Chapter 1

Mathematical Concepts

1.1 Basic concepts on probability

Without describing in detail the formalism used by Probability Theory, we will simply remind the reader of some useful concepts. However we advise the reader to consult some of the many books with authority on the subject [1].

Definition 1.1 (Discrete random variable) *A random variable X is said to be discrete if the set of its possible values is, at the most, countable. If $\{a_0, \ldots, a_n, \ldots\}$, where $n \in \mathbb{N}$, is the set of its values, the probability distribution of X is characterized by the sequence:*

$$p_X(n) = \Pr(X = a_n) \tag{1.1}$$

representing the probability that X is equal to the element a_n. These values are such that $0 \leq p_X(n) \leq 1$ and $\sum_{n\geq 0} p_X(n) = 1$.

This leads us to the probability for the random variable X to belong to the interval $]a, b]$. It is given by:

$$\Pr(X \in]a, b]) = \sum_{n\geq 0} p_X(n) \mathbb{1}(a_n \in]a, b])$$

The function defined for $x \in \mathbb{R}$ by:

$$\begin{aligned} F_X(x) &= \Pr(X \leq x) = \sum_{\{n:a_n \leq x\}} p_X(n) \\ &= \sum_{n\geq 0} p_X(n) \mathbb{1}(a_n \in]-\infty, x]) \end{aligned} \tag{1.2}$$

is called the *cumulative distribution function* (*cdf*) of the random variable X. It is a monotonic increasing function, and verifies $F_X(-\infty) = 0$ and $F_X(+\infty) = 1$.

Its graph resembles that of a staircase function, the jumps of which are located at x-coordinates a_n and have an amplitude of $p_X(n)$.

Definition 1.2 (Two discrete random variables) *Let X and Y be two discrete random variables, with possible values $\{a_0, \ldots, a_n, \ldots\}$ and $\{b_0, \ldots, b_k, \ldots\}$ respectively. The joint probability distribution is characterized by the sequence of positive values:*

$$p_{XY}(n,k) = \Pr(X = a_n, Y = b_k) \tag{1.3}$$

with $0 \leq p_{XY}(n,k) \leq 1$ and $\sum_{n\geq 0}\sum_{k\geq 0} p_{XY}(n,k) = 1$.

$\Pr(X = a_n, Y = b_k)$ represents the probability to *simultaneously* have $X = a_n$ and $Y = b_k$. This definition can easily be extended to the case of a finite number of random variables.

Property 1.1 (Marginal probability distribution) *Let X and Y be two discrete random variables, with possible values $\{a_0, \ldots, a_n, \ldots\}$ and $\{b_0, \ldots, b_k, \ldots\}$ respectively, and with their joint probability distribution characterized by $p_{XY}(n,k)$. We have:*

$$\begin{aligned} p_X(n) &= \Pr(X = a_n) = \sum_{k=0}^{+\infty} p_{XY}(n,k) \\ p_Y(k) &= \Pr(Y = b_k) = \sum_{n=0}^{+\infty} p_{XY}(n,k) \end{aligned} \tag{1.4}$$

$p_X(n)$ and $p_Y(k)$ denote the marginal probability distribution of X and Y respectively.

Definition 1.3 (Continuous random variable) *A random variable is said to be continuous*[1] *if its values belong to $\mathbb{R}$ and if, for any real numbers a and b, the probability that X belongs to the interval $]a,b]$ is given by:*

$$\Pr(X \in]a,b]) = \int_a^b p_X(x)dx = \int_{-\infty}^{\infty} p_X(x)\mathbb{1}(x \in]a,b])dx \tag{1.5}$$

where $p_X(x)$ is a function that must be positive or equal to zero such that $\int_{-\infty}^{+\infty} p_X(x)dx = 1$. $p_X(x)$ is called the probability density function (pdf) of X.

[1]The exact expression says that the probability distribution of X is *absolutely continuous with respect to the Lebesgue measure.*

The function defined for any $x \in \mathbb{R}$ by:

$$F_X(x) = \Pr(X \leq x) = \int_{-\infty}^{x} p_X(u)du \tag{1.6}$$

is called the *cumulative distribution function* (*cdf*) of the random variable X. It is a monotonic increasing function and it verifies $F_X(-\infty) = 0$ and $F_X(+\infty) = 1$. Notice that $p_X(x)$ also represents the derivative of $F_X(x)$ with respect to x.

Definition 1.4 (Two continuous random variables) *Let X and Y be two random variables with possible values in $\mathbb{R} \times \mathbb{R}$. They are said to be continuous if, for any domain Δ of $\mathbb{R}^2$, the probability that the pair (X, Y) belongs to Δ is given by:*

$$\Pr((X, Y) \in \Delta) = \iint_{\Delta} p_{XY}(x, y)dxdy \tag{1.7}$$

where the function $p_{XY}(x, y) \geq 0$, and is such that:

$$\iint_{\mathbb{R}^2} p_{XY}(x, y)dxdy = 1$$

$p_{XY}(x, y)$ is called the joint probability density function of the pair (X, Y).

Property 1.2 (Marginal probability distributions) *Let X and Y be two continuous random variables with a joint probability distribution characterized by $p_{XY}(x, y)$. The probability distributions of X and Y have the following marginal probability density functions:*

$$\begin{aligned} p_X(x) &= \int_{-\infty}^{+\infty} p_{XY}(x, y)dy \\ p_Y(y) &= \int_{-\infty}^{+\infty} p_{XY}(x, y)dx \end{aligned} \tag{1.8}$$

An example involving two real random variables (X, Y) is the case of a complex random variable $Z = X + jY$.

