound says.

However, you will not fail to recognize that Theorem 5.9.1 is mathematically interesting for another reason, which has been left implicit from the beginning:

> *the bound does not depend on either the distribution of the data* $(U_1, Y_1), \cdots, (U_N, Y_N)$*, or the way in which* Y_i *depends on* U_i*.*

The result is, in other words, *distribution-free.* Many statistical methods in widespread use assume hypotheses like Gaussianity like they were just obvious, while they are in fact rather restrictive, if not sometimes unrealistic, and rely on the estimation of parameters like mean and variance of Gaussian variables. Instead, the above result makes the least possible assumptions on the data (the most demanding one being that they actually come in *independent* pairs); and common sense tells us that the less assumptions we make on our data, the broader will be the area in which our theory can be applied. This is not to say that inferential statistics based on the Gaussian, Fisher, and Student's distributions should be avoided (the central limit theorem stands there as a milestone suggesting just the opposite), but that we should refrain as much as possible from pretending hypotheses "for free" just because the consequent theses are mathematically more appealing, or simpler, or traditional. Indeed, more and more research is being done, nowadays, on statistical methods that do not assume that the data belong to restricted classes of distributions, parameterizable with a few numbers (mean, variance, and so on). "Non-parametric statistics" is the generic term encompassing such methods; in the following chapters, we will study other two of them.

5.10 Interval classifiers

We propose some arguments, analogous to those of the previous section, applied to a family of *interval classifiers* similar to those that we have already encountered. The family is the following:

$$\mathcal{F}_I = \{\mathbb{1}_{(a,b]}(u)\}.$$

They are the indicator functions of intervals *open on the left and closed on the right*, parameterized by a pair $c = (a, b) \in C$, where $C = \{(a, b) \in \mathbb{R}^2 \mid a < b\}$.) A classifier in this family has the form

$$\hat{f}_c(u) = \begin{cases} 1, & \text{if } a < u \leq b; \\ 0, & \text{otherwise.} \end{cases}$$

Let us define, as before,

$$\begin{aligned} F^0(u) &= \mathsf{P}\left[U_i \leq u, Y_i = 0\right] \\ G^1(u) &= \mathsf{P}\left[V_i \leq u, Y_i = 1\right] \end{aligned}$$

etc.

Now the only subtle point to notice is that, exactly like for a distribution function it holds

$$\mathsf{P}\left[a < X \leq b\right] = F(b) - F(a),$$

for the "marginal" distribution functions it holds

$$\mathsf{P}\left[a < U_i \leq b, Y_i = 0\right] = F^0(b) - F^0(a)$$

The rest relies on tedious, but straightforward computations, which we summarize briefly:

$$\begin{aligned} \bar{J}(c) &= \bar{J}((a, b)) \\ &= \mathsf{P}\left[(a < U_i \leq b, Y_i = 0) \vee (U_i \leq a, Y_i = 1) \vee (U_i > b, Y_i = 1)\right] \\ &= F^0(b) - F^0(a) + F^1(a) + G^1(-b); \\ \hat{J}_N(c) &= \hat{J}_N((a, b)) \\ &= \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{a<U_i \leq b, Y_i=0} + \mathbb{1}_{U_i \leq a, Y_i=1} + \mathbb{1}_{U_i > b, Y_i=1} \\ &= \hat{F}^0_N(b) - \hat{F}^0_N(a) + \hat{F}^1_N(a) + \hat{G}^1_N(-b) + \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{V_i=-b, Y_i=1}; \end{aligned}$$

Therefore

$$\begin{aligned}\left|\hat{J}_N(c)-\bar{J}(c)\right| &= \left|\hat{J}_N((a,b))-\bar{J}((a,b))\right| \\ &\leq \left|\hat{F}_N^0(a)-F^0(a)\right| + \left|\hat{F}_N^0(b)-F^0(b)\right| \\ &+ \left|\hat{F}_N^1(a)-F^1(a)\right| + \left|\hat{G}_N^1(-b)-G^1(-b)\right| \\ &+ \left|\frac{1}{N}\sum_{i=1}^N \mathbb{1}_{U_i=b,Y_i=1}\right|.\end{aligned}$$

With these observations, using only the Glivenko/Cantelli theorem (although now $C \subset \mathbb{R}^2$), one can prove the following result, analogous of Theorem 5.9.1:

Theorem 5.10.1 *Let $\mathcal{F}_I = \{\mathbb{1}_{(a,b]}(u)\}$, parameterized by $c \in C$, where $C = \{(a,b) \in \mathbb{R}^2 \mid a < b\}$. Suppose that $(U_1,Y_1),\cdots,(U_N,Y_N)$ are independent and identically distributed, where U_i has continuous distribution $F(u)$, $Y_i \in \{0,1\}$, and $\mathsf{P}[Y_i = 0] = \alpha$. Define $\bar{J}(c)$, $\hat{J}_N(c)$, $\bar{c}$, and $\hat{c}_N$ as usual, and assume that $\bar{c}$ exists. Then:*

1. *almost surely, $\hat{J}_N \to \bar{J}$ uniformly;*
2. *almost surely, $\bar{J}(\hat{c}_N) \to \bar{J}(\bar{c})$;*
3. *for fixed $\varepsilon > 0$ and N, it holds*

$$\mathsf{P}\left[\sup_{c\in C}|\hat{J}_N(c)-\bar{J}(c)| \geq \varepsilon\right] \leq \frac{16}{\varepsilon}e^{-N\varepsilon^2/32}.$$

(As you can guess from the example in the previous section, the bound is now rather loose.)

5.11 Exercises for Chapter 5

Problem 1 (complaint telephone calls).
A big company receives N complaint calls $\{t_i\}$, $i = 1,\cdots,N$, and for each call it records the region $\{r_i\}$, $i = 1,\cdots,N$, of the caller. Suppose that the $\{r_i\}$ are independent and identically distributed random variables taking values, say, in the set of the 20 Italian regions $\{\text{Piedmont}, \text{Lombardy},\cdots,\text{Sicily}\}$ with respective probabilities $P = \{p(\text{Piedmont}), p(\text{Lombardy}),\cdots,p(\text{Sicily})\}$ (which depend, in general, on the region's population, on the quality of service in the region, etc.). Using Hoeffding's inequality, compute how many telephone calls should be recorded in order to estimate the "mass distribution" P so that, with confidence at least $1 - 10^{-4}$, the estimation error of the probability is at most $\varepsilon = 1\%$ *at all the regions simultaneously.*

Problem 2 (finitely many classifiers).
Prove the following

Theorem 5.11.1 *Let $\mathcal{F}$ be a family of classifiers, parameterized by $c \in C$, where C* is a finite of $\mathbb{R}^p$*, namely $|C| = K$. Suppose that $(U_1, Y_1), \cdots, (U_N, Y_N)$ are independent and identically distributed, where U_i has continuous distribution $F(u)$ and $Y_i \in \{0, 1\}$. Define $\bar{J}(c)$, $\hat{J}_N(c)$, $\bar{c}$, and $\hat{c}_N$ as usual. Now the points $\hat{c}_N, \bar{c} \in C$ trivially exist, since C is finite; assume that $\bar{c}$ is unique. Then:*

1. *almost surely, $\hat{c}_N \to \bar{c}$;*

2. *for fixed $\varepsilon > 0$ and N, it holds*

$$\mathsf{P}\left[\max_{c \in C} |\hat{J}_N(c) - \bar{J}(c)| \geq \varepsilon\right] \leq 2Ke^{-2N\varepsilon^2}.$$

6 The LSCR method

6.1 Introduction and motivation

Suppose that a sample $\{y_1, \cdots, y_N\}$ of independent and identically distributed random variables is drawn from a Gaussian distribution $\mathcal{N}(\theta^o, \sigma^2)$ whose mean θ^o is not known. The goal of this chapter (and the leitmotif of this course) is to extract information about θ^o from the data.
In the spirit of Chapter 1, the sample can be seen as N measures

$$y_i = \theta^o + \varepsilon_i, \quad i = 1, \cdots, N,$$

to be "explained" in terms of the "true" parameter θ^o, corrupted by some errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. The approach of Chapter 1 would be to provide a *point estimate* of θ^o minimizing some cost criterion (the sum of the squared residuals); as we know, the estimate provided by the least squares method is

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

(we also know that this estimate is unbiased, consistent, and a lot of other stuff).
However, such an estimate, for finite N, comes with no guarantee at all. The only thing that can be said from the probabilistic standpoint, *on the sole basis of* $\hat{\theta}$, is that $\mathsf{P}[\hat{\theta} = \theta^o] = 0$, since the distribution of the $\{y_i\}$ is continuous, and this is not valuable information. Instead, a "guarantee" on the usefulness of $\hat{\theta}$ should be a "certificate" that $\hat{\theta}$ and θ^o are *probably* close. In parametric statistics such "certificate" usually comes in two forms, that in the present case are strictly related to each other:

- the variance of $\hat{\theta}$. Since $\mathsf{E}[\hat{\theta}] = \theta^o$, the smaller its variance, the closer it is, on average, to θ^o;

- a statement like "the probability $\mathsf{P}[|\hat{\theta} - \theta^o| \geq d]$ is α", for a certain threshold d and significance α, say $\alpha = 5\% = 0.05$. If α is fixed, a smaller threshold d implies that with a fixed probability the two quantities are closer (this is the usual way to pose the problem). If, on the other hand, d is fixed, a smaller significance α implies that the two quantities are close with higher probability.

How can we provide such "certificates"?

6.2 Confidence intervals, the standard way

If σ^2 is known, there is no problem. Indeed,

$$\begin{aligned}\mathsf{Var}\left[\hat{\theta}\right] &= \mathsf{E}\left[(\hat{\theta}-\theta^o)^2\right] \\ &= \mathsf{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i\right)^2\right] = \frac{1}{N^2}\mathsf{E}\left[\left(\sum_{i=1}^{N}\varepsilon_i\right)^2\right] = \frac{1}{N^2}\sum_{i=1}^{N}\mathsf{E}\left[\varepsilon_i^2\right] \\ &= \frac{\sigma^2}{N}.\end{aligned}$$

Moreover, since a sum of independent Gaussian variables is itself Gaussian, it follows that $\hat{\theta} \sim N\left(\theta^o, \frac{\sigma^2}{N}\right)$. Hence, the random variable

$$z := \frac{\hat{\theta}-\theta^o}{\sigma/\sqrt{N}}$$

is $\mathcal{N}(0,1)$, i.e. with density $\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$. Now, in statistical books or using any statistical software, we can find a percentile z_α such that

$$\mathsf{P}\left[\left|\frac{\hat{\theta}-\theta^o}{\sigma/\sqrt{N}}\right| \geq z_\alpha\right] = \mathsf{P}\left[|z| \geq z_\alpha\right] = 1 - \int_{-z_\alpha}^{z_\alpha}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dz = \alpha.$$

But then also

$$\mathsf{P}\left[\left|\hat{\theta}-\theta^o\right| \geq \frac{\sigma z_\alpha}{\sqrt{N}}\right] = \alpha,$$

hence $d = \frac{\sigma z_\alpha}{\sqrt{N}}$ provides a "certificate" of the second kind. Another way to say the same thing is that the interval

$$I_\alpha := \left[\hat{\theta} - \frac{\sigma z_\alpha}{\sqrt{N}},\ \hat{\theta} + \frac{\sigma z_\alpha}{\sqrt{N}}\right]$$

contains θ^o with *probability* $1-\alpha$. We should be careful, here, about the use of the word "probability". *Before the sample is drawn*, the interval I_α is itself random, and whether or not it will contain θ^o depends on the outcome of the experiment; namely, it will happen with *probability* $1-\alpha$. But *after the sample has been drawn* the interval becomes deterministic. Either it contains θ^o or it does not; in other words, the probability that $\theta^o \in I_\alpha$ is either 1 or 0. Thus, $1-\alpha$ is not anymore the "probability" of anything; the usual name for it is *confidence*. Since it aims at what can be said *after* a sample is drawn, the notion of confidence is often introduced from the "frequentist" point of view; in this perspective, it means more or less the following: if we run $M \gg 0$ experiments in parallel, that is, we draw M samples

$$\{y_1^{(1)},\cdots,y_N^{(1)}\},\{y_1^{(2)},\cdots,y_N^{(2)}\},\cdots,\{y_1^{(M)},\cdots,y_N^{(M)}\},$$

and we construct the respective intervals

$$I_\alpha^{(1)}, I_\alpha^{(2)}, \cdots, I_\alpha^{(M)},$$

then

$$\frac{\text{number of intervals that contain } \theta^o}{M} \simeq 1 - \alpha.$$

The above reasoning can be resumed as follows:

Lemma 6.2.1 *Suppose that $\{y_1, \cdots, y_N\}$ are independent and identically distributed random variables drawn from a Gaussian distribution $\mathcal{N}(\theta^o, \sigma^2)$ whose variance σ^2 is known but whose mean θ^o is not. Let $\hat{\theta} = \frac{1}{N}\sum_{i=1}^N y_i$. Then*

$$I_\alpha = \left[\hat{\theta} - \frac{\sigma z_\alpha}{\sqrt{N}},\ \hat{\theta} + \frac{\sigma z_\alpha}{\sqrt{N}}\right]$$

is a $(1-\alpha)$-confidence interval *for θ^o.*

For example, if $\alpha = 5\%$, this means that if we re-sampled and computed the interval a lot of times under the same assumptions, the interval would contain the mean about 95% of the times.
However, this is not really a useful result. The problem with the above construction is that it gives for granted that we know the variance σ^2, whereas in general this is not the case. There is a standard way, though, to construct a $(1-\alpha)$-confidence interval for θ^o, which does *not* rely on the knowledge of any parameter (σ^2). Intuitively, since *less* knowledge is assumed, the result will be less precise (and notably so for little N), i.e. the confidence interval will be larger.

The trick is, of course, to estimate σ^2 from the dispersion of the data; the following is the standard estimator, called the *sample variance* of the $\{y_i\}$:

$$\bar{s}^2 = \frac{1}{N-1}\sum_{i=1}^N (y_i - \hat{\theta})^2$$

It can be shown that $\bar{s}^2$ is a consistent estimator of σ^2 (that is, it converges to σ^2 almost surely as $N \to \infty$). Note that the sum is weighted with $\frac{1}{N-1}$, not $\frac{1}{N}$ as one would expect. In this way, $\bar{s}^2$ becomes also an *unbiased* estimator (i.e. $\mathsf{E}[\bar{s}^2] = \sigma^2$).
Let us recall here some results which are also stated in the Appendix:

Lemma 6.2.2 *Let $y_1, \cdots, y_N \sim \mathcal{N}(0,1)$ be independent random variables. Then*

$$\sum_{i=1}^N y_i^2 \sim \chi^2(N)$$

In words: a sum of N squared independent "standard" ($\mathcal{N}(0,1)$) Gaussian variables has a chi-square distribution with N degrees of freedom. A chi-square is a continuous random variable taking values in $[0, \infty)$.

Theorem 6.2.1 *Let $y_1, \cdots, y_N \sim \mathcal{N}(\theta^o, \sigma^2)$ be independent random variables. Then*

- $\hat{\theta} \sim \mathcal{N}(\theta^o, \frac{\sigma^2}{N})$;
- $(N-1)\frac{\bar{s}^2}{\sigma^2} \sim \chi^2(N-1)$;
- $\hat{\theta}$ *and* $\bar{s}^2$ *are independent.*

The reason why $(N-1)\frac{\bar{s}^2}{\sigma^2}$ has $N-1$ degrees of freedom is because it happens to be the sum of N terms of the form $\left(\frac{y_i - \hat{\theta}}{\sigma}\right)^2$; this is more or less what happens in Lemma 6.2.2, but in this case the terms are not independent: the sum has indeed "one less degree of freedom", namely due to the constraint that links $y_1, \cdots, y_N$ to the value of $\hat{\theta}$.
On the other hand, that $\hat{\theta}$ and $\bar{s}^2$ are independent is crucial, because this allows us to apply the following

Theorem 6.2.2 *Let $Z \sim \mathcal{N}(0,1)$, $V \sim \chi^2(n)$ be independent random variables. Then*

$$T = \frac{Z}{\sqrt{V/n}} \sim t(n)$$

In words, the ratio of two independent variables, one a standardized Gaussian, and the other the square root of a chi-square, divided by its own degrees of freedom, is distributed as a Student's t with the same number of degrees of freedom.

The "Student's" distribution t, and the test of hypothesis named "Student's" t-test based on it were found and published in 1908 by William S. Gosset, a statistician working for the Guinness brewery in Dublin. Since the brewery management did not want the concurrence to know that the company was running statistical tests on its products to ensure quality, Gosset published its result somewhat anonymously, under the pseudonym "Student". The shape of a $t(n)$ density resembles that of a Gaussian $\mathcal{N}(0,1)$, namely it is symmetric around 0 and is close to the Gaussian one, but it has *fatter tails.* As $n \to \infty$, the $t(n)$ density converges to the $\mathcal{N}(0,1)$ density.
An immediate consequence of Theorem 6.2.2 is the following:

Theorem 6.2.3 *Let $y_1, \cdots, y_N \sim \mathcal{N}(\theta^o, \sigma^2)$ be independent random variables. Then*

$$T := \frac{\hat{\theta} - \theta^o}{\bar{s}/\sqrt{N}} \sim t(N-1)$$

Proof.

$$T = \frac{\hat{\theta} - \theta^o}{\bar{s}/\sqrt{N}} = \frac{\frac{\hat{\theta}-\theta^o}{\sigma/\sqrt{N}}}{\frac{\bar{s}}{\sigma}} = \frac{\frac{\hat{\theta}-\theta^o}{\sigma/\sqrt{N}}}{\sqrt{\frac{(N-1)\bar{s}^2}{\sigma^2}/(N-1)}} = \frac{Z}{\sqrt{V/(N-1)}}$$

The last quantity is the ratio of a standardized Gaussian and the square root of a chi-square divided by its own degrees of freedom, independent of each other. The claim follows by Theorem 6.2.2. □

Theorem 6.2.3 is a powerful tool in inferential statistics: it says that the distribution of the statistic T, depending on a normal sample $\{y_1, \cdots, y_N\}$ and containing as much information about θ^o as the sample itself, has a distribution $t(N-1)$ that depends only on the size N of the sample. Since the cumulative distribution of a $t(N-1)$ variable, and its inverse, are tabulated in books and available in every statistical software, one can use T to make inferences about θ^o.
Any statistic of the data, which does not depend on the parameters of their distribution, but only on their number, is called a "pivotal" statistic. Here, $t(N-1)$ is indeed a "pivotal" distribution, that allows to make inferences about θ^o without any knowledge on σ^2.

For example, suppose that we have reasons to believe that $\theta^o \neq 0$, and that we want to provide a statistical justification to this claim. Making the opposite hypothesis $\theta^o = 0$ (called, in statistical jargon, the *null hypothesis*), the statistic T becomes

$$T = \frac{\hat{\theta}}{\bar{s}/\sqrt{N}} \sim t(N-1).$$

We find on books or through software, the percentile t_α such that

$$\int_{-t_\alpha}^{t_\alpha} f(x)\ dx = 1 - \alpha,$$

where $f(x)$ is the density of a $t(N-1)$ random variable. Now we compute T for the sample at hand, and if $|T| > t_\alpha$, we can "reject the hypothesis that $\theta^o = 0$, with confidence $1-\alpha$". In other words, if α is, say, 5%, we can affirm that "the mean is not 0", and if we repeat the whole procedure a lot of times, we will be wrong in only 5% of the cases. This is a typical Student's t-test of hypothesis.

Suppose, on the other hand, that we do not make hypotheses on θ^o. Since, anyway, $\mathsf{P}\left[|T| \leq t_\alpha\right] = 1 - \alpha$, with probability $1-\alpha$ it still holds

$$-t_\alpha \leq \frac{\hat{\theta} - \theta^o}{\bar{s}/\sqrt{N}} \leq t_\alpha$$

$$\hat{\theta} - t_\alpha \frac{\bar{s}}{\sqrt{N}} \leq \theta^o \leq \hat{\theta} + t_\alpha \frac{\bar{s}}{\sqrt{N}}$$

In other terms, the *random* interval

$$I_\alpha = \left[\hat{\theta} - t_\alpha \frac{\bar{s}}{\sqrt{N}},\ \hat{\theta} + t_\alpha \frac{\bar{s}}{\sqrt{N}}\right]$$

contains θ^o with probability $1-\alpha$, *before the sample is drawn.* And *after the sample has been drawn*, we have finally

Theorem 6.2.4 *Suppose that $\{y_1, \cdots, y_N\}$ are independent and identically distributed random variables drawn from a Gaussian distribution $\mathcal{N}(\theta^o, \sigma^2)$ whose parameters θ^o and σ^2 are not known. Let $\hat{\theta} = \frac{1}{N}\sum_{i=1}^N y_i$ and $\bar{s}^2 = \frac{1}{N-1}\sum_{i=1}^N (y_i - \hat{\theta})^2$. Then*

$$I_\alpha = \left[\hat{\theta} - \frac{\bar{s}t_\alpha}{\sqrt{N}},\ \hat{\theta} + \frac{\bar{s}t_\alpha}{\sqrt{N}}\right]$$

is a $(1-\alpha)$-confidence interval *for θ^o.*

We have now a beautiful "pivotal" result, very popular in applied statistics, which extracts all the possible information from a sample to yield a certified, reliable confidence interval for a parameter, and does not assume the knowledge of the other one. In our search for results with the least possible hypotheses on the data, Theorem 6.2.4 is definitely in the right direction. It has, still, a drawback, that has been left implicit from the beginning of this chapter:

it still depends crucially on the Gaussianity of data.

Theorem 6.2.4 is still based on the knowledge that the "true" distribution belongs to a precise family, parameterized by two numbers θ^o, σ^2. Such result is not *distribution free.*

Now that you are familiar with the concept of confidence interval, you are ready for the main question of this chapter. Suppose that the measures

$$y_i = \theta^o + \varepsilon_i, \quad i = 1, \cdots, N,$$

are "explained" in terms of the "true" parameter θ^o, corrupted by some errors ε_i, which are independent random variables drawn from a continuous

distribution. Is it possible to provide a confidence interval for θ^o without restricting the distribution of the errors to some parametric family?
A positive answer has been provided in recent years by a clever method named LSCR (*Leave-out Sign-dominant Correlation Regions*) due to Marco Campi and Erik Weyer, under the only further hypothesis, quite reasonable, that the density of ε_i is symmetric around 0. We will approach the method by means of an example; its starting point may be quite surprising.

6.3 Groups

Definition 6.3.1 *A* group $(G, *)$ *is a set* G *endowed with an operation* $*$*, i.e. a function* $* : G \times G \to G$ *(whose values are usually denoted* $a * b$ *in place of* $*(a, b)$*), such that the following properties hold:*

- *for all* $a, b, c \in G$*,* $((a * b) * c) = (a * (b * c))$ *(associativity);*
- *there exist an element* $e \in G$*, called the identity element, such that* $e * a = a * e = a$ *for all* $a \in G$*;*
- *for all* $a \in G$*, there exist an element* $a^{-1} \in G$ *such that* $a * a^{-1} = a^{-1} * a = e$*; such* a^{-1} *is called the* inverse *of* a*.*

You already know some groups; for example:

- the set of invertible functions of a set S onto itself, where $*$ denotes composition of functions (i.e. $f * g$ is the function such that $f * g(s) = f(g(s))$ for all $s \in S$), and e is the identity function $e(s) = s$;
- the symmetries of a regular polygon, that is, those rigid movements of the plane (rotations, reflections) that map the polygon onto itself; again, $*$ denotes composition (applying one movement *after* another), and e denotes “no movement”;
- the set of all the invertible matrices in $\mathbb{R}^{n\times n}$, where $*$ is matrix multiplication, $e = I$, and a^{-1} is the inverse matrix;
- the set of *sequences of moves* that can be done on a Rubik cube, where $*$ means applying a sequence of moves *after* another, and e is “make no move”.

In none of the above examples it happens, in general, that any two elements commute (that is, $a * b = b * a$).

Definition 6.3.2 *A group* $(G, *)$ *in which the following further property holds:*

- *for all* $a, b \in G$*,* $a * b = b * a$ *(commutativity)* *is called a* commutative*, or* Abelian *group.*

In Abelian groups, usually, the operation $*$ is denoted $+$ (plus), the identity element is denoted 0 (zero), and the inverse of a is denoted $-a$. Of course you know *a lot* of Abelian groups: $(\mathbb{Z},+)$, $(\mathbb{Q},+)$, $(\mathbb{R},+)$, $(\mathbb{C},+)$, $(\mathbb{R}^n,+)$, $(\mathbb{R}^{m\times n},+)$, etc.; any vector space, in particular, must by definition be an Abelian group with respect to its own $+$ operation.

Definition 6.3.3 *A* subgroup *of* $(G,*)$ *is a set* $H \subseteq G$ *which is closed with respect to the operation* $*$ *(if* $a,b \in H$*, then also* $a*b \in H$*). In other words, a subgroup of* $(G,*)$ *is a pair* $(H,*)$*, where* $H \subseteq G$ *and* $*$ *is the same operation restricted to* H*, such that* $(H,*)$ *is a group in its own right.*

Any subgroup of $(G,*)$ has the same identity element of $(G,*)$. For example, $(\mathbb{Q},+)$ is a subgroup of $(\mathbb{R},+)$, and has the same identity element 0.
Here is one of the simplest possible *finite* Abelian groups:

$$B = (\{\circ,\bullet\},+),$$

where addition is defined in the following way:

$+$	$\circ$	$\bullet$
$\circ$	$\circ$	$\bullet$
$\bullet$	$\bullet$	$\circ$

(its identity element is $\circ$, and the inverse of each element is the element itself; it is nothing else than addition of digits in binary, i.e. modulo 2, arithmetics). And here follows a group that will serve us for our "canonical" example:

$G_7 =$

	1	2	3	4	5	6	7
I_1	$\bullet$	$\bullet$	$\circ$	$\bullet$	$\bullet$	$\circ$	$\circ$
I_2	$\bullet$	$\circ$	$\bullet$	$\bullet$	$\circ$	$\bullet$	$\circ$
I_3	$\circ$	$\bullet$	$\bullet$	$\circ$	$\bullet$	$\bullet$	$\circ$
I_4	$\bullet$	$\bullet$	$\circ$	$\circ$	$\circ$	$\bullet$	$\bullet$
I_5	$\bullet$	$\circ$	$\bullet$	$\circ$	$\bullet$	$\circ$	$\bullet$
I_6	$\circ$	$\bullet$	$\bullet$	$\bullet$	$\circ$	$\circ$	$\bullet$
I_7	$\circ$	$\circ$	$\circ$	$\bullet$	$\bullet$	$\bullet$	$\bullet$
I_8	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$	$\circ$

It is a group with 8 elements, namely the rows $I_1,\cdots,I_8$ of the above table; addition is defined component-wise[25], for example

$$\begin{aligned} I_1 + I_2 &= \begin{bmatrix} \bullet & \bullet & \circ & \bullet & \bullet & \circ & \circ \end{bmatrix} \\ &+ \begin{bmatrix} \bullet & \circ & \bullet & \bullet & \circ & \bullet & \circ \end{bmatrix} \\ &= \begin{bmatrix} \circ & \bullet & \bullet & \circ & \bullet & \bullet & \circ \end{bmatrix} = I_3, \end{aligned}$$

[25] Think at the bitwise XOR between two strings of 7 bits.

It is, in particular, a subgroup of $(B^7, +)$, the set of all the $2^7 = 128$ possible 7-tuples of elements in $\{\circ, \bullet\}$ with component-wise addition. Any two rows in the above table have another row as their sum; the identity element is I_8, and the inverse of a certain row is the row itself.

6.4 Confidence intervals revisited

Now comes our prototypical example. We suppose that we are given 7 measures (very few, indeed)

$$y_i = \theta^o + \varepsilon_i, \quad i = 1, \cdots, 7.$$

We suppose that $\{\varepsilon_i\}$ are continuous variables with a density centered around zero (with mean zero), independent *but not necessarily identically distributed.* The standard way to estimate θ^o would be to compute

$$\hat{\theta} = \frac{1}{7}\sum_{i=1}^{7} y_i,$$

but here we are *not* interested in an estimate; our aim is a confidence interval. Consider, for each measure, an affine function in the (θ, z) plane:

$$f_i(\theta) = y_i - \theta.$$

It intersects the θ-axis at exactly one point. Now, for each of the first seven elements, $I_i = I_1, \cdots, I_7$, of the group in section 6.3, consider a point-wise average of 4 such functions, namely the 4 ones whose indices are marked $\bullet$ in the corresponding row:

$$\begin{aligned}
g_1(\theta) &= \frac{1}{4}\sum_{k\in\{1,2,4,5\}} f_k(\theta) \\
&= \frac{1}{4}\sum_{k\in\{1,2,4,5\}} y_k - \theta = \frac{1}{4}\sum_{k\in\{1,2,4,5\}} \theta^o + \varepsilon_k - \theta = (\theta^o - \theta) + \frac{1}{4}\sum_{k\in\{1,2,4,5\}} \varepsilon_k; \\
&\ \vdots \\
g_i(\theta) &= \frac{1}{4}\sum_{I_i[k]=\bullet} f_k(\theta) = (\theta^o - \theta) + \frac{1}{4}\sum_{I_i[k]=\bullet} \varepsilon_k; \\
&\ \vdots \\
g_7(\theta) &= \frac{1}{4}\sum_{k\in\{4,5,6,7\}} f_k(\theta) = (\theta^o - \theta) + \frac{1}{4}\sum_{k\in\{4,5,6,7\}} \varepsilon_k;
\end{aligned}$$

The average corresponding to I_8 is trivial, namely

$$g_8(\theta) = \frac{1}{4}\sum_{I_8[k]=\bullet} f_k(\theta) = \frac{1}{4}\sum_{k\in\varnothing} f_k(\theta) = 0.$$

Each of the $g_i(\cdot)$ $i = 1, \cdots, 7$, has exactly one intersection with the θ-axis; let us call it θ_i:

$$\theta_i := \text{the only } \theta \text{ such that } g_i(\theta) = 0.$$

The θ_i are *random* points coming in some order on the θ-axis; since the errors ε_i are independent and have continuous distributions, almost surely no two of the intersections θ_i coincide, and none of them equals θ^o. Furthermore, let $\bar{\theta}_1, \cdots, \bar{\theta}_7$ denote the same seven points, but *sorted*[26], i.e. such that $\bar{\theta}_1 < \bar{\theta}_2 < \cdots < \bar{\theta}_7$. These points split the θ-axis in 8 intervals (the outermost ones are semi-infinite, namely $(-\infty, \bar{\theta}_1)$ and $(\bar{\theta}_7, \infty)$).
Now note that at $\theta = \theta^o$ it holds

$$g_1(\theta^o) = \frac{1}{4} \sum_{k \in \{1,2,4,5\}} \varepsilon_k,$$
$$\vdots$$
$$g_7(\theta^o) = \frac{1}{4} \sum_{k \in \{4,5,6,7\}} \varepsilon_k.$$

Each average on the right-hand side is a sum of independent variables, *with density symmetric around* 0. *Hence, the average has also a density symmetric around* 0, and any $g_i(\theta^o)$ has equal probabilities of lying above or below the θ-axis, i.e. positive or negative sign. However, any $g_i(\theta^o)$ is averaging 4 noise variables, thus is intuitively more concentrated towards zero than them.
The ingenious idea of the LSCR method is to compare the signs of each $g_i(\theta^o)$, depending on *where* is θ^o, namely in which one of the 8 intervals. If $\theta^o \in (-\infty, \bar{\theta}_1)$, the leftmost interval, then $g_1(\theta^o) > 0, g_2(\theta^o) > 0, \cdots, g_7(\theta^o) > 0$; intuitively, that all the $g_i(\theta^o)$ have positive sign has little probability (but what probability, indeed?). On the other hand if $\theta^o \in (\bar{\theta}_7, \infty)$, the rightmost interval, then $g_1(\theta^o) < 0, g_2(\theta^o) < 0, \cdots, g_7(\theta^o) < 0$. If θ^o belongs to the second interval, $(\bar{\theta}_1, \bar{\theta}_2)$, then one and only one among $g_1(\theta^o), g_2(\theta^o), \cdots, g_7(\theta^o)$ has negative sign, etc.