It is also possible to have a mixed situation, where one of the two variables is discrete and the other is continuous. This leads to the following:

Definition 1.5 (Mixed random variables) *Let X be a discrete random variable with possible values $\{a_0, \ldots, a_n, \ldots\}$ and Y a continuous random variable*

with possible values in $\mathbb{R}$. *For any value* a_n, *and for any real numbers* a *and* b, *the probability:*

$$\Pr(X = a_n, Y \in]a, b]) = \int_a^b p_{XY}(n, y)dy \tag{1.9}$$

where the function $p_{XY}(n, y)$, *with* $n \in \{0, \ldots, k, \ldots\}$ *and* $y \in \mathbb{R}$, *is* ≥ 0 *and verifies* $\sum_{n\geq 0} \int_{\mathbb{R}} p_{XY}(n, y)dy = 1$.

Definition 1.6 (Two independent random variables) *Two random variables* X *and* Y *are said to be independent if and only if their joint probability distribution is the product of the marginal probability distributions. This can be expressed:*

- *for two discrete random variables:*

$$p_{XY}(n, k) = p_X(n)p_Y(k)$$

- *for two continuous random variables:*

$$p_{XY}(x, y) = p_X(x)p_Y(y)$$

- *for two mixed random variables:*

$$p_{XY}(n, y) = p_X(n)p_Y(y)$$

where the marginal probability distributions are obtained with formulae (1.4) *and* (1.8).

It is worth noting that, knowing $p_{XY}(x, y)$, we can tell whether or not X and Y are independent. To do this, we need to calculate the marginal probability distributions and to check that $p_{XY}(x, y) = p_X(x)p_Y(y)$. If that is the case, then X and Y are independent.

The following definition is more general.

Definition 1.7 (Independent random variables) *The random variables* $(X_1, \ldots, X_n)$ *are jointly independent if and only if their joint probability distribution is the product of their marginal probability distributions. This can be expressed:*

$$p_{X_1X_2\ldots X_n}(x_1, x_2, \ldots, x_n) = p_{X_1}(x_1)p_{X_2}(x_2)\ldots p_{X_n}(x_n) \tag{1.10}$$

where the marginal probability distributions are obtained as integrals with respect to $(n-1)$ *variables, calculated from* $p_{X_1X_2\ldots X_n}(x_1, x_2, \ldots, x_n)$.

For example, the marginal probability distribution of X_1 has the expression:

$$p_{X_1}(x_1) = \underbrace{\int \dots \int}_{\mathbb{R}^{n-1}} p_{X_1 X_2 \dots X_n}(x_1, x_2, \dots, x_n) dx_2 \dots dx_n$$

In practice, the following result is a simple method for determining whether or not random variables are independent: if $p_{X_1 X_2 \dots X_n}(x_1, x_2, \dots, x_n)$ is a product of n positive functions of the type $f_1(x_1) f_2(x_2) \dots f_n(x_n)$, then the variables are independent.

It should be noted that if n random variables are independent of one another, it does not necessarily mean that they are jointly independent.

Definition 1.8 (Mathematical expectation) *Let X be a random variable and $f(x)$ a function. The mathematical expectation of $f(X)$ – respectively $f(X, Y)$ – is the value, denoted by $\mathbb{E}\{f(X)\}$ – respectively $\mathbb{E}\{f(X, Y)\}$ – defined:*

– *for a discrete random variable, by:*

$$\mathbb{E}\{f(X)\} = \sum_{n \geq 0} f(a_n) p_X(n)$$

– *for a continuous random variable, by:*

$$\mathbb{E}\{f(X)\} = \int_{\mathbb{R}} f(x) p_X(x) dx$$

– *for two discrete random variables, by:*

$$\mathbb{E}\{f(X, Y)\} = \sum_{n \geq 0} \sum_{k \geq 0} f(a_n, b_k) p_{XY}(n, k)$$

– *for two continuous random variables, by:*

$$\mathbb{E}\{f(X, Y)\} = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x, y) p_{XY}(x, y) dx dy$$

provided that all expressions exist.

Property 1.3 *If $\{X_1, X_2, \dots, X_n\}$ are jointly independent, then for any integrable functions $f_1, f_2, \dots, f_n$:*

$$\mathbb{E}\left\{\prod_{k=1}^{n} f_k(X_k)\right\} = \prod_{k=1}^{n} \mathbb{E}\{f_k(X_k)\} \tag{1.11}$$

Definition 1.9 (Characteristic function) *The characteristic function of the probability distribution of the random variables* $X_1, \ldots, X_n$ *is the function of* $(u_1, \ldots, u_n) \in \mathbb{R}^n$ *defined by:*

$$\phi_{X_1 \ldots X_n}(u_1, \ldots, u_n) = \mathbb{E}\left\{e^{ju_1X_1 + \cdots + ju_nX_n}\right\} = \mathbb{E}\left\{\prod_{k=1}^{n} e^{ju_kX_k}\right\} \tag{1.12}$$

Because $\left|e^{juX}\right| = 1$, the characteristic function exists and is continuous even if the moments $\mathbb{E}\left\{X^k\right\}$ do not exist. The Cauchy probability distribution, for example, the probability density function of which is $p_X(x) = 1/\pi(1 + x^2)$, has no moment and has the characteristic function $e^{-|u|}$. Let us notice $|\phi_{X_1 \ldots X_n}(u_1, \ldots, u_n)| \leq \phi_X(0, \ldots, 0) = 1$.