Here is the point: *what is the probability of these events*?
The first one ($\theta^o \in (-\infty, \bar{\theta}_1)$, hence $g_1(\theta^o) > 0, \cdots, g_7(\theta^o) > 0$) happens

[26] The random variables $\bar{\theta}_1, \cdots, \bar{\theta}_7$ are the so-called *order statistics* of $\theta_1, \cdots, \theta_7$.

exactly when

$$\begin{array}{cccccccl}
\varepsilon_1 & +\varepsilon_2 & & +\varepsilon_4 & +\varepsilon_5 & & & > 0, \\
\varepsilon_1 & & +\varepsilon_3 & +\varepsilon_4 & & +\varepsilon_6 & & > 0, \\
 & \varepsilon_2 & +\varepsilon_3 & & +\varepsilon_5 & +\varepsilon_6 & & > 0, \\
\varepsilon_1 & +\varepsilon_2 & & & & +\varepsilon_6 & +\varepsilon_7 & > 0, \\
\varepsilon_1 & & +\varepsilon_3 & & +\varepsilon_5 & & +\varepsilon_7 & > 0, \\
 & \varepsilon_2 & +\varepsilon_3 & +\varepsilon_4 & & & +\varepsilon_7 & > 0, \\
 & & & \varepsilon_4 & +\varepsilon_5 & +\varepsilon_6 & +\varepsilon_7 & > 0.
\end{array} \tag{38}$$

If $\theta^o \in (\bar{\theta}_1, \bar{\theta}_2)$, then exactly one $g_i(\theta^o)$ among $g_1(\theta^o), \cdots, g_7(\theta^o)$ has negative sign, hence all the other values at θ^o (including $g_8(\theta^o) = 0$) are greater than it. Suppose, to fix ideas, that the incriminated g_i is g_1. Then

$$\begin{array}{cccccccccccc}
 & & & 0 & & & & > & \varepsilon_1 & +\varepsilon_2 & +\varepsilon_4 & +\varepsilon_5, \\
\varepsilon_1 & & +\varepsilon_3 & +\varepsilon_4 & & +\varepsilon_6 & & > & \varepsilon_1 & +\varepsilon_2 & +\varepsilon_4 & +\varepsilon_5, \\
 & \varepsilon_2 & +\varepsilon_3 & & +\varepsilon_5 & +\varepsilon_6 & & > & \varepsilon_1 & +\varepsilon_2 & +\varepsilon_4 & +\varepsilon_5, \\
\varepsilon_1 & +\varepsilon_2 & & & & +\varepsilon_6 & +\varepsilon_7 & > & \varepsilon_1 & +\varepsilon_2 & +\varepsilon_4 & +\varepsilon_5, \\
\varepsilon_1 & & +\varepsilon_3 & & +\varepsilon_5 & & +\varepsilon_7 & > & \varepsilon_1 & +\varepsilon_2 & +\varepsilon_4 & +\varepsilon_5, \\
 & \varepsilon_2 & +\varepsilon_3 & +\varepsilon_4 & & & +\varepsilon_7 & > & \varepsilon_1 & +\varepsilon_2 & +\varepsilon_4 & +\varepsilon_5, \\
 & & & \varepsilon_4 & +\varepsilon_5 & +\varepsilon_6 & +\varepsilon_7 & > & \varepsilon_1 & +\varepsilon_2 & +\varepsilon_4 & +\varepsilon_5.
\end{array}$$

Simplifying and bringing every term to the left-hand side we obtain

$$\begin{array}{cccccccl}
-\varepsilon_1 & -\varepsilon_2 & & -\varepsilon_4 & -\varepsilon_5 & & & > 0, \\
 & -\varepsilon_2 & +\varepsilon_3 & & -\varepsilon_5 & +\varepsilon_6 & & > 0, \\
-\varepsilon_1 & & +\varepsilon_3 & -\varepsilon_4 & & +\varepsilon_6 & & > 0, \\
 & & & -\varepsilon_4 & -\varepsilon_5 & +\varepsilon_6 & +\varepsilon_7 & > 0, \\
 & -\varepsilon_2 & +\varepsilon_3 & -\varepsilon_4 & & & +\varepsilon_7 & > 0, \\
-\varepsilon_1 & & +\varepsilon_3 & & -\varepsilon_5 & & +\varepsilon_7 & > 0, \\
-\varepsilon_1 & -\varepsilon_2 & & & & +\varepsilon_6 & +\varepsilon_7 & > 0.
\end{array} \tag{39}$$

And now you see that the sets of inequalities (38) and (39) are very similar, the only differences being that the inequalities appear in different orders, and that some of the signs are changed. That the sets of inequalities contain terms with the same indices happens precisely *due to the group structure according to which the averages have been built.* Moreover, comparing the corresponding inequalities, say the first ones, you see that events like

$$\{\varepsilon_1 + \varepsilon_2 + \varepsilon_4 + \varepsilon_5 > 0\},$$
$$\{-\varepsilon_1 - \varepsilon_2 - \varepsilon_4 - \varepsilon_5 > 0\}$$

happen exactly with the same probability, *because all the ε_i have symmetric densities.* The same reasoning applies to the entire sets of inequalities. Hence

the probabilities that $\theta^o \in (-\infty, \bar{\theta}_1)$ and that $\theta^o \in (\bar{\theta}_1, \bar{\theta}_2)$ are equal.

But the very same procedure could have been applied to any of the intervals; hence θ^o belongs to any of the intervals with the same probability; the intervals are 8, therefore such probability is $\frac{1}{8}$. And finally, the probability that θ^o belongs to any of the two outermost intervals is $\frac{1}{8} + \frac{1}{8} = \frac{1}{4}$, hence the probability that the converse happens is $\frac{3}{4}$. What we have shown is an informal proof of the following

Theorem 6.4.1 *(Campi, Weyer). Let*

$$y_i = \theta^o + \varepsilon_i, \quad i = 1, \cdots, 7;$$

suppose that $\{\varepsilon_i\}$ are continuous variables with a density centered around zero (with mean zero), independent but not necessarily identically distributed. Construct the functions $g_1(\cdot), \cdots, g_7(\cdot)$ and sort their intersections with the θ-axis, $\bar{\theta}_1 < \cdots < \bar{\theta}_7$, as above. Then

$$I := \left[\bar{\theta}_1,\ \bar{\theta}_7\right]$$

is a $\frac{3}{4} = 75\%$-confidence interval for θ^o.

Some remarks are in order. The applicability of the method is of course not limited to 7 measures. The case for 7 measures was presented because it is sufficiently instructive, yet it does not involve overwhelming computations; with more measures, the proof transports without significant changes. The only real issue with the general, N-measures case is how is one supposed to construct a suitable subgroup of $(B^N, +)$. Such group should be "balanced" in the sense that it should be fairly small (i.e. its cardinality should not grow exponentially with N), that the number of bullets (●) should be the same in each row, and it should be approximately half the size of a row. This construction is actually easy when $N = 2^n - 1$ for some n. For $N = 3 = 2^2 - 1$ we have

$G_3 =$

	1	2	3
I_1	●	●	○
I_2	●	○	●
I_3	○	●	●
I_4	○	○	○

For $N = 7 = 2^3 - 1$ we have the group of the above example:

$G_7 =$

	1	2	3	4	5	6	7
I_1	●	●	○	●	●	○	○
I_2	●	○	●	●	○	●	○
I_3	○	●	●	○	●	●	○
I_4	●	●	○	○	○	●	●
I_5	●	○	●	○	●	○	●
I_6	○	●	●	●	○	○	●
I_7	○	○	○	●	●	●	●
I_8	○	○	○	○	○	○	○

$=$

<table>
<tr><th></th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th></tr>
<tr><td>I_1</td><td colspan="3" rowspan="3">G_3</td><td colspan="3" rowspan="3">G_3</td><td>○</td></tr>
<tr><td>I_2</td><td>○</td></tr>
<tr><td>I_3</td><td>○</td></tr>
<tr><td>I_4</td><td colspan="3" rowspan="3">G_3</td><td colspan="3" rowspan="3">$\bar{G}_3$</td><td>●</td></tr>
<tr><td>I_5</td><td>●</td></tr>
<tr><td>I_6</td><td>●</td></tr>
<tr><td>I_7</td><td>○</td><td>○</td><td>○</td><td>●</td><td>●</td><td>●</td><td>●</td></tr>
<tr><td>I_8</td><td>○</td><td>○</td><td>○</td><td>○</td><td>○</td><td>○</td><td>○</td></tr>
</table>

where $\bar{G}_3$ is the complement of the matrix representing G_3 ($\bullet \leftrightarrow \circ$). Balanced groups for the next powers of 2 can be constructed recursively in this fashion; you can find the precise construction procedure in the appendix of [7] or, for a deeper discussion, in the paper [13] cited therein.

Note that, as N increases and the number of measures in each average is kept approximately equal to $\frac{N}{2}$, as happens in the above construction, two things happen:

- the number of intervals remains equal to $N+1$; hence, all of them having the same probability $\frac{1}{N+1}$, some more intervals can be discarded, other than the two outermost ones, in order to obtain a smaller interval (with of course a smaller confidence). In this way, the confidence is "tunable", although not continuously as in the Gaussian/Student case. The fact that the confidence interval is obtained discarding the outermost intervals where the signs of the functions $g_i(\theta)$ are equal, or anyway a set of intervals where a vast majority of them has the same sign, explains three letters in the acronym LSCR: *Leave-out Sign-dominant … Regions.*

- Since approximately $\frac{N}{2}$ noise terms are averaged in each intersection, we can expect the final confidence interval to shrink towards θ^o as N increases; under fairly general conditions, it could be shown that the extremes of the interval converge to θ^o almost surely as $N \to \infty$; for example, if we make the further assumption that the ε_i are identically distributed, this is a simple consequence of the strong law of large numbers.

6.5 The case with inputs

The construction of a confidence interval for θ^o can be extended to the case where an input (i.e. an "explanatory variable") is involved. Let

$$y_i = \theta^o u_i + \varepsilon_i, \quad i = 1, \cdots, 7.$$

As usual, we suppose that $\{\varepsilon_i\}$ are independent continuous variables with a density symmetric around zero. The inputs u_i can be deterministic or random, but if we want the *same* construction of the previous section to be useful, and the confidence interval to "shrink" towards θ^o for big N, we should pretend, in some sense, that they "stay away from zero":

- if they are deterministic, they are nonzero, and they do not tend to zero as $N \to \infty$;

- if they are random (to simplify, suppose also identically distributed), they are independent of ε_i, and they do *not* have zero mean.

This said, we can develop the theory of section 6.4 again, with minor changes. Consider, for each measure, an affine function:

$$f_i(\theta) = y_i - \theta u_i.$$

Differently from before, these functions may have different slopes. But exactly as before, they intersect the θ-axis at one point (*unless* $u_i = 0$, *but we have assumed above that this is not the case*). For each $I_i = I_1, \cdots, I_7$ consider an average of 4 different f_k, according to the group structure:

$$\begin{aligned}
g_1(\theta) &= \frac{1}{4} \sum_{k \in \{1,2,4,5\}} f_k(\theta) \\
&= \frac{1}{4} \sum_{k \in \{1,2,4,5\}} y_k - \theta u_k = \frac{1}{4} \sum_{k \in \{1,2,4,5\}} \theta^o u_k + \varepsilon_k - \theta u_k \\
&= \frac{\theta^o - \theta}{4} \sum_{k \in \{1,2,4,5\}} u_k + \frac{1}{4} \sum_{k \in \{1,2,4,5\}} \varepsilon_k; \\
&\vdots \\
g_7(\theta) &= \frac{1}{4} \sum_{k \in \{4,5,6,7\}} f_k(\theta) \\
&= \frac{\theta^o - \theta}{4} \sum_{k \in \{4,5,6,7\}} u_k + \frac{1}{4} \sum_{k \in \{4,5,6,7\}} \varepsilon_k;
\end{aligned}$$

These are also affine functions with different slopes, intersecting the θ-axis at one point. Then we can build $\bar{\theta}_1 < \cdots < \bar{\theta}_7$ as before, and since

$$\begin{aligned}
g_1(\theta^o) &= \frac{1}{4} \sum_{k \in \{1,2,4,5\}} \varepsilon_k, \\
&\vdots \\
g_7(\theta^o) &= \frac{1}{4} \sum_{k \in \{4,5,6,7\}} \varepsilon_k,
\end{aligned}$$

the reasoning of Section 6.4 applies without any other change. $\left[\bar{\theta}_1,\ \bar{\theta}_7\right]$ is a 75%-confidence interval for θ^o.

As already mentioned, this result is useful because the inputs u_i "stay away from zero". Now, what happens if the inputs $\{u_i\}$ are random, independent of the $\{\varepsilon_i\}$, but their mean is indeed 0? Consider the intersection of (say)

$g_1(\theta)$ with the θ-axis. It is the point θ_1 such that

$$
\begin{aligned}
0 = g_1(\theta_1) &= (\theta^o - \theta_1)\left(\frac{1}{4}\sum_{k\in\{1,2,4,5\}} u_k\right) + \frac{1}{4}\sum_{k\in\{1,2,4,5\}} \varepsilon_k; \\
\theta_1 &= \theta^o + \frac{\frac{1}{4}\sum_{k\in\{1,2,4,5\}}\varepsilon_k}{\frac{1}{4}\sum_{k\in\{1,2,4,5\}}u_k}.
\end{aligned}
$$

The issue with this case is evident: as more measures (and bigger groups) come into the picture, the term $\frac{1}{(N+1)/2}\sum_k u_k$ at the denominator of the last expression, which is also the coefficient of θ in the expression of $g_1(\theta)$, tends to 0. The same phenomenon happens for all the g_i; this means that the straight lines corresponding to each of these is approximately horizontal, the intersections tend to be very distant from each other, and the confidence interval gets *enormous*, i.e. practically useless.

To this issue, there is a remedy. Suppose, to fix ideas, that the inputs are independent and identically distributed, that $\mathsf{E}[u_i] = 0$, and $\mathsf{E}[u_i^2] = \sigma^2 > 0$ (if $\mathsf{E}[u_i^2]$ was also equal to 0, then it would be $u_i = 0$ almost surely, and we would be pretending too much from the theory). We consider, now, instead of the functions g_i above, the following ones:

$$
f_i(\theta) = (y_i - \theta u_i)u_i = y_i u_i - \theta u_i^2, \qquad i = 1, \cdots, 7;
$$

$$
\begin{aligned}
g_1(\theta) &= \frac{1}{4}\sum_{k\in\{1,2,4,5\}} f_k(\theta) \\
&= \frac{1}{4}\sum_{k\in\{1,2,4,5\}} \theta^o u_k^2 + \varepsilon_k u_k - \theta u_k^2 \\
&= (\theta^o - \theta)\left(\frac{1}{4}\sum_{k\in\{1,2,4,5\}} u_k^2\right) + \frac{1}{4}\sum_{k\in\{1,2,4,5\}} \varepsilon_k u_k; \\
&\vdots \\
g_7(\theta) &= (\theta^o - \theta)\left(\frac{1}{4}\sum_{k\in\{4,5,6,7\}} u_k^2\right) + \frac{1}{4}\sum_{k\in\{4,5,6,7\}} \varepsilon_k u_k.
\end{aligned}
$$

You see? Now the intersection of (say) $g_1(\theta)$ with the θ-axis is given by

$$
\begin{aligned}
0 = g_1(\theta_1) &= (\theta^o - \theta_1)\left(\frac{1}{4}\sum_{k\in\{1,2,4,5\}} u_k^2\right) + \frac{1}{4}\sum_{k\in\{1,2,4,5\}} \varepsilon_k u_k; \\
\theta_1 &= \theta^o + \frac{\frac{1}{4}\sum_{k\in\{1,2,4,5\}}\varepsilon_k u_k}{\frac{1}{4}\sum_{k\in\{1,2,4,5\}}u_k^2}.
\end{aligned}
$$

The denominator poses no problem: as more measures come into the picture, the term $\frac{1}{(N+1)/2}\sum_k u_k^2$ tends almost surely to $\mathsf{E}[u_i^2] > 0$ (by the strong law of large numbers), so that the functions g_i have slopes with comparable magnitude. On the other hand, as far as the $\{u_i\}$ are independent of the $\{\varepsilon_i\}$, *the terms $\varepsilon_k u_k$ still have densities symmetric around* 0, so that the fundamental idea of LSCR (the manipulation of the sets of inequalities exploiting the group structure) applies without changes. Thus, with 7 measures, $\left[\bar{\theta}_1,\ \bar{\theta}_7\right]$ is once again a 75%-confidence interval for θ^o.
Moreover, the numerator has the form $\frac{1}{(N+1)/2}\sum_k \varepsilon_k u_k$ which, under fairly general conditions, converges at least weakly to $\mathsf{E}[\varepsilon_i u_i] = \mathsf{E}[\varepsilon_i]\mathsf{E}[u_i] = 0$; hence, we expect smaller and smaller confidence intervals as the number N of measures increases.

6.6 Leave-out Sign-dominant Correlation Regions

Finally, the LSCR method generalizes to (and actually was conceived for) system identification. Consider, for example, the simple autoregressive (AR) process

$$y_i = \theta^o y_{i-1} + \varepsilon_i,$$

where i is a discrete *time* index. Along the lines of the previous section, one may be tempted to start defining

$$f_i(\theta) = y_i - \theta y_{i-1}$$

(which is a *prediction error* of y_i given y_{i-1}; note that now, to develop the canonical example, 8 measures of the process are needed instead of 7). But after the above considerations ($\mathsf{E}[u_i] = 0$) this choice reveals hopeless. Indeed, in the typical case when the process is *stationary*, that is, when the transfer function $W(z) = \frac{z}{z-\theta^o}$ is BIBO-stable[27], it is unavoidable that $\mathsf{E}[y_{i-1}] = 0$. Therefore, Campi and Weyer went for the following functions instead:

$$f_i(\theta) = (y_{i+1} - \theta y_i)(y_i - \theta y_{i-1}).$$

[27] This in turn means that its only pole θ^o is in the open unit disc; hence, since it is real, it holds $-1 < \theta^o < 1$.

The remaining construction is similar:

$$\begin{aligned}
g_1(\theta) &= \frac{1}{4} \sum_{k \in \{1,2,4,5\}} f_k(\theta) \\
&= \frac{1}{4} \sum_{k \in \{1,2,4,5\}} (y_{k+1} - \theta y_k)(y_k - \theta y_{k-1}) \\
&= \frac{1}{4} \sum_{k \in \{1,2,4,5\}} ((\theta^o - \theta) y_k + \varepsilon_{k+1}) \left((\theta^o - \theta) y_{k-1} + \varepsilon_k\right) \\
&= (\theta^o - \theta)^2 \left(\frac{1}{4} \sum_{k \in \{1,2,4,5\}} y_k y_{k-1} \right) + (\theta^o - \theta) \left(\frac{1}{4} \sum_{k \in \{1,2,4,5\}} y_k \varepsilon_k \right) \\
&\quad + (\theta^o - \theta) \left(\frac{1}{4} \sum_{k \in \{1,2,4,5\}} y_{k-1} \varepsilon_{k+1} \right) + \frac{1}{4} \sum_{k \in \{1,2,4,5\}} \varepsilon_{k+1} \varepsilon_k
\end{aligned}$$

(and analogously for $g_2, \cdots, g_7$). The functions g_i now are not anymore affine — they are parabolas — hence their intersections with the θ-axis are not anymore supposed to be unique. Nevertheless, they still split the θ-axis in a finite number of intervals. Or better, they split the θ-axis in of 8 *regions*, that are unions of one or more disjoint intervals, where one out of eight situations happen: 1) *all the g_i are positive*; 2) *exactly* 1 *of the g_i is negative*, 3) *exactly* 2 *of the g_i are negative*, ..., 8) *all the g_i are negative.*

For each such region we can repeat the familiar reasoning: *what happens if θ^o belongs to this region?* Note that since ε_{k+1} and ε_k are independent, each with symmetric density, their product $\varepsilon_{k+1}\varepsilon_k$ also has symmetric density. Note, also, that $g_1(\theta^o) = \frac{1}{4}\sum_{k \in \{1,2,4,5\}} \varepsilon_{k+1}\varepsilon_k$; in words: at $\theta = \theta^o$, $g_1(\theta)$ is *an average of terms with symmetric density* (the same happens for $g_2(\theta^o), \cdots, g_7(\theta^o)$).

Then we can repeat the proof of section 6.4, which is based *only* on the symmetry and the group structure, to conclude that the regions have equal probability $\frac{1}{8}$. Discarding the regions where the g_i have all the same sign, we obtain, again, a 75%-confidence *region* for θ^o. [28]

We can also "tune" the confidence discarding some other regions, typically those where at most one among the functions g_i has its sign opposite to the others; in this way we obtain a 50%-confidence region. We note, finally, that a function like

$$g_1(\theta) = \frac{1}{4} \sum_{k \in \{1,2,4,5\}} (y_{k+1} - \theta y_k)(y_k - \theta y_{k-1}) = \frac{1}{4} \sum_{k \in \{1,2,4,5\}} \epsilon_k(\theta)\epsilon_{k-1}(\theta)$$

[28] In their paper [7], Campi and Weyer exploit some more *a-priori* information, namely that the process is stationary (i.e. $-1 < \theta^o < 1$) to reduce the set of intervals where θ^o may actually belong; in this way, they come up with a confidence *region* which happens to be an *interval.*

is indeed an empirical 1-*lag correlation* between two prediction errors; these observations, at last, should make every word explicit in the name of the method: Leave-out Sign-dominant *Correlation* Regions.

The LSCR method has been applied to the identification of way more general system structures; extending it to a multi-dimensional setting ($\theta^o \in \mathbb{R}^p$) one must cope with new *numerical* problems, but the essence of the theory remains the same. For more details, see [7] and the references therein[29].

6.7 Exercises for Chapter 6

Problem 1 (discrete distribution, wrong confidence).
Suppose that three measures are available:

$$\begin{aligned} y_1 &= \theta^o + \varepsilon_1, \\ y_2 &= \theta^o + \varepsilon_2, \\ y_3 &= \theta^o + \varepsilon_3, \end{aligned}$$

where ε_1, ε_2, and ε_3 are independent and identically distributed with *discrete distribution*, each taking only the values 1 or -1 with equal probabilities:

$$\varepsilon_1 = \begin{cases} 1 & \text{with probability } \frac{1}{2}, \\ -1 & \text{with probability } \frac{1}{2}, \end{cases}$$
$$\varepsilon_2 = \begin{cases} 1 & \text{with probability } \frac{1}{2}, \\ -1 & \text{with probability } \frac{1}{2}, \end{cases}$$
$$\varepsilon_3 = \begin{cases} 1 & \text{with probability } \frac{1}{2}, \\ -1 & \text{with probability } \frac{1}{2}. \end{cases}$$

Let $\theta^o = 1$. We employ LSCR method with the following group:

	1	2	3
I_1	●	●	○
I_2	●	○	●
I_3	○	●	●
I_4	○	○	○

and select the interval $[\bar{\theta}_1, \bar{\theta}_3]$ which, according to the LSCR theory, should have a 50%-confidence interval. Show that the confidence of such interval is *not* 50% (this may be the case if the distribution is discrete). Hint: see what happens for every possible value of ε_1, ε_2, and ε_3.

Problem 2 (discrete distribution, correct confidence).

[29] The method has also been applied to the identification of *nonlinear* systems; see [4].

Suppose that three measures are available:

$$\begin{aligned} y_1 &= \theta^o + \varepsilon_1, \\ y_2 &= \theta^o + \varepsilon_2, \\ y_3 &= \theta^o + \varepsilon_3, \end{aligned}$$

where ε_1, ε_2, and ε_3 are independent but with discrete distribution, each taking only two possible values symmetrically around 0. Namely:

$$\varepsilon_1 = \begin{cases} 1 & \text{with probability } \frac{1}{2}, \\ -1 & \text{with probability } \frac{1}{2}, \end{cases}$$
$$\varepsilon_2 = \begin{cases} \frac{1}{2} & \text{with probability } \frac{1}{2}, \\ -\frac{1}{2} & \text{with probability } \frac{1}{2}, \end{cases}$$
$$\varepsilon_3 = \begin{cases} 2 & \text{with probability } \frac{1}{2}, \\ -2 & \text{with probability } \frac{1}{2}. \end{cases}$$

Let $\theta^o = 1$, and verify that the LSCR method with the following group:

	1	2	3
I_1	●	●	○
I_2	●	○	●
I_3	○	●	●
I_4	○	○	○

correctly provides a 50%-confidence interval even though the errors do not have a density. In order to do this, see what happens for every possible value of ε_1, ε_2, and ε_3.

It's not what we don't know that hurts us, it's what we know for certain that just ain't so.
Mark Twain.

Le doute n'est pas un état bien agréable, mais l'assurance est un état ridicule.
Voltaire, in a letter to Frederick the Great, 1770.

7 Interval predictor models

7.1 Introduction and motivation

In Chapter 1 we have seen how to construct a model

$$y = \hat{a}_{\mathrm{LS}} + \hat{b}_{\mathrm{LS}}\ x \tag{40}$$

given some measures $(x_1, y_1), \cdots, (x_N, y_N)$ of explanatory variables x and explained variables y. It was stated there, that one of the objectives of such a model is to *predict* the value of a future variable y_{N+1} when the corresponding explanatory variable x_{N+1} will be available. The predicted value is

$$\hat{y}_{N+1} = \hat{a}_{\mathrm{LS}} + \hat{b}_{\mathrm{LS}}\ x_{N+1}.$$

But what does it really mean that $\hat{y}_{N+1}$ *predicts* the value of y_{N+1}?

To start with, if y_{N+1} is a random variable, in general it is impossible that the two values are equal (the probability that this happens is 0, if y_{N+1} has a density); and in general, if the density is not bounded (for example if it is Gaussian), it is not even possible to say, *with certainty*, that it will be close to the predicted value. Indeed, unless *very* strong assumptions are made on the law that generates data, very little can be hold for sure about the future observation (recall the above quotes by Mark Twain and Voltaire)[30].
Thus, $\hat{y}_{N+1}$ is just a "plausible" value for the future observation, that comes with the hope that the future observation be close to it. Our way to quantify "hope" is probability, hence we will retain that the prediction is meaningful if we can prove that the density of y_{N+1} is concentrated around $\hat{y}_{N+1}$. Usually, this means two things:

[30] Here are some examples of strong assumptions:

- the pairs are generated by a differentiable function $y = g(x)$, where $|g'(x)| \leq 1$ for all x;
- the pairs are generated by the model $y = a^o + b^o x + \varepsilon$, where ε is a *bounded* random variable, namely $\varepsilon \in [-\delta, \delta]$ almost surely, for a *known* small constant $\delta \in \mathbb{R}$.

This is the kind of illusory knowledge that we do not want to assume.

- that the means of y_{N+1} and $\hat{y}_{N+1}$ are equal, and the variance of their difference is small;
- that there exists an interval $[a, b]$, preferably small, that contains both $\hat{y}_{N+1}$ and y_{N+1} with known probability (the higher, the better).

In this chapter we will focus on the second one. Thus, our goal is now to construct, given the data $(x_1, y_1), \cdots, (x_N, y_N)$ and a future observation x_{N+1}, an interval that contains y_{N+1} with certified probability. We will *not* start from a model like (40) and build the interval around a point estimate $\hat{y}_{N+1}$, although in the end we will recover something similar.

How can we build a "prediction interval" for y_{N+1}? As we shall now see, the problem is not trivial even if, to simplify the problem as much as possible, we discard the explanatory data $\{x_i\}$ and build such interval with the sole knowledge of the past observations $y_1, \cdots, y_N$. We will develop our construction in four steps:

1. we will show, just for comparison, how a prediction interval for y_{N+1} is built in the "classical" way, under the (strong) assumption that the $\{y_i\}$ are Gaussian;
2. we will show how a radical change of perspective lets us build a prediction interval for y_{N+1} without any knowledge on the underlying density;
3. we will change again our point of view, and obtain the same prediction interval as a solution to a convex optimization problem;
4. finally, we will easily extend the convex optimization approach reintroducing the explanatory data. This will end up in a model that, given x_{N+1}, yields an entire prediction interval for y_{N+1}, with certified probability, instead of a single-point prediction $\hat{y}_{N+1}$.

7.2 Prediction intervals, the standard way

Let us start, then, by making an assumption on the distribution of the data, rather strong for our purposes, but common in applied statistics. Suppose that we extract a sample $\{y_1, \cdots, y_N\}$ from a *Gaussian* distribution $\mathcal{N}(\theta^o, \sigma^2)$. Can we provide an interval $[a, b]$ in which a future observation y_{N+1} will fall with probability $1 - \alpha$, say 95%?
If we knew θ^o and σ^2 the answer would be trivial and would not rely on the sample at all, because in that case

$$\frac{y_{N+1} - \theta^o}{\sigma} \sim \mathcal{N}(0, 1),$$

and since we can find the cumulative distribution function of a $\mathcal{N}(0,1)$ variable, and its inverse, tabulated in any book on statistics or computed in any statistical software, we can easily find the percentile z_α such that

$$\int_{-z_\alpha}^{z_\alpha} f(z)\ dz = 1-\alpha,$$

where $f(z)$ is the Gaussian density $\mathcal{N}(0,1)$. Then

$$\mathsf{P}\left[\left|\frac{y_{N+1}-\theta^o}{\sigma}\right| \le z_\alpha\right] = 1-\alpha,$$

or

$$\theta^o - z_\alpha\sigma \le y_{N+1} \le \theta^o + z_\alpha\sigma$$

with probability $1-\alpha = 0.95$. The interval

$$[a,b] = [\theta^o - z_\alpha\sigma,\ \theta^o + z_\alpha\sigma]$$

is called a *prediction interval* for y_{N+1}.

Anyway, it is seldom, if at all, the case that we know θ^o and σ^2, and we wish to construct a prediction interval based on the observed data. The usual result is very similar to the one regarding the construction of a confidence interval for θ^o, based on the sample average $\hat{\theta}$ and the sample variance $\bar{s}^2$, along the lines of Section 6.2.
Note that, if all the variables are independent and Gaussian, the variable $y_{N+1}-\hat{\theta}$ is also Gaussian, with mean and variance:

$$\begin{aligned}\mathsf{E}\left[y_{N+1}-\hat{\theta}\right] &= 0,\\ \mathsf{Var}\left[y_{N+1}-\hat{\theta}\right] &= \sigma^2\left(1+\frac{1}{N}\right).\end{aligned}$$

We have

Theorem 7.2.1 *Let $y_1,\cdots,y_N \sim \mathcal{N}(\theta^o,\sigma^2)$ independent and one more measure $y_{n+1} \sim \mathcal{N}(\theta^o,\sigma^2)$. Then*

$$\frac{y_{N+1}-\hat{\theta}}{\bar{s}\sqrt{1+\frac{1}{N}}} \sim t(N-1)$$

Proof.

$$\frac{y_{N+1}-\hat{\theta}}{\bar{s}\sqrt{1+\frac{1}{N}}} = \frac{\frac{y_{N+1}-\hat{\theta}}{\sigma\sqrt{1+\frac{1}{N}}}}{\frac{\sqrt{N-1}\bar{s}}{\sigma}/\sqrt{N-1}},$$

where $\frac{y_{N+1}-\hat{\theta}}{\sigma\sqrt{1+\frac{1}{N}}}$ is $\mathcal{N}(0,1)$ and independent of $\frac{\sqrt{N-1}\bar{s}}{\sigma}/\sqrt{N-1}$, which is the square root of a $\chi^2(N-1)$ variable, divided by the number of its degrees of freedom. The result follows from Theorem D.9.2. □

Now we can find on books or through software, the percentile t_α such that

$$\int_{-t_\alpha}^{t_\alpha} f(t)\ dt = 1-\alpha,$$

where $f(t)$ is the density of a $t(N-1)$ random variable. It holds

$$\mathsf{P}\left[\left|\frac{y_{N+1}-\hat{\theta}}{\bar{s}\sqrt{1+\frac{1}{N}}}\right| \leq t_\alpha\right] = 1-\alpha.$$

Repeating the steps in the construction of a confidence interval for θ^o, finally we find the $(1-\alpha)$-probability prediction interval for y_{N+1}:

$$[a,b] = \left[\hat{\theta} - t_\alpha\bar{s}\sqrt{1+\frac{1}{N}},\ \ \hat{\theta} + t_\alpha\bar{s}\sqrt{1+\frac{1}{N}}\right].$$

As usual, we must pay attention to what we mean with the word "probability". *Before the sample is drawn*, $1-\alpha = 95\%$ is the probability that a random interval ($[a,b]$) covers an independent random variable (y_{N+1}). *After the sample is drawn*, it is the probability that y_{N+1}, which is still random, falls in the *deterministic* interval $[a,b]$. It is still a probability (we will not use the word "confidence" here), but conditioned to the values of $y_1,\cdots,y_N$. To distinguish the two, one usually calls the first "*a priori* probability".
This result is quite nice, because it exploits all the available information and does not depends on parameters other than the size of the sample, but it still has a fundamental flaw:

it depends crucially on the Gaussianity of data,

whereas it is seldom known, if at all, that the data are distributed according to a precise parametric family of distributions.

7.3 Prediction intervals, a different perspective

Here we sketch a procedure to build a prediction interval in a radically different way; the construction will *not* depend on the distribution of the data and, above all, it will generalize nicely to more complex prediction problems. Let us consider a "chain" of examples, that get closer and closer to the main point.

Example. An urn contains $N + 1$ balls, of which 2 are red and the others are white. We extract one ball from the urn. What is the probability that the extracted ball is red? The answer is, according to classical probability, $\frac{2}{N+1}$. □

Example. An urn contains $N + 1$ balls, of which 2 are labeled "extreme point" and the others are white. We extract a ball from the urn. What is the probability that the extracted ball has the label? Of course the answer is again $\frac{2}{N+1}$. □

Example. We have a list of $N + 1$ different numbers. We write each number on a ball, and mark the two balls with the minimum and the maximum number as "extreme point". Now we put the balls in an urn, shake, and extract one ball. What is the probability that the extracted ball is an extreme? Same story: $\frac{2}{N+1}$. □

Example. We repeat the previous experiment, but this time we extract *all* the balls in random order. What is the probability that the *last* ball is an extreme? It is immediate to convince ourselves that if all the different $(N + 1)!$ orders have the same probability, the selection of the last ball is equivalent to the extraction of a single ball. Hence the requested probability is again $\frac{2}{N+1}$.
More explicitly, the permutations of the balls that have either the minimum or the maximum at the last position are $N! + N! = 2N!$; therefore, all the permutations having the same probability, the probability of extracting a permutation with an extreme at the last position is $\frac{2N!}{(N+1)!} = \frac{2}{N+1}$. □

Example. We extract $N + 1$ independent samples $y_1, y_2, \cdots, y_{N+1}$ of a random variable having density $f(y)$. What is the probability that y_{N+1} is either the maximum or the minimum? Denote the event "the last number y_{N+1} is an extreme" as $\mathcal{E}$, and the event "any two numbers among $y_1, \cdots, y_{N+1}$ are equal" as $\mathcal{N}$. Due to the hypothesis that the random variables have a density, the probability of $\mathcal{N}$ is 0, and the probability of $\mathcal{E}$ conditioned to the extraction of $N + 1$ different numbers is always the same.