Theorem 1.1 (Fundamental) $(X_1, \ldots, X_n)$ *are independent if and only if for any point* $(u_1, u_2, \ldots, u_n)$ *of* $\mathbb{R}^n$*:*

$$\phi_{X_1 \ldots X_n}(u_1, \ldots, u_n) = \prod_{k=1}^{n} \phi_{X_k}(u_k)$$

Notice that the characteristic function $\phi_{X_k}(u_k)$ of the marginal probability distribution of X_k can be directly calculated using (1.12). We have $\phi_{X_k}(u_k) = \mathbb{E}\left\{e^{ju_kX_k}\right\} = \phi_{X_1 \ldots X_n}(0, \ldots, 0, u_k, 0, \ldots, 0)$.

Definition 1.10 (Mean, variance) *The mean of the random variable* X *is defined as the first order moment, that is to say* $\mathbb{E}\{X\}$*. If the mean is equal to zero, the random variable is said to be centered. The variance of the random variable* X *is the quantity defined by:*

$$\operatorname{var}(X) = \mathbb{E}\left\{(X - \mathbb{E}\{X\})^2\right\} = \mathbb{E}\left\{X^2\right\} - (\mathbb{E}\{X\})^2 \tag{1.13}$$

The variance is always positive, and its square root is called the standard deviation.

As an exercise, we are going to show that, for any constants a and b:

$$\mathbb{E}\{aX + b\} = a\mathbb{E}\{X\} + b \tag{1.14}$$

$$\operatorname{var}(aX + b) = a^2 \operatorname{var}(X) \tag{1.15}$$

Expression (1.14) is a direct consequence of the integral's linearity. We assume that $Y = aX + b$, then $\operatorname{var}(Y) = \mathbb{E}\left\{(Y - \mathbb{E}\{Y\})^2\right\}$. By replacing $\mathbb{E}\{Y\} = a\mathbb{E}\{X\} + b$, we get $\operatorname{var}(Y) = \mathbb{E}\left\{a^2(X - \mathbb{E}\{X\})^2\right\} = a^2 \operatorname{var}(X)$.

A generalization of these two results to random vectors (their components are random variables) will be given by property (1.6).

Definition 1.11 (Covariance, correlation) *Let (X, Y) be two random variables*[2]*. The covariance of X and Y is the quantity defined by:*

$$\begin{aligned} \operatorname{cov}(X,Y) &= \mathbb{E}\{(X - \mathbb{E}\{X\})(Y^* - \mathbb{E}\{Y^*\})\} \\ &= \mathbb{E}\{XY^*\} - \mathbb{E}\{X\}\,\mathbb{E}\{Y^*\} \end{aligned} \tag{1.16}$$

In what follows, the variance of the random variable X will be noted $\operatorname{var}(X)$. $\operatorname{cov}(X)$ *or* $\operatorname{cov}(X,X)$ *have exactly the same meaning.*

X and Y are said to be uncorrelated if $\operatorname{cov}(X,Y) = 0$ *that is to say if* $\mathbb{E}\{XY^*\} = \mathbb{E}\{X\}\,\mathbb{E}\{Y^*\}$. *The correlation coefficient is the quantity defined by:*

$$\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)}\sqrt{\operatorname{var}(Y)}} \tag{1.17}$$

Applying the Schwartz inequality gives us $|\rho(X,Y)| \leq 1$.

Definition 1.12 (Mean vector and covariance matrix) *Let $\{X_1, \ldots, X_n\}$ be n random variables with the respective means $\mathbb{E}\{X_i\}$. The mean vector is the n dimension vector with the means $\mathbb{E}\{X_i\}$ as its components. The $n \times n$ covariance matrix $\boldsymbol{C}$ is the matrix with the generating element $C_{ij} = \operatorname{cov}(X_i, X_j)$ for $1 \leq i \leq n$ and $1 \leq j \leq n$.*

Matrix notation: if we write

$$\boldsymbol{X} = \begin{bmatrix} X_1 & \ldots & X_n \end{bmatrix}^T$$

to refer to the random vector with the random variable X_k as its k-th component, the mean-vector can be expressed:

$$\mathbb{E}\{\boldsymbol{X}\} = \begin{bmatrix} \mathbb{E}\{X_1\} & \ldots & \mathbb{E}\{X_n\} \end{bmatrix}^T$$

the covariance matrix:

$$\begin{aligned} \boldsymbol{C} &= \mathbb{E}\left\{(\boldsymbol{X} - \mathbb{E}\{\boldsymbol{X}\})(\boldsymbol{X} - \mathbb{E}\{\boldsymbol{X}\})^H\right\} \\ &= \mathbb{E}\left\{\boldsymbol{X}\boldsymbol{X}^H\right\} - \mathbb{E}\{\boldsymbol{X}\}\,\mathbb{E}\{\boldsymbol{X}\}^H \end{aligned} \tag{1.18}$$

and the *correlation matrix*

$$\boldsymbol{R} = \boldsymbol{D}\boldsymbol{C}\boldsymbol{D} \tag{1.19}$$

[2]Except in some particular cases, the random variables considered from now on will be real. However, the definitions involving the mean and the covariance can be generalized with no exceptions to complex variables by conjugating the second variable. This is indicated by a star $(*)$ in the case of scalars and by the exponent H in the case of vectors.