Hence the requested probability is

$$\begin{aligned}
p &= \int_{\mathbb{R}^{N+1}} \mathsf{P}\left[\mathcal{E} \mid y_1, \cdots, y_{N+1}\right] f(y_1)\cdots f(y_{N+1})\ dy_1\ \cdots\ dy_{N+1} \\
&= \int_{\mathbb{R}^{N+1}\setminus\mathcal{N}} \mathsf{P}\left[\mathcal{E} \mid y_1, \cdots, y_{N+1}\right] f(y_1)\cdots f(y_{N+1})\ dy_1\ \cdots\ dy_{N+1} \\
&= \int_{\mathbb{R}^{N+1}\setminus\mathcal{N}} \frac{2}{N+1}\ f(y_1)\cdots f(y_{N+1})\ dy_1\ \cdots\ dy_{N+1} \\
&= \frac{2}{N+1}\ \int_{\mathbb{R}^{N+1}\setminus\mathcal{N}} f(y_1)\cdots f(y_{N+1})\ dy_1\ \cdots\ dy_{N+1} = \frac{2}{N+1}.
\end{aligned}$$

□

Example. We extract N independent samples $y_1, y_2, \cdots, y_N$, from a random variable with density $f(y)$. Let a and b be their minimum and maximum respectively. We ask now what is the probability that a "future" sample y_{N+1} falls outside the interval $[a, b]$. Guess what? The probability is $\frac{2}{N+1}$, because the above procedure is nothing more than a rephrasing of "we extract $N+1$ independent samples and ask about the last one", and y_{N+1} is outside $[a, b]$ exactly whenever it is either the maximum or the minimum of the whole sample. □

The above chain of examples is actually an informal proof of the following

Lemma 7.3.1 *Let $y_1, y_2, \cdots, y_N$ be independent and identically distributed variables with density $f(y)$. Let a and b be their minimum and maximum respectively. Then $[a, b]$ is a prediction interval for a future sample y_{N+1}, with probability $1 - \frac{2}{N+1} = \frac{N-1}{N+1}$.*

Some remarks are in order:

- the probability $1 - \frac{2}{N+1}$ with which the future sample will fall inside the prediction interval is *exact* (i.e. exactly as happens for the LSCR method, and unlike the Hoeffding-type theory we have developed in machine learning, it is not an upper-bound);

- note the radical change in the point of view: in the "standard way" (Section 7.2), the key aspect of randomness that was taken into account was the common distribution of the sample; now it is the *order* in which the sample is drawn;

- the great advantage of this method is that the prediction interval is *distribution-free*, that is, completely independent of the density of the $\{y_i\}$;

- comparing with the "standard way", we have some less freedom in "tuning" the probability. It could be shown that other prediction intervals, with different *a priori* probability can be built selecting not the minimum and the maximum, but some other points, close to the extremes (in other words, discarding some "outliers"); thus the *a priori* probability could still be tuned, but not continuously (90%, 95.2%, 99.9% etc.) as in the Gaussian case. In this chapter we will maintain the simplest possible choice; thus, the probability is $1 - \frac{2}{N+1}$, that is it.

7.4 Convex problems and Helly's theorem

Definition 7.4.1 *We will call* convex problem *an optimization problem of the form:*

$$\begin{aligned} \textit{minimize} \quad & f(\theta) \\ \textit{subject to} \quad & f_1(\theta) \leq 0, \\ & f_2(\theta) \leq 0, \\ & \vdots \\ & f_n(\theta) \leq 0, \\ & \theta \in \Theta \subseteq \mathbb{R}^d, \end{aligned}$$

where $f, f_1, \cdots, f_n$ are convex functions and Θ is a convex set. Each of the inequalities $f_i(\theta) \leq 0$ is called a constraint. *The problem is said to be* feasible *if there exists at least a point $\bar{\theta} \in \Theta$ that satisfies all the constraints, that is, such that $f_i(\bar{\theta}) \leq 0$ for all $i = 1, \cdots, n$.*

Note that a constraint of the form $f_i(\theta) \leq 0$ is the 0-sublevel set of a convex function, which is a convex set; hence, the problem is asking for the minimum value that a convex function f attains in the intersection of $n+1$ convex sets, which is again a convex set. Usually the definition of a convex problem includes equality constraints, of the form $h_i(\theta) = 0$; we will not need any such constraint here.

The fundamental fact about convex problems is that if a point is *locally* a minimum, i.e. it satisfies some minimality condition in the neighborhood of a point, then it is a minimum also *globally*, hence it is a solution to the problem; thus, a local test of optimality is sufficient to establish global optimality. Convex problems are considered "easy", because we have readily available efficient and robust algorithms to solve them numerically (interior point methods & co.). Recognizing that a particular optimization problem can be formulated as a convex problem may be hard; but once this has been done, you can consider it as "practically solved". In the words of Stephen Boyd, convex optimization is "almost a technology", that is, it is approaching a stage of maturity in which "it can be reliably used by many people who do not know, and do not need to know, the details". In comparison, according to him the method of least squares is a mature technology. See the standard reference [2] for further observations.

A convex problem may or may not have a solution, and a solution may not be unique.

Example. The following scalar problem:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{\theta} \\ \text{subject to} \quad & \theta \in [1, \infty) \end{aligned}$$

does not have a solution, because $\inf_{\theta \in [1,\infty)} \frac{1}{\theta} = 0$, but $\frac{1}{\theta} \neq 0$ for all $\theta \in [1, \infty)$. The following scalar problem:

$$\begin{aligned} \text{minimize} \quad & \theta^2 \\ \text{subject to} \quad & \theta \in [1, \infty) \end{aligned}$$

has the unique solution $\theta^* = 1$. The following *linear* problem:

$$\begin{aligned} \text{minimize} \quad & \theta_1 + \theta_2 \\ \text{subject to} \quad & -\theta_1 - \theta_2 \leq 0, \\ & \theta_2 - \theta_1 - 1 \leq 0, \\ & \theta_1 - \theta_2 - 1 \leq 0, \\ & (\theta_1, \theta_2) \in \mathbb{R}^2, \end{aligned}$$

has infinitely many solutions, namely all the points (θ_1^*, θ_2^*) belonging to the line segment that joins the points $(-\frac{1}{2}, \frac{1}{2})$ and $(\frac{1}{2}, -\frac{1}{2})$. (To see why, draw a picture with all the constraints and think at the direction in which $f(\theta_1, \theta_2) = \theta_1 + \theta_2$ decreases.) □

In the following, we will consider $(d+1)$-dimensional problems of this form:

$$\begin{aligned} \text{minimize} \quad & \gamma \\ \text{subject to} \quad & g_1(\theta) - \gamma \leq 0 \\ & g_2(\theta) - \gamma \leq 0, \\ & \vdots \\ & g_n(\theta) - \gamma \leq 0, \\ & (\theta, \gamma) \in \Theta \times \mathbb{R}, \end{aligned}$$

where $g_1, \cdots, g_n$ are convex functions and $\Theta \subseteq \mathbb{R}^d$. You can easily recognize that it is a particular case of Definition 7.4.1 (yet sufficiently general for our purposes). Indeed the goal function $f(\theta,\gamma) = \gamma$ is trivially convex, and if $g_i(\theta)$ is convex, then the constraint function $f_i(\theta,\gamma) = g_i(\theta) - \gamma$ is also convex[31]. We will assume that all the problems under consideration are feasible and admit a unique solution; or better, since the constraints will arise from random data, we will assume that this happens *almost surely.* We will denote the minimizing solution (θ^*,γ^*).

Definition 7.4.2 *Consider the following convex problem:*

$$\begin{aligned} \textit{minimize} \quad & \gamma \\ \textit{subject to} \quad & g_1(\theta) - \gamma \leq 0, \\ & \vdots \\ & g_n(\theta) - \gamma \leq 0, \\ & (\theta,\gamma) \in \Theta \times \mathbb{R}. \end{aligned}$$

Let (θ^*,γ^*) be its solution, and consider one of its constraints, $g_i(\theta) - \gamma \leq 0$. We will call the latter a support constraint *if the solution $(\theta^{**},\gamma^{**})$ to the problem obtained by removing the constraint,*

$$\begin{aligned} \textit{minimize} \quad & \gamma \\ \textit{subject to} \quad & g_1(\theta) - \gamma \leq 0, \\ & \vdots \\ & g_{i-1}(\theta) - \gamma \leq 0, \\ & g_{i+1}(\theta) - \gamma \leq 0, \\ & \vdots \\ & g_n(\theta) - \gamma \leq 0, \\ & (\theta,\gamma) \in \Theta \times \mathbb{R}. \end{aligned}$$

is strictly "better" than (θ^*,γ^*), meaning that it attains $\gamma^{**} < \gamma^*$.

In words, removing a support constraint the solution "falls". In particular, such a constraint is *active*, meaning that the inequality "≤ 0" actually works as an equality "$= 0$" for the solution.

[31] Indeed:

$$\begin{aligned} & f_i(\lambda\theta^1 + (1-\lambda)\theta^2,\ \lambda\gamma^1 + (1-\lambda)\gamma^2) \\ & \quad = g_i(\lambda\theta^1 + (1-\lambda)\theta^2) - \lambda\gamma^1 - (1-\lambda)\gamma^2 \\ & \quad \leq \lambda g_i(\theta^1) + (1-\lambda)g_i(\theta^2) - \lambda\gamma^1 - (1-\lambda)\gamma^2 \\ & \quad = \lambda f_i(\theta^1,\gamma^1) + (1-\lambda)f_i(\theta^2,\gamma^2). \end{aligned}$$

An obvious but crucial observation is that if a constraint is *not* a support constraint, then it can be removed from the problem without any consequence (the solution is the same). Another crucial observation is that if we *add* a constraint to a problem, and it happens to become a support constraint in the new problem, then the new γ^* of the solution must *increase* with respect to the old solution (indeed, removing it again γ^* must decrease).
The main result of this chapter is that the support constraints of our $d+1$-dimensional convex problems are at most $d+1$. In order to prove it, we need a classical result, which we state here without proof:

Theorem 7.4.1 *(Helly). Let $S_1, \cdots, S_i, \cdots, S_N$ be convex subsets of $\mathbb{R}^n$. If the intersection of any $n+1$ of these subsets is nonempty, that is,*

$$\bigcap_{k=1}^{n+1} S_{i_k} \neq \varnothing \quad \textit{for any choice of} \;\; \{i_1, \cdots, i_{n+1}\} \subset \{1, \cdots, N\},$$

then the intersection of all *the subsets is nonempty, that is,*

$$\bigcap_{i=1}^{N} S_i \neq \varnothing.$$

Proof. See [25]. □

The following result is a simplified version of one of the cornerstones in a general theory about robust optimization, called "scenario approach", due to Marco C. Campi, Giuseppe Calafiore, Simone Garatti, and others.

Theorem 7.4.2 *Any feasible convex problem like the following:*

$$\begin{aligned}
\textit{minimize} \quad & \gamma \\
\textit{subject to} \quad & g_1(\theta) - \gamma \leq 0 \\
& g_2(\theta) - \gamma \leq 0, \\
& \vdots \\
& g_n(\theta) - \gamma \leq 0, \\
& \theta \in \mathbb{R}^d, \gamma \in \mathbb{R},
\end{aligned}$$

has at most $d+1$ support constraints.

(Note that the number $d+1$ is due to the particular structure of the problem under consideration, which is $(d+1)$-dimensional; a general statement should indeed read "any d-dimensional feasible convex problem has at most d support constraints".)

Proof. Let (θ^*, γ^*) be the solution to the problem, and define

$$\begin{aligned} S_1 &= \left\{ (\theta, \gamma) \in \mathbb{R}^{d+1} \mid g_1(\theta) \leq \gamma \right\}, \\ &\vdots \\ S_n &= \left\{ (\theta, \gamma) \in \mathbb{R}^{d+1} \mid g_n(\theta) \leq \gamma \right\}, \\ Z &= \left\{ (\theta, \gamma) \in \mathbb{R}^{d+1} \mid \gamma < \gamma^* \right\}. \end{aligned}$$

The sets $S_1, \cdots S_n \subset \mathbb{R}^{d+1}$ are the epigraphs of the convex functions $g_1(\cdot), \cdots, g_n(\cdot)$, hence they are convex sets; a point belongs to S_i if and only if it satisfies the i-th constraint. The set $Z \subset \mathbb{R}^{d+1}$ is an open half-plane, hence of course another convex set. Any point belonging to Z is "super-optimal" $(\gamma < \gamma^*)$; there cannot exist a point that satisfies all the constraints and belongs to Z, otherwise (θ^*, γ^*) would not be the minimizing solution.
For the sake of contradiction, assume now that the support constraints of the problem are (without loss of generality) $d + 2$. Extract an arbitrary collection of $d + 2$ sets from $S_1, \cdots S_n, Z$.
If Z happens to be in the collection, then the collection contains exactly $d+1$ (epigraphs of) constraints, of which at most $d + 1$ are support constraints. Since the support constraints are assumed to be $d + 2$, at least one of them has been "removed", and the solution "falls", namely there exists a point $(\theta^{**}, \gamma^{**})$ which satisfies the $d + 1$ constraints (meaning that it belongs to the $d + 1$ sets $\{S_i\}$) and attains $\gamma^{**} < \gamma^*$ (meaning that it belongs to Z).
On the other hand, if Z is not in the collection, the latter contains just $d+2$ constraints, and since the problem is feasible, there exists at least a point satisfying all of them.
Summing up, for any choice of $d + 2$ sets in $S_1, \cdots S_n, Z$, their intersection is non-empty. Applying Helly's theorem we obtain that

$$S_1 \cap \cdots \cap S_n \cap Z \neq \varnothing,$$

but this is clearly in contradiction with the hypothesis that (θ^*, γ^*) is the solution to the problem. The contradiction stems from the assumption that the support constraints were more than $d+1$, and is enough to establish the claim. □

7.5 Prediction intervals revisited

A simple question in the style of Section 7.3 will clarify what is our final objective:

Example. An urn contains $N + 1$ balls, of which $d + 1$ are labeled "support constraint" and the others are white. We extract a ball from the urn. What is the probability that the extracted ball has the label? The answer is of course $\frac{d+1}{N+1}$. □

Let $y_1, y_2, \cdots, y_N$ a random sample drawn from the density $f(y)$, and consider the following convex problem (the constraints are now random):

$$\begin{aligned} \text{minimize} \quad & \gamma \\ \text{subject to} \quad & |\theta - y_1| - \gamma \leq 0 \\ & \vdots \\ & |\theta - y_N| - \gamma \leq 0 \\ & \theta \in \mathbb{R}, \gamma \in \mathbb{R}, \end{aligned}$$

Note that a constraint of the form $g(\theta, \gamma) = |\theta - y| - \gamma \leq 0$ is convex in the variables (θ, γ). Indeed, any set of the form $|\theta - y| - \gamma \leq 0$ is the intersection of two half-spaces:

$$\begin{aligned} \theta - y &\leq \gamma, \\ \theta - y &\geq -\gamma. \end{aligned}$$

Hence it is the intersection of two convex sets, which is itself convex. Therefore, the problem is in turn convex.
The solution to the problem is the pair (θ^*, γ^*) attaining the *minimum* γ such that θ *has distance at most* γ *from all the points* $\{y_i\}$. It is not difficult to convince ourself that the solution to this problem is such that

$$\begin{aligned} \theta^* - \gamma^* &= \min_i \{y_i\}, \\ \theta^* + \gamma^* &= \max_i \{y_i\}. \end{aligned}$$

Recall that the support constraints of the problem are at most $d + 1 = 2$. It may actually be the case that they are *less* than 2; this happens if and only if there are more than one point equal to the minimum or to the maximum (for example, if two measures are equal to the maximum, removing one of them, that is removing the corresponding constraint, the maximum does not change). But since the random variables have a density, this happens with probability 0. Hence, *almost surely* the support constraints are exactly 2 (we say that the problem is "fully supported"), and the constraints corresponding to the minimum and the maximum are precisely the 2 support constraints of this 2-dimensional problem.
The solution yields the prediction interval that we had obtained before:

$$[a, b] = [\theta^* - \gamma^*,\ \theta^* + \gamma^*] = \left[\min_i \{y_i\},\ \max_i \{y_i\}\right].$$

The question now arises as whether or not the solution to the above problem remains the same if another constraint, corresponding to a new measure y_{N+1} drawn from the same density, is added:

$$
\begin{aligned}
\text{minimize} \quad & \gamma \\
\text{subject to} \quad & |\theta - y_1| - \gamma \leq 0 \\
& \vdots \\
& |\theta - y_N| - \gamma \leq 0 \\
& |\theta - y_{N+1}| - \gamma \leq 0 \\
& \theta \in \mathbb{R}, \gamma \in \mathbb{R},
\end{aligned}
$$

If $y_{N+1} \in [a, b]$, then the last constraint does not change anything, and the solutions to the two problems are identical.
On the other hand, if $y_{N+1} \notin [a, b]$ (the new measure falls outside the prediction interval), then the last constraint $|\theta - y_{N+1}| - \gamma \leq 0$ becomes a support constraint for the new problem, and the solution must change (the solution must increase when the new support constraint is added, indeed because it must decrease when the constraint is removed). In this case we say that the new constraint *violates* the previous solution.

What is the probability that the new constraint violates the old solution? Consider the second problem, with $N + 1$ constraints, and let the event

$$
\begin{aligned}
\mathcal{E} = \{ & \text{the new constraint } |\theta - y_{N+1}| - \gamma \leq 0 \text{ is a support constraint} \\
& \text{for the problem with } N + 1 \text{ constraints} \} .
\end{aligned}
$$

Since the support constraints are almost surely $d + 1$, conditioning to the extraction $y_1, y_2, \cdots, y_N, y_{N+1}$, the reasoning of Section 7.3 applies without any substantial change:

$$
\begin{aligned}
p &= \int_{\mathbb{R}^{N+1}} \mathsf{P}\left[\mathcal{E} \mid y_1, \cdots, y_{N+1}\right] f(y_1) \cdots f(y_{N+1}) \, dy_1 \, \cdots \, dy_{N+1} \\
&= \int_{\mathbb{R}^{N+1} \setminus \mathcal{N}} \mathsf{P}\left[\mathcal{E} \mid y_1, \cdots, y_{N+1}\right] f(y_1) \cdots f(y_{N+1}) \, dy_1 \, \cdots \, dy_{N+1} \\
&= \int_{\mathbb{R}^{N+1} \setminus \mathcal{N}} \frac{d+1}{N+1} \, f(y_1) \cdots f(y_{N+1}) \, dy_1 \, \cdots \, dy_{N+1} \\
&= \frac{d+1}{N+1} \int_{\mathbb{R}^{N+1} \setminus \mathcal{N}} f(y_1) \cdots f(y_{N+1}) \, dy_1 \, \cdots \, dy_{N+1} \\
&= \frac{d+1}{N+1}.
\end{aligned}
$$

Summing up: we have considered a convex problem with N constraints, and its solution has yielded the prediction interval that we knew from before. A new measure y_{N+1} falls outside this prediction interval exactly when, adding to the problem a new constraint corresponding to y_{N+1}, this becomes a support constraint, in other words it violates the old solution. By virtue of the same reasoning on the ordering of measurements that we have considered in Section 7.3, the probability of violation is $\frac{d+1}{N+1}$. Finally, since in this example $d =$ dimension of $\theta = 1$, the violation probability is $\frac{2}{N+1}$, as we had found before.

7.6 Interval predictor models

What in the previous section may have looked like a tricky way to build a prediction interval for a scalar variable y_{N+1}, reveals its true power when explanatory variables are added.
Let $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N), (x_{N+1}, y_{N+1})$ be random pairs, independent and identically distributed according to an unknown density, where $x_i, y_i \in \mathbb{R}$. Consider the optimization problem with the constraints corresponding to the first N measures:

$$\begin{aligned}
\text{minimize} \quad & \gamma \\
\text{subject to} \quad & |\theta_1 + \theta_2 x_1 - y_1| - \gamma \leq 0 \\
& \vdots \\
& |\theta_1 + \theta_2 x_N - y_N| - \gamma \leq 0 \\
& (\theta_1, \theta_2) \in \mathbb{R}^2, \gamma \in \mathbb{R}.
\end{aligned}$$

Note that a constraint of the form $g(\theta_1, \theta_2, \gamma) = |\theta_1 + \theta_2 x - y| - \gamma \leq 0$ is convex in the variables $(\theta_1, \theta_2, \gamma)$, because any set of the form $|\theta_1 + \theta_2 x - y| - \gamma \leq 0$ is the intersection of two half-spaces:

$$\begin{aligned}
\theta_1 + \theta_2 x - y &\leq \gamma, \\
\theta_1 + \theta_2 x - y &\geq -\gamma,
\end{aligned}$$

hence it is the intersection of two convex sets, which is itself convex. Therefore, the problem is convex. The solution to the problem is a certain triple $(\theta_1^*, \theta_2^*, \gamma^*)$. The parameters θ_1^*, θ_2^* yield a linear model:

$$y = \theta_1^* + \theta_2^* \, x \tag{41}$$

which is "closest to the data $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$ as much as possible", in the sense that the quantity

$$\max_{i=1,\cdots,N} |\theta_1^* + \theta_2^* \, x_i - y_i| = \gamma^*$$

is the minimum that can be obtained letting θ_1, θ_2 vary. The linear model (41) is of course *not* the least squares model, because what is minimized here is the maximum among the moduli of the errors[32], not the sum of their squares, and in general

$$\min \max_{i=1,\cdots,N} |\theta_1^* + \theta_2^*\ x_i - y_i| \neq \min \sum_{i=1}^{N} (\theta_1^* + \theta_2^*\ x_i - y_i)^2$$

$$\text{interval predictor models} \neq \text{least squares};$$

however, for well-behaved data one can expect that the two models are not too different from each other. Now, as the "next" explanatory variable x_{N+1} comes, we can *predict* the value of y_{N+1} as follows:

$$\hat{y}_{N+1} = \theta_1^* + \theta_2^*\ x_{N+1}$$

and append a "certificate" to this prediction: it will hold

$$|\hat{y}_{N+1} - y_{N+1}| = |\theta_1^* + \theta_2^*\ x_{N+1} - y_{N+1}| > \gamma^*$$

with little probability, namely the same probability with which the new constraint

$$|\theta_1 + \theta_2 x_{N+1} - y_{N+1}| - \gamma \leq 0 \tag{42}$$

would violate the solution $(\theta_1^*, \theta_2^*, \gamma^*)$ if added to the problem. Since here d = dimension of $\theta = 2$, the maximum number of support constraints in this problem is $d+1 = 3$. Further, it could be shown that the number of support constraints is less than 3 exactly when more than three points belong to the straight lines

$$y = \theta_1^* + \theta_2^* x - \gamma^*$$
$$y = \theta_1^* + \theta_2^* x + \gamma^*$$

that delimit the "stripe" containing all the data, and that this happens with probability 0. Therefore, almost surely the support constraints are exactly 3, and according to the same reasoning of the previous section the violation probability is

$$p = \frac{d+1}{N+1} = \frac{3}{N+1}.$$

In other terms, the interval

$$[a, b] = [\theta_1^* + \theta_2^*\ x_{N+1} - \gamma^*, \quad \theta_1^* + \theta_2^*\ x_{N+1} + \gamma^*]$$

is a prediction interval for y_{N+1} with probability $1 - \frac{3}{N+1}$.

More in general, if we consider the whole pair (x_{N+1}, y_{N+1}) as the future observation, we can predict that it will fall outside the region

$$\left\{ (x, y) \in \mathbb{R}^2 \ \middle|\ |\theta_1^* + \theta_2^*\ x - y| > \gamma^* \right\}$$

[32] Problems like the one at hand are usually called "min-max problems".

with the same probability of violation of the constraint (42), which is again $\frac{d+1}{N+1} = \frac{3}{N+1}$. In other words, such region is a $\left(1 - \frac{3}{N+1}\right)$-probability "prediction set" for the new observation.

The method can be generalized to multi-dimensional explanatory variables and to nonlinear regressors:

$$\begin{aligned}
\text{minimize} \quad & \gamma \\
\text{subject to} \quad & \left|\varphi(x_1)^\top \theta - y_1\right| - \gamma \leq 0 \\
& \vdots \\
& \left|\varphi(x_N)^\top \theta - y_N\right| - \gamma \leq 0 \\
& \theta \in \mathbb{R}^d, \gamma \in \mathbb{R};
\end{aligned}$$

the problem remains convex because the expressions within moduli are linear in the parameter θ, and the results about the violation probability transport without much effort.

The only subtle point remains the *exact* number of support constraints: when the data (x_i, y_i) are *continuous*, i.e. distributed according to a density, it could be shown that for a large family of regressor functions φ the support constraint are *almost surely* $d+1$, hence the violation probability is *exactly* $\frac{d+1}{N+1}$. On the other hand, if the data are not distributed according to a density (for example if y_i takes a certain value $\bar{y}$ with non-zero probability), then the support constraints may be *less* than $d+1$, and the violation probability is not exact. However, it satisfies an inequality in the "good" direction:

$$\mathsf{P}\left[\left|\varphi(x_{N+1})^\top \theta^* - y_{N+1}\right| > \gamma^*\right] \leq \frac{d+1}{N+1}$$

so that, given x_{N+1}, the following

$$[a, b] = \left[\varphi(x_{N+1})^\top \theta^* - \gamma^*,\ \varphi(x_{N+1})^\top \theta^* + \gamma^*\right]$$

is a prediction interval for y_{N+1} with probability *at least* $1 - \frac{d+1}{N+1}$.

It is also possible to construct prediction intervals with exact probability when certain unrealistic samples are discarded as "outliers". For a broad and rigorous exposition of the subject, you can refer to the research paper [6] and to the references therein.

A Brief reminder of linear algebra

This appendix is a collection of definitions and facts in no particular order; it is only meant as a refresher, and most of the material covered should already belong to the background of any student in information engineering. There is absolutely no pretension of mathematical rigor here.
References for linear algebra: [16] (intermediate), [10] (intermediate), [29] (intermediate to advanced), [14] (advanced).

A.1 Subspaces

Definition A.1.1 *We say that a set $V \subset \mathbb{R}^p$ is a* subspace *of $\mathbb{R}^p$ if V is closed with respect to the operations of addition and multiplication by a scalar that hold in $\mathbb{R}^p$. In other terms, V is a subspace if*

$$v_1, v_2 \in V \Rightarrow \alpha v_1 + \beta v_2 \in V$$

for all $\alpha, \beta \in \mathbb{R}$.

Definition A.1.2 *Let $v_1, \cdots, v_n \in \mathbb{R}^p$. We denote*

$$\text{span } \{v_1, \cdots, v_n\} = \left\{ \sum_{i=1}^{n} a_i v_i \;\middle|\; a_1, \cdots, a_n \in \mathbb{R} \right\}$$

the subset of $\mathbb{R}^p$ made of all the linear combinations of $v_1, \cdots, v_n$.

As an exercise, you should prove that span $\{v_1, \cdots, v_n\}$ is actually a *subspace* of $\mathbb{R}^p$. If $V =$ span $\{v_1, \cdots, v_n\}$, we also say that the vectors $v_1, \cdots, v_n$ *generate* V.

Definition A.1.3 *Let $v_1, \cdots, v_n \in \mathbb{R}^p$. We say that $v_1, \cdots, v_n$ are* linearly independent *if the* only *linear combination that yields zero,*

$$\sum_{i=1}^{n} a_i v_i = 0,$$

is the one with zero coefficients $a_1 = \cdots = a_n = 0$. If $v_1, \cdots, v_n$ are not linearly independent, we call them linearly dependent*.*

Proposition A.1.1 *Any $p + 1$ vectors in $\mathbb{R}^p$ are linearly dependent. In other words, if $v_1, \cdots, v_n$ are* linearly independent *then $n \leq p$.*

Definition A.1.4 *If $v_1, \cdots, v_n \in \mathbb{R}^p$ are* linearly independent *and $V =$* span $\{v_1, \cdots, v_n\}$*, we say that the set $\{v_1, \cdots, v_n\}$ is a* basis *of the subspace V. In particular, if $n = p$ and* span $\{v_1, \cdots, v_n\} = \mathbb{R}^p$*, the set $\{v_1, \cdots, v_n\}$ is a basis of $\mathbb{R}^p$.*

Proposition A.1.2 *If $v_1, \cdots, v_n \in \mathbb{R}^p$ are linearly independent, then there exist $v_{n+1}, \cdots, v_p \in \mathbb{R}^p$ such that $v_1, \cdots, v_p$ is a basis of $\mathbb{R}^p$.*

In words, every linearly independent subset of $\mathbb{R}^p$ can be extended to form a basis.

These definitions (span, linear independence, basis) generalize quite naturally to abstract vector spaces V, real or complex. Then one can show that if V has a finite basis $\{v_1, \cdots, v_n\}$ with n elements, then every other basis of V has the same number of elements. V is then said to be *finite-dimensional*; the number n is called the *dimension* of V, and denoted $\dim V$. In particular, $\dim \mathbb{R}^p = p$, because $\mathbb{R}^p$ admits at least the following basis of p vectors:

$$\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \cdots, \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix},$$

called the *canonical basis*.
Exercise: prove that the above canonical "basis" is really a basis of $\mathbb{R}^p$.

A.2 Scalar products and orthogonality

Definition A.2.1 *A* scalar product *in a real vector space V is a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ that satisfies these properties:*

- $\langle x, x \rangle \geq 0$ *for all $x \in V$, and* $\langle x, x \rangle = 0$ *if and only if* $x = 0$ *(positive definiteness);*
- $\langle x, \alpha y + \beta z \rangle = \alpha \langle x, y \rangle + \beta \langle x, z \rangle$ *for all $x, y, z \in V$ and $\alpha, \beta \in \mathbb{R}$ (linearity);*
- $\langle x, y \rangle = \langle y, x \rangle$ *for all $x, y \in V$ (symmetry).*

The following is the prototypical scalar product in $\mathbb{R}^p$:

$$\langle x, y \rangle = x^\top y, \qquad x, y \in \mathbb{R}^p$$

Definition A.2.2 *Similarly, a scalar product in a complex vector space V is a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{C}$ that satisfies these properties:*

- *for all $x \in V$ it holds $\langle x, x \rangle \in \mathbb{R}$, $\langle x, x \rangle \geq 0$, and $\langle x, x \rangle = 0$ if and only if $x = 0$;*
- $\langle x, \alpha y + \beta z \rangle = \alpha \langle x, y \rangle + \beta \langle x, z \rangle$ *for all $x, y, z \in V$ and $\alpha, \beta \in \mathbb{C}$;*
- $\langle x, y \rangle = \overline{\langle y, x \rangle}$ *for all $x, y \in V$.*

The only difference with the real case is the conjugation in the "symmetry" property. The following is the prototypical scalar product in $\mathbb{C}^p$:

$$\langle x, y\rangle = x^* y, \qquad x, y \in \mathbb{C}^p$$

where $*$ denotes transpose-conjugate. Of course, when both x and y happen to be real, the scalar product in $\mathbb{C}^p$ coincides with the previous one in $\mathbb{R}^p$.

Definition A.2.3 *A* norm *in a vector space V (real or complex) is a function $\|\cdot\| : V \to \mathbb{R}$ that satisfies these properties:*

- $\|x\| \geq 0$ *for all* $x \in V$, *and* $\|x\| = 0$ *if and only if* $x = 0$*;*
- $\|\alpha x\| = |\alpha|\ \|x\|$ *for all* $x \in V$ *and* $\alpha \in \mathbb{R}$*;*
- $\|x + y\| \leq \|x\| + \|y\|$ *for all* $x, y \in V$ *(triangular inequality).*

Proposition A.2.1 *Let $\langle\cdot,\cdot\rangle$ be a scalar product in a vector space V (real or complex). Then the following:*

$$\|x\|_2 = \sqrt{\langle x, x\rangle}$$

is a well-defined norm. In the vector spaces $\mathbb{R}^p$ and $\mathbb{C}^p$, if the scalar product is defined respectively as $\langle x, y\rangle = x^\top y$ or $\langle x, y\rangle = x^* y$, this norm is called the Euclidean *norm.*

Definition A.2.4 *Two vectors $v, w \in \mathbb{R}^p$ are* orthogonal *if $\langle v, w\rangle = 0$. This is denoted $v \perp w$.*

Theorem A.2.1 *(Pythagoras). If v and w are orthogonal vectors, then*

$$\|v + w\|^2 = \|v\|^2 + \|w\|^2.$$

Proof.

$$\begin{aligned} \|v + w\|^2 &= \langle v + w, v + w\rangle = \langle v, v\rangle + \langle v, w\rangle + \langle w, v\rangle + \langle w, w\rangle \\ &= \langle v, v\rangle + \langle w, w\rangle = \|v\|^2 + \|w\|^2. \end{aligned}$$

□

Proposition A.2.2 *If the nonzero vectors $\{v_1, \cdots, v_n\}$ are orthogonal to each other, then they are linearly independent.*

Proof. Let $v_i \perp v_j$ for all $i, j = 1 \cdots n, i \neq j$. Suppose that

$$a_1 v_1 + \cdots + a_i v_i + \cdots + a_n v_n = 0.$$

Then, taking the scalar product with v_i,

$$\begin{aligned} 0 &= a_1 \langle v_i, v_1 \rangle + \cdots + a_i \langle v_i, v_i \rangle + \cdots + a_n \langle v_i, v_n \rangle \\ &= a_1 \cdot 0 + \cdots + a_i \|v_i\|^2 + \cdots + a_n \cdot 0 \\ &= a_i \|v_i\|^2. \end{aligned}$$

Since v_i is nonzero, it must be $a_i = 0$. Repeating for $i = 1, \cdots, n$ we get $a_1 = \cdots = a_n = 0$. Therefore the only linear combination yielding zero is the one with all the coefficients equal to zero, hence $\{v_1, \cdots, v_n\}$ are linearly independent by definition. □

Definition A.2.5 *Given a subset $S \subset \mathbb{R}^p$, the* orthogonal complement *of S in $\mathbb{R}^p$ is the set*

$$S^\perp = \{v \in \mathbb{R}^p \mid v \perp w \textit{ for all } w \in S\}.$$

Whatever set is S, $S^\perp$ is a subspace of $\mathbb{R}^p$; indeed if $v_1, v_2 \in S^\perp$, then for all $w \in S$, $\langle w, av_1 + bv_2 \rangle = a \langle w, v_1 \rangle + b \langle w, v_2 \rangle = 0$, so that $av_1 + bv_2 \in S^\perp$ as well.

Proposition A.2.3 *Let V be a subspace of $\mathbb{R}^p$. Then*[33]

$$\left(V^\perp\right)^\perp = V.$$

Theorem A.2.2 *Let W be a subspace of $\mathbb{R}^p$. Then every vector $v \in \mathbb{R}^p$ can be expressed in an unique way as*

$$v = w + w^\perp$$

where $w \in W$ and $w^\perp \in W^\perp$ [34].

The vector w is called the *orthogonal projection* of v on the subspace W (similarly, $w^\perp$ is the orthogonal projection of v on the subspace $W^\perp$).

A.3 Range and null space of a matrix

Definition A.3.1 *The* range *of a matrix $A \in \mathbb{R}^{n \times p}$ is the set*[35]

$$\text{range } A = \{v \in \mathbb{R}^n \mid \textit{there exists } w \in \mathbb{R}^p \textit{ such that } v = Aw\}.$$

Note that

$$\text{range } A = \text{span } \{\text{columns of } A\}.$$

[33] This is not true for a subspace V of any vector space H. It is essential that H is finite-dimensional, like $H = \mathbb{R}^p$.

[34] In other words, $\mathbb{R}^p$ is the *direct sum* of W and $W^\perp$.

[35] The range of A is the image of $\mathbb{R}^p$ under A understood as a linear mapping from $\mathbb{R}^p$ to $\mathbb{R}^n$. Hence it is also denoted Im A in the literature.