with

$$\boldsymbol{D} = \begin{bmatrix} C_{11}^{-1/2} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & C_{nn}^{-1/2} \end{bmatrix} \tag{1.20}$$

$\boldsymbol{R}$ is obtained by dividing each element C_{ij} of $\boldsymbol{C}$ by $\sqrt{C_{ii}C_{jj}}$, provided that $C_{ii} \neq 0$. Therefore $R_{ii} = 1$ and $|R_{ij}| \leq 1$.

Notice that the diagonal elements of a covariance matrix represent the respective variances of the n random variables. They are therefore positive. *If the n random variables are uncorrelated, their covariance matrix is diagonal and their correlation matrix is the identity matrix.*

Property 1.4 (Positivity of the covariance matrix) *Any covariance matrix is positive, meaning that for any vector $\boldsymbol{a} \in \mathbb{C}^n$, we have $\boldsymbol{a}^H \boldsymbol{C} \boldsymbol{a} \geq 0$.*

Property 1.5 (Bilinearity of the covariance) *Let $X_1, \dots, X_m$, $Y_1, \dots, Y_n$ be random variables, and $v_1, \dots, v_m$, $w_1, \dots, w_n$ be constants. Hence:*

$$\operatorname{cov}\left(\sum_{i=1}^{m} v_i^* X_i, \sum_{j=1}^{n} w_j^* Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} v_i^* w_j \operatorname{cov}(X_i, Y_j) \tag{1.21}$$

Let $\boldsymbol{V}$ and $\boldsymbol{W}$ be the vectors of components v_i and w_j respectively, and $A = \boldsymbol{V}^H \boldsymbol{X}$ and $B = \boldsymbol{W}^H \boldsymbol{Y}$. By definition, $\operatorname{cov}(A, B) = \{(A - \mathbb{E}\{A\})(B - \mathbb{E}\{B\})^*\}$. Replacing A and B by their respective expressions and using $\mathbb{E}\{A\} = \boldsymbol{V}^H \mathbb{E}\{\boldsymbol{X}\}$ and $\mathbb{E}\{B\} = \boldsymbol{W}^H \mathbb{E}\{\boldsymbol{Y}\}$, we obtain, successively:

$$\begin{aligned} \operatorname{cov}(A, B) &= \mathbb{E}\left\{\boldsymbol{V}^H (\boldsymbol{X} - \mathbb{E}\{\boldsymbol{X}\})(\boldsymbol{Y} - \mathbb{E}\{\boldsymbol{Y}\})^H \boldsymbol{W}\right\} \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n} v_i^* w_j \operatorname{cov}(X_i, Y_j) \end{aligned}$$

thus demonstrating expression (1.21). Using matrix notation, this is written:

$$\operatorname{cov}\left(\boldsymbol{V}^H \boldsymbol{X}, \boldsymbol{W}^H \boldsymbol{Y}\right) = \boldsymbol{V}^H \boldsymbol{C} \boldsymbol{W} \tag{1.22}$$

where $\boldsymbol{C}$ designates the covariance matrix of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

Property 1.6 (Linear transformation of a random vector) *Let* $\{X_1, \ldots, X_n\}$ *be* n *random variables with* $\mathbb{E}\{\boldsymbol{X}\}$ *as their mean vector and* $\boldsymbol{C}_X$ *as their covariance matrix, and let* $\{Y_1, \ldots, Y_q\}$ *be* q *random variables obtained by the linear transformation:*

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_q \end{bmatrix} = \boldsymbol{A} \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} + \boldsymbol{b}$$

where $\boldsymbol{A}$ *is a matrix and* $\boldsymbol{b}$ *is a non-random vector with the adequate sizes. We then have:*

$$\begin{aligned} \mathbb{E}\{\boldsymbol{Y}\} &= \boldsymbol{A}\mathbb{E}\{\boldsymbol{X}\} + \boldsymbol{b} \\ \boldsymbol{C}_Y &= \boldsymbol{A}\boldsymbol{C}_X\boldsymbol{A}^H \end{aligned}$$

Definition 1.13 (White sequence) *Let* $\{X_1, \ldots, X_n\}$ *be a set of* n *random variables. They are said to form a white sequence if* $\operatorname{var}(X_i) = \sigma^2$ *and if* $\operatorname{cov}(X_i, X_j) = 0$ *for* $i \neq j$*. Hence their covariance matrix can be expressed:*

$$\boldsymbol{C} = \sigma^2 \boldsymbol{I}_n$$

where $\boldsymbol{I}_n$ *is the* $n \times n$ *identity matrix.*

Property 1.7 (Independence $\Rightarrow$ non-correlation) *The random variables* $\{X_1, \ldots, X_n\}$ *are independent, then uncorrelated, and hence their covariance matrix is diagonal. Usually the converse statement is false.*