Definition A.3.2 *The* null space *of a matrix $A \in \mathbb{R}^{n\times p}$ is the set*[36]

$$\text{null } A = \{v \in \mathbb{R}^p \mid Av = 0\}.$$

Of course null A is a subspace of $\mathbb{R}^p$; indeed if $v_1, v_2 \in$ null A, then $A(av_1 + bv_2) = aAv_1 + bAv_2 = 0$, hence $av_1 + bv_2 \in$ null A as well. In other terms, the null space of A is orthogonal complement in $\mathbb{R}^p$ of the set of the columns of $A^\top$ (transposes of the rows of A):

$$\text{null } A = \left\{\text{columns of } A^\top\right\}^\perp.$$

Theorem A.3.1 *Let $A \in \mathbb{R}^{n\times p}$, understood as a linear mapping from $\mathbb{R}^p$ to $\mathbb{R}^n$. Then*

$$p = \dim \text{range } A + \dim \text{null } A.$$

Proposition A.3.1 *Let $A \in \mathbb{R}^{n\times p}$, understood as a linear mapping from $\mathbb{R}^p$ to $\mathbb{R}^n$. Then*

1. range $A = \left(\text{null } A^\top\right)^\perp$;
2. null $A = \left(\text{range } A^\top\right)^\perp$;
3. range $A^\top = (\text{null } A)^\perp$;
4. null $A^\top = (\text{range } A)^\perp$.

As an exercise, you should give an interpretation to each of these properties. For example, the first goes as follows. Let r_i be the i-th row of $A^\top$. Then null $A^\top = \{v \mid r_i v = 0 \text{ for all } i\} = \{v \mid c_i^\top v = \langle c_i, v\rangle = 0 \text{ for all } i\}$, where now c_i is the i-th column of A. Then null $A^\top$ is the orthogonal complement of span $\{$columns of $A\}$ = range A. Taking its orthogonal complement, we obtain range A.

Corollary A.3.1 *For any matrix $A \in \mathbb{R}^{m\times n}$,*

$$\text{range } A = \text{range } AA^\top$$

Proof. Suppose that $v \in$ null $A^\top$. This means $A^\top v = 0$, hence also $AA^\top v = 0$ and $v \in$ null $AA^\top$. Suppose, on the other hand, that $v \in$ null $AA^\top$. Then $AA^\top v = 0$, hence also $\|A^\top v\|_2^2 = (A^\top v)^\top A^\top v = v^\top AA^\top v = 0$. This implies that $A^\top v = 0$ and $v \in$ null $A^\top$. Hence null $A^\top$ = null $AA^\top$, and by the previous Proposition,

$$\text{range } A = (\text{null } A^\top)^\perp = (\text{null } AA^\top)^\perp = \text{range } AA^\top.$$

Note that this proof is nothing more than a compact form of the proof of Lemma 1.4.1, if we let $A = \Phi^\top = \left[\begin{array}{ccc} \varphi_1 & \cdots & \varphi_N \end{array}\right]$. □

[36] In the literature, the null space of A is also called the *kernel* of A, and denoted Ker A.

Definition A.3.3 *Let $A \in \mathbb{R}^{n \times p}$. The* rank *of A is the maximum dimension of a square matrix, obtained from A by suppressing some rows and/or columns, with nonzero determinant.*

The following characterizations are more intuitive and more useful:

Proposition A.3.2 *The rank of A is equal to:*

- *the dimension of the subspace of $\mathbb{R}^n$ generated by its columns:*

$$\begin{aligned} \text{rank } A &= \dim \operatorname{span} \{\text{columns of } A\} \\ &= \dim \operatorname{range} A; \end{aligned}$$

- *the dimension of the subspace of $\mathbb{R}^p$ generated by its rows:*

$$\begin{aligned} \text{rank } A &= \dim \operatorname{span} \{\text{rows of } A\} \\ &= \dim \operatorname{span} \{\text{columns of } A^\top\} \\ &= \dim \operatorname{range} A^\top. \end{aligned}$$

If $A \in \mathbb{R}^{n \times p}$, where $n \geq p$ ("tall" matrix, i.e. more rows than columns), we say that A has *full rank* if rank $A = p =$ number of columns. Then the columns of A are linearly independent, and the subspace of $\mathbb{R}^n$ generated by them has dimension p (the maximum possible).

Conversely, if $n \leq p$ ("flat" matrix, i.e. more columns than rows), we say that A has *full rank* if rank $A = n =$ number of rows. Then the *rows* of A are linearly independent, and their span has dimension n.

In particular, if $A \in \mathbb{R}^{p \times p}$ (square), the following statements are equivalent:

- A has full rank ($= p$);
- the columns of A are linearly independent, and form a basis of $\mathbb{R}^p$;
- the columns of $A^\top$ (transposes of the rows of A) are linearly independent, and form a basis of $\mathbb{R}^p$;
- A is invertible, i.e. non-singular, its determinant is nonzero, etc.

A.4 Eigenvalues, eigenvectors, and diagonalization

Definition A.4.1 *A real square matrix $A \in \mathbb{R}^{p \times p}$ is called:*

- symmetric*, if it coincides with its transpose: $A = A^\top$. The rows of a symmetric matrix, taken in order, are the transposes of its columns, taken in the same order.*

- orthogonal*, if it is invertible and its inverse coincides with its transpose: $AA^\top = A^\top A = I_p$. The columns of an orthogonal matrix form an orthonormal basis of $\mathbb{R}^p$, and so do its rows. Orthogonal transformations preserve scalar products and Euclidean norms in $\mathbb{R}^p$: indeed*

  $$\langle Ax, Ay\rangle = (Ax)^\top Ay = x^\top A^\top Ay = x^\top y = \langle x, y\rangle$$
  $$\|Ax\|_2 = \sqrt{\langle Ax, Ax\rangle} = \sqrt{\langle x, x\rangle} = \|x\|_2$$

  (In geometric language, one says that orthogonal mappings, such as rotations and reflections, preserve angles and lengths.)

- normal*, if it commutes with its transpose: $AA^\top = A^\top A$. In particular, symmetric and orthogonal matrices are normal.*

Correspondingly, a complex square matrix $A \in \mathbb{C}^{p\times p}$ is called:

- Hermitian*, if it coincides with its transpose-conjugate: $A = A^*$. The rows of a Hermitian matrix, taken in order, are the transpose-conjugates of its columns, taken in the same order.*

- unitary*, if it is invertible and its inverse coincides with its transpose-conjugate: $AA^* = A^*A = I_p$. The columns of a unitary matrix form an orthonormal basis of $\mathbb{C}^p$, and so do its rows. Unitary transformations preserve scalar products and Euclidean norms in $\mathbb{C}^p$.*

- normal*, if it commutes with its transpose-conjugate: $AA^* = A^*A$. In particular, Hermitian and unitary matrices are normal in the complex sense.*

Definition A.4.2 *Let $A \in \mathbb{C}^{p\times p}$. If there exist a nonzero vector $v \in \mathbb{C}^p$ and a complex number λ such that*

$$Av = \lambda v,$$

then λ is called an eigenvalue *of A, and v an* eigenvector *of A corresponding to that eigenvalue.*

The above requirement about $v \neq 0$ is the same as the following: the linear system

$$(\lambda I - A)v = 0$$

admits a nonzero solution v. For this to hold, the matrix $\lambda I - A$ must be non-invertible (or “singular”), i.e.

$$\det(\lambda I - A) = 0.$$

The function $\chi(\lambda) = \det(\lambda I - A)$ is a polynomial in the variable λ, called the *characteristic polynomial of A*; its (complex) roots are the eigenvalues of A.

Definition A.4.3 *A complex matrix $A \in \mathbb{C}^{p\times p}$ is called* diagonalizable *if it admits a decomposition*

$$A = M\Lambda M^{-1}$$

where M is an invertible matrix (in general, complex) and Λ is a diagonal matrix (in general, with complex entries on the diagonal).

The same terminology applies to *real* matrices $A \in \mathbb{R}^{p\times p}$, but while complex eigenvalues and complex eigenvectors always exist (in particular eigenvalues always exist because any polynomial of degree ≥ 1 has at least one root, by the fundamental theorem of algebra), such eigenvalues and eigenvectors are real only in particular cases. In the same fashion, it may very well happen that a real matrix A is diagonalizable with complex M and Λ, but not with real M and Λ. For example,

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -i & i \end{bmatrix} \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -i & i \end{bmatrix}^{-1}$$

admits a complex diagonalization but not a real one, because its eigenvalues are complex.
Anyway, any such diagonal decomposition is of paramount importance. Indeed, suppose that $A \in \mathbb{R}^{p\times p}$ is diagonalizable with real M and Λ. If we let $m_1, \cdots, m_p$ be the columns of M and we multiply on the right-hand side by M,

$$AM = M\Lambda = M \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_p \end{bmatrix}$$

which, read column by column, means

$$Am_i = \lambda_i m_i$$

In other words, any diagonal decomposition yields a basis of $\mathbb{R}^p$ made of eigenvectors of A. Any column m_i is an eigenvector of A, and λ_i is the corresponding eigenvalue.

Definition A.4.4 *We call a complex matrix $A \in \mathbb{C}^{p\times p}$* unitarily diagonalizable *if it is diagonalizable with M unitary. Since the inverse of an unitary matrix is its transpose-conjugate, the decomposition reads:*

$$A = M\Lambda M^*.$$

One of the most fundamental results in finite-dimensional linear algebra is the following:

Theorem A.4.1 *(Spectral theorem). A matrix $A \in \mathbb{C}^{p \times p}$ is unitarily diagonalizable if and only if it is normal.*

As a particular case, then, Hermitian matrices are unitarily diagonalizable; but in this case we can say more. Indeed, if A is Hermitian, v is an eigenvector, and λ is the corresponding eigenvalue, then

$$\lambda \|v\|^2 = \lambda v^* v = v^* A v = v^* A^* v = (Av)^* v = (\lambda v)^* v = \bar{\lambda} v^* v = \bar{\lambda} \|v\|^2$$

Since any eigenvector is supposed to be nonzero, $\|v\|^2$ is a positive quantity, hence $\bar{\lambda} = \lambda$, that is, λ is real. Thus, the eigenvalues of a Hermitian matrix are real.
An analogous result holds in the real case. The eigenvalues of a symmetric matrix $A \in \mathbb{R}^{n \times n}$ are real; in this case, moreover, there exist an orthonormal basis of $\mathbb{R}^n$ made of *real* eigenvectors of A, therefore A admits a decomposition

$$A = M \Lambda M^\top$$

where Λ is diagonal with real entries, and M is real and orthogonal.

A.5 Positive semi-definite matrices

Definition A.5.1 *A symmetric matrix $P = P^\top \in \mathbb{R}^{p \times p}$ is called positive semi-definite if for all $v \in \mathbb{R}^p$ it holds $v^\top P v \geq 0$; this property is denoted $P \geq 0$.*
If, in addition, $v^\top P v > 0$ for all $v \neq 0$, then P is called positive definite; this property is denoted $P > 0$.

A positive semi-definite matrix P is diagonalizable, because it is supposed to be symmetric; moreover, the entries on the diagonal matrix (that is, the eigenvalues of P) are greater than or equal to zero. Indeed, if v is an eigenvector and λ the corresponding eigenvalue,

$$\lambda \|v\|^2 = \lambda v^\top v = v^\top \lambda v = v^\top A v \geq 0$$

Since any eigenvector is supposed to be nonzero, $\|v\|^2$ is a positive quantity, hence $\lambda \geq 0$.
A similar proof shows that the eigenvalues of a positive-definite matrix must be strictly positive. In this case, the diagonal decomposition reads

$$P = M \Lambda M^\top$$

where the diagonal entries of Λ are strictly positive, hence Λ is invertible. Consequently, any positive definite matrix P is invertible.

For any positive semi-definite matrix $P = M\Lambda M^\top$, where

$$\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{bmatrix},$$

we define the *square root of* P as follows:

$$P^{1/2} := M \begin{bmatrix} \sqrt{\lambda_1} & & \\ & \ddots & \\ & & \sqrt{\lambda_p} \end{bmatrix} M^\top.$$

Given P, $P^{1/2}$ is the unique positive semi-definite matrix such that $\left(P^{1/2}\right)^2 = P$. If P is positive *definite*, hence invertible, then $P^{1/2}$ is also invertible, and we denote its inverse as $P^{-1/2}$ (as an exercise, define $P^{-1/2}$ in terms of $M, M^\top$, and $\lambda_1, \cdots, \lambda_p$).

Positive semi-definiteness induces a partial ordering between square matrices; for any $A, B \in \mathbb{R}^{p\times p}$ we write $A \geq B$ (equivalently, $B \leq A$) whenever $A - B \geq 0$, and $A > B$ (equivalently, $B < A$) whenever $A - B > 0$. This ordering is transitive:

- if $A \geq B$ and $B \geq C$, then $A \geq C$;
- if $A \geq B$ and $B > 0$, then $A > 0$ (hence both A and B are invertible);

Proposition A.5.1 *If $A \geq B$ and $B > 0$, then $A^{-1} \leq B^{-1}$.*

Proof. Suppose, first, that $B = I$, that is $A \geq I$. Multiplying both sides by $A^{-1/2}$, we obtain

$$B^{-1} = I = A^{-1/2}AA^{-1/2} \geq A^{-1/2}IA^{-1/2} = A^{-1}.$$

Now to the general case: if $A \geq B$, then multiplying by $B^{-1/2}$ on both sides,

$$B^{-1/2}AB^{-1/2} \geq I.$$

Hence, applying the previous case,

$$\left(B^{-1/2}AB^{-1/2}\right)^{-1} = B^{1/2}A^{-1}B^{1/2} \leq I,$$

and multiplying on each side by $B^{-1/2}$ we obtain the claim. □

A.6 Other matrix computations

Lemma A.6.1 *(matrix inversion lemma). Let A and C be square, invertible matrices. Then*

$$(A+BCD)^{-1} = A^{-1} - A^{-1}B\left(C^{-1}+DA^{-1}B\right)^{-1}DA^{-1}$$

for any two matrices B, D with compatible dimensions.

Proof. The simplest way to prove the assertion is to show that the right-hand side, multiplied by $A+BCD$, yields the identity. Indeed:

$$\begin{aligned}
&\Big(A^{-1} - A^{-1}B\left(C^{-1}+DA^{-1}B\right)^{-1}DA^{-1}\Big)(A+BCD)\\
&= A^{-1}A + A^{-1}BCD - A^{-1}B\left(C^{-1}+DA^{-1}B\right)^{-1}DA^{-1}A\\
&\qquad - A^{-1}B\left(C^{-1}+DA^{-1}B\right)^{-1}DA^{-1}BCD\\
&= I + A^{-1}BCD - A^{-1}B\Big(\left(C^{-1}+DA^{-1}B\right)^{-1} + \left(C^{-1}+DA^{-1}B\right)^{-1}DA^{-1}BC\Big)D\\
&= I + A^{-1}BCD - A^{-1}B\Big(\left(C^{-1}+DA^{-1}B\right)^{-1}C^{-1}C + \left(C^{-1}+DA^{-1}B\right)^{-1}DA^{-1}BC\Big)D\\
&= I + A^{-1}BCD - A^{-1}B\Big(\left(C^{-1}+DA^{-1}B\right)^{-1}\left(C^{-1}+DA^{-1}B\right)\Big)CD\\
&= I + A^{-1}BCD - A^{-1}BCD = I.
\end{aligned}$$

□

B Review of some concepts in mathematical analysis

References for mathematical analysis: [28] (intermediate).

B.1 Uniform convergence

Let C be a set, $f : C \to \mathbb{R}$ a function and $\{f_n\}_{n=1}^{\infty}$ a sequence of functions $f_n : C \to \mathbb{R}$.

Definition B.1.1 *We say that f_n converges to f* point-wise *if*

$$\lim_{n\to\infty} f_n(x) = f(x) \quad \textit{for all } x \in C.$$

Definition B.1.2 *We say that f_n converges to f* uniformly *if*

$$\lim_{n\to\infty} \sup_{x\in C} |f_n(x) - f(x)| = 0.$$

This means that for all $\varepsilon > 0$ there exists $N \geq 0$ such that, for all $n \geq N$, $|f_n(x) - f(x)| \leq \varepsilon$ over all the domain C (in words, the f_n are all "uniformly ε-close" to f after a certain index N). If the f_n converge to f uniformly, then they do so also point-wise (in general the converse is false). Indeed,

$$\lim_{n\to\infty} |f_n(x) - f(x)| \leq \lim_{n\to\infty} \sup_{x\in C} |f_n(x) - f(x)| = 0$$

Uniform convergence plays an important role in classical analysis, where it is often required to establish results on continuity and Riemann-integrability. For example, recall these properties from calculus courses:

Theorem B.1.1 *Let f_n be continuous on the interval $[a,b] \subset \mathbb{R}$, and suppose that $f_n \to f$ uniformly on $[a,b]$. Then*

- *f is continuous on $[a,b]$;*
- $\lim_{n\to\infty} \int_a^b f_n(x)dx = \int_a^b f(x)\ dx$.

B.2 Compactness

Definition B.2.1 *A subset C of a metric space X is called* compact *if from any sequence of points $x_n \subseteq C$ it is possible to extract a sub-sequence that converges to a point belonging to C.*

Here, “metric space” means any space in which there exist the notion of a distance, e.g. the vector space $\mathbb{R}^p$ endowed with the Euclidean distance $d(x, y) = \|x - y\|_2$ is a metric space.

Example. The closed interval $\bar{C} = [0, 1] \subset \mathbb{R}$ is compact. For instance, from the sequence $\frac{1}{2}, \frac{1}{2}, \frac{1}{3}, \frac{2}{3}, \frac{1}{4}, \frac{3}{4}, \cdots, \frac{1}{k}, \frac{k-1}{k}, \cdots$ we can extract the subsequence $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \cdots, \frac{1}{k}, \cdots$ which converges to $0 \in \bar{C}$. The open interval $C = (0, 1) \subset \mathbb{R}$ is *not* compact, because any subsequence extracted from the sequence $x_n = \frac{1}{n+1}$ converges to $0 \notin C$. The set $\mathbb{N} = \{0, 1, 2, 3, \cdots\} \subset \mathbb{R}$ is *not* compact, because any subsequence extracted from the sequence $x_n = 2n$ diverges. □

Theorem B.2.1 *(Heine/Borel). A subset C of a* finite-dimensional *vector space is compact if and only if it is closed and bounded.*

The hypothesis of finite dimensionality (think of $\mathbb{R}^p$) is crucial. For example, the set $\{x \mid \|x\| \leq 1\}$, which is compact in $\mathbb{R}^p$ by the Heine/Borel theorem, is *never* compact in any infinite-dimensional space, by a theorem of Riesz. Compactness is often required to establish the *existence* of a solution to some problem. For example, the following fundamental result generalizes a property that you already know from calculus courses:

Theorem B.2.2 *(Weierstrass). Let C be a subset of a metric space and $f : C \to \mathbb{R}$ a function. If C is compact and f is continuous, then there exist $\underline{x}, \bar{x} \in C$ such that*

$$f(\underline{x}) = \min_{x \in C} f(x),$$
$$f(\bar{x}) = \max_{x \in C} f(x).$$

(Here $\underline{\mathrm{x}}$ and $\bar{x}$ are, respectively, the so-called “arg min” and “arg max” of f over C.)

B.3 Convexity

Definition B.3.1 *A subset S of a vector space is called* convex *if, whenever the points x, y belong to S, the point $z_\lambda = \lambda x + (1 - \lambda)y$ also belongs to S for all $\lambda \in [0, 1]$. (Any such z_λ is called a* convex combination *of x and y.)*

Intuitively, this means that whenever $x, y \in S$, the whole line segment joining x and y is contained in S. For example:

- subspaces and their translations, hyperplanes etc. are convex sets;
- closed and open balls (i.e. sets of the form $\{x \in \mathbb{R}^p \mid \|x - c\| \leq r\}$ or $\{x \in \mathbb{R}^p \mid \|x - c\| < r\}$) are convex sets.

Lemma B.3.1 *An arbitrary intersection of convex sets is itself convex.*

Proof. Let $\{S_i\}_{i\in I}$ be convex sets, $S = \cap_{i\in I} S_i$, and let $\lambda \in [0,1]$. If $x, y \in S$, then x and y belong to each S_i; since each S_i is convex, $\lambda x + (1-\lambda)y$ also belongs to S_i for all i, and therefore it must belong to their intersection, which is S. Hence, S is convex. □

Definition B.3.2 *A function $f : S \to \mathbb{R}$ defined on a convex set S is called* convex *if, for all $x, y \in S$ and all $\lambda \in [0,1]$, it holds*

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y).$$

The function f is called strictly convex *if it is convex and the above inequality is strict ($<$) whenever $x \neq y$ and $0 < \lambda < 1$.*

For example:

- linear and affine functions ($f(x) = c^\top x + d$) are convex, indeed

$$\begin{aligned} f(\lambda x + (1-\lambda)y) &= c^\top(\lambda x + (1-\lambda)y) + d \\ &= c^\top(\lambda x + (1-\lambda)y) + (\lambda d + (1-\lambda)d) \\ &= \lambda(c^\top x + d) + (1-\lambda)(c^\top y + d) \\ &= \lambda f(x) + (1-\lambda) f(y) \end{aligned}$$

  (the inequality "$\le$" holds as "$=$");

- any norm is a convex function, indeed

$$\|\lambda x + (1-\lambda)y\| \le \|\lambda x\| + \|(1-\lambda)y\| = \lambda\|x\| + (1-\lambda)\|y\|.$$

- the square function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$, is strictly convex;

- any composition of a convex function f with an *affine* function $g(x) = c^\top x + d$ is convex. Indeed:

$$\begin{aligned} f\left(c^\top(\lambda x + (1-\lambda)y) + d\right) &= f\left(\lambda(c^\top x + d) + (1-\lambda)(c^\top y + d)\right) \\ &\le \lambda f(c^\top x + d) + (1-\lambda) f(c^\top y + d); \end{aligned}$$

- any sum of [strictly] convex functions is [strictly] convex.

Definition B.3.3 *The* epigraph *of a function $f : S \to \mathbb{R}$ is the subset of $S \times \mathbb{R}$ defined as follows:*

$$\text{Epi } f = \left\{(x, y) \in S \times \mathbb{R} \;\middle|\; f(x) \le y\right\}.$$

In words, the epigraph of f is the set of all points lying over its graph.

Lemma B.3.2 *A function $f : S \to \mathbb{R}$ is convex if and only if* Epi f *is a convex set.*

Proof. Suppose that f is convex, and let $(x_1, y_1), (x_2, y_2) \in \text{Epi } f$. This means that

$$\begin{aligned} f(x_1) &\le y_1, \\ f(x_2) &\le y_2; \end{aligned}$$

but then, for all $\lambda \in [0, 1]$,

$$\begin{aligned} f(\lambda x_1 + (1 - \lambda)x_2) &\le \lambda f(x_1) + (1 - \lambda) f(x_2) \\ &\le \lambda y_1 + (1 - \lambda) y_2, \end{aligned}$$

and this means that $\lambda(x_1, y_1) + (1 - \lambda)(x_2, y_2) \in \text{Epi } f$. Hence Epi f is convex.

Suppose, on the other hand, that Epi f is convex, and let $x_1, x_2 \in S$, $\lambda \in [0, 1]$. Then of course $(x_1, f(x_1)), (x_2, f(x_2)) \in \text{Epi } f$. Since Epi f is convex, then $\lambda(x_1, f(x_1)) + (1 - \lambda)(x_2, f(x_2)) \in \text{Epi } f$ or, which is the same,

$$(\lambda x_1 + (1 - \lambda)x_2,\ \lambda f(x_1) + (1 - \lambda) f(x_2)) \in \text{Epi } f,$$

and this in turn means that

$$f(\lambda x_1 + (1 - \lambda)x_2) \le \lambda f(x_1) + (1 - \lambda) f(x_2).$$

Thus, f is convex. □

Definition B.3.4 *The k-*sublevel set *of a function $f : S \to \mathbb{R}$, where $k \in \mathbb{R}$, is the subset of S defined as follows:*

$$S_k = \left\{ x \in S \;\middle|\; f(x) \le k \right\}.$$

Lemma B.3.3 *The sublevel sets of a convex function $f : S \to \mathbb{R}$ are convex sets.*

The converse is in general false (find a counterexample).

Proof. Suppose that f is convex and $k \in \mathbb{R}$, and let S_k be the k-sublevel set of f. If $x_1, x_2 \in S_k$ and $\lambda \in [0, 1]$, then

$$\begin{aligned} f(\lambda x_1 + (1 - \lambda)x_2) &\le \lambda f(x_1) + (1 - \lambda) f(x_2) \\ &\le \lambda k + (1 - \lambda) k = k, \end{aligned}$$

hence also $\lambda x_1 + (1 - \lambda)x_2 \in S_k$, and S_k is convex. □

Theorem B.3.1 *Suppose that $f : S \to \mathbb{R}$ is convex and differentiable over S (its gradient $\nabla f(\bar{x})$ exists at each point $\bar{x} \in S$). Then*

$$f(x) \geq f(\bar{x}) + \nabla f(\bar{x})^\top (x - \bar{x})$$

for all $x, \bar{x} \in S$.

Proof. See [2, p. 70]. □

Corollary B.3.1 *If $f : S \to \mathbb{R}$ is convex and differentiable over S, and a certain $\bar{x} \in S$ is such that $\nabla f(\bar{x}) = 0$, then $f(\bar{x}) \leq f(x)$ for all $x \in S$. In other words, $\bar{x}$ is a global minimum point for f.*

C Review of discrete-time linear, time-invariant systems

References for discrete-time systems: [22] (intermediate).

C.1 Discrete-time signals and systems

A *deterministic* discrete-time signal $\{u(t)\} = \{u(t)\}_{-\infty}^{+\infty}$ is a sequence of numbers $\cdots, u(-2), u(-1), u(0), u(1), u(2), \cdots$, infinite in both directions[37]. A signal $\{u(t)\}$ is called *bounded* if there exists a constant $K_u \geq 0$ such that

$$|u(t)| \leq K_u \quad \text{for all } t \in \mathbb{Z}.$$

The bounded signal

$$\delta(t) = \begin{cases} 1, & \text{if } t = 0, \\ 0, & \text{otherwise,} \end{cases}$$

is called the (discrete-time) *impulse*.

A signal $\{u(t)\}$ is called *summable* if

$$\sum_{t=-\infty}^{+\infty} |u(t)| = M < \infty.$$

The *convolution* of two signals $\{u(t)\}$, $\{v(t)\}$ is the signal $\{u * v(t)\}$ defined as follows:

$$u * v(t) := \sum_{\tau=-\infty}^{+\infty} u(t-\tau)v(\tau) = \lim_{T\to\infty} \sum_{\tau=-T}^{T} u(t-\tau)v(\tau); \tag{43}$$

the above expression is well-defined only if the series converges. The convolution '$*$' is an *operation* between sequences: it maps a pair of sequences into another sequence; as an operation, it has both the associative property ($(u * v) * w = u * (v * w)$) and the commutative property ($u * v = v * u$). Moreover, it possesses an *identity element*, which is precisely the impulse sequence:

$$\delta * u(t) = u * \delta(t) = \sum_{\tau=-\infty}^{+\infty} u(t-\tau)\delta(\tau) = \sum_{\tau=0} u(t-\tau) = u(t) \quad \text{for all } t,$$

that is $\delta * u = u * \delta = u$.

[37] The numbers can be real or complex; at one point, namely C.4, we will consider *complex* sequences of the form $u(t) = e^{j\omega t}$; in real-world applications, signals are usually real.

A *discrete-time linear system* is, basically, a linear map φ from a vector space $\mathcal{U}$ of discrete-time signals, called the *input* set, to another vector space $\mathcal{Y}$ of discrete-time signals, called the *output* set. To any sequence $\{u(t)\} \in \mathcal{U}$, φ associates another sequence $\{y(t)\} \in \mathcal{Y}$:

$$\varphi : \{\cdots, u(-1), u(0), u(1), \cdots\} \mapsto \{\cdots, y(-1), y(0), y(1), \cdots\}$$

The law governing some systems impose that the output sample $y(t)$ is a function of the sole input sample $u(t)$, and possibly of the time t, for all t:

$$y(t) = f(t, u(t)).$$

These systems are called *instantaneous transformations.* Despite the fact that they are indeed useful models for a lot of phenomena considered in control engineering and telecommunications (e.g. quantizers, saturations, "dead zones", etc.), they are not much interesting from our point of view.
In the cases that we are going to consider, the output sample $y(t)$ is in general a function of the whole input signal $\{u(t)\}$, or of part of it, and possibly of the time t; for instance, with loose notation,

$$y(t) = f(t, \cdots, u(t+1), u(t), u(t-1), \cdots) \quad \text{(time-varying, non-causal system)},$$
$$y(t) = f(u(t), u(t-1), \cdots) \quad \text{(time-invariant causal system)}.$$

These are called *dynamical* systems (from the Greek word "dynamis", *force*), because they are suitable models of physical systems subject to forces, accelerations and so on.
Among these, *linear dynamical systems* are of paramount importance in practically every branch of science and engineering. Recall that linearity means that the *superposition principle* holds: if

$$\varphi : \{u_1(t)\} \mapsto \{y_1(t)\} \qquad \text{and} \qquad \varphi : \{u_2(t)\} \mapsto \{y_2(t)\},$$

then for any constants a and b,

$$\varphi : \{au_1(t) + bu_2(t)\} \mapsto \{ay_1(t) + by_2(t)\}$$

("to the sum of the causes corresponds the sum of the effects"). The system is called *time-invariant* if to the translation in time of an input corresponds the translation in time of the output, with the same time lag τ: if

$$\varphi : \{u(t)\} \mapsto \{y(t)\}$$

then for any $\tau \in \mathbb{Z}$,

$$\varphi : \{u(t+\tau)\} \mapsto \{y(t+\tau)\}$$

A time-invariant system has the following property: the output corresponding to a certain input $\{u(t)\}$ is the convolution between $\{u(t)\}$ and the output $\{w(t)\} = \varphi\left[\{\delta(t)\}\right]$ corresponding to $\{\delta(t)\}$. Indeed, with loose notation,

$$\begin{aligned} y(t) &= \varphi\left[\{u(t)\}\right] = \varphi\left[\{\delta * u(t)\}\right] = \varphi\left[\left\{\sum_{\tau=-\infty}^{+\infty} \delta(t-\tau)u(\tau)\right\}\right] \\ &= \sum_{\tau=-\infty}^{+\infty} \varphi\left[\{\delta(t-\tau)\}\right]u(\tau) = \sum_{\tau=-\infty}^{+\infty} w(t-\tau)u(\tau) = w * u(t), \end{aligned} \tag{44}$$

where the fourth equality is an application of linearity that deliberately ignores convergence details (the sum is infinite!), and the fifth one is due to time-invariance. The sequence $\{w(t)\}$ is called the *impulse response* of the system. In what follows we will always refer to discrete-time systems that are both linear and time-invariant, and we will call them LTI systems for short.

C.2 Stability and causality

An LTI system is called *externally stable*, or *BIBO-stable*, if any bounded input signal is mapped to an output signal which is also bounded[38]. More precisely, if $|u(t)| \leq K_u$ for all t, then there exists K_y such that $|y(t)| \leq K_y$ for all t. It is easy to show that if the impulse response is summable, then the system is BIBO-stable[39]. Indeed, suppose that $\{w(t)\}$ is such that $\sum_{t=-\infty}^{+\infty} |w(t)| = M < \infty$ and that $|u(t)| \leq K_u$ for all t; then

$$\begin{aligned} |y(t)| &= \left|\sum_{\tau=-\infty}^{+\infty} u(\tau)w(t-\tau)\right| \leq \sum_{\tau=-\infty}^{+\infty} |u(\tau)|\,|w(t-\tau)| \\ &\leq K_u \sum_{\tau=-\infty}^{+\infty} |w(t-\tau)| = K_u M := K_y \end{aligned}$$

for all t, so that $\{y(t)\}$ is a bounded signal as well.

An BIBO-stable LTI system is called *causal* if whenever two input sequences satisfy

$$u_1(\tau) = u_2(\tau), \quad \tau = \cdots, t-2, t-1, t;$$

the corresponding output sequences satisfy

$$y_1(\tau) = y_2(\tau), \quad \tau = \cdots, t-2, t-1, t.$$

This means that the output $y(t)$ at a certain time t depends on the past samples of the input signal $\{u(\tau)\}, \tau = \cdots, t-2, t-1, t$, but not on its future samples $\{u(\tau)\}, \tau = t+1, t+2, \cdots$.

[38] BIBO stands for Bounded Input ⇒ Bounded Output.

[39] The converse is also true, but details are omitted here.

This definition implies, in particular, that the output corresponding to an input signal $\{u(t)\}$ such that $u(\tau) = 0$ tor all $\tau < t_0$ also satisfies $y(\tau) = 0$ tor all $\tau < t_0$: thus, if the input "starts" at t_0, so does the output. In view of time-invariance, it is customary to let always $t_0 = 0$, and to call *causal* also those signals that "start at 0". Thus, a causal LTI system has a causal *impulse response* $\{w(t)\}$, because the impulse is causal in the first place ($\delta(t) = 0$ for all $t < 0$). The response of a causal system to an arbitrary input $\{u(t)\}$ is

$$y(t) = w * u(t) = \sum_{\tau=-\infty}^{+\infty} w(t-\tau)u(\tau) = \sum_{\tau=-\infty}^{t} w(t-\tau)u(\tau), \tag{45}$$

because $w(t-\tau) = 0$ for $\tau > t$. Here $y(t)$ is well defined if and only if the series converges. If, moreover, the input sequence $\{u(t)\}$ is also causal, then

$$y(t) = \sum_{\tau=-\infty}^{t} w(t-\tau)u(\tau) = \sum_{\tau=0}^{t} w(t-\tau)u(\tau), \tag{46}$$

because $u(\tau) = 0$ for $\tau < 0$. Differently from (43), (44), and (45), the convolution (46) is *always* well-defined for all t, because the sum is finite.

C.3 Transforms and transfer functions

The Fourier transform of a signal $\{u(t)\}_{-\infty}^{+\infty}$ is the power series:

$$\begin{aligned}\hat{U}(\omega) = \mathcal{F}\left[\{u(t)\}\right](\omega) &:= \sum_{t=-\infty}^{+\infty} u(t)e^{-j\omega t} \\ &= \lim_{T\to\infty} \sum_{t=-T}^{T} u(t)e^{-j\omega t},\end{aligned} \tag{47}$$

where $\omega \in [-\pi, \pi]$. Such series may very well not converge, and the Fourier transform may not exist; a sufficient condition for the existence of $\hat{U}(\omega)$ for all ω is that the sequence is (absolutely) summable: if $\sum_{t=-\infty}^{+\infty} |u(t)| < \infty$, then

$$\sum_{t=-\infty}^{+\infty} |u(t)e^{-j\omega t}| = \sum_{t=-\infty}^{+\infty} |u(t)|\ |e^{-j\omega t}| = \sum_{t=-\infty}^{+\infty} |u(t)|,$$

and (47) converges at all ω, since it converges also absolutely.

A Fourier transform, even that of a real signal, is in general a complex function of ω. It can therefore be expressed as $\hat{U}(\omega) = |\hat{U}(\omega)|\ e^{j\angle\hat{U}(\omega)}$.

However, if $\{u(t)\}$ is real its transform enjoys the following property,

$$\begin{aligned}\hat{U}(-\omega) &= \sum_{t=-\infty}^{+\infty} u(t)e^{j\omega t} = \sum_{t=-\infty}^{+\infty} \overline{u(t)}\ \overline{e^{-j\omega t}} \\ &= \overline{\sum_{t=-\infty}^{+\infty} u(t)e^{-j\omega t}} = \overline{\hat{U}(\omega)},\end{aligned}$$

called *Hermitian symmetry*. It follows at once that

$$\begin{aligned}|\hat{U}(-\omega)| &= |\hat{U}(\omega)|; \\ \angle\hat{U}(-\omega) &= -\angle\hat{U}(\omega).\end{aligned} \tag{48}$$

In words, the absolute value of the Fourier transform of a real signal is an even function, and its phase is an odd one. Also, it is immediate to show that its real part is even, and its imaginary part is odd.

The so-called $\mathscr{Z}$-transform of a signal $\{u(t)\}$ is the power series:

$$U(z) = \mathscr{Z}\left[\{u(t)\}\right](z) := \sum_{t=-\infty}^{+\infty} u(t)z^{-t}, \tag{49}$$

where $z \in \mathbb{C}$. As happens for the Fourier transform, the $\mathscr{Z}$-transform may not converge for any $z \in \mathbb{C}$ (take for example the sequence $u(t) = |t|!$); if, however, $\{u(t)\}$ is (absolutely) summable, then it converges at least on the *unit circle* $\{z \in \mathbb{C} \text{ s.t. } |z| = 1\} = \{e^{j\omega} \mid \omega \in [-\pi, \pi]\}$, and there it coincides with the Fourier transform, i.e. $\hat{U}(\omega) = U\left(e^{j\omega}\right)$.

The $\mathscr{Z}$-transform of a *causal* signal $\{u(t)\}$ is

$$U(z) = \mathscr{Z}\left[\{u(t)\}\right](z) = \sum_{t=0}^{+\infty} u(t)z^{-t}. \tag{50}$$

In this case, if the series converges for a certain $\bar{z} \in \mathbb{C}$, then it converges absolutely for all $z \in \mathbb{C}$ such that $|z| > |\bar{z}|$. Indeed, if $\sum_{t=0}^{+\infty} u(t)\bar{z}^{-t}$ converges, then the sequence $\{u(t)\bar{z}^{-t}\}$ must be bounded, that is $|u(t)\bar{z}^{-t}| \leq K$ for all t. But then, for all $|z| > |\bar{z}|$,

$$\sum_{t=0}^{+\infty} |u(t)z^{-t}| = \sum_{t=0}^{+\infty} |u(t)\bar{z}^{-t}| \left|\frac{z^{-t}}{\bar{z}^{-t}}\right| \leq K \sum_{t=0}^{+\infty} \left|\frac{\bar{z}}{z}\right|^{t} = \frac{K}{1 - |\bar{z}/z|} < \infty.$$

Hence, either the series does not converge for any $z \in \mathbb{C}$ (example: $u(t) = t!$), or it converges at least on an open region outside a disc, i.e. on a set of the form $\{z \in \mathbb{C} \text{ s.t. } |z| > R\}$. The minimum R for which this happens is called

convergence radius. If, in particular, $R < 1$, then (49) converges on the unit circle and the Fourier transform can be recovered as $\hat{U}(\omega) = U\left(e^{j\omega}\right)$.
The $\mathcal{Z}$-transform of the impulse response of an LTI causal system is called the *transfer function* of that system. Its Fourier transform is also called sometimes the transfer function or, depending on the context, the *frequency response* of the system.

Some fundamental facts about $\mathcal{Z}$-transforms follow:

1. The operator $\mathcal{Z}$ that maps sequences to transforms is *linear*: if $\{u_1(t)\}$ and $\{u_2(t)\}$ are signals, and a_1, a_2 are real constants, then

$$\begin{aligned}\mathcal{Z}\left[\{a_1u_1(t) + a_2u_2(t)\}\right](z) &= \sum_{t=-\infty}^{+\infty} (a_1u_1(t) + a_2u_2(t))z^{-t} \\ &= a_1 \sum_{t=-\infty}^{+\infty} u_1(t)z^{-t} + a_2 \sum_{t=-\infty}^{+\infty} u_2(t)z^{-t} \\ &= a_1U_1(z) + a_2U_2(z),\end{aligned}$$

provided that both $U_1(z)$ and $U_2(z)$ exist for some $z \in \mathbb{C}$; the transform of the linear combination exists at all such z. More generally,

$$\mathcal{Z}\left[\left\{\sum_{\tau=0}^{T} a_\tau u_\tau(t)\right\}\right](z) = \sum_{\tau=0}^{T} a_\tau U_\tau(z),$$

for all $z \in \mathbb{C}$ such that $U_\tau(z)$ exists for $\tau = 0, \cdots, T$. Even more generally

$$\begin{aligned}\mathcal{Z}\left[\left\{\sum_{\tau=0}^{+\infty} a_\tau u_\tau(t)\right\}\right](z) &= \sum_{t=-\infty}^{+\infty}\sum_{\tau=0}^{+\infty} a_\tau u_\tau(t)z^{-t} \\ &= \sum_{\tau=0}^{+\infty} a_\tau U_\tau(z),\end{aligned}$$

provided that the two limits in the series can be interchanged. We omit further details; we will use this fact in the following point, without worrying about convergence issues.

2. The transform of a convolution is the product of the respective transforms. Consider for example the input-output relation of a causal LTI system with impulse response $\{w(t)\}$ (so that $y(t) = w * u(t)$, where $\{u(t)\}$ is the input signal and $\{y(t)\}$ the output signal). Ignoring tech-

nical details about convergence,

$$
\begin{aligned}
y(t) &= \sum_{\tau=0}^{+\infty} w(t-\tau)u(\tau) \\
Y(z) &= \sum_{t=0}^{+\infty} \left(\sum_{\tau=0}^{+\infty} w(t-\tau)u(\tau) \right) z^{-t} \\
&= \sum_{\tau=0}^{+\infty} \sum_{t=0}^{+\infty} w(t-\tau)u(\tau) z^{-(t-\tau+\tau)} \\
&= \sum_{\tau=0}^{+\infty} u(\tau) z^{-\tau} \sum_{t=0}^{+\infty} w(t-\tau) z^{-(t-\tau)};
\end{aligned}
$$

since $\{w(t)\}$ is causal all the terms in the inner sum for which $t-\tau < 0$ vanish, hence the inner sum starts from τ, and

$$
\begin{aligned}
Y(z) &= \sum_{\tau=0}^{+\infty} u(\tau) z^{-\tau} \sum_{t=\tau}^{+\infty} w(t-\tau) z^{-(t-\tau)} \\
&= \left(\sum_{\tau=0}^{+\infty} u(\tau) z^{-\tau} \right) \left(\sum_{t'=0}^{+\infty} w(t') z^{-t'} \right) \qquad (51) \\
&= W(z)U(z).
\end{aligned}
$$

3. If $U(z)$ is the transform of a causal signal $\{u(t)\}$, then $z^{-1}U(z)$ is the transform of its *delayed* version $\{\bar{u}(t)\}$, defined by $\bar{u}(t) := u(t-1)$ for all t. Indeed,

$$
\begin{aligned}
\bar{U}(z) &= \sum_{t=0}^{+\infty} \bar{u}(t) z^{-t} = \sum_{t=0}^{+\infty} u(t-1) z^{-t} = z^{-1} \sum_{t=0}^{+\infty} u(t-1) z^{-(t-1)} \\
&= z^{-1} \sum_{t'=0}^{+\infty} u(t') z^{-(t')} = z^{-1} U(z),
\end{aligned}
$$

where the fourth equality holds because $u(-1) = 0$. This fact tells us that, despite being a complex number in the original definition, z^{-1} can be interpreted as a *delay operator* acting on $\mathscr{Z}$-transforms. With a slight abuse of notation, we will write "z^{-1}" to denote a delay also when dealing with sequences, e.g. $\bar{u}(t) := z^{-1}u(t)$. [40]

[40] This is the discrete-time counterpart of the customary interpretation, in continuous-time models, of the complex variable s of Laplace transforms as a representative of the *derivative* operator.

C.4 The response to harmonic signals

Consider a BIBO-stable system. Its impulse response $\{w(t)\}$ must be summable ($\sum_{t=-\infty}^{+\infty} |w(t)| < \infty$), hence its frequency response $\hat{W}(\omega) = W\left(e^{j\omega}\right)$ must exist for every $\omega \in [-\pi, \pi]$.
A fundamental property of such a systems is that its response to *harmonic* signals (i.e. sinusoids, or sums of sinusoids) are also harmonic. To see this, consider first the response $\{y(t)\}_{-\infty}^{+\infty}$ to a complex input $\{u(t)\}_{-\infty}^{+\infty}$ of the form $u(t) = e^{j\omega t}$:

$$\begin{aligned}
y(t) = w * u(t) &= \sum_{\tau=-\infty}^{+\infty} w(t-\tau) e^{j\omega\tau} \\
&= \sum_{\tau=-\infty}^{+\infty} w(t-\tau) e^{-j\omega(t-\tau)} e^{j\omega t} \\
&= e^{j\omega t} \sum_{\tau'=-\infty}^{+\infty} w(\tau') e^{-j\omega\tau'} = \hat{W}(\omega)\ e^{j\omega t}.
\end{aligned}$$

This property is too important to let it pass without re-stating it in the proper, magnificent, *linear-algebraic* language. A BIBO-stable LTI system is a linear operator φ mapping sequences to sequences. For all the sequences $u(t) = e^{j\omega t}$ it holds, with loose notation,

$$\varphi\left[e^{j\omega t}\right] = \hat{W}(\omega)\ e^{j\omega t};$$

and here is the statement:

any harmonic sequence of the form $e^{j\omega t}$ is an eigenvector (or "eigenfunction") of φ, having $\hat{W}(\omega)$ as the corresponding eigenvalue.

Since the property holds for any $\omega \in [-\pi, \pi]$, it also does for $-\omega$; if $u(t) = e^{-j\omega t} = e^{j(-\omega)t}$, then:

$$y(t) = \hat{W}(-\omega) e^{-j\omega t}.$$

Now let $\{u(t)\}_{-\infty}^{+\infty}$ be a sinusoidal signal with frequency ω:

$$\begin{aligned}
u(t) &= A\cos(\omega t + \varphi) \\
&= A\frac{e^{j(\omega t+\varphi)} + e^{-j(\omega t+\varphi)}}{2} = \frac{A}{2} e^{j\varphi} e^{j\omega t} + \frac{A}{2} e^{-j\varphi} e^{-j\omega t};
\end{aligned}$$

in view of linearity, the corresponding response is

$$\begin{aligned}
y(t) &= \frac{A}{2}e^{j\varphi}\ \hat{W}(\omega)\ e^{j\omega t} + \frac{A}{2}e^{-j\varphi}\ \hat{W}(-\omega)\ e^{-j\omega t} \\
&= |\hat{W}(\omega)|\ \frac{A}{2}\left(e^{j\varphi}\ e^{j\angle\hat{W}(\omega)}\ e^{j\omega t} + Ae^{-j\varphi}\ e^{-j\angle\hat{W}(\omega)}\ e^{-j\omega t}\right) \\
&= |\hat{W}(\omega)|\ A\ \frac{e^{j(\omega t+\varphi+\angle\hat{W}(\omega))} + e^{-j(\omega t+\varphi+\angle\hat{W}(\omega))}}{2} \\
&= |\hat{W}(\omega)|\ A\cos(\omega t+\varphi+\angle\hat{W}(\omega)).
\end{aligned}$$

Thus, to a sinusoidal signal with frequency ω (non-causal, i.e. infinite in both directions), a BIBO-stable system responds with the same sinusoidal signal, amplified by $|\hat{W}(\omega)|$ and anticipated by $\angle\hat{W}(\omega)$.

What happens if the system is causal and the sinusoid is causal too, i.e. it "starts at 0"? Consider now a truncated exponential signal,

$$u(t) = \begin{cases} 0, & t < 0; \\ e^{j\omega t}, & t \geq 0. \end{cases}$$

It holds

$$\begin{aligned}
y(t) = w * u(t) &= \sum_{\tau=0}^{t} w(t-\tau)e^{j\omega\tau} = \sum_{\tau=0}^{t} w(t-\tau)e^{-j\omega(t-\tau)}e^{j\omega t} \\
&= e^{j\omega t}\sum_{\tau'=0}^{t} w(\tau')e^{-j\omega\tau'} = e^{j\omega t}\left(\hat{W}(\omega) - \sum_{\tau'=t+1}^{+\infty} w(\tau')e^{-j\omega\tau'}\right);
\end{aligned}$$

note that since $\hat{W}(\omega) = \sum_{\tau'=0}^{+\infty} w(\tau')e^{-j\omega\tau'}$ must exist finite, the infinite sum within parentheses must tend to 0 as $t \to \infty$; hence

$$y(t) = \hat{W}(\omega)e^{j\omega t} + s(t),$$

where $s(t)$ is a "transient" term, which tends to 0 as $t \to \infty$. The same reasoning can be done for the truncated version of $u(t) = e^{-j\omega t}$, hence the response to a sinusoidal signal "starting at 0"

$$u(t) = \begin{cases} 0, & t < 0, \\ A\cos(\omega t+\varphi), & t \geq 0, \end{cases}$$

is, by similar computations to the above ones,

$$y(t) = \begin{cases} 0, & t < 0; \\ |\hat{W}(\omega)|\ A\cos(\omega t+\varphi+\angle\hat{W}(\omega)) + \bar{s}(t), & t \geq 0, \end{cases}$$

where $\bar{s}(t)$ is another transient term.

In conclusion, now we have four distinct interpretations for the frequency response of a LTI BIBO-stable system:

1. $\hat{W}(\omega)$ is the Fourier transform of the impulse response of the system; provided that the transfer function $W(z)$ converges in a region of the complex plane that includes the unit circle, the frequency response is $\hat{W}(\omega) = W\left(e^{j\omega}\right)$;

2. provided that the $\mathscr{Z}$-transforms of both the input and the output of the system converge in a region of the complex plane including the unit circle (this happens if they are summable), in that region the transfer function is a proportionality factor linking them:

$$Y(z) = W(z)U(z),$$

and similarly

$$\hat{Y}(\omega) = \hat{W}(\omega)\hat{U}(\omega);$$

3. the values $\hat{W}(\omega)$ are the eigenvalues of the system φ, corresponding to the "eigenfunctions" $e^{j\omega t}$;

4. the response of the system to a sinusoid with frequency ω is the same sinusoid, amplified by the modulus $|\hat{W}(\omega)|$ of the frequency response and anticipated by its phase $\angle\hat{W}(\omega)$; if the system is causal and the sinusoid is fed at the input only starting from a certain time, the response "starts" at that time and approaches the amplified and anticipated sinusoid after a transient.

C.5 Difference equations

The causal LTI systems that are used in practice to model filters, sampled version of continuous-time systems etc., are usually denoted by so-called *difference equations.* These are equalities written in one of the following equivalent forms, depending on which is more convenient for ease of notation:

$$a_0 y(t) + a_1 y(t-1) + \cdots + a_n y(t-n) = b_0 u(t) + b_1 u(t-1) + \cdots + b_m u(t-m)$$
$$a_0 y(t) - a_1 y(t-1) - \cdots - a_n y(t-n) = b_0 u(t) + b_1 u(t-1) + \cdots + b_m u(t-m),$$

where $m \leq n$ and $a_0 \neq 0$. Without loss of generality, it is also customary to divide everything by a_0, obtaining a model where the first coefficient is 1; this being convenient for our purposes, we will work with the following model:

$$y(t) - a_1 y(t-1) - \cdots - a_n y(t-n) = b_0 u(t) + b_1 u(t-1) + \cdots + b_m u(t-m). \quad (52)$$

The linear model (52) "represents" a causal system if the system imposes that (52) holds at all times t. In particular, given a *causal* input, the corresponding output can be defined recursively:

$$y(t) = 0 \quad \text{for all } t < 0 \text{ by assumption (causality);}$$
$$y(t) = a_1 y(t-1) + \cdots + a_n y(t-n) + b_0 u(t) + \cdots + b_m u(t-m), \quad t \geq 0.$$

Here, $y(t)$ is well defined because we *assume* that the system is causal; hence $y(t)$ is a function of "past" samples $y(t-1), \cdots, y(t-n), u(t), \cdots, u(t-m)$, and not of future ones; but *nowhere* does an equation like (52) imply that the represented system is indeed *causal.* In other words, (52) is *not* "the system": it is a *property* of the system, and causality is another; the equation per se, taken alone, could very well represent a non-causal system.
Supposing that the output $\{y(t)\}$ possesses a $\mathscr{Z}$-transform as well as the input, $\{u(t)\}$, let transform both sides:

$$\begin{aligned} Y(z) - a_1 z^{-1} Y(z) - \cdots - a_n z^{-n} Y(z) &= b_0 U(z) + b_1 z^{-1} U(z) + \cdots + b_m z^{-m} U(z), \\ (1 - a_1 z^{-1} - \cdots - a_n z^{-n}) Y(z) &= (b_0 + b_1 z^{-1} + \cdots + b_m z^{-m}) U(z), \\ Y(z) &= \frac{b_0 + b_1 z^{-1} + \cdots + b_m z^{-m}}{1 - a_1 z^{-1} - \cdots - a_n z^{-n}} \, U(z). \end{aligned} \tag{53}$$

The function

$$\bar{W}(z) = \frac{b_0 + b_1 z^{-1} + \cdots + b_m z^{-m}}{1 - a_1 z^{-1} - \cdots - a_n z^{-n}} = \frac{b_0 z^n + b_1 z^{n-1} + \cdots + b_m z^{n-m}}{z^n - a_1 z^{n-1} - \cdots - a_n} \tag{54}$$

exists in the region where both the transforms $U(z)$ and $Y(z)$ converge (the open region outside a disc), and comparing with (51), it must coincide there with the transfer function $W(z)$ of the system. It can be rewritten as follows:

$$\bar{W}(z) = \frac{B(z)}{A(z)} = \frac{b_0 (z - z_1)(z - z_2) \cdots (z - z_n)}{(z - p_1)(z - p_2) \cdots (z - p_n)}, \tag{55}$$

The (complex) roots $p_1, \cdots, p_n$ of the polynomial $A(z) = z^n - a_1 z^{n-1} - \cdots - a_n$ are called the *poles* of $\bar{W}(z) = W(z)$, and the roots $z_1, \cdots, z_n$ of the polynomial $B(z) = b_0 z^n + b_1 z^{n-1} + \cdots + b_m z^{n-m}$ are called its *zeros.* It so happens that the region in which $W(z)$ converges is the set $\{z \in \mathbb{C} \mid |z| > |p_i|\}$, where p_i is the pole with maximum modulus; in other words, the maximum modulus among poles is the convergence radius. Consequently, the system is BIBO-stable if and only if the poles of $\bar{W}(z)$ belong to the open unit disc, that is, if all the poles p_i satisfy $|p_i| < 1$.
Again, the expression of $W(z)$ as a function of a complex variable is a *property* of the system, and it is not sufficient to describe it completely unless causality is assumed apart. Indeed, it is causality that dictates the region where $\bar{W}(z)$ is defined (and where $W(z)$ converges)[41].

[41] All of these concepts have their counterpart in the world of *causal* continuous-time systems, which you should remember from control courses: the difference equation (52) corresponds to an ordinary linear differential equation with constant coefficients; the left- and right-hand sides of the differential equation can be transformed according to Laplace, and the rational function that one obtains dividing the right-hand polynomial by the left-hand one happens to be the transfer function of the system. Such transfer function exists on a right half-plane *assuming that the system is causal*; and the continuous-time system

Models like (52) are called *finite-dimensional*, because they can be realized with a recursive algorithm using only a finite amount of memory. In view of the reasoning above, it is obvious that any causal LTI system satisfying (52) admits a *rational* transfer function. However,

- it is not at all true that every LTI causal system can be described by a finite-dimensional model. This is rather intuitive, because the true representative of an LTI system is its impulse response, an infinite sequence which cannot, in general, be reconstructed by an algorithm with a finite amount of memory. Stated another way, the impulse response may very well have a non-rational $\mathscr{Z}$-transform. Anyway, finite-dimensional models are of paramount importance in engineering, because they are far handier than other models for computation, estimation, prediction, identification, and closed-loop control. Moreover, the transfer function of any BIBO-stable system can be approximated by a rational one with arbitrary accuracy.

- to any finite-dimensional model there corresponds a unique rational transfer function, but the converse is false. Consider indeed $\bar{W}(z)$ written in the form (55): zero/pole cancellations may happen, and they correspond to a "hidden" dynamics of the system that either is not affected by the input, or is not visible at the output, or both. The proper way to understand such dynamics is through state-space system theory; however, the point here is that adding, so to say, "a zero and a pole at the same arbitrary position", two at a time, one leaves the transfer function unchanged, but obtaining larger and larger models (52): hence, to the same transfer function there correspond infinitely many models[42].

C.6 A simple example

Consider a causal LTI system satisfying to the following first-order difference equation:

$$y(t) = ay(t-1) + u(t) \tag{56}$$

is BIBO stable if and only if the poles of its transfer function lie in the open left-hand plane having the imaginary axis as its boundary. Indeed the unit circle plays for discrete-time systems the role that the imaginary axis plays for continuous-time ones, i.e. a sort of frontier between stable systems and unstable ones.

[42] Common sense tells that the smaller the model, the more useful it is for applications. There is, indeed, a model with minimum order n, called a "minimal realization", corresponding to any rational transfer function $W(z)$; namely, it is obtained from the rational function reduced to the lowest terms, all possible cancellations carried out; without loss of generality one can also assume that $a_0 = 1$: then the model is also unique. Such model has no "hidden" dynamics.

The impulse response of the system is

$$w(t) = \begin{cases} 0, & t < 0, \\ a^t, & t \geq 0. \end{cases}$$

This claim can be shown by induction. $w(t) = 0$ for $t < 0$ is by definition, since the system is causal. The base of the induction is $y(0) = y(-1)+\delta(0) = 1 = a^0$, because $y(-1) = 0$, the system being causal, and this is correct. Suppose now that $y(t) = a^t$ for a certain $t \geq 0$; then $y(t+1) = a \cdot a^t + \delta(t+1) = a^{t+1}$, and this concludes the proof.
The transfer function the system is

$$W(z) := \mathscr{Z}[w](z) = \sum_{t=0}^{\infty} a^t z^{-t} = \sum_{t=0}^{\infty} (az^{-1})^t = \lim_{t\to\infty} \frac{1-(az^{-1})^t}{1-az^{-1}};$$

for $|z| > |a|$ this series converges absolutely:

$$W(z) = \frac{1}{1-az^{-1}} = \frac{z}{z-a}, \qquad |z| > |a|. \tag{57}$$

$W(z)$ is a rational transfer function; its only pole is a. If, in particular, $|a| < 1$, then the impulse response $\{a^t\}_{t=0}^{\infty}$ is summable, the system is BIBO-stable, and the convergence region of W includes the unit circle. In this case, the Fourier transform of $\{w(t)\}$, i.e. the frequency response, can be recovered as $\hat{W}(\omega) = W\left(e^{j\omega}\right)$. Note that the expression of the transfer function (but *not* its convergence region) can be read promptly from (56), interpreting z^{-1} as a delay operator and rewriting the difference equation in symbolic form:

$$\begin{aligned} y(t) &= ay(t-1) + u(t) = az^{-1}y(t) + u(t); \\ y(t)&(1-az^{-1}) = u(t); \\ y(t) &= \frac{1}{1-az^{-1}}u(t) = W(z)u(t). \end{aligned}$$

To demonstrate that a difference equation per se cannot tell anything about causality, suppose that $a \neq 0$ and let us "flip" equation (56),

$$\begin{gathered} y(t-1) = \frac{1}{a}y(t) - \frac{1}{a}u(t) \\ \text{or} \\ y(t) = \frac{1}{a}y(t+1) - \frac{1}{a}u(t+1), \end{gathered} \tag{58}$$

and re-interpret it as an *anti-causal*, i.e. backward system: it is formally the same equation as (56), but now the present value of the output $y(t)$ is expressed as a function of the *future* values of the input $u(t+1)$ and the output $y(t+1)$ respectively. Its impulse response is

$$w(t) = \begin{cases} 0, & t > -1, \\ -a^t, & t \leq -1. \end{cases}$$

This can also be proven by (backward) induction. $w(t) = 0$ for $t > -1$ is by definition, since the system is anti-causal. The base of the induction is $w(-1) = \frac{1}{a}w(0) - \frac{1}{a}\delta(0) = -\frac{1}{a} = -a^{-1}$, which is correct. Now suppose that the expression is correct for a certain $t \leq -1$; then $w(t-1) = \frac{1}{a}w(t) - \frac{1}{a}\delta(t) = -\frac{a^t}{a} = -a^{t-1}$, and this concludes the proof.
The transfer function of the system is

$$\begin{aligned} W(z) &:= \sum_{t=-\infty}^{\infty} w(t)z^{-t} = \sum_{t=-\infty}^{-1} -a^t z^{-t} = -\sum_{t=-\infty}^{-1} (a^{-1}z)^{-t} \\ &= -\sum_{t=1}^{\infty}(a^{-1}z)^t = 1 - \sum_{t=0}^{\infty}(a^{-1}z)^t = 1 - \lim_{t\to\infty} \frac{1-(a^{-1}z)^t}{1-z/a} \end{aligned}$$

For $|z| < |a|$ the series converges absolutely, and

$$W(z) = 1 - \frac{1}{1-z/a} = \frac{-z/a}{1-z/a} = \frac{z}{z-a}, \qquad |z| < |a|.$$

As you can see, the expression of $W(z)$ is exactly the same as in (57), but its convergence region is complementary. In particular, the convergence region now includes the unit circle, and the Fourier transform of $w(t)$ exists, only if $|a| > 1$.
This example shows that neither a difference equation nor the expression of a transfer function determines the behavior of an LTI system; what truly characterizes the system is indeed only its impulse response. More complex examples would show that to the same rational transfer function can correspond *many* systems, many impulse responses, and many convergence regions; in general, a convergence region may be the complement of a disc (in this case it corresponds to a causal system, having a causal impulse response), or to a disc centered at the origin (anti-causal system/impulse response), or to a *ring* of the form $\{z \in \mathbb{C} \mid |a| < |z| < |b|\}$, like in a so-called *Laurent series expansion*; the boundaries of such regions are determined by the positions of the poles (e.g. a and b). If many convergence regions are possible, and one of them includes the unit circle, that region is the only one corresponding to a summable impulse response, i.e. to a BIBO-stable system.

D Brief reminder of probability theory

This appendix is a collection of definitions and facts in no particular order; it is only meant as a refresher, and most of the material covered should already belong to the background of any student in information engineering. There is absolutely no pretension of mathematical rigor here.
References on probability theory: [26] (basic to intermediate), [15] (intermediate to advanced), [8] and [9] (intermediate to advanced, and a true classic).

References on statistics: [21] (basic), [11] (basic to intermediate).

D.1 Probability

D.1.1 Notation

Let Ω be a set. For any subset $S \subseteq \Omega$, we denote S^c the *complement* of S with respect to Ω:

$$S^c := \Omega \setminus S = \{\omega \in \Omega \mid \omega \notin S\}.$$

For any subset $S \subseteq \Omega$, the *indicator function* of S is the function $\mathbb{1}_S : \Omega \to \{0, 1\}$ defined as follows:

$$\mathbb{1}_S(\omega) := \begin{cases} 1, & \omega \in S; \\ 0, & \text{otherwise.} \end{cases}$$

D.1.2 Probability spaces

Definition D.1.1 *Let Ω be a set. A family $\mathcal{F}$ of subsets of Ω is called a σ-algebra if it satisfies the following properties:*

- *$\varnothing \in \mathcal{F}$;*
- *if $S \in \mathcal{F}$, then $S^c \in \mathcal{F}$. In words, $\mathcal{F}$ must be closed with respect to complements;*
- *if $S_1, S_2, \dots \in \mathcal{F}$, then $\cup_{i=1}^{\infty} S_i \in \mathcal{F}$. In words, $\mathcal{F}$ must be closed with respect to countable unions.*

It follows from these axioms that a σ-algebra $\mathcal{F}$ is closed also with respect to countable *intersections* ($\cap_{i=1}^{\infty} S_i \in \mathcal{F}$). Also, taking $S_i = \varnothing$ for all $i > n$, it follows that $\mathcal{F}$ is closed with respect to *finite* unions and intersections ($\cup_{i=1}^{n} S_i \in \mathcal{F}$, $\cap_{i=1}^{n} S_i \in \mathcal{F}$). The set equipped with a σ-algebra, that is the pair $(\Omega, \mathcal{F})$, is called a *measurable space.*

Definition D.1.2 *Let $(\Omega, \mathcal{F})$ be a measurable space. A function $\mu : \mathcal{F} \to [0, \infty)$ is called a* measure *if, for any sequence of* pairwise disjoint *sets $S_1, S_2, \dots \in \mathcal{F}$, it holds*

$$\mu\left[\bigcup_{i=1}^{\infty} S_i\right] = \sum_{i=1}^{\infty} \mu[S_i].$$

This property is called σ-additivity*.*

Think of a measure as a generalization of the concept of "area" in the plane (or "volume" in the 3-dimensional space). The defining property tells us that the "area" of the union of disjoint subsets (i.e. "figures", even infinitely many) is the sum of the respective areas. Indeed, an "area" could be defined as a very particular measure μ on $(\Omega, \mathcal{F})$, where $\Omega = \mathbb{R}^2$ (the plane) and $\mathcal{F}$ is an adequately large class of subsets (for example $\mathcal{B}^2$ defined in Section D.2), such that $\mu(A)$ remains the same if the set $A \in \mathcal{F}$ is translated, rotated, or reflected through any axis.

Proposition D.1.1 *Let μ be a measure on $(\Omega, \mathcal{F})$. For any sequence of sets $S_1, S_2, \dots \in \mathcal{F}$ (not necessarily disjoint), it holds*

$$\mu\left[\bigcup_{i=1}^{\infty} S_i\right] \leq \sum_{i=1}^{\infty} \mu[S_i].$$

This property is called sub-additivity*.*

Definition D.1.3 *A measure $\mathsf{P} : \mathcal{F} \to [0, 1]$ such that $\mathsf{P}[\Omega] = 1$ is called a* probability*. A measurable space equipped with a probability, denoted $(\Omega, \mathcal{F}, \mathsf{P})$, is called a* probability space*. The set Ω is called the* sample space*, the subsets of Ω in $\mathcal{F}$ are called* events*, and if $S \in \mathcal{F}$, $\mathsf{P}[S]$ is called* the probability of the event *S.*

In a probability space $(\Omega, \mathcal{F}, \mathsf{P})$, for any sequence of *events* $S_1, S_2, \dots \in \mathcal{F}$ sub-additivity reads

$$\mathsf{P}\left[\bigcup_{i=1}^{\infty} S_i\right] \leq \sum_{i=1}^{\infty} \mathsf{P}\left[S_i\right].$$

Proposition D.1.2 *Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space.*

- *If $S_1 \subseteq S_2 \subseteq \cdots$ is an increasing sequence of events in $\mathcal{F}$, then*

$$\lim_{i \to \infty} \mathsf{P}[S_i] = \mathsf{P}\left[\bigcup_{i=1}^{\infty} S_i\right];$$

- *If* $S_1 \supseteq S_2 \supseteq \cdots$ *is a decreasing sequence of events in* $\mathcal{F}$*, then*

$$\lim_{i\to\infty} \mathsf{P}[S_i] = \mathsf{P}\left[\bigcap_{i=1}^{\infty} S_i\right].$$

Definition D.1.4 *If a certain property holds in a set* S *such that* $\mathsf{P}[S] = 1$ *(i.e.* $\mathsf{P}[\Omega \setminus S] = 0$*), we say that the property holds* almost surely *in* Ω*.*

D.2 Random variables and vectors

D.2.1 Measurable functions

Proposition D.2.1 *Let* $\Omega = \mathbb{R}$*. There exists a smallest* σ*-algebra* $\mathcal{B}$*, called the* Borel σ-algebra*, that contains all the open subsets of* $\mathbb{R}$ *(including* $\mathbb{R}$ *itself). Here, "smallest" means that* $\mathcal{B}$ *is a subset of every other* σ*-algebra containing the open sets. Finite sets, closed sets, countable unions of intervals of any kind, and much more complicated sets, all belong to* $\mathcal{B}$*. Thus,* $(\mathbb{R}, \mathcal{B})$ *is a measurable space.*

Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space, and let $X : \Omega \to \mathbb{R}$ be a function. For a certain subset A of $\mathbb{R}$, we denote $X^{-1}(A)$ the inverse image of A under X, i.e. the set

$$X^{-1}(A) := \{\omega \in \Omega \mid X(\omega) \in A\}.$$

In particular, $X^{-1}(\varnothing) = \varnothing$ and $X^{-1}(\mathbb{R}) = \Omega$.

Proposition D.2.2 *Inverse images preserve complements, countable unions, and countable intersections. If* $A_1, A_2, \dots \in \mathcal{B}$*,*

$$\begin{aligned} X^{-1}\left(A_i^c\right) &= X^{-1}\left(A_i\right)^c \\ X^{-1}\left(\bigcup_{i=1}^{\infty} A_i\right) &= \bigcup_{i=1}^{\infty} X^{-1}\left(A_i\right) \\ X^{-1}\left(\bigcap_{i=1}^{\infty} A_i\right) &= \bigcap_{i=1}^{\infty} X^{-1}\left(A_i\right) \end{aligned}$$

Therefore, the inverse images of all the sets belonging to $\mathcal{B}$ form a σ-algebra in Ω (verify!). A very important case is when the latter is contained in $\mathcal{F}$.