1.2 Conditional expectation

Definition 1.14 (Conditional expectation) *We consider a random variable* X *and a random vector* $\boldsymbol{Y}$ *taking values respectively in* $\mathcal{X} \subset \mathbb{R}$ *and* $\mathcal{Y} \subset \mathbb{R}^q$ *with joint probability density* $p_{X\boldsymbol{Y}}(x, \boldsymbol{y})$*. The conditional expectation of* X *given* $\boldsymbol{Y}$*, is a (measurable) real valued function* $g(\boldsymbol{Y})$ *such that for any other real valued function* $h(\boldsymbol{Y})$ *we have:*

$$\mathbb{E}\left\{|X - g(\boldsymbol{Y})|^2\right\} \leq \mathbb{E}\left\{|X - h(\boldsymbol{Y})|^2\right\} \tag{1.23}$$

$g(\boldsymbol{Y})$ is commonly denoted by $\mathbb{E}\{X|\boldsymbol{Y}\}$*.*

Property 1.8 (Conditional probability distribution) *We consider a random variable* X *and a random vector* $\boldsymbol{Y}$ *taking values respectively in* $\mathcal{X} \subset \mathbb{R}$ *and* $\mathcal{Y} \subset \mathbb{R}^q$ *with joint probability density* $p_{X\boldsymbol{Y}}(x, \boldsymbol{y})$*. Then* $\mathbb{E}\{X|\boldsymbol{Y}\} = g(\boldsymbol{Y})$ *where:*

$$g(\boldsymbol{y}) = \int_{\mathcal{X}} x\, p_{X|\boldsymbol{Y}}(x, \boldsymbol{y}) dx$$

with

$$p_{X|\boldsymbol{Y}}(x,\boldsymbol{y}) = \frac{p_{X\boldsymbol{Y}}(x,\boldsymbol{y})}{p_{\boldsymbol{Y}}(\boldsymbol{y})} \quad \textit{and} \quad p_{\boldsymbol{Y}}(\boldsymbol{y}) = \int_{\mathcal{X}} p_{X\boldsymbol{Y}}(x,\boldsymbol{y})dx \tag{1.24}$$

$p_{X|\boldsymbol{Y}}(x,\boldsymbol{y})$ is known as the conditional probability distribution of X given $\boldsymbol{Y}$.

Property 1.9 *The conditional expectation verifies the following properties:*

1. *linearity:* $\mathbb{E}\{a_1X_1 + a_2X_2|\mathbf{Y}\} = a_1\mathbb{E}\{X_1|\mathbf{Y}\} + a_2\mathbb{E}\{X_2|\mathbf{Y}\}$;
2. *orthogonality:* $\mathbb{E}\{(X - \mathbb{E}\{X|\mathbf{Y}\})h(\mathbf{Y})\} = 0$ *for any function* $h : \mathcal{Y} \mapsto \mathbb{R}$;
3. $\mathbb{E}\{h(\mathbf{Y})f(X)|\mathbf{Y}\} = h(\mathbf{Y})\mathbb{E}\{f(X)|\mathbf{Y}\}$, *for all functions* $f : \mathcal{X} \mapsto \mathbb{R}$ *and* $h : \mathcal{Y} \mapsto \mathbb{R}$;
4. $\mathbb{E}\{\mathbb{E}\{f(X,\mathbf{Y})|\mathbf{Y}\}\} = \mathbb{E}\{f(X,\mathbf{Y})\}$ *for any function* $f : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$; *specifically*

$$\mathbb{E}\{\mathbb{E}\{X|\boldsymbol{Y}\}\} = \mathbb{E}\{X\}$$

5. *refinement by conditioning: it can be shown (see page 13) that*

$$\mathrm{cov}\left(\mathbb{E}\{X|\boldsymbol{Y}\}\right) \leq \mathrm{cov}\left(X\right) \tag{1.25}$$

   The variance is therefore reduced by conditioning;

6. *if X and $\boldsymbol{Y}$ are independent, then* $\mathbb{E}\{f(X)|\boldsymbol{Y}\} = \mathbb{E}\{f(X)\}$. *Specifically,* $\mathbb{E}\{X|\boldsymbol{Y}\} = \mathbb{E}\{X\}$. *The reverse is not true;*
7. $\mathbb{E}\{X|\boldsymbol{Y}\} = X$, *if and only if X is a function of $\boldsymbol{Y}$.*

1.3 Projection theorem

Definition 1.15 (Scalar product) *Let $\mathcal{H}$ be a vector space constructed over $\mathbb{C}$. The scalar product is an application*

$$X, Y \in \mathcal{H} \times \mathcal{H} \mapsto (X,Y) \in \mathbb{C}$$

which verifies the following properties:

- $(X,Y) = (Y,X)^*$;
- $(\alpha X + \beta Y, Z) = \alpha(X,Z) + \beta(Y,Z)$;
- $(X,X) \geq 0$. *The equality occurs if, and only if,* $X = 0$.