Definition D.2.1 *Let* $(\Omega, \mathcal{F}, \mathsf{P})$ *be a probability space. If* $X : \Omega \to \mathbb{R}$ *is such that*

$$X^{-1}(A) \in \mathcal{F}$$

for all $A \in \mathcal{B}$*, then* X *is said to be a* measurable function *or, in the language of probability, a* random variable*. The probability* $\mathsf{P}[X^{-1}(A)]$ *will be denoted, for short,* $\mathsf{P}[X \in A]$*. If we define for all* $A \in \mathcal{B}$

$$\mathsf{P}_X[A] := \mathsf{P}[X \in A],$$

then $(\mathbb{R}, \mathcal{B}, \mathsf{P}_X)$ *becomes a probability space.*

Indeed, what truly "matters" about random variables is not the set Ω on which they are defined, but the sets S in the σ-algebra $\mathcal{F}$ and their probabilities $\mathsf{P}[S]$. A random variable X "translates" the probabilities of the relevant events from $\mathcal{F}$ to $\mathcal{B}$, effectively turning $\mathbb{R}$ into a probability space standing by its own. For example, in the random experiment of tossing a coin, we let $\Omega = \{\text{head}, \text{tail}\}$, $\mathcal{F} = \{\varnothing, \{\text{head}\}, \{\text{tail}\}, \Omega\}$, and then $\mathsf{P}[\varnothing] = 0$, $\mathsf{P}[\{\text{head}\}] = \mathsf{P}[\{\text{tail}\}] = \frac{1}{2}$, and $\mathsf{P}[\Omega] = 1$. But then we may define the function $X : \Omega \to \mathbb{R}$ such that $X(\text{head}) = 0, X(\text{tail}) = 1$. X is measurable, that is, a random variable: indeed for any $A \in \mathcal{B}$

$$X^{-1}(A) = \begin{cases} \varnothing, & \text{if } 0 \notin A \text{ and } 1 \notin A \\ \{\text{head}\}, & \text{if } 0 \in A \text{ and } 1 \notin A \\ \{\text{tail}\}, & \text{if } 0 \notin A \text{ and } 1 \in A \\ \{\text{head}, \text{tail}\} = \Omega, & \text{if } 0, 1 \in A. \end{cases}$$

so that $X^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{B}$. It follows that, for any $A \in \mathcal{B}$, P_X is well defined, as follows:

$$\mathsf{P}_X[A] = \begin{cases} 0, & \text{if } 0 \notin A \text{ and } 1 \notin A \\ \frac{1}{2}, & \text{if } 0 \in A \text{ and } 1 \notin A \\ \frac{1}{2}, & \text{if } 0 \notin A \text{ and } 1 \in A \\ 1, & \text{if } 0, 1 \in A. \end{cases}$$

This definition makes $(\mathbb{R}, \mathcal{B}, \mathsf{P}_X)$ a probability space, which contains enough information about the coin-tossing experiment irrespective of what was the initial space Ω of heads and tails, which can indeed be "forgotten" for all practical purposes. (Please study this example and check the details until you get the idea.)

Proposition D.2.3 *Let X and Y be random variables on $(\Omega, \mathcal{F}, \mathsf{P})$, and $a \in \mathbb{R}$. Define point-wise the functions $X + Y$ and aX as follows, for all $\omega \in \Omega$:*

$$\begin{aligned} [X + Y](\omega) &:= X(\omega) + Y(\omega); \\ [aX](\omega) &:= aX(\omega). \end{aligned}$$

Then $X + Y$ and aX are also random variables. Thus, the set of all the random variables on $(\Omega, \mathcal{F}, \mathsf{P})$, endowed with the above operations, is a vector space.

Proposition D.2.4 *A necessary and sufficient condition for X to be a random variable is that $X^{-1}(A) \in \mathcal{F}$ for all the subsets $A \subset \mathbb{R}$ of the form $(-\infty, x]$.*

This happens because all the open sets in $\mathbb{R}$, and therefore all the Borel sets, can be obtained by means of finite or countable unions and intersections, or complements, of sets of the form $(-\infty, x]$.

Definition D.2.2 *Let X be a random variable on the probability space $(\Omega, \mathcal{F}, \mathsf{P})$. The function $F_X : \mathbb{R} \to [0,1]$ defined as follows:*

$$F_X(x) := \mathsf{P}[X^{-1}((-\infty, x])] = \mathsf{P}[X \in (-\infty, x]] = \mathsf{P}[X \leq x] \quad \text{(for short)}$$

is called the distribution *of the random variable X.*

Proposition D.2.5 *For any random variable X,*

- *F_X is nondecreasing;*
- *F_X is continuous from the right;*
- *$\lim_{x\to-\infty} F_X(x) = 0$;*
- *$\lim_{x\to+\infty} F_X(x) = 1$.*

Exercise: find the distribution of X in the coin-tossing experiment.

Definition D.2.3 *If there exists a function $f_X : \mathbb{R} \to [0, \infty)$ such that*

$$F_X(x) = \mathsf{P}[X \leq x] = \int_{-\infty}^{x} f_X(\xi)\ d\xi,$$

then X is called a continuous *random variable*[43]*, and f_X is called its* density*.*

This happens, *in particular*, if F_X is differentiable at each $x \in \mathbb{R}$; in this case, of course, f_X is its derivative. For example Gaussian variables, with which you are surely familiar, are continuous variables with an everywhere differentiable distribution, and with a density of the form

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{59}$$

[43] This usage of the word "continuous" is different from the one commonly adopted in mathematical analysis. Even if X is a continuous random variable, in no way this means that X is continuous as a function $\Omega \to \mathbb{R}$. Specifically, it does not mean anything like "$\lim_{\omega\to\bar{\omega}} X(\omega) = X(\bar{\omega})$", above all because Ω is any set without further (topological) structure, hence "$\omega \to \bar{\omega}$" means nothing. Instead, that X is continuous means that its *distribution function* enjoys a strong form of continuity (called absolute continuity) as a function $\mathbb{R} \to [0,1]$.

D.2.2 Random vectors

Proposition D.2.6 *In $\mathbb{R}^k$, there exists a smallest σ-algebra $\mathcal{B}^k$, called again the* Borel σ-algebra*, that contains all the open subsets of $\mathbb{R}^k$. Thus, $(\mathbb{R}^k, \mathcal{B}^k)$ is a measurable space.*

Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space, and let $X : \Omega \to \mathbb{R}^k$ be a function (it can be understood as a k-tuple of real functions $(X_1, \cdots, X_k)$, where $X_i : \Omega \to \mathbb{R}$ for all $i = 1 \cdots k$).

Definition D.2.4 *If $X : \Omega \to \mathbb{R}^k$ is such that $X^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{B}^k$, then X is said to be a* random vector*. The probability $\mathsf{P}[X^{-1}(A)]$ will be denoted, as before, $\mathsf{P}[X \in A]$.*

Proposition D.2.7 *A necessary and sufficient condition for X to be a random vector is that $X^{-1}(A) \in \mathcal{F}$ for all the subsets $A \subset \mathbb{R}$ of the form $(-\infty, x_1] \times \cdots \times (-\infty, x_k] \subset \mathbb{R}^k$.*

Definition D.2.5 *Let $X = (X_1, \cdots, X_k)$ be a random vector on $(\Omega, \mathcal{F}, \mathsf{P})$. The function $F_X : \mathbb{R}^k \to [0, 1]$ defined as follows:*

$$F_X(x_1, \cdots, x_k) := \mathsf{P}[X \in (-\infty, x_1] \times \cdots \times (-\infty, x_k]] = \mathsf{P}[X_1 \le x_1, \cdots, X_k \le x_k]$$

is called the distribution *of X, or the* joint distribution *of $X_1, \cdots, X_k$. The function*

$$F_i(x) := \lim_{x_j \to \infty,\ j=1\cdots i-1, i+1\cdots k} F(x_1, \cdots, x_{i-1}, x, x_{i+1}, \cdots, x_k)$$

is the distribution of the random variable X_i, also called, in this context, the marginal distribution *of X_i.*

Definition D.2.6 *If there exists a function $f_X : \mathbb{R}^k \to [0, \infty)$ such that*

$$\begin{aligned} F_X(x_1, \cdots, x_k) &= \mathsf{P}[X_1 \le x_1, \cdots, X_k \le x_k] \\ &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_k} f_X(\xi_1, \cdots, \xi_k)\ d\xi_1 \cdots d\xi_k, \end{aligned}$$

then X is called a continuous *random vector, and f_X is called its* density.

In what follows, whenever we say something like “let X and Y be random variables defined on the same probability space” we implicitly assume that they have a joint distribution, i.e. that (X, Y) is a random vector.

Definition D.2.7 *As we know, $(\mathbb{R}^k, \mathcal{B}^k)$ and $(\mathbb{R}^m, \mathcal{B}^m)$ are measurable spaces for any $k, m \ge 1$. A function $g : \mathbb{R}^k \to \mathbb{R}^m$ is said to be measurable (more precisely,* Borel-*measurable, if $g^{-1}(B) \in \mathcal{B}^k$ for all $B \in \mathcal{B}^m$.*

The measurability of $g : \mathbb{R}^k \to \mathbb{R}^m$ is exactly the same concept as the measurability of a random vector $X : \Omega \to \mathbb{R}^k$, except that there is still no measure attached to $(\mathbb{R}^k, \mathcal{B}^k)$.

Proposition D.2.8 *If $g : \mathbb{R}^k \to \mathbb{R}^m$ is a* continuous *function, then it is Borel-measurable.*

Proposition D.2.9 *If $X : \Omega \to \mathbb{R}^k$ is a random vector in $(\Omega, \mathcal{F}, \mathsf{P})$ and $g : \mathbb{R}^k \to \mathbb{R}^m$, then the composition $g(X)$ is a random vector (taking values in $\mathbb{R}^m$).*

In words, the composition of measurable functions is measurable. In particular ($k = m = 1$), if X is a random variable and $g : \mathbb{R} \to \mathbb{R}$ is continuous, then $g(X)$ is a random variable. So, for example, $|X|$, X^2, $(X-3)^2$, etc., are all random variables.

D.3 Expectation

D.3.1 Definition

Definition D.3.1 *Let X be a random variable on the probability space $(\Omega, \mathcal{F}, \mathsf{P})$. The* expectation *of X, denoted $\mathsf{E}[X]$, is defined as the "abstract integral"*

$$\mathsf{E}[X] := \int_\Omega X(\omega)\ d\mathsf{P}(\omega).$$

Do not worry if you do not know what an "abstract integral" is[44], i.e. if you have not had any class on modern (Lebesgue) integration, because in many

[44] It is not a big deal, after all. Let the sets $S_1, \cdots, S_n \in \mathcal{F}$. Any finite sum of weighted indicator functions $s(\omega) = \sum_{i=1}^n a_i \mathbb{1}_{S_i}(\omega)$, where $a_1, \cdots, a_n \in \mathbb{R}$, is called a *simple* function. The abstract integral of a simple function with respect to the measure P is defined as $\int_\Omega s(\omega)\ d\mathsf{P}(\omega) := \sum_{i=1}^n a_i \mathsf{P}[S_i]$. If $X : \Omega \to \mathbb{R}$ is a nonnegative measurable function ($X(\omega) \geq 0$ almost surely), then its integral is defined as $\int_\Omega X(\omega)\ d\mathsf{P}(\omega) := \sup_s \int_\Omega s(\omega)\ d\mathsf{P}(\omega)$, where the supremum is taken over all the simple functions s such that $s(\omega) \leq X(\omega)$ almost surely. For an arbitrary measurable function $X : \Omega \to \mathbb{R}$ we define $X^+(\omega) = \max\{X(\omega), 0\}$ and $X^-(\omega) = -\min\{X(\omega), 0\}$ (both X^+ and X^- are nonnegative measurable functions). Then, if both the integrals $I^+ = \int_\Omega X^+(\omega)\ d\mathsf{P}(\omega)$ and $I^- = \int_\Omega X^-(\omega)\ d\mathsf{P}(\omega)$ are finite, the integral of X is defined as $\mathsf{E}[X] = \int_\Omega X(\omega)\ d\mathsf{P}(\omega) := I^+ - I^-$, and X is called an *integrable* function. Usually one allows the integral of X to be $\pm\infty$, if one and only one among I^+ and I^- is infinite; on the other hand, if both I^+ and I^- are infinite, $\int_\Omega X(\omega)\ d\mathsf{P}(\omega)$ is not defined. For any set $S \in \mathcal{F}$, the integral over S is defined as $\int_S X(\omega)\ d\mathsf{P}(\omega) := \int_\Omega \mathbb{1}_S(\omega) X(\omega)\ d\mathsf{P}(\omega)$. The definition of abstract integral is the same for arbitrary measures (i.e. not such that $\mu(\Omega) = 1$); in particular, on $(\mathbb{R}, \mathcal{B})$ there exists a "translation-invariant" measure $\mu : \mathcal{B} \to [0, \infty)$ such that for all continuous functions f it holds $\int_{[a,b]} f(x)\ d\mu(x) = \int_a^b f(x)\ dx$, where the integral on the right-hand side is the Riemann integral taught in calculus courses. μ is called the *Lebesgue* measure. In other terms, any function which has integral in the Riemann sense has also an integral in the Lebesgue sense. The converse is definitely false; the "abstract" integral is indeed a far-reaching generalization of the Riemann integral.

standard cases the situation is simpler: for example if X is continuous, a "change of variable" (made possible by the fact that X "translates" events and probabilities from Ω to $\mathbb{R}$) guarantees that the expectation is

$$\begin{aligned}\mathsf{E}[X] &= \int_\Omega X(\omega)\ d\mathsf{P}(\omega) \quad \text{(abstract integral)}\\ &= \int_\mathbb{R} x\ d\mathsf{P}_X(x) \quad \text{(another abstract integral)}\\ &= \int_{-\infty}^{+\infty} x\ f_X(x)\ dx.\end{aligned}$$

In many applications the last one is a Riemann integral, i.e. the familiar integral taught in calculus courses. For example, if X is Gaussian with density (59), then

$$\begin{aligned}\mathsf{E}[X] &= \int_{-\infty}^{+\infty} x\ \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-\frac{(x-\mu)^2}{2\sigma^2}}\ dx\\ &\qquad \text{make the substitution } u = \frac{x-\mu}{\sigma}\\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{+\infty} (\sigma u + \mu)\ e^{-\frac{u^2}{2}}\ \sigma\ du\\ &= \sigma\cdot\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} u\ e^{-\frac{u^2}{2}}\ du + \mu\cdot\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-\frac{u^2}{2}}\ du\\ &= \sigma\cdot 0 + \mu\cdot 1 = \mu.\end{aligned}$$

Note that the expectation may not even exist. For example, in the case of continuous variables, if the integrals $\int_{-\infty}^{0} x\ f_X(x)\ dx$ and $\int_{0}^{+\infty} x\ f_X(x)\ dx$ both diverge, then $\mathsf{E}[X]$ is not defined.
The terms *mean* and *average* are synonyms for expectation.

D.3.2 Properties

Proposition D.3.1 *Let X and Y be random variables on the same probability space $(\Omega, \mathcal{F}, \mathsf{P})$.*

- *If $X = Y$ almost surely, and if the expectations of X and Y exist, then $\mathsf{E}[X] = \mathsf{E}[Y]$. If $X = c$ almost surely, where $c \in \mathbb{R}$ is a constant, then $\mathsf{E}[X] = c$. (For instance, $\mathsf{E}[17] = 17$.)*

- *If the expectations of X and Y exist, then for any $a, b \in \mathbb{R}$*

  $$\mathsf{E}[aX + bY] = a\mathsf{E}[X] + b\mathsf{E}[Y].$$

  In words, the expectation $\mathsf{E}[\cdot]$ is a linear *operator.*

- *If the expectations of X and Y exist, and if $X \leq Y$ almost surely, then*
$$\mathsf{E}[X] \leq \mathsf{E}[Y].$$
- *If the expectation of X exists, then the expectation of $|X|$ also exists, and*
$$|\mathsf{E}[X]| \leq \mathsf{E}\left[|X|\right]$$
If, in particular, $\mathsf{E}\left[|X|\right] = 0$, then $X = 0$ almost surely.
- *For all $S \in \mathcal{F}$, the indicator function $\mathbb{1}_S : \Omega \to \{0,1\}$ is a random variable, and*
$$\mathsf{E}[\mathbb{1}_S] = \mathsf{P}[S].$$
For all $A \in \mathcal{B}$, the indicator function $\mathbb{1}_A(X(\cdot)) : \Omega \to \{0,1\}$ is a random variable, and
$$\mathsf{E}[\mathbb{1}_A(X)] = \mathsf{E}[\mathbb{1}_{\{\omega \in \Omega \ | \ X(\omega) \in A\}}] = \mathsf{P}[X \in A].$$

Do not get confused by the two different notations regarding indicator functions, $\mathbb{1}_A(X)$ and $\mathbb{1}_S$, one with parentheses and the other without. In both cases the indicator is a *function*, but in the first it is applied to X, and in the second to ω. If we were to make everything explicit, the first indicator would read $\mathbb{1}_A(X(\cdot))$, and the second $\mathbb{1}_S(\cdot)$, and for a fixed ω they would take the values $\mathbb{1}_A(X(\omega))$ and $\mathbb{1}_S(\omega)$ respectively; however, since the sample space Ω does not really matter, as we have mentioned above, in probability theory the dependence on ω is almost always left implicit (hence we denote random variables X, not $X(\cdot)$, despite the fact that all of them are functions; and $\mathbb{1}_S$ is indeed a random variable).

Recall that if X is a random variable (or vector) and $g : \mathbb{R} \to \mathbb{R}$ is a measurable function, then $g(X)$ is also a random variable (or vector).

Proposition D.3.2 *If X is a continuous random variable/vector, $g : \mathbb{R} \to \mathbb{R}$ is a measurable function, and $\mathsf{E}[g(X)]$ exists, then*
$$\mathsf{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) \ f_X(x) \ dx.$$

For our purposes, the most interesting functions are $g(X) = X^2$ and $g(X) = (X - \mathsf{E}[X])^2$ (or, if X is a vector, $g(X) = XX^\top$ and $g(X) = (X - \mathsf{E}[X])(X - \mathsf{E}[X])^\top$ respectively). Since $\mathsf{E}[X]$ is a constant, it does not pose any problem if it appears inside $\mathsf{E}[\cdot]$ as the parameter of a function.

Definition D.3.2
If X is a random variable, then

- $\mathsf{E}\left[X^2\right]$, *if it exists, is called the* second-order moment *of* X*; by Proposition D.3.1, it always holds* $\mathsf{E}\left[X^2\right] \geq 0$*. If, in particular,* $\mathsf{E}\left[X^2\right] = 0$*, then* $X = 0$ *almost surely.*

- $\mathsf{E}\left[(X - \mathsf{E}[X])^2\right]$*, that is the second-order moment of* $X - \mathsf{E}[X]$*, is called the* variance *of* X*. It is usually denoted* σ^2 *or* σ_X^2 *or* $\mathsf{Var}\,[X]$*, and used as a measure of* dispersion *of* X *around its mean* $\mu = \mathsf{E}[X]$*.*

- *The square root of* $\mathsf{E}\left[(X - \mathsf{E}[X])^2\right]$*, denoted* σ *or* σ_X*, is called the* standard deviation *of* X*, and also used as a measure of dispersion.*

If X *and* Y *are random variables defined on the same probability space, then*

- $\mathsf{E}\left[(X - \mathsf{E}[X])(Y - \mathsf{E}[Y])\right]$ *is called the* covariance *of* X *and* Y*, and it is usually denoted* σ_{XY} *or* $\mathsf{Cov}\,[X, Y]$*. By the so-called* Cauchy-Schwarz inequality*, it holds* $|\sigma_{XY}| \leq \sigma_X \sigma_Y$*. If, in particular,* $\mathsf{Cov}\,[X, Y] = 0$*, then* X *and* Y *are said to be* uncorrelated*.*

- *The quantity* $\rho := \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$ *is called the* correlation coefficient *of* X *and* Y*. It is used as a statistical measure of how much the behavior of one of the variables influences the behavior of the other. By the Cauchy-Schwarz inequality, it holds* $-1 \leq \rho \leq 1$*. If* $\rho = 0$*, then* X *and* Y *are uncorrelated.*

If X *is a random vector, then*

- $\mathsf{E}\left[(X - \mathsf{E}[X])(X - \mathsf{E}[X])^\top\right]$ *is called the* covariance matrix *of* X *(usually denoted* Σ*). It is always a symmetric matrix. Moreover, it is always at least positive-semidefinite because, for all constant vectors* v*,* $v^\top(X - \mathsf{E}[X])$ *is a random variable, hence*

$$\begin{aligned} v^\top \Sigma v &= v^\top \mathsf{E}\left[(X - \mathsf{E}[X])(X - \mathsf{E}[X])^\top\right] v = \mathsf{E}\left[v^\top (X - \mathsf{E}[X])(X - \mathsf{E}[X])^\top v\right] \\ &= \mathsf{E}\left[(v^\top (X - \mathsf{E}[X]))^2\right] \geq 0. \end{aligned}$$

For example, if $X = \begin{bmatrix} Y & Z \end{bmatrix}^\top$*,*

$$\begin{aligned} &\mathsf{E}\left[\left(\begin{bmatrix} Y \\ Z \end{bmatrix} - \mathsf{E}\left[\begin{bmatrix} Y \\ Z \end{bmatrix}\right]\right)\left(\begin{bmatrix} Y & Z \end{bmatrix} - \mathsf{E}\left[\begin{bmatrix} Y & Z \end{bmatrix}\right]\right)\right] \\ &= \begin{bmatrix} \mathsf{E}\left[(Y - \mathsf{E}[Y])^2\right] & \mathsf{E}\left[(Y - \mathsf{E}[Y])(Z - \mathsf{E}[Z])\right] \\ \mathsf{E}\left[(Z - \mathsf{E}[Z])(Y - \mathsf{E}[Y])\right] & \mathsf{E}\left[(Z - \mathsf{E}[Z])^2\right] \end{bmatrix} \\ &= \begin{bmatrix} \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{YZ} & \sigma_Z^2 \end{bmatrix} \end{aligned}$$

Proposition D.3.3
If X and Y are random variables and $a \in \mathbb{R}$ is a constant, then

- $\mathsf{Var}\,[aX] = a^2\ \mathsf{Var}\,[X]$. *Indeed,*

$$\begin{aligned}\mathsf{Var}\,[aX] &= \mathsf{E}\left[(aX - \mathsf{E}[aX])^2\right] = \mathsf{E}\left[(aX - a\mathsf{E}[X])^2\right]\\ &= \mathsf{E}\left[a^2\,(X - \mathsf{E}[X])^2\right] = a^2\ \mathsf{E}\left[(X - \mathsf{E}[X])^2\right].\end{aligned}$$

- $\mathsf{Var}\,[X+Y] = \mathsf{Var}\,[X] + \mathsf{Var}\,[Y] + 2\mathsf{Cov}\,[X,Y]$. *Indeed*

$$\begin{aligned}\mathsf{Var}\,[X+Y] &= \mathsf{E}\left[(X+Y-\mathsf{E}[X+Y])^2\right] = \mathsf{E}\left[(X-\mathsf{E}[X]+Y-\mathsf{E}[Y])^2\right]\\ &= \mathsf{E}\left[(X-\mathsf{E}[X])^2\right] + \mathsf{E}\left[(Y-\mathsf{E}[Y])^2\right]\\ &\quad + \mathsf{E}\,[(X-\mathsf{E}[X])\,(Y-\mathsf{E}[Y])] + \mathsf{E}\,[(Y-\mathsf{E}[Y])\,(X-\mathsf{E}[X])].\end{aligned}$$

- *If X and Y are uncorrelated,* $\mathsf{Var}\,[X+Y] = \mathsf{Var}\,[X] + \mathsf{Var}\,[Y]$. *This follows immediately from the preceding point.*

- *In general, if $X_1, X_2, \cdots, X_N$ are pairwise uncorrelated, then*

$$\mathsf{Var}\left[\sum_{i=1}^{N} X_i\right] = \sum_{i=1}^{N} \mathsf{Var}\,[X_i].$$

Example. Suppose that $X_1, X_2, \cdots, X_N$ are pairwise uncorrelated variables with the same mean $\mathsf{E}\,[X_i] = \mu$ and the same variance $\mathsf{Var}\,[X_i] = \sigma^2$. Then

$$\begin{aligned}\mathsf{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] &= \frac{1}{N}\sum_{i=1}^{N}\mathsf{E}\,[X_i] = \frac{1}{N}\sum_{i=1}^{N}\mu = \mu;\\ \mathsf{Var}\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] &= \frac{1}{N^2}\mathsf{Var}\left[\sum_{i=1}^{N} X_i\right] = \frac{1}{N^2}\sum_{i=1}^{N}\mathsf{Var}\,[X_i] = \frac{1}{N^2}N\sigma^2 = \frac{\sigma^2}{N}.\end{aligned}$$

□

D.3.3 The space of square-summable random variables

Definition D.3.3 *We denote $L^2(\Omega, \mathcal{F}, \mathsf{P})$ the set of all those random variables X defined on $(\Omega, \mathcal{F}, \mathsf{P})$ such that $\mathsf{E}[X^2] < \infty$. All such variables are called* square summable[45]. *Further, in $L^2(\Omega, \mathcal{F}, \mathsf{P})$ we identify all those variables which are equal almost surely ($X = Y$ almost surely) as if they were "the same" variable ($X(\omega) = Y(\omega)$ for all ω).*

[45] Because the "abstract integral" of the square of the function, which is $\mathsf{E}[X^2]$, is finite.

Lemma D.3.1 $L^2(\Omega, \mathcal{F}, \mathsf{P})$ *is a vector space.*

Proof. That the set of *all* the random variables on $(\Omega, \mathcal{F}, \mathsf{P})$ is a vector space is the statement of Proposition D.2.3. Thus, all we need to check is that $L^2(\Omega, \mathcal{F}, \mathsf{P})$ is a well-defined *subspace* of the space of all random variables. More explicitly, we must check that if $X, Y \in L^2(\Omega, \mathcal{F}, \mathsf{P})$ ($\mathsf{E}\left[X^2\right] < \infty$, and analogously for Y) and $a \in \mathbb{R}$, then aX and $X + Y$ also belong to $L^2(\Omega, \mathcal{F}, \mathsf{P})$. Indeed

$$\mathsf{E}\left[(aX)^2\right] = \mathsf{E}\left[a^2X^2\right] = a^2\mathsf{E}\left[X^2\right] < \infty,$$

and since

$$\begin{aligned} 0 &\leq (X - Y)^2 = X^2 - 2XY + Y^2, \\ 2XY &\leq X^2 + Y^2, \\ (X + Y)^2 &= X^2 + 2XY + Y^2 \leq 2X^2 + 2Y^2, \end{aligned}$$

then

$$\mathsf{E}\left[(X + Y)^2\right] \leq 2\mathsf{E}\left[X^2\right] + 2\mathsf{E}\left[Y^2\right] < \infty.$$

□

We will denote $L_0^2(\Omega, \mathcal{F}, \mathsf{P})$ the set of all those random variables $X \in L^2(\Omega, \mathcal{F}, \mathsf{P})$ such that $\mathsf{E}[X] = 0$. Of course, $X - \mathsf{E}[X] \in L_0^2(\Omega, \mathcal{F}, \mathsf{P})$ for all $X \in L^2(\Omega, \mathcal{F}, \mathsf{P})$. Also in $L_0^2(\Omega, \mathcal{F}, \mathsf{P})$ we identify all those variables which are equal almost surely as if they were "the same" variable.
Exercise: show that $L_0^2(\Omega, \mathcal{F}, \mathsf{P})$ is a subspace of $L^2(\Omega, \mathcal{F}, \mathsf{P})$.

Now comes the nice point of this section: in $L_0^2(\Omega, \mathcal{F}, \mathsf{P})$ we define

$$\langle X, Y\rangle := \mathsf{E}\left[XY\right] = \mathsf{Cov}\left[X, Y\right],$$

where the second equality stems from the fact that X and Y have mean zero. In words, the *covariance* of X and Y can be interpreted as a *scalar product* between X and Y. The only subtle point in showing that this is actually a well-defined scalar product is the fact that it should hold $\langle X, X\rangle = 0$ if, *and only if*, $X = 0$. This property relies crucially on the (rather vague) statement "we identify all those variables which are equal almost surely as if they were the same variable", which could indeed be made precise[46].

If a scalar product is available, then a *norm* is available as well:

$$\|X\|_2 := \sqrt{\langle X, X\rangle} = \sqrt{\mathsf{E}\left[X^2\right]} = \sqrt{\mathsf{Var}\left[X\right]} = \sigma$$

[46] The property of the scalar product here becomes: "$\langle X, X\rangle = 0$ if and only if $X = 0$ almost surely".

The standard deviation of a variable is its length!

And if we have a notion of scalar product, we have the corresponding notion of *orthogonality*:

$$X \perp Y \quad \text{if } \langle X, Y \rangle = \mathsf{Cov}\,[X, Y] = 0.$$

Uncorrelated variables are orthogonal!

And the "angle between X and Y"? It is the arc cosine of the correlation coefficient ρ. Indeed $\rho = \frac{\langle X,Y \rangle}{\|X\|_2 \|Y\|_2}$, that is $\langle X, Y \rangle = \rho\ \|X\|_2 \|Y\|_2$, and $-1 \leq \rho \leq 1$; hence, recalling what you know about the correspondence between angles and scalar products in geometry, you see that ρ resembles very much the cosine of an angle.

And we could go on translating a good lot of mathematical statistics into linear algebra and geometry. For instance, the *orthogonal projection* of a random variable Y on the *subspace generated* by other random variables $X_1, \cdots, X_n$, would turn out to be the "best linear statistical estimate" of Y based on the sole knowledge of $X_1, \cdots, X_n$, where "best" means that the variance of the error is the minimum possible. This nice construction of Euclidean geometry into the world of random variables is essentially due to Kolmogorov[47].

D.4 Probability inequalities

Lemma D.4.1 *(Markov's inequality). Let X be a nonnegative random variable (that is, $X \geq 0$ almost surely). Then for all $\varepsilon > 0$*

$$\mathsf{P}\,[X \geq \varepsilon] \leq \frac{\mathsf{E}\,[X]}{\varepsilon}.$$

Proof. Consider the function $\mathbb{1}_{[\varepsilon,\infty)}(x)$ that takes the values 1 when $x \geq \varepsilon$, and 0 otherwise. Then, for all $x \geq 0$, $\mathbb{1}_{[\varepsilon,\infty)}(x) \leq x/\varepsilon$ (verify this!). Consequently,

$$\mathsf{P}\,[X \geq \varepsilon] = \mathsf{E}\,[\mathbb{1}_{X \geq \varepsilon}] = \mathsf{E}\left[\mathbb{1}_{[\varepsilon,\infty)}(X)\right] \leq \mathsf{E}\,[X/\varepsilon] = \frac{\mathsf{E}\,[X]}{\varepsilon}.$$

□

[47] Of course this section is far from rigorous, and relies only on your intuition. This is more than enough for the purposes of these lecture notes, but be aware that there is a frontier where your intuition on orthogonality should refrain: it is *not* true that, if V is a subspace of $L^2_0(\Omega, \mathcal{F}, \mathsf{P})$, then $\left(V^\perp\right)^\perp = V$, because $L^2_0(\Omega, \mathcal{F}, \mathsf{P})$ is *not* finite-dimensional as $\mathbb{R}^p$. Therefore, in this context propositions like A.3.1 must be rephrased a bit taking into account the topological properties of infinite-dimensional spaces; the resulting geometry is the theory of so-called Hilbert spaces. Start from [18] for more information on this subject.

Lemma D.4.2 *(Čebyšev's inequality). Let X be any real random variable with mean μ and variance σ^2. Then*

$$\mathsf{P}\left[|X-\mu| \geq \varepsilon\right] \leq \frac{\sigma^2}{\varepsilon^2}.$$

Proof. Let us apply the Markov inequality to the nonnegative random variable $(X-\mu)^2$:

$$\mathsf{P}\left[|X-\mu| \geq \varepsilon\right] = \mathsf{P}\left[(X-\mu)^2 \geq \varepsilon^2\right] \leq \frac{\mathsf{E}\left[(X-\mu)^2\right]}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}.$$

□

Lemma D.4.3 *(Chernoff's bound). Let X be any real random variable. Then for any $s > 0$*

$$\mathsf{P}\left[X \geq \varepsilon\right] \leq \frac{\mathsf{E}\left[e^{sX}\right]}{e^{s\varepsilon}}.$$

Proof. We have, by Markov's inequality,

$$\mathsf{P}\left[X \geq \varepsilon\right] = \mathsf{P}\left[e^{sX} \geq e^{s\varepsilon}\right] \leq \frac{\mathsf{E}\left[e^{sX}\right]}{e^{s\varepsilon}}.$$

□

Note that, since the Chernoff bound holds for all $s > 0$, the inequality must hold for the infimum over all $s > 0$ as well:

$$\mathsf{P}\left[X \geq \varepsilon\right] \leq \inf_{s>0} \frac{\mathsf{E}\left[e^{sX}\right]}{e^{s\varepsilon}}.$$

D.5 Independence

What we have seen so far, regarding properties of the probability, expectations, inequalities and so on, falls naturally in the modern theory of measure and abstract integration. Even the "distinctive trait" of probability, that is $\mathsf{P}[\Omega] = 1$, may appear as nothing more than a convenient rescaling; other additive scalar quantities, for example *charge* or *mass*, enjoy similar properties. For example, consider a linear mass distribution summing up to a finite mass. The *center of mass* of the distribution is nothing more than what we have called the expectation, and the *moment of inertia* computed with respect to the center of mass is a rescaled version of the variance.

What truly sets probability theory apart from the theory of integration, making it a branch of mathematics standing by its own, is the concept of *independence.*

Definition D.5.1 *Let* $(\Omega, \mathcal{F}, \mathsf{P})$ *be a probability space. The events* $S_1, S_2, \cdots, S_n \in \mathcal{F}$ *are called* independent *if, for any sub-family* $\{i_1, i_2, \cdots, i_k\} \subseteq \{1, 2, \cdots, n\}$ *it holds*

$$\mathsf{P}\left[\bigcap_{j=1}^{k} S_{i_j}\right] = \prod_{j=1}^{k} \mathsf{P}\left[S_{i_j}\right]. \tag{60}$$

In words, Equation (60) says that, for any choice of events extracted from $S_1, S_2, \cdots, S_n$, the probability that they happen simultaneously is the product of the respective probabilities.
In particular, two events $S_1, S_2 \in \mathcal{F}$ are independent if $\mathsf{P}[S_1 \cap S_2] = \mathsf{P}[S_1]\mathsf{P}[S_2]$, and if any two events S_i, S_j chosen from $S_1, S_2, \cdots, S_n \in \mathcal{F}$ are independent, we say that $S_1, \cdots, S_n$ are *pairwise* independent. Now, if $S_1, \cdots, S_n$ are independent, of course they are also pairwise independent, but *in general the converse is false.*

Example. Let $\Omega = \{[A], [B], [C], [ABC]\}$ be a set of cards containing the letters A, B, C, or all the three letters; let $\mathcal{F}$ be the family of all possible subsets of Ω, and for each $S \in \mathcal{F}$ let $\mathsf{P}[S] = \frac{\text{the number of cards in } S}{4}$ (this amounts to say that every card has the same probability $\frac{1}{4}$ of being extracted). Now consider the random experiment of extracting a card, and define

$$\begin{aligned}
S_A &= \{\text{the card contains the letter } A\} = \{[A], [ABC]\};\\
S_B &= \{\text{the card contains the letter } B\} = \{[B], [ABC]\};\\
S_C &= \{\text{the card contains the letter } C\} = \{[C], [ABC]\}.
\end{aligned}$$

These events are *pairwise* independent, because

$$\begin{aligned}
&\mathsf{P}[S_A \cap S_B] = \mathsf{P}[\{[ABC]\}] = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \mathsf{P}[S_A]\mathsf{P}[S_B],\\
&\qquad\qquad \text{and in the same way}\\
&\mathsf{P}[S_B \cap S_C] = \mathsf{P}[S_B]\mathsf{P}[S_C],\\
&\mathsf{P}[S_A \cap S_C] = \mathsf{P}[S_A]\mathsf{P}[S_C];
\end{aligned}$$

on the other hand S_A, S_B, and S_C are *not* independent, because

$$\mathsf{P}[S_A \cap S_B \cap S_C] = \mathsf{P}[\{[ABC]\}] = \frac{1}{4} \neq \frac{1}{8} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \mathsf{P}[S_A]\mathsf{P}[S_B]\mathsf{P}[S_C].$$

This classic counterexample is due to the Russian mathematician S. Bernstein. □

Definition D.5.2 *Let* $X_1, X_2, \cdots, X_n$ *be random variables defined on the same probability space* $(\Omega, \mathcal{F}, \mathsf{P})$. $X_1, \cdots, X_n$ *are called* independent *if, for* *any choice of* $\{i_1, i_2, \cdots, i_k\} \subseteq \{1, 2, \cdots, n\}$*, and for any choice of Borel sets* $B_{i_1}, B_{i_2}, \cdots, B_{i_k} \in \mathcal{B}$*, it holds*

$$\mathsf{P}\left[\bigcap_{j=1}^{k} X_{i_j} \in B_{i_j}\right] = \prod_{j=1}^{k} \mathsf{P}\left[X_{i_j} \in B_{i_j}\right]. \qquad (61)$$

In particular, two random variables X, Y are independent if $\mathsf{P}[X \in A, Y \in B] = \mathsf{P}[X \in A]\mathsf{P}[Y \in B]$.
An immediate corollary of the definition is the following

Lemma D.5.1 *If* $Z = \begin{bmatrix} X & Y \end{bmatrix}^\top$ *is a random vector on* $(\Omega, \mathcal{F}, \mathsf{P})$*, and if* X, Y *are independent, then*

$$F_Z\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = F_X(x)F_Y(y).$$

In words, the (joint) distribution function of independent variables factorizes into the product of the respective distribution functions. The lemma generalizes in an obvious way to k-dimensional random vectors.

If the random vector is continuous, then its density factorizes as well:

Lemma D.5.2 *If* $Z = \begin{bmatrix} X & Y \end{bmatrix}^\top$ *is a continuous random vector on* $(\Omega, \mathcal{F}, \mathsf{P})$ *with density* f_Z*, and if* X, Y *are independent, then they are also continuous, and*

$$f_Z\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = f_X(x)f_Y(y).$$

The most important consequence of Lemma D.5.1 is that expectations factorize as well:

Lemma D.5.3 *Suppose that* X, Y *are independent; then*

$$\mathsf{E}[XY] = \mathsf{E}[X]\mathsf{E}[Y],$$

provided that the above expectations exist.

Corollary D.5.1 *Independent variables are uncorrelated.*

(The converse is, in general, false.)
Proof. Let $\mathsf{E}[X] = \mu_X, \mathsf{E}[Y] = \mu_Y$. Then

$$\begin{aligned}\mathsf{Cov}[X, Y] &= \mathsf{E}[(X - \mu_X)(Y - \mu_Y)] \\ &= \mathsf{E}[XY] - \mu_X\mathsf{E}[Y] - \mathsf{E}[X]\mu_Y + \mu_X\mu_Y \\ &= \mathsf{E}[X]\mathsf{E}[Y] - \mu_X\mu_Y - \mu_X\mu_Y + \mu_X\mu_Y = 0.\end{aligned}$$

□

D.6 Stochastic processes

A broad discussion arises, and the deep results in probability theory start to show up, when we consider, instead of a single random variable or a k-tuple of random variables, an *infinite sequence* of random variables on the same probability space.

Definition D.6.1 *If* $\{X_i\}_{i=1}^{+\infty} = \{X_1, X_2, \cdots, X_i, \cdots\}$ *is a sequence of random variables on* $(\Omega, \mathcal{F}, \mathsf{P})$*, so that any finite sub-sequence* $(X_{i_1}, \cdots, X_{i_n})$ *(for any* n*) is a random vector on* $(\Omega, \mathcal{F}, \mathsf{P})$*, such sequence is called a* stochastic process*.*

We will not delve into the subtleties implicit in such definition. But note that, in the same way as the value $X(\omega)$ that a random variable takes for fixed ω is a real number, and the value $(X_1(\omega), \cdots, X_k(\omega))$ that a random vector takes for fixed ω is a k-tuple of real numbers, the "value" $X_1(\omega), X_2(\omega), \cdots$ that a stochastic process takes for fixed ω is an *infinite sequence of real numbers.*

Thus, in the same way as random variables "translate" events and probabilities from $(\Omega, \mathcal{F}, \mathsf{P})$ to $(\mathbb{R}, \mathcal{B})$, and random vectors "translate" events and probabilities from $(\Omega, \mathcal{F}, \mathsf{P})$ to $(\mathbb{R}^k, \mathcal{B}^k)$, one expects a stochastic process to "translate" events and probabilities from $(\Omega, \mathcal{F}, \mathsf{P})$ to *the set of all infinite sequences of real numbers* (let us denote it $\mathbb{R}^{\mathbb{N}}$), equipped with a certain σ-algebra $\mathcal{F}^{\mathbb{N}}$, in such a way that

$$\left(\mathbb{R}^{\mathbb{N}}, \mathcal{F}^{\mathbb{N}}, \mathsf{P}^{\mathbb{N}}\right)$$

becomes a probability space with a certain probability $\mathsf{P}^{\mathbb{N}}$. For the whole thing to make any sense, such probability has to be compatible with the variables $X_1, X_2, \cdots$ in the sense that, for example,

$$\mathsf{P}^{\mathbb{N}}\left[X_2 \leq 5, \text{ whatever the values of } X_1, X_3, X_4, \cdots\right] = \mathsf{P}[X_2 \leq 5].$$

That such $\mathcal{F}^{\mathbb{N}}$ and $\mathsf{P}^{\mathbb{N}}$ actually exist is a remarkable result due to Kolmogorov, and we shall stop our discussion here. However, you should be aware that, besides events regarding the first n variables of the process, which fall within the theory of random vectors, *other* kinds of events, like

$$\{\omega \in \Omega \mid X_i(\omega) = 0 \text{ for infinitely many } i\} \quad \text{in the prob. space } (\Omega, \mathcal{F}, \mathsf{P})$$
$$\left\{(X_1, X_2, \cdots) \in \mathbb{R}^{\mathbb{N}} \mid X_i = 0 \text{ for infinitely many } i\right\} \quad \text{in the prob. space } (\mathbb{R}^{\mathbb{N}}, \mathcal{F}^{\mathbb{N}}, \mathsf{P}^{\mathbb{N}}),$$

or

$$\left\{\omega \in \Omega \mid \lim_{i\to\infty} X_i(\omega) = 0\right\} \quad \text{in the prob. space } (\Omega, \mathcal{F}, \mathsf{P})$$
$$\left\{(X_1, X_2, \cdots) \in \mathbb{R}^{\mathbb{N}} \mid \lim_{i\to\infty} X_i = 0\right\} \quad \text{in the prob. space } (\mathbb{R}^{\mathbb{N}}, \mathcal{F}^{\mathbb{N}}, \mathsf{P}^{\mathbb{N}}),$$

can, and will, be considered. These fall inevitably within the theory of stochastic processes, and have a more natural interpretation in the "translated" probability space $\left(\mathbb{R}^{\mathbb{N}}, \mathcal{F}^{\mathbb{N}}, \mathsf{P}^{\mathbb{N}}\right)$.

Definition D.6.2 *The stochastic process $\{X_i\}_{i=1}^{+\infty}$ is said to form an* independent sequence *if, for any* finite *choice of indices $\{i_1, i_2, \cdots, i_k\} \subset \mathbb{N}$, the corresponding variables $X_{i_1}, X_{i_2}, \cdots, X_{i_k}$ are independent. If, moreover, every random variable of the process shares the same distribution function $F = F_{X_1} = F_{X_2} = \cdots$, the variables are called* independent and identically distributed *(i.i.d. for short).*

D.7 Convergence of random variables

D.7.1 Many notions of convergence

Let $\{X_i\}_{i=1}^{\infty}$ be a stochastic process, where X_i is distributed according to the distribution function $F_i(x) = \mathsf{P}[X_i \leq x]$, and let X be another random variable distributed according to $F(x)$.

Definition D.7.1 *X_i is said to converge to X* in distribution[48] *if*

$$\lim_{i\to\infty} F_i(x) = F(x)$$

at all those points x at which F is continuous.

The classical result on convergence in distribution is also the most important and elegant theorem in the theory of probability, and the very reason why the Gaussian distribution shows up everywhere. There are different versions of this result, but the most standard one goes as follows:

Theorem D.7.1 *(Central limit theorem). Let $\{X_i\}_{i=1}^{\infty}$ be independent and identically distributed random variables, each with mean μ and variance σ^2. Then the variable*

$$\frac{\frac{1}{N}\sum_{i=1}^{N} X_i - \mu}{\sigma/\sqrt{N}}$$

(which is the sample average, centered on its own expectation μ, and normalized by its own standard deviation $\sigma/\sqrt{N}$) converges in distribution, as $N \to \infty$, to a Gaussian variable with mean 0 and variance 1. In other terms, for all $x \in \mathbb{R}$ it holds

$$\lim_{N\to\infty} \mathsf{P}\left[\frac{\frac{1}{N}\sum_{i=1}^{N} X_i - \mu}{\sigma/\sqrt{N}} \leq x\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt.$$

[48] With respect to random variables, the terms "convergence in distribution", "convergence in law" and "weak convergence", usually found in the literature, have exactly the same meaning.

In words, the central limit theorem says that any sum of many independent and identically distributed random variables is approximately Gaussian.

Definition D.7.2 *X_i is said to converge to X* in probability *if for all $\varepsilon > 0$*

$$\lim_{i\to\infty} \mathsf{P}\left[|X_i - X| \geq \varepsilon\right] = 0.$$

Definition D.7.3 *X_i is said to converge to X* almost surely *if the event*

$$\left\{\omega \in \Omega \mid \lim_{i\to\infty} X_i(\omega) = X(\omega)\right\} \quad \text{in the prob. space } (\Omega, \mathcal{F}, \mathsf{P})$$

$$\left\{\lim_{i\to\infty} X_i = X\right\} \quad \text{in the prob. space } (\mathbb{R}^{\mathbb{N}}, \mathcal{F}^{\mathbb{N}}, \mathsf{P}^{\mathbb{N}})$$

has probability 1*. In other terms, the set of trajectories $X_1(\omega), X_2(\omega), X_3(\omega), \cdots$ which do* not *converge to X has probability* 0*.*

Among these, almost sure convergence is the strongest form of convergence, and the most desirable; convergence in probability is an intermediate property, and convergence in distribution is the weakest one. Here, “strong” and “weak” are meant in the sense that the stronger property implies the weaker:

$$\begin{gathered} X_i \text{ converges almost surely to } X \\ \Downarrow \\ X_i \text{ converges in probability to } X \\ \Downarrow \\ X_i \text{ converges in distribution to } X \end{gathered}$$

Definition D.7.4 *X_i is said to converge to X* in the mean-square *if*

$$\lim_{i\to\infty} \mathsf{E}\left[|X_i - X|^2\right] = 0.$$

If $\{X_i\}$ and X belong to the space $L^2(\Omega, \mathcal{F}, \mathsf{P})$ defined in Section D.3.3, mean-square convergence is none other than convergence with respect to the Euclidean distance:

$$\lim_{i\to\infty} \|X_i - X\|_2 = 0.$$

Mean-square convergence is also a kind-of-strong form of convergence, in the sense that

$$\begin{gathered} X_i \text{ converges in mean-square to } X \\ \Downarrow \\ X_i \text{ converges in probability to } X \\ \Downarrow \\ X_i \text{ converges in distribution to } X \end{gathered}$$

However, no implication exists between almost sure convergence and convergence in the mean-square.

D.7.2 Convergence theorems

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of random variables having the same mean $\mathsf{E}[X_i] = \mu$. A theorem is called a "law of large numbers" if it has the form

$$\text{[Some more hypotheses]} \quad \Rightarrow \quad \frac{1}{N}\sum_{i=1}^{N} X_i \to \mu,$$

where the arrow $\to$ means some kind of probabilistic convergence happening as $N \to \infty$, and by μ on the right-hand side is meant the "pseudo-random" variable that takes the value $\mu \in \mathbb{R}$ with probability 1. Here is a "weak" form of the theorem (we do not actually need it in the notes, but it is an instructive application of Čebyšev's inequality):

Lemma D.7.1 *(Weak law of large numbers, Čebyšev). Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of uncorrelated random variables such that for all i*

$$\begin{aligned} \mathsf{E}[X_i] &= \mu; \\ \mathsf{Var}[X_i] &= \mathsf{E}\left[(X_i-\mu)^2\right] = \sigma^2; \end{aligned}$$

then

$$\frac{1}{N}\sum_{i=1}^{N} X_i \to \mu \quad \textit{in probability.}$$

Proof. We already know that, since $X_1, X_2, \cdots$ are uncorrelated,

$$\mathsf{E}\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \mu, \qquad \mathsf{Var}\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] = \frac{\sigma^2}{N}.$$

Therefore for all $\varepsilon > 0$, by Čebyšev's inequality,

$$\mathsf{P}\left[\left|\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right| \geq \varepsilon\right] \leq \frac{\sigma^2/N}{\varepsilon^2} \to 0$$

as $N \to \infty$. □

Here are, instead, two "strong" laws, in the sense that they both establish almost sure convergence. You can find the proofs in [9].

Theorem D.7.2 *(Strong law of large numbers, Kolmogorov). Let $\{X_i\}_{i=1}^{\infty}$ be independent and identically distributed random variables with mean $\mathsf{E}[X_i] = \mu$. Then*

$$\frac{1}{N}\sum_{i=1}^{N} X_i \to \mu \quad \textit{almost surely.}$$

Theorem D.7.3 *(Strong law of large numbers, also by Kolmogorov). Let $\{X_i\}_{i=1}^{\infty}$ be independent random variables with arbitrary distributions but having the same mean $\mathsf{E}[X_i] = \mu$. If*

$$\sum_{i=1}^{\infty} \frac{1}{i^2}\mathsf{E}\left[X_i^2\right] < \infty,$$

then

$$\frac{1}{N}\sum_{i=1}^{N} X_i \to \mu \quad \textit{almost surely.}$$

In the notes, unless otherwise stated, by "strong law of large numbers" we mean Theorem D.7.2.

D.8 Estimators

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of independent and identically distributed random variables, distributed according to a distribution $F(x;\theta)$ which depends on a parameter θ. In mathematical statistics a finite "chunk" $X_1, X_2, \cdots, X_N$ of the infinite sequence of variables is usually called a *random sample* extracted from the distribution $F(x;\theta)$, and a function (or better, a family of functions depending on the number N)

$$f : (X_1, X_2, \cdots, X_N) \mapsto \hat{\theta}$$

used to extract information on the parameter θ from the random sample, is called a *point estimator*, or simply an *estimator*, of θ. Note that since the $\{X_i\}$ are random quantities, $f(X_1, \cdots, X_N)$ is also a random quantity, until a "realization" of $X_1, X_2, \cdots, X_N$ has been drawn. Some "good" properties are usually required of an estimator. Among the most popular are the following ones:

Definition D.8.1 *$f(X_1, \cdots, X_N)$ is called an* unbiased *estimator of θ if*

$$\mathsf{E}[f(X_1, \cdots, X_N)] = \theta.$$

If, on the other hand, $\mathsf{E}[f(X_1, \cdots, X_N)] = \theta + \beta$, with $\beta \neq 0$, the estimator is called *biased*, and the number β is called its *bias*.

Definition D.8.2 *$f(X_1, \cdots, X_N)$ is called a* consistent *estimator of θ if*

$$\lim_{N\to\infty} f(X_1, \cdots, X_N) = \theta \quad \textit{almost surely.}$$

Example. The sample average $\bar{X} = \mathsf{M}[X] = \frac{1}{N}\sum_{i=1}^N X_i$ is a good estimator of the mean $\mu = \mathsf{E}[X_i]$, being

- unbiased: indeed

$$\mathsf{E}\left[\bar{X}\right] = \mathsf{E}\left[\frac{1}{N}\sum_{i=1}^N X_i\right] = \frac{1}{N}\sum_{i=1}^N \mathsf{E}[X_i] = \frac{1}{N}\sum_{i=1}^N \mu = \mu;$$

- consistent: this is precisely the statement of the strong law of large numbers (Theorem D.7.2).

□

Example. The sample variance $s^2 = \frac{1}{N}\sum_{i=1}^N (X_i - \bar{X})^2$ is a “kind-of-good” estimator of the variance $\sigma^2 = \mathsf{E}\left[(X_i - \mu)^2\right]$:

- it is *biased*, indeed

$$\begin{aligned}\sum_{i=1}^N (X_i - \bar{X})^2 &= \sum_{i=1}^N (X_i - \mu + \mu - \bar{X})^2 \\ &= \sum_{i=1}^N (X_i - \mu)^2 + \sum_{i=1}^N (\mu - \bar{X})^2 + 2\sum_{i=1}^N (X_i - \mu)(\mu - \bar{X}) \\ &= \sum_{i=1}^N (X_i - \mu)^2 + N(\bar{X} - \mu)^2 - 2(\bar{X} - \mu)\sum_{i=1}^N (X_i - \mu) \\ &= \sum_{i=1}^N (X_i - \mu)^2 + N(\bar{X} - \mu)^2 - 2(\bar{X} - \mu)N(\bar{X} - \mu) \\ &= \sum_{i=1}^N (X_i - \mu)^2 - N(\bar{X} - \mu)^2;\end{aligned}$$

hence

$$\begin{aligned}\mathsf{E}\left[s^2\right] = \mathsf{E}\left[\frac{1}{N}\sum_{i=1}^N (X_i - \bar{X})^2\right] &= \mathsf{E}\left[\frac{1}{N}\sum_{i=1}^N (X_i - \mu)^2\right] - \mathsf{E}\left[(\bar{X} - \mu)^2\right] \\ = \frac{1}{N}\sum_{i=1}^N \mathsf{Var}[X_i] - \mathsf{Var}\left[\bar{X}\right] &= \sigma^2 - \frac{\sigma^2}{N} = \frac{N-1}{N}\sigma^2.\end{aligned}$$

For big N, the bias is not much relevant; for small N, instead, usually the *unbiased* estimator $\bar{s}^2 = \frac{N}{N-1}s^2 = \frac{1}{N-1}\sum_{i=1}^N (X_i - \bar{X})^2$ is preferred.

- it is, however, consistent. Indeed

  $$s^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2 - (\bar{X} - \mu)^2;$$

  by the strong law of large numbers, the first term on the right-hand side converges to σ^2, and the second to 0 almost surely.

□

We mention on the fly other two popular "good" properties that an estimator of θ may or may not have:

- that of being, among all the unbiased estimators of θ, the one with the minimum possible variance: this property is called *efficiency*;
- very loosely speaking, that of exploiting all the possible information about θ that can be extracted from the data. If the estimator has this property, it is called a *sufficient statistic* for θ.

Example. Suppose that $X_1, \cdots, X_N, \cdots$ are independent and identically distributed Gaussian variables, with $\mathsf{E}[X_i] = \mu$ and $\mathsf{Var}[X_i] = \sigma^2$. One could prove that the sample average $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ is the minimum variance unbiased estimator of the mean $\mathsf{E}[X_i] = \mu$, since $\mathsf{Var}\left[\bar{X}\right] = \frac{\sigma^2}{N}$ is actually the minimum possible, and also that it is a sufficient statistic for μ.
On the other hand, suppose that $N = 2k$ (even), and consider the following estimator:

$$\tilde{X} = \frac{X_2 + X_4 + X_6 + \cdots + X_{2k}}{k} = \frac{1}{k}\sum_{i=1}^{k} X_{2i}.$$

Now, $\tilde{X}$ is both unbiased and consistent, by the same reasons for which $\bar{X}$ is so; nevertheless, it is neither an efficient estimator, since $\mathsf{Var}\left[\tilde{X}\right] = \frac{\sigma^2}{k} > \frac{\sigma^2}{2k} = \frac{\sigma^2}{N}$ ($\bar{X}$ is better), nor a sufficient statistic for μ, for it deliberately discards all the information about μ contained in the samples with odd indexes $X_1, X_3, \cdots$. □

D.9 Three continuous distributions used in inferential statistics

Definition D.9.1 *A random variable X is said to be* normal*, or* Gaussian*, with parameters μ and σ^2 (this is denoted $X \sim \mathcal{N}(\mu, \sigma^2)$), if it has the density:*

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}.$$

You can easily verify that $\mathsf{E}[X] = \mu, \mathsf{Var}[X] = \sigma^2$.

Lemma D.9.1 *Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then*

$$Z := \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

Definition D.9.2 *A nonnegative random variable X is said to be distributed according to a* chi-square distribution with N degrees of freedom *(denoted $X \sim \chi^2(N)$), if it has the density:*

$$f_X(x) = \frac{1}{\Gamma(N/2)2^{N/2}}\ x^{N/2-1}e^{-x/2}, \quad x \geq 0,$$

where Γ is Euler's Gamma function.

(The constant $\frac{1}{\Gamma(N/2)2^{N/2}}$ is not relevant to our discussion; it is just a normalizing factor such that $\int_{-\infty}^{\infty} f_X(x)\ dx = 1$.)

Lemma D.9.2 *Let $X_1, \cdots, X_N \sim \mathcal{N}(0, 1)$ be independent random variables. Then*

$$\sum_{i=1}^{N} X_i^2 \sim \chi^2(N).$$

Let $X_1, \cdots, X_N$ be independent random variables with mean μ and variance σ^2. Recall that the *sample average* $\frac{1}{N}\sum_{i=1}^{N} X_i$ is an unbiased estimator of μ with variance $\mathsf{Var}\left[\bar{X}\right] = \frac{\sigma^2}{N}$, and that the *sample variance* $s^2 := \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2$ is a *biased* estimator of σ^2, so that the unbiased estimator $\bar{s}^2 = \frac{N}{N-1}s^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2$ is usually preferred.
If the variables are Gaussian, we have the following

Theorem D.9.1 *Let $X_1, \cdots, X_N \sim \mathcal{N}(\mu, \sigma)$ be independent random variables. Then*

- $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{N})$*;*
- $(N-1)\frac{\bar{s}^2}{\sigma^2} = N\frac{s^2}{\sigma^2} \sim \chi^2(N-1)$*;*
- $\bar{X}$ *and* $\bar{s}^2$ *are independent.*

Definition D.9.3 *A random variable T is said to be distributed according to a* Student distribution with n degrees of freedom *(denoted $T \sim t(n)$), if it has the density*

$$f_T(x) = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{\pi n}}\frac{1}{(1 + x^2/n)^{(n+1)/2}}.$$

Theorem D.9.2 *Let $W \sim \mathcal{N}(0, 1)$, $V \sim \chi^2(n)$ be independent. Then*

$$T = \frac{W}{\sqrt{V/n}} \sim t(n).$$

Theorem D.9.3 *Let $X_1, \cdots, X_N \sim \mathcal{N}(\mu, \sigma^2)$ be independent. Then*

$$T := \frac{\bar{X} - \mu}{\bar{s}/\sqrt{N}} \sim t(N-1).$$

Proof.

$$\frac{\bar{X} - \mu}{\bar{s}/\sqrt{N}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{N}}}{\frac{\sqrt{N-1}\bar{s}}{\sigma}/\sqrt{N-1}} = \frac{W}{\sqrt{V/(N-1)}},$$

where W is $\mathcal{N}(0,1)$ and independent of $V/\sqrt{N-1}$, which is the square root of a $\chi^2(N-1)$ variable divided by the number of its degrees of freedom. The result follows from Theorem D.9.2. □

Theorem D.9.3 is a powerful tool in inferential statistics: it says that the distribution of the statistic T, depending on a normal sample $\{X_1, \cdots, X_N\}$ (recall that the X_i's are supposed to be i.i.d.) and containing as much information about θ^o as the sample itself, has a distribution $t(N-1)$ that depends only on the size N of the sample. Since the cumulative distribution of a $t(N-1)$ variable, and its inverse, are tabulated in books and available in every statistical software, one can use T to make inferences about $\mathsf{E}[X_i]$.

For example, suppose that the X_i's are independent measures of a quantity involved in a phenomenon that we, as experimenters, are studying, and that some remarkable conclusion about the phenomenon could be drawn if $\mathsf{E}[X_i]$ was *different* from a certain number $\hat{\mu} \in \mathbb{R}$. We want to test whether or not this holds, hence we make the *statistical hypothesis* $\mathsf{E}[X_i] = \hat{\mu}$; under this hypothesis, the statistic T reads

$$T = \frac{\bar{X} - \hat{\mu}}{\bar{s}/\sqrt{N}}.$$

We find, on books or through software, the percentile t_{95} such that

$$\int_{-t_{95}}^{t_{95}} f(x)\ dx = 0.95,$$

where $f(x)$ is the density of a $t(N-1)$ random variable. Before the sample $\{X_1, \cdots, X_N\}$ is drawn, $\{|T| > t_{95}\}$ is an unlikely event, with probability $0.05 = 5\%$. When the sample has been drawn, we can actually compute T, and if it happens that $|T| > t_{95}$ it is customary to hold that *there is enough evidence to reject the hypothesis* $\mathsf{E}[X_i] = \hat{\mu}$ *and to support the conclusion.* This is called a *test of hypothesis with significance level* 5%.

The idea behind this method is that if we run a lot of experiments, make a hypothesis for each experiment, and reject the hypothesis all the times

that an event happens, being “5%-unlikely” under the hypothesis, we will be wrong in only 5% of the cases. Hence, the conclusion of *any* “test with significance 5%” is plausible and reliable in the following sense: exactly the 5% of *all* the “tests with significance 5%”, that is, only a small fraction of them, yields wrong conclusions.
Note that if, in the above setting, it happens instead that $|T| \leq t_{95}$, we are *not* supposed to accept the hypothesis “the mean is equal to $\hat{\mu}$” and claim that the conclusion is false: a test of hypothesis is designed to reject a hypothesis under an unlikely event, not to accept it under a likely one.

Suppose, on the other hand, that we do not make hypotheses on $\mu = \mathsf{E}[X_i]$. Since, anyway, $\mathsf{P}[|T| \leq t_{95}] = 0.95$, with probability 0.95 it still holds

$$-t_{95} \leq \frac{\bar{X} - \mu}{\bar{s}/\sqrt{N}} \leq t_{95}$$
$$\bar{X} - t_{95}\frac{\bar{s}}{\sqrt{N}} \leq \mu \leq \bar{X} + t_{95}\frac{\bar{s}}{\sqrt{N}}$$

In other terms, the *random* interval

$$[a, b] = \left[\bar{X} - t_{95}\frac{\bar{s}}{\sqrt{N}},\ \bar{X} + t_{95}\frac{\bar{s}}{\sqrt{N}}\right]$$

contains μ with probability 0.95. Note that after the sample $\{X_1, \cdots, X_N\}$ has been drawn, the interval can be computed exactly and is not random anymore, hence we cannot still use the word *probability*: we say instead that $[a, b]$ is a *confidence interval* for the mean, with confidence 95%. This means that, if we re-sampled and computed the interval a lot of times under the same assumptions, the interval would contain the mean about 95% of the times.

E Matlab/Octave code for Chapter 3

E.1 Identification of an ARX process

```
% Model: y(t) = a y(t-1) + b u(t-1) + e(t)
% n(t) is a white noise; u(t) and y(t) are measured without errors

nruns = 100;   % Number of identification experiments
a_estimates = zeros(1, nruns);
b_estimates = zeros(1, nruns);

for run=1:nruns,
   % Construction of the process
   N = 1000;                   % Time horizon
   a = 0.8;                    % Parameter
   b = 0.2;                    % Parameter
   u = randn(N,1);             % Some input signal with persistent excitation
   e = 0.3*randn(N+1,1);       % Process noise
   y = zeros(N+1, 1);
   y(1) = 4;                   % Initial condition
   for t=2:N+1,
      y(t) = a*y(t-1) + b*u(t-1) + e(t);
   end

   % Data for least squares
   ypast    = y(1:N);
   ypresent = y(2:N+1);
   Phi = [ypast, u];           % Regressors

   % Least squares estimation
   theta_LS = pinv(Phi)*ypresent;    % Pseudoinverse. Same as: inv(Phi'*Phi)*Phi'*ypresent
a_estimates(run) = theta_LS(1);
b_estimates(run) = theta_LS(2);
end

disp(sprintf('LS estimate of a over %d runs: average %7.5f, variance %7.5f',
             nruns, mean(a_estimates), var(a_estimates)));
disp(sprintf('LS estimate of b over %d runs: average %7.5f, variance %7.5f',
             nruns, mean(b_estimates), var(b_estimates)));
```

The following was the output:

```
LS estimate of a over 100 runs: average 0.80272, variance 0.00020
LS estimate of b over 100 runs: average 0.20072, variance 0.00008
```

E.2 Identification with regressors correlated to noise

```
% Model: y(t) = a y(t-1) + b u(t-1)
% u(t) is measured without errors
% y(t) is measured with an error, ym(t) = y(t) + e(t)

nruns = 100;   % Number of identification experiments
a_estimates = zeros(1, nruns);
b_estimates = zeros(1, nruns);

for run=1:nruns,
   % Construction of the process
   N = 1000;                 % Time horizon
   a = 0.8;                  % Parameter
   b = 0.2;                  % Parameter
   u = randn(N,1);           % Some input signal with persistent excitation
   e = 0.3*randn(N+1,1);     % Some random noise
   y = zeros(N+1, 1);
   y(1) = 4;                 % Initial condition
   for t=2:N+1,
      y(t) = a*y(t-1) + b*u(t-1);
   end
   ym = y + e;               % Measured output

   % Data for least squares
   ypast    = ym(1:N);
   ypresent = ym(2:N+1);
   Phi = [ypast, u];         % Regressors

   % Least squares estimation
   theta_LS = pinv(Phi)*ypresent;
a_estimates(run) = theta_LS(1);
b_estimates(run) = theta_LS(2);
end

disp(sprintf('LS estimate of a over %d runs: average %7.5f, variance %7.5f',
             nruns, mean(a_estimates), var(a_estimates)));
disp(sprintf('LS estimate of b over %d runs: average %7.5f, variance %7.5f',
             nruns, mean(b_estimates), var(b_estimates)));
```

The following was the output:

```
LS estimate of a over 100 runs: average 0.50657, variance 0.00051
LS estimate of b over 100 runs: average 0.20026, variance 0.00010
```

The estimate of b is just fine, but the estimate of a is completely wrong.