A vector space constructed over $\mathbb{C}$ has a Hilbert space structure if it possesses a scalar product and if it is complete[3]. The norm of X is defined by $\|X\| = \sqrt{(X,X)}$ and the distance between two elements by $d(X_1,X_2) = \|X_1 - X_2\|$. Two elements X_1 and X_2 are said to be orthogonal, noted $X_1 \perp X_2$, if and only if $(X_1,X_2) = 0$. The demonstration of the following properties is trivial:

- Schwarz inequality:

$$|(X_1,X_2)| \leq \|X_1\| \, \|X_2\| \tag{1.26}$$

  the equality occurs if and only if λ exists such that $X_1 = \lambda X_2$;

- triangular inequality:

$$| \, \|X_1\| - \|X_2\| \, | \leq \|X_1 - X_2\| \leq \|X_1\| + \|X_2\| \tag{1.27}$$

- parallelogram identity:

$$\|X_1 + X_2\|^2 + \|X_1 - X_2\|^2 = 2\|X_1\|^2 + 2\|X_2\|^2 \tag{1.28}$$

In a Hilbert space, the projection theorem enables us to associate any given element from the space with its best quadratic approximation contained in a closed vector sub-space:

Theorem 1.2 (Projection theorem) *Let $\mathcal{H}$ be a Hilbert space defined over $\mathbb{C}$ and $\mathcal{C}$ a closed vector sub-space of $\mathcal{H}$. Each vector of $\mathcal{H}$ may then be associated with a unique element X_0 of $\mathcal{C}$ such that $\forall Y \in \mathcal{C}$ we have $d(X,X_0) \leq d(X,Y)$. Vector X_0 verifies, for any $Y \in \mathcal{C}$, the relationship $(X - X_0) \perp Y$.*

The relationship $(X - X_0) \perp Y$ constitutes the *orthogonality principle.*

A geometric representation of the orthogonality principle is shown in Figure 1.1. The element of $\mathcal{C}$ closest in distance to X is given by the *orthogonal projection* of X onto $\mathcal{C}$. In practice, this is the relationship which allows us to find the solution X_0.

This result is used alongside the expression of the norm of $X - X_0$, which is written:

$$\begin{aligned} \|X - X_0\|^2 &= (X, X - X_0) - (X_0, X - X_0) \\ &= \|X\|^2 - (X, X_0) \end{aligned} \tag{1.29}$$

The term $(X_0, X - X_0)$ is null due to the orthogonality principle.

[3]A definition of the term "complete" in this context may be found in mathematical textbooks. In the context of our presentation, this property plays a concealed role, for example in the existence of the orthogonal projection in theorem 1.2.

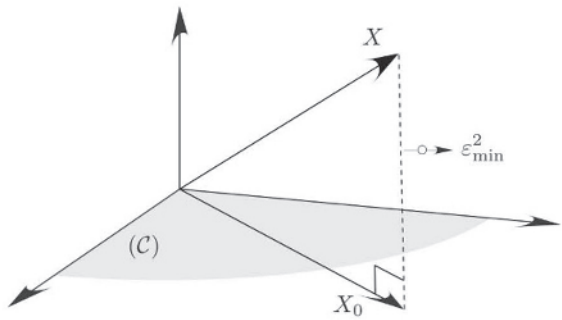

Figure 1.1 – *Orthogonality principle: the point X_0 which is the closest to X in $\mathcal{C}$ is such that $X - X_0$ is orthogonal to $\mathcal{C}$*

In what follows, the vector X_0 will be noted $(X|\mathcal{C})$, or $(X|Y_{1:n})$ when the sub-space onto which projection occurs is spanned by the linear combinations of vectors $Y_1, \ldots, Y_n$.

The simplest application of theorem 1.2 provides that for any $X \in \mathcal{C}$ and any $\varepsilon \in \mathcal{C}$:

$$(X|\varepsilon) = \frac{(X,\varepsilon)}{(\varepsilon,\varepsilon)}\varepsilon \tag{1.30}$$

The projection theorem leads us to define an application associating element X with element $(X|\mathcal{C})$. This application is known as the *orthogonal projection* of X onto $\mathcal{C}$. The orthogonal projection verifies the following properties:

1. linearity: $(\lambda X_1 + \mu X_2|\mathcal{C}) = \lambda(X_1|\mathcal{C}) + \mu(X_2|\mathcal{C})$;
2. contraction: $\|(X|\mathcal{C})\| \leq \|X\|$;
3. if $\mathcal{C}' \subset \mathcal{C}$, then $((X|\mathcal{C})|\mathcal{C}') = (X|\mathcal{C}')$;
4. if $\mathcal{C}_1 \perp \mathcal{C}_2$, then $(X|\mathcal{C}_1 \oplus \mathcal{C}_2) = (X|\mathcal{C}_1) + (X|\mathcal{C}_2)$.

The following result is fundamental:

$$(X|Y_{1:n+1}) = (X|Y_{1:n}) + (X|\varepsilon) = (X|Y_{1:n}) + \frac{(X,\varepsilon)}{(\varepsilon,\varepsilon)}\varepsilon \tag{1.31}$$

where $\varepsilon = Y_{n+1} - (Y_{n+1}|Y_{1:n})$. Because the sub-space spanned by $Y_{1:n+1}$ coincides with the sub-space spanned by $(Y_{1:,n}, \varepsilon)$ and because ε is orthogonal to the sub-space generated by $(Y_{1:n})$, then property (4) applies. To complete the proof we use (1.30).

Formula (1.31) is the basic formula used in the determination of many recursive algorithms, such as Kalman filter or Levinson recursion.

Theorem 1.3 (Square-integrable r.v.) *Let $\mathcal{L}_P^2$ be the vector space of square-integrable random variables, defined over the probability space $(\Omega, \mathcal{A}, P)$. Using the scalar product $(X, Y) = \mathbb{E}\{XY^*\}$, $\mathcal{L}_P^2$ has a Hilbert space structure.*

Conditional expectation

The conditional expectation $\mathbb{E}\{X|Y\}$ may be seen as the orthogonal projection of X onto sub-space $\mathcal{C}$ of all measurable functions of Y. Similarly, $\mathbb{E}\{X\}$ may be seen as the orthogonal projection of X onto the sub-space $\mathcal{D}$ of the constant random variables. These vectors are shown in Figure 1.2. Because $\mathcal{D} \subset \mathcal{C}$, using Pythagoras's theorem, we deduce that:

$$\mathrm{var}(X) = \|X - \mathbb{E}\{X\}\|^2 = \|X - \mathbb{E}\{X|Y\}\|^2 + \underbrace{\|\mathbb{E}\{X|Y\} - \mathbb{E}\{X\}\|^2}_{=\mathrm{var}(\mathbb{E}\{X|Y\})}$$

demonstrating $\mathrm{var}(\mathbb{E}\{X|Y\}) \leq \mathrm{var}(X)$. This can be extended to random vectors, giving the inequality (1.25) i.e. $\mathrm{cov}(\mathbb{E}\{X|\boldsymbol{Y}\}) \leq \mathrm{cov}(X)$.

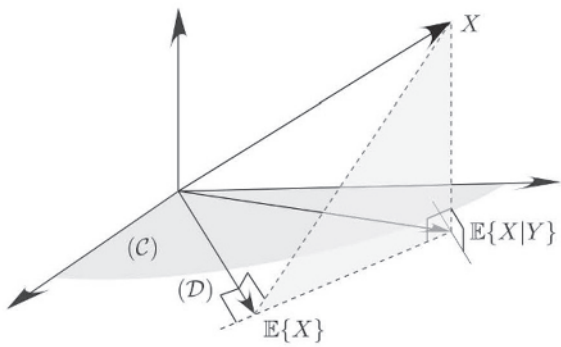

Figure 1.2 – *The conditional expectation $\mathbb{E}\{X|Y\}$ is the orthogonal projection of X onto the set $\mathcal{C}$ of measurable functions of Y. The expectation $\mathbb{E}\{X\}$ is the orthogonal projection of X onto the set $\mathcal{D}$ of constant functions. Clearly, $\mathcal{D} \subset \mathcal{C}$*

1.4 Gaussianity

Real Gaussian random variable

Definition 1.16 *A random variable X is said to be Gaussian, or normal, if all its values belong to $\mathbb{R}$ and if its characteristic function (see definition (1.9)) has the expression:*

$$\phi_X(u) = \exp\left(jmu - \frac{1}{2}\sigma^2 u^2\right) \tag{1.32}$$

where m is a real parameter and σ is a positive parameter. We check that its mean is equal to m and its variance to σ^2.

If $\sigma \neq 0$, it can be shown that the probability distribution has a probability density function with the expression:

$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) \tag{1.33}$$

Complex Gaussian random variable

In some problems, and particularly in the field of communications, the complex notation $X = U + jV$ is used, where U and V refer to two *real*, *Gaussian*, *centered*, *independent* random variables with the same variance $\sigma^2/2$. Because of independence (definition (1.7)), the joint probability distribution of the pair (U, V) has the following probability density:

$$\begin{aligned} p_{UV}(u,v) &= p_U(u)p_V(v) = \frac{1}{\sigma\sqrt{\pi}} \exp\left(-\frac{u^2}{\sigma^2}\right) \times \frac{1}{\sigma\sqrt{\pi}} \exp\left(-\frac{v^2}{\sigma^2}\right) \\ &= \frac{1}{\pi\sigma^2} \exp\left(-\frac{u^2+v^2}{\sigma^2}\right) \end{aligned}$$

If we notice that $|x|^2 = u^2 + v^2$, and if we introduce the notation $p_X(x) = p_{UV}(u, v)$, we can also write:

$$p_X(x) = \frac{1}{\pi\sigma^2} \exp\left(-\frac{|x|^2}{\sigma^2}\right) \tag{1.34}$$

Expression (1.34) is called the probability density of a *complex Gaussian* random variable. The word *circular* is sometimes added as a reminder that the isodensity contours are the circles $u^2 + v^2 =$ constant.

Note that:

$$\begin{aligned} \mathbb{E}\left\{X^2\right\} &= \mathbb{E}\left\{(U+jV)(U+jV)\right\} = 0 \\ &\text{and} \\ \mathbb{E}\left\{|X|^2\right\} &= \mathbb{E}\left\{XX^*\right\} = \mathbb{E}\left\{(U+jV)(U-jV)\right\} \\ &= \mathbb{E}\left\{U^2\right\} + \mathbb{E}\left\{V^2\right\} = \sigma^2 \end{aligned}$$

Gaussian random vectors

Definition 1.17 (Gaussian vector) *$X_1, \ldots, X_n$ are said to be n jointly Gaussian variables, or that the length n vector $\begin{bmatrix} X_1 & \ldots & X_n \end{bmatrix}^T$ is Gaussian, if any linear combination of its components, that is to say $Y = \boldsymbol{a}^H \boldsymbol{X}$ for any $\boldsymbol{a} = \begin{bmatrix} a_1 & \ldots & a_n \end{bmatrix}^T \in \mathbb{C}^n$, is a Gaussian random variable.*

This definition is applicable for vectors with real or complex components.