E.3 Instrumental variable

```
% Model: y(t) = a y(t-1) + b u(t-1)
% u(t) is measured without errors
% y(t) is measured with an error, ym(t) = y(t) + e(t)
% This time we apply an instrumental variable

nruns = 100;   % Number of identification experiments
a_estimates = zeros(1, nruns);
b_estimates = zeros(1, nruns);

for run=1:nruns,
   % Construction of the process
   N = 1000;                  % Time horizon
   a = 0.8;                   % Parameter
   b = 0.2;                   % Parameter
   u = randn(N+1,1);          % Some input signal
   e = 0.3*randn(N+2,1);      % Some random noise
   y = zeros(N+2, 1);
   y(1) = 4;                  % Initial condition
   for t=2:N+1,
      y(t) = a*y(t-1) + b*u(t-1);
   end
   ym = y + e;                % Measured output

   % Data for least squares
   upast     = u(2:N+1);
   upastpast = u(1:N);
ypast     = ym(2:N+1);
   ypresent  = ym(3:N+2);
   Phi = [ypast, upast];      % Regressors
Psi = [upast, upastpast];   % Instrumental variables

   % Least squares estimation
   theta_LS = inv(Psi'*Phi)*Psi'*ypresent;
a_estimates(run) = theta_LS(1);
b_estimates(run) = theta_LS(2);
end

disp(sprintf('LS estimate of a over %d runs: average %7.5f, variance %7.5f',
             nruns, mean(a_estimates), var(a_estimates)));
disp(sprintf('LS estimate of b over %d runs: average %7.5f, variance %7.5f',
             nruns, mean(b_estimates), var(b_estimates)));
```

The following was the output:

```
LS estimate of a over 100 runs: average 0.80565, variance 0.00364
LS estimate of b over 100 runs: average 0.19820, variance 0.00013
```

E.4 Wolf's sunspot numbers

```
% Octave/Matlab code:
% Autoregressive fit of Wolf's sunspot numbers
% Comparison of the power spectral densities estimates obtained from DFT and AR fit

%-------------------------------------------------------------------
% Load data
sunspot_data = load('sunspots.txt');
N = size(sunspot_data,1);
yrs = sunspot_data(:, 1);     % Years
ss  = sunspot_data(:, 2);     % Wolf's sunspot numbers
ssd = ss - mean(ss);          % Detrend the time series

%-------------------------------------------------------------------
% Plot sunspot numbers
figure(1);
clf();
plot(yrs, ss, 'r');
axis([1749, 1927, 0, 160]);
xlabel('year');
ylabel('Wolf''s number');
%print -color -depslatexstandalone wolf.eps

%-------------------------------------------------------------------
% Estimate the power spectrum by PEM (ordinary least squares)
% Model: y(t) = a y(t-1) + b y(t-2) + e(t)
Y = ssd(3:N);
Phi = [ssd(2:N-1), ssd(1:N-2)];         % Build the 'data matrix'
theta = pinv(Phi)*Y;                    % Estimate the parameters of the model
resid_var = var(Y - Phi*theta);         % Variance of the residuals
a = theta(1);
b = theta(2);

disp(sprintf('LS estimate of model parameters: a = %5.2f, b = %7.2f', a, b));
disp(sprintf('Variance of the residuals: %5.2f', resid_var));

%-------------------------------------------------------------------
% Estimate the power spectrum with the periodogram
dft = fft(ssd);                         % Discrete Fourier transform of the signal
dft = [ dft((N/2+1):N); dft(1:N/2); ];  % Interpret second chunk of the DFT as negative frequencies
dft_spectrum = abs(dft).^2./N;          % Periodogram

%-------------------------------------------------------------------
% Plot the spectrum estimates for comparison
frequencies = linspace(-pi, pi, N);     % Frequency scale for plotting
ar_spectrum = resid_var./( 1 + a^2 + b^2 + 2*a*(b-1)*cos(frequencies) - 2*b*cos(2*frequencies) );

figure(2);
clf();
plot(frequencies, dft_spectrum, 'r');
hold on;
plot(frequencies, ar_spectrum, 'k');
legend('Spectrum estimated with DFT', 'Spectrum estimated with AR fitting');
axis([-pi, pi]);
xlabel('Frequency');
ylabel('Power spectral density');
hold off;
%print -color -depslatexstandalone wolfspec.eps

%-------------------------------------------------------------------
% Compute the peak frequency and the phase of the (conjugate) poles of W(z),
% and the corresponding periods in years

freq_peak = acos(a*(b-1)/(4*b));
r = roots([1, -a, -b]);
freq_root = atan2( imag(r(1)), real(r(1)) );

disp(sprintf('Peak frequency: %5.2f', freq_peak));
disp(sprintf('Corresponding period (in years): %5.1f', 2*pi/freq_peak));
disp(sprintf('Phase of the poles: %5.2f', freq_root));
disp(sprintf('Corresponding period (in years): %5.1f', 2*pi/freq_root));
```

The following was the output:

```
LS estimate of model parameters: a =  1.34, b =   -0.65
Variance of the residuals: 239.31
Peak frequency:  0.56
Corresponding period (in years):  11.2
Phase of the poles:  0.59
Corresponding period (in years):  10.6
```

F Proofs

F.1 Almost sure convergence of the empirical distribution, proof with Hoeffding's inequality

For arbitrary ε and N, the probability of the region B_N^ε of the sample space where $\hat{F}_N(x)$ does not belong to $[F(x)-\varepsilon, F(x)+\varepsilon]$ is, by Hoeffding's inequality, at most

$$\mathsf{P}\left[B_N^\varepsilon\right] = \mathsf{P}\left[|\hat{F}_N(x) - F(x)| \geq \varepsilon\right] \leq 2e^{-2N\varepsilon^2}$$

In general, the probability of the region B_{N+k}^ε where $\hat{F}_{N+k}(x)$ does not belong to $[F(x)-\varepsilon, F(x)+\varepsilon]$ is at most

$$\mathsf{P}\left[B_{N+k}^\varepsilon\right] = \mathsf{P}\left[|\hat{F}_{N+k}(x) - F(x)| \geq \varepsilon\right] \leq 2e^{-2(N+k)\varepsilon^2}$$

Now the probability of the region $B^{\varepsilon,N}$ where $\hat{F}_n(x)$ falls outside $[F(x)-\varepsilon, F(x)+\varepsilon]$ for at least one $n \geq N$ is

$$\begin{aligned}
\mathsf{P}\left[B^{\varepsilon,N}\right] &= \mathsf{P}\left[\bigcup_{k=0}^{\infty} B_{N+k}^\varepsilon\right] \\
&\leq \sum_{k=0}^{\infty} \mathsf{P}\left[B_{N+k}^\varepsilon\right] \quad \text{(by the sub-additivity of } \mathsf{P}\text{)} \\
&\leq \sum_{k=0}^{\infty} 2e^{-2(N+k)\varepsilon^2} \quad \text{(by Hoeffding's inequality)} \\
&= 2e^{-2N\varepsilon^2} \sum_{k=0}^{\infty} e^{-2\varepsilon^2 k} = \frac{2e^{-2N\varepsilon^2}}{1 - e^{-2\varepsilon^2}}
\end{aligned}$$

which tends to 0 as $N \to \infty$. Note that $B^{\varepsilon,N+1} \subset B^{\varepsilon,N}$ for all N. Then, the region B^ε of the sample space in which $\hat{F}_n(x)$ is not always in $[F(x)-\varepsilon, F(x)+\varepsilon]$ after a certain N has probability

$$\mathsf{P}\left[B^\varepsilon\right] = \mathsf{P}\left[\bigcap_{N=1}^{\infty} B^{\varepsilon,N}\right] = \lim_{N\to\infty} \mathsf{P}\left[B^{\varepsilon,N}\right] = 0$$

Note, now, that $B^{\varepsilon_1} \subset B^{\varepsilon_2}$ whenever $\varepsilon_2 < \varepsilon_1$. Finally, the region B where there exists $\varepsilon > 0$ such that $\hat{F}_n(x)$ is not always in $[F(x)-\varepsilon, F(x)+\varepsilon]$ after a certain N has probability

$$\mathsf{P}\left[B\right] = \mathsf{P}\left[\bigcup_{\varepsilon>0} B^\varepsilon\right] = \lim_{\varepsilon\to 0} \mathsf{P}\left[B^\varepsilon\right] = 0$$

The realizations of the sequence $\{\hat{F}_n(x)\}$ that belong to B are precisely those that do not converge to $F(x)$, and they form a set with probability 0. Hence, $\hat{F}_n(x)$ converges to $F(x)$ almost surely. $\square$

Note that the crucial fact of this proof is that the sum $\sum_{k=0}^{\infty} 2e^{-2(N+k)\varepsilon^2}$ converges, and the sum tends to zero as $N \to \infty$. This fact is due to the exponential bound; it would not follow from, say, Čebyšev's inequality.

G Solutions to exercises

G.1 Solutions to exercises for Chapter 1

Solution 1 (price of train tickets).
The model reads

$$p_i = a + bd_i + \varepsilon_i,$$

where p_i is the price of the ticket in €(the "explained" variable), a the fixed price in €, b the proportionality coefficient in €/km, d_i is the distance in km (the "explanatory" variable), and ε_i is a quantization error (in €). The regressors are $\varphi_1(d) = 1$ and $\varphi_2(d) = d$, and a, b are the parameters to be estimated. To pose the problem in compact form, we let

$$Y = \begin{bmatrix} p_1 \\ \vdots \\ p_4 \end{bmatrix}, \quad \Phi = \begin{bmatrix} 1 & d_1 \\ \vdots & \vdots \\ 1 & d_4 \end{bmatrix}, \quad \theta = \begin{bmatrix} a \\ b \end{bmatrix},$$

the normal equations read

$$\Phi^\top \Phi \; \theta = \Phi^\top Y,$$

and the least squares solution is

$$\hat{\theta}_{\text{LS}} = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \arg\min_{\theta} \|\Phi\theta - Y\|^2 = \left(\Phi^\top \Phi\right)^{-1} \Phi^\top Y.$$

Once $\hat{a}$ and $\hat{b}$ are known, the estimated price of a ticket from Milan to Vicenza is $\hat{p} = \hat{a} + \hat{b} \cdot (199.138 \text{ km})$. The Matlab code

```
% Example: estimation of ticket prices
Y = [ 7.00 ;  11.55 ;  15.65 ;  18.35 ];
Phi = [ 1,   82.842 ;
        1,  147.480 ;
        1,  229.408 ;
        1,  266.341  ];
thetaLS = pinv(Phi)*Y
priceToVicenza = thetaLS(1) + thetaLS(2)*199.138
```

yields the estimates $\hat{a} = 2.254566$ €, $\hat{b} = 0.059955$ €/km, and $\hat{p} \simeq 14.20$ €. For comparison, the actual price of a ticket to Vicenza was 14.30 €[49].

Solution 2 (amplitude and phase of a sinusoid).
For brevity, let $\omega = 2\pi F$. The measurement model is then

$$y_i = A \sin(\omega t_i + \phi) + \varepsilon_i.$$

[49] Retrieved from http://www.trenitalia.com on February 25, 2013.

Since $\sin(x+y) = \sin(x)\cos(y) + \cos(x)\sin(y)$, we have

$$y_i = A\cos(\phi)\sin(\omega t_i) + A\sin(\phi)\cos(\omega t_i) + \varepsilon_i$$

Letting $a = A\cos(\phi)$ and $b = A\sin(\phi)$, this becomes a linear model:

$$y_i = \begin{bmatrix} \sin(\omega t_i) & \cos(\omega t_i) \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \varepsilon_i,$$

where of course the explanatory data are the t_i, and the regressors are $\varphi_1(t) = \sin(\omega t), \varphi_2(t) = \cos(\omega t)$. To pose the problem in compact form, we let

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_{10} \end{bmatrix}, \quad \Phi = \begin{bmatrix} \sin(\omega t_1) & \cos(\omega t_1) \\ \vdots & \vdots \\ \sin(\omega t_{10}) & \cos(\omega t_{10}) \end{bmatrix}, \quad \theta = \begin{bmatrix} a \\ b \end{bmatrix},$$

and the least squares solution is, as usual,

$$\hat{\theta}_{\text{LS}} = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \arg\min_{\theta} \|\Phi\theta - Y\|^2 = \left(\Phi^\top \Phi\right)^{-1} \Phi^\top Y.$$

Note that

$$\sqrt{a^2+b^2} = A\sqrt{\cos^2(\phi) + \sin^2(\phi)} = A;$$
$$\frac{b}{a} = \frac{A\sin(\phi)}{A\cos(\phi)} = \tan(\phi).$$

Therefore, once $\hat{a}$ and $\hat{b}$ are known, we can recover an estimate of A and ϕ as follows:

$$\hat{A} = \sqrt{\hat{a}^2 + \hat{b}^2}$$
$$\hat{\phi} = \arctan(\hat{b}/\hat{a})$$

(or $\hat{\phi} = \arctan(\hat{b}/\hat{a}) + \pi$, depending on the signs of $\hat{a}$ and $\hat{b}$). The Matlab code

```
% Example: amplitude and phase of a sinusoid
F = 2;
omega = 2*pi*F;
T = [  2.188; 3.043;  4.207; 4.937;  5.675; 6.104;  6.260;  7.150; 8.600;  9.655 ];
Y = [ -1.112; 2.358; -1.807; 1.202; -0.814; 1.298; -2.520; -0.132; 1.421; -0.302 ];
Phi = [sin(omega*T), cos(omega*T)];
thetaLS = pinv(Phi)*Y;
Ahat    = sqrt(thetaLS(1)^2 + thetaLS(2)^2)
phihat  = atan2(thetaLS(2), thetaLS(1))
```

yields the estimates $\hat{A} = 2.5036$ and $\hat{\phi} = 1.2938$. For comparison, the true values were $A = 2.5$ and $\phi = 1.3$ radians, and ε_i were Gaussian with mean 0 and variance 0.01 (i.e. standard deviation 0.1).

Solution 3 (weighted least squares).

1. To find $\arg\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^N w_i \, (y_i - \varphi_i^\top \theta)^2$, we set equal to zero the derivative with respect to θ in the very same way as we do for ordinary least squares:

$$\begin{aligned}
\frac{\partial}{\partial \theta} \sum_{i=1}^N w_i \, (y_i - \varphi_i^\top \theta)^2 &= \sum_{i=1}^N w_i \, 2(y_i - \varphi_i^\top \theta)(-\varphi_i^\top) \\
&= \sum_{i=1}^N 2 \, w_i \, (y_i - \theta^\top \varphi_i)(-\varphi_i^\top) = 0,
\end{aligned}$$

After grouping terms and transposing, we find

$$\left(\sum_{i=1}^N w_i \, \varphi_i \varphi_i^\top \right) \theta = \sum_{i=1}^N w_i \, \varphi_i y_i. \tag{62}$$

This is the weighted version of the normal equations. Another way, but nicer, to get to the same result, is to bring the weights inside the squares before minimizing:

$$\sum_{i=1}^N w_i \, (y_i - \varphi_i^\top \theta)^2 = \sum_{i=1}^N (\sqrt{w_i} y_i - \sqrt{w_i} \varphi_i^\top \theta)^2;$$

defining $\bar{y}_i = \sqrt{w_i} y_i$ and $\bar{\varphi}_i = \sqrt{w_i} \varphi_i$, the problem becomes

$$\arg\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^N (\bar{y}_i - \bar{\varphi}_i^\top \theta)^2, \tag{63}$$

which is a least squares problem in standard form. The corresponding normal equations are

$$\left(\sum_{i=1}^N \bar{\varphi}_i \bar{\varphi}_i^\top \right) \theta = \sum_{i=1}^N \bar{\varphi}_i \bar{y}_i,$$

which of course are the same as (62), once the coefficients $\sqrt{w_i}$ are extracted back from $\bar{y}_i$ and $\bar{\varphi}_i$.

2. In the same spirit as in (63), we define

$$\bar{Y} = \begin{bmatrix} \sqrt{w_1} y_1 \\ \vdots \\ \sqrt{w_N} y_N \end{bmatrix} = \begin{bmatrix} \sqrt{w_1} & & \\ & \ddots & \\ & & \sqrt{w_N} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = W^{1/2} Y,$$
$$\bar{\Phi} = \begin{bmatrix} \sqrt{w_1} \varphi_1^\top \\ \vdots \\ \sqrt{w_1} \varphi_N^\top \end{bmatrix} = \begin{bmatrix} \sqrt{w_1} & & \\ & \ddots & \\ & & \sqrt{w_N} \end{bmatrix} \begin{bmatrix} \varphi_1^\top \\ \vdots \\ \varphi_N^\top \end{bmatrix} = W^{1/2} \Phi,$$

where $W = \operatorname{diag}(w_1, \cdots, w_N) \in \mathbb{R}^{N \times N}$, and $W^{1/2}$ denotes its square root. The problem then reads

$$\arg\min_{\theta \in \mathbb{R}^p} \left\| \bar{\Phi}\theta - \bar{Y} \right\|_2^2,$$

the corresponding normal equations are

$$\bar{\Phi}^\top \bar{\Phi} \theta = \bar{\Phi}^\top \bar{Y};$$
$$\Phi^\top \left(W^{1/2} \right)^\top W^{1/2} \Phi \theta = \Phi^\top \left(W^{1/2} \right)^\top W^{1/2} Y;$$
$$\Phi^\top W \Phi \theta = \Phi^\top W Y,$$

and finally

$$\hat{\theta}_{\text{WLS}} = \left(\Phi^\top W \Phi \right)^{-1} \Phi^\top W Y.$$

Solution 4 (ranges).
Suppose that $v \in \text{null } \Phi$. This means $\Phi v = 0$, hence also $\Phi^\top \Phi v = 0$ and $v \in \text{null } \Phi^\top \Phi$. Suppose, on the other hand, that $v \in \text{null } \Phi^\top \Phi$. Then $\Phi^\top \Phi v = 0$, hence also $\|\Phi v\|_2^2 = (\Phi v)^\top \Phi v = v^\top \Phi^\top \Phi v = 0$. This implies that $\Phi v = 0$ and $v \in \text{null } \Phi$. Hence $\text{null } \Phi^\top \Phi = \text{null } \Phi$.
Now, since all the spaces in consideration are subspaces of finite-dimensional vector spaces,

$$\text{range } \Phi^\top \Phi = \left(\text{null } \left(\Phi^\top \Phi \right)^\top \right)^\perp = \left(\text{null } \Phi^\top \Phi \right)^\perp = (\text{null } \Phi)^\perp = \text{range } \Phi^\top,$$

and consequently

$$\text{rank } \Phi^\top \Phi = \dim \text{range } \Phi^\top \Phi = \dim \text{range } \Phi^\top = \text{rank } \Phi^\top.$$

Solution 5 (systematic errors).
Since ε_i are Gaussian with mean μ and variance σ^2, we can write $\varepsilon_i = \bar{\varepsilon}_i + \mu$, where $\bar{\varepsilon}_i$ are Gaussian with mean zero. The variables $\bar{\varepsilon}_i$ are still independent of each other and independent of φ_i.
Consider the normal equations, with the substitution $y_i = \varphi_i^\top \theta^o + \varepsilon_i$ and divided by N:

$$\left(\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varphi_i^\top\right)\hat{\theta}_{\mathrm{LS}} = \left(\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varphi_i^\top\right)\theta^o + \frac{1}{N}\sum_{i=1}^{N}\varphi_i\varepsilon_i$$

By a strong law of large numbers, $\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varphi_i^\top \to \Sigma$ almost surely, hence for big N the matrix $\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varphi_i^\top$ is invertible, and

$$\begin{aligned}
\hat{\theta}_{\mathrm{LS}} &= \theta^o + \left(\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varphi_i^\top\right)^{-1}\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varepsilon_i \\
&= \theta^o + \left(\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varphi_i^\top\right)^{-1}\frac{1}{N}\sum_{i=1}^{N}\varphi_i\bar{\varepsilon}_i + \left(\frac{1}{N}\sum_{i=1}^{N}\varphi_i\varphi_i^\top\right)^{-1}\frac{1}{N}\sum_{i=1}^{N}\varphi_i\mu
\end{aligned}$$

Now, since $\bar{\varepsilon}_i$ and φ_i are independent, and $\mathsf{E}[\bar{\varepsilon}_i] = 0$, the second term converges to zero almost surely by a strong law of large numbers. And since μ is a constant (we can bring it outside the sum), the third term also converges almost surely, namely to $\Sigma^{-1}\bar{\varphi}\mu$. Hence,

$$\hat{\theta}_{\mathrm{LS}} \to \theta^o + \Sigma^{-1}\bar{\varphi}\mu \quad \text{almost surely.}$$

The take-home message is that, in general, you cannot pretend the least squares method to be consistent in the presence of a *systematic error* (μ).

G.2 Solutions to exercises for Chapter 2

Solution 1 (SVD and pseudo-inverse).

1. We check that U and V are orthogonal; indeed

$$\begin{aligned}
\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} &= \begin{bmatrix} 1/\sqrt{5} & -2/\sqrt{5} \\ 2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix}\begin{bmatrix} 1/\sqrt{5} & 2/\sqrt{5} \\ -2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix} \\
&= \begin{bmatrix} 1/\sqrt{5} & 2/\sqrt{5} \\ -2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix}\begin{bmatrix} 1/\sqrt{5} & -2/\sqrt{5} \\ 2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix}; \\
\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} &= \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \\
&= \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}\begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}.
\end{aligned}$$

Moreover, it holds

$$\begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{5} & -2/\sqrt{5} \\ 2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix} \begin{bmatrix} \sqrt{10} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}.$$

Thus, $A = U\Sigma V^\top$ as required. Note that the eigenvalues of both $AA^\top$ and $A^\top A$ are 10 and 0.

2. From the previous point, we have

$$\begin{aligned} A^+ = V\Sigma^+ U^\top &= \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 1/\sqrt{10} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{5} & 2/\sqrt{5} \\ -2/\sqrt{5} & 1/\sqrt{5} \end{bmatrix} \\ &= \begin{bmatrix} 1/10 & 1/5 \\ 1/10 & 1/5 \end{bmatrix}. \end{aligned}$$

Solution 2 (orthogonal projector).
Any $v \in \mathbb{R}^m$ can be decomposed in a unique way as

$$v = v^c + v^\perp,$$

where $v^c \in \text{span}\,\{\text{columns of } A\} = \text{range}\, A$ (v^c is the requested orthogonal projection) and $v^\perp \in \text{span}\,\{\text{columns of } A\}^\perp = (\text{range}\, A)^\perp = \text{null}\, A^\top$. Specifically,

$$\begin{aligned} v^c &= Ax \quad \text{for some } x \in \mathbb{R}^n; \\ A^\top v^\perp &= 0. \end{aligned}$$

Therefore, recalling the defining properties of the pseudo-inverse,

$$\begin{aligned} \Pi_A v &= AA^+(v^c + v^\perp) \\ &= \left(AA^+A\right)x + \left(AA^+\right)v^\perp \\ &= Ax + \left(AA^+\right)^\top v^\perp \\ &= v^c + \left(A^+\right)^\top A^\top v^\perp \\ &= v^c. \end{aligned}$$

G.3 Solutions to exercises for Chapter 5

Solution 1 (complaint telephone calls).

Let $r \in R = \{\text{Piedmont}, \text{Lombardy}, \cdots, \text{Sicily}\}$ denote regions. We estimate the probability $p(r) = \mathsf{P}[r_i = r]$ of a call from r with the "empirical mass distribution" (i.e. frequency)

$$\hat{p}(r) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\{r_i = r\}} = \frac{\text{number of calls received from } r}{\text{total calls}}$$

Of course, since $\mathsf{E}[\mathbb{1}_{\{r_i=r\}}] = \mathsf{P}[r_i = r] = p(r)$,

$$\mathsf{E}[\hat{p}(r)] = p(r),$$

hence, using Hoeffding's inequality, the probability that $|\hat{p}(r) - p(r)| > \varepsilon$ at *any* of the 20 regions is

$$\begin{aligned}\mathsf{P}\left[\bigcup_{r\in R} \{|\hat{p}(r) - p(r)| > \varepsilon\}\right] &\leq \sum_{r\in R} \mathsf{P}\left[|\hat{p}(r) - p(r)| > \varepsilon\right] \\ &\leq \sum_{r\in R} 2e^{-2N\varepsilon^2} \\ &= 40e^{-2N\varepsilon^2}.\end{aligned}$$

The problem asks precisely to find N such that $40e^{-2N\varepsilon^2} \leq 10^{-4}$. Solving for N,

$$\begin{aligned} e^{-2N\left(\frac{1}{100}\right)^2} &\leq 25 \cdot 10^{-7}; \\ -2N\frac{1}{100^2} &\leq \log(25) - 7\log(10) \simeq -12.9; \\ N &\geq 5000 \cdot 12.9 = 64500. \end{aligned}$$

Thus, any $N \geq 64500$ will do.

Solution 2 (finitely many classifiers).
First, we prove that almost surely, $\hat{J}_N \to \bar{J}$ uniformly. Remembering the proof of Glivenko/Cantelli's theorem, this is quite easy; indeed by the strong law of large numbers, at all points $c \in C$ it holds $\hat{J}_N(c) \to \bar{J}(c)$ almost surely. Hence, almost surely for all $\varepsilon > 0$ there exists N_c such that $|\hat{J}_N(c) - \bar{J}(c)| \leq \varepsilon$ for all $N \geq N_c$. Since the c are finitely many, it is well defined the index $\bar{N} := \max_{c\in C} N_c$, such that for all $N \geq \bar{N}$ the inequalities

$$|\hat{J}_N(c) - \bar{J}(c)| \leq \varepsilon, \quad c \in C$$

hold simultaneously; this is enough to establish uniform convergence.
Now we can invoke the lemma on uniform convergence, *exploiting all the hypotheses*, because

- any *finite* subset $C \subset \mathbb{R}^p$ is automatically compact (because it is closed and bounded — recall the Heine/Borel theorem);
- any function defined on a finite set $C \subset \mathbb{R}^p$ is automatically continuous;
- $\bar{c}$ is unique by assumption.

(If you are not at ease with the second claim, see the remarks after Definition 4.5 in [28]; alternatively, you may develop an ad-hoc version of the lemma on uniform convergence.)
It follows that $\hat{c}_N \to \bar{c}$ almost surely.
Finally, recalling the definitions of $\hat{J}_N$ and $\bar{J}$, and exploiting Hoeffding's inequality,

$$\begin{aligned}
\mathsf{P}\left[\max_{c\in C} |\hat{J}_N(c) - \bar{J}(c)| \geq \varepsilon\right] &= \mathsf{P}\left[\bigcup_{c\in C} \left\{|\hat{J}_N(c) - \bar{J}(c)| \geq \varepsilon\right\}\right] \\
&\leq \sum_{c\in C} \mathsf{P}\left[|\hat{J}_N(c) - \bar{J}(c)| \geq \varepsilon\right] \\
&\leq \sum_{c\in C} 2e^{-2N\varepsilon^2} \\
&= 2Ke^{-2N\varepsilon^2}.
\end{aligned}$$

G.4 Solutions to exercises for Chapter 6

Problem 1 (discrete distribution, wrong confidence).
The LSCR method works by considering 3 partial-average functions:

$$\begin{aligned}
g_1(\theta) &= \frac{y_1 + y_2}{2} - \theta = (\theta^o - \theta) + \frac{\varepsilon_1 + \varepsilon_2}{2}; \\
g_2(\theta) &= \frac{y_1 + y_3}{2} - \theta = (\theta^o - \theta) + \frac{\varepsilon_1 + \varepsilon_3}{2}; \\
g_3(\theta) &= \frac{y_2 + y_3}{2} - \theta = (\theta^o - \theta) + \frac{\varepsilon_2 + \varepsilon_3}{2}.
\end{aligned}$$

Their respective intersections with the θ-axis are

$$\begin{aligned}
\theta_1 &= \frac{y_1 + y_2}{2}, \\
\theta_2 &= \frac{y_1 + y_3}{2}, \\
\theta_3 &= \frac{y_2 + y_3}{2},
\end{aligned}$$

and they split it in 4 intervals (the outermost two being semi-infinite), where θ^o falls with equal probability. Thus, a reasonable choice is to choose $[\bar{\theta}_1, \bar{\theta}_3]$ as a 50%-confidence interval, where $\bar{\theta}_1 = \min(\theta_1, \theta_2, \theta_3)$ and $\bar{\theta}_3 = \max(\theta_1, \theta_2, \theta_3)$.

Let us tabulate y_i and θ_i for each possible value of ε_1, ε_2, and ε_3:

ε_1	ε_2	ε_3	y_1	y_2	y_3	θ_1	θ_2	θ_3	interval	contains θ^o?
1	1	1	2	2	2	2	2	2	[2, 2]	no
1	1	−1	2	2	0	2	1	1	[1, 2]	yes
1	−1	1	2	0	2	1	2	1	[1, 2]	yes
1	−1	−1	2	0	0	1	1	0	[0, 1]	yes
−1	1	1	0	2	2	1	1	2	[1, 2]	yes
−1	1	−1	0	2	0	1	0	1	[0, 1]	yes
−1	−1	1	0	0	2	0	1	1	[0, 1]	yes
−1	−1	−1	0	0	0	0	0	0	[0, 0]	no

(Of course, an "interval" like $[2,2]$ means the set $\{2\}$.) As you can see, in 6 cases out of 8 the interval $[\bar{\theta}_1, \bar{\theta}_3]$ computed in the last-but-one column contains θ^o; the confidence of the interval $[\bar{\theta}_1, \bar{\theta}_3]$ is 75%, not 50% (thus, in this case the conclusions of the LSCR theory are *conservative*).

Problem 2 (discrete distribution, correct confidence).
The LSCR method works by considering 3 partial-average functions:

$$\begin{aligned} g_1(\theta) &= \frac{y_1+y_2}{2} - \theta = (\theta^o - \theta) + \frac{\varepsilon_1+\varepsilon_2}{2}; \\ g_2(\theta) &= \frac{y_1+y_3}{2} - \theta = (\theta^o - \theta) + \frac{\varepsilon_1+\varepsilon_3}{2}; \\ g_3(\theta) &= \frac{y_2+y_3}{2} - \theta = (\theta^o - \theta) + \frac{\varepsilon_2+\varepsilon_3}{2}. \end{aligned}$$

Their respective intersections with the θ-axis are

$$\begin{aligned} \theta_1 &= \frac{y_1+y_2}{2}, \\ \theta_2 &= \frac{y_1+y_3}{2}, \\ \theta_3 &= \frac{y_2+y_3}{2}, \end{aligned}$$

and they split it in 4 intervals (the outermost two being semi-infinite), where θ^o falls with equal probability. Thus, a reasonable choice is to choose $[\bar{\theta}_1, \bar{\theta}_3]$ as a 50%-confidence interval, where $\bar{\theta}_1 = \min(\theta_1, \theta_2, \theta_3)$ and $\bar{\theta}_3 = \max(\theta_1, \theta_2, \theta_3)$.
We tabulate y_i and θ_i for each possible value of ε_1, ε_2, and ε_3 (all the θ_i are written as multiples of $\frac{1}{4}$ for ease of reading):

ε_1	ε_2	ε_3	y_1	y_2	y_3	θ_1	θ_2	θ_3	interval	contains θ^o?
1	1/2	2	2	3/2	3	7/4	10/4	9/4	[7/4, 10/4]	no
1	1/2	−2	2	3/2	−1	7/4	2/4	1/4	[1/4, 7/4]	yes
1	−1/2	2	2	1/2	3	5/4	10/4	7/4	[5/4, 10/4]	no
1	−1/2	−2	2	1/2	−1	5/4	2/4	−1/4	[−1/4, 5/4]	yes
−1	1/2	2	0	3/2	3	3/4	6/4	9/4	[3/4, 9/4]	yes
−1	1/2	−2	0	3/2	−1	3/4	−2/4	1/4	[−2/4, 3/4]	no
−1	−1/2	2	0	1/2	3	1/4	6/4	7/4	[1/4, 7/4]	yes
−1	−1/2	−2	0	1/2	−1	1/4	−2/4	−1/4	[−2/4, 1/4]	no

As you can see, in 4 cases out of 8 the interval $[\bar{\theta}_1, \bar{\theta}_3]$ computed in the last-but-one column contains θ^o; thus, $[\bar{\theta}_1, \bar{\theta}_3]$ is indeed a 50%-confidence interval for θ^o. In this case, the result is exactly what the LSCR theory claims, despite the fact that the distributions are discrete (in this particular example, no two intersections θ_i coincide, whatever the values of $\varepsilon_1, \varepsilon_2, \varepsilon_3$).

References

[1] Patrick Billingsley, *Probability and Measure 3rd. ed.*, Wiley, 1995.

[2] Stephen Boyd, Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[3] Marco C. Campi, *Selected topics in probability*, see `http://www.ing.unibs.it/~campi/`.

[4] Marco Dalai, Erik Weyer, Marco C. Campi, *Parameter identification for nonlinear systems: Guaranteed confidence regions through LSCR*, Automatica, vol. 43, pp. 1418-1425, 2007.

[5] Marco C. Campi, *What can be learned from data? (Identification and the intrinsic limits in learning from data).* (Contact Prof. Campi to obtain a copy.)

[6] Marco C. Campi, Giuseppe Calafiore, and Simone Garatti, *Interval Predictor Models: Identification and Reliability*, Automatica, 45:382-392, 2009.

[7] Marco C. Campi, Erik Weyer, *Identification with finitely many data points: the LSCR approach*, Proceedings of the 14th Symposium on System Identification (SYSID), Newcastle, Australia, 2006.

[8] William Feller, *An Introduction to Probability Theory and its Applications vol. I*, Wiley, 1968.

[9] William Feller, *An Introduction to Probability Theory and its Applications vol. II*, Wiley, 1971.

[10] Enrico Gregorio, Luigi Salce, *Algebra lineare*, Libreria Progetto, Padova, 2005.

[11] Robert V. Hogg, Allen T. Craig, *Introduction to Mathematical Statistics, 5th ed.*, Prentice Hall, 1995.

[12] Gene H. Golub, Charles F. Van Loan, *Matrix Computations 3rd ed.*, Johns Hopkins University Press, Baltimore, 1996.

[13] L. Gordon, *Completely separating groups in subsampling*, Annals of Statistics, vol. 2, pp. 572-578, 1974.

[14] Roger A. Horn, Charles R. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1985.

[15] Jean Jacod, Philip Protter, *Probability Essentials*, Springer, 2004.

[16] Serge Lang, *Linear Algebra*, Addison-Wesley, 1966. Trad. it. *Algebra lineare*, Bollati Boringhieri, 2000.

[17] Lennart Ljung, *System identification - Theory for the user, 2nd ed.*, Prentice Hall, 1999.

[18] David G. Luenberger, *Optimization by Vector Space Methods*, Wiley, 1969.

[19] K. Madsen, H.B. Nielsen, O. Tingleff, *Methods for Non-Linear Least Squares Problems*, Informatics and Mathematical Modelling, Technical University of Denmark, 2004. See `http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/3215/pdf/imm3215.pdf`

[20] MathWorks Inc., Matlab's System Identification Toolbox manual, see `http://www.mathworks.com/products/sysid/index.html`

[21] Douglas C. Montgomery, George C. Runger, Norma Faris Hubele, *Engineering Statistics 3rd ed.*, Wiley, 2004. Trad. it. *Statistica per ingegneria*, EGEA, 2004.

[22] A.V. Oppenheim, R.W. Schafer, *Digital signal processing*, Prentice-Hall, 1975.

[23] Giorgio Picci, *Metodi statistici per l'identificazione di sistemi lineari* (lecture notes). See `http://www.dei.unipd.it/~picci`

[24] Lawrence R. Rabiner, Ronald W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, 1978.

[25] R. Tyrrell Rockafellar, *Convex analysis*, Princeton University Press, 1997.

[26] Sheldon M. Ross, *A first course in probability, 7th ed.*, Pearson Education/Prentice Hall, 2006. Trad. it. *Calcolo delle probabilità*, APOGEO, 2007.

[27] Riccardo Rovatti, *Elementi di teoria statistica dei segnali*, Zanichelli, 2005.

[28] Walter Rudin, *Principles of mathematical analysis 3rd ed.*, McGraw-Hill, 1976.

[29] Luigi Salce, *Lezioni sulle matrici*, Decibel/Zanichelli, Padova, 1993.

[30] Torsten Söderström, Petre Stoica, *System Identification*, Prentice Hall UK, 1989. Available for download at `http://www.it.uu.se/research/syscon/Ident`

[31] G. Udny Yule, *On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers*, Phil. Trans. R. Soc. of London, 1927.