Theorem 1.4 (Distribution of a real Gaussian vector) *It can be shown that the probability distribution of a n Gaussian vector, with a length n mean vector* $\boldsymbol{m}$ *and an* $(n \times n)$ *covariance matrix* $\boldsymbol{C}$*, has the characteristic function:*

$$\phi_{\boldsymbol{X}}(u_1, \ldots, u_n) = \exp\left(j\boldsymbol{m}^T\boldsymbol{u} - \frac{1}{2}\boldsymbol{u}^T\boldsymbol{C}\boldsymbol{u}\right) \tag{1.35}$$

where $\boldsymbol{u} = (u_1, \ldots, u_n)^T \in \mathbb{R}^n$*. Let* $\boldsymbol{x} = (x_1, \ldots, x_n)^T$*. If* $\det\{\boldsymbol{C}\} \neq 0$*, the probability distribution's density has the expression:*

$$p_{\boldsymbol{X}}(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}\sqrt{\det\{\boldsymbol{C}\}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{m})^T\boldsymbol{C}^{-1}(\boldsymbol{x} - \boldsymbol{m})\right) \tag{1.36}$$

Theorem 1.5 (Distribution of a complex Gaussian vector) *We consider a length n complex Gaussian vector, with a length n mean vector* $\boldsymbol{m}$ *and an* $(n \times n)$ *covariance matrix* $\boldsymbol{C}$*. If* $\det\{\boldsymbol{C}\} \neq 0$*, the probability distribution's density has the expression:*

$$p_{\boldsymbol{X}}(x_1, \ldots, x_n) = \frac{1}{\pi^n \det\{\boldsymbol{C}\}} \exp\left(-(\boldsymbol{x} - \boldsymbol{m})^H\boldsymbol{C}^{-1}(\boldsymbol{x} - \boldsymbol{m})\right) \tag{1.37}$$

We have

$$\mathbb{E}\left\{(\boldsymbol{x} - \boldsymbol{m})(\boldsymbol{x} - \boldsymbol{m})^H\right\} = \boldsymbol{C} \tag{1.38}$$

$$\mathbb{E}\left\{(\boldsymbol{x} - \boldsymbol{m})(\boldsymbol{x} - \boldsymbol{m})^T\right\} = \boldsymbol{0}_n \tag{1.39}$$

where $\boldsymbol{0}_n$ *is the* $(n \times n)$ *null-matrix.*

Below, the real and complex Gaussian distributions will be noted $\mathcal{N}(\boldsymbol{m}, \boldsymbol{C})$ and $\mathcal{N}_c(\boldsymbol{m}, \boldsymbol{C})$ respectively.

Theorem 1.6 (Gaussian case: non-correlation $\Rightarrow$ independence) *If n jointly Gaussian variables are uncorrelated,* $\boldsymbol{C}$ *is diagonal; then they are independent.*

Theorem 1.7 (Linear transformation of a Gaussian vector) *Let* $[X_1 \ldots X_n]^T$ *be a Gaussian vector with a mean vector* $\boldsymbol{m}_X$ *and a covariance matrix* $\boldsymbol{C}_X$*. The random vector* $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}$*, where* $\boldsymbol{A}$ *and* $\boldsymbol{b}$ *are a matrix and a vector respectively, with the ad hoc length, is Gaussian and we have:*

$$\boldsymbol{m}_Y = \boldsymbol{A}\boldsymbol{m}_X + \boldsymbol{b} \quad \text{and} \quad \boldsymbol{C}_Y = \boldsymbol{A}\boldsymbol{C}_X\boldsymbol{A}^H \tag{1.40}$$

In other words, the Gaussian nature of a vector is untouched by linear transformations.

Equations (1.40) are a direct consequence of definition (1.17) and of property (1.6).

More specifically, if $\boldsymbol{X}$ is a random Gaussian vector $\mathcal{N}(\boldsymbol{m}, \boldsymbol{C})$, then the random variable $\boldsymbol{Z} = \boldsymbol{C}^{-1/2}(\boldsymbol{X} - \boldsymbol{M})$ follows a Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Another way of expressing this is to say that if $\boldsymbol{Z}$ has the distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ then $\boldsymbol{X} = \boldsymbol{M} + \boldsymbol{C}^{1/2}\boldsymbol{Z}$ has the distribution $\mathcal{N}(\boldsymbol{m}, \boldsymbol{C})$.

Note that, if $\boldsymbol{C}$ denotes a positive matrix, a square root of $\boldsymbol{C}$ is a matrix $\boldsymbol{M}$ which verifies:

$$\boldsymbol{C} = \boldsymbol{M}\boldsymbol{M}^H \tag{1.41}$$

Hence, if $\boldsymbol{M}$ is a square root of $\boldsymbol{C}$, then for any unitary matrix $\boldsymbol{U}$, i.e. such that $\boldsymbol{U}\boldsymbol{U}^H = \boldsymbol{I}$, matrix $\boldsymbol{M}\boldsymbol{U}$ is also a square root of $\boldsymbol{C}$. Matrix $\boldsymbol{M}$ is therefore defined to within a unitary matrix. One of the square roots is positive, and is obtained in MATLAB® using the function `sqrtm`.

The Gaussian distribution is defined using the first and second order moments, i.e. the mean and the covariance. Consequently, all moments of an order greater than 2 are expressed as a function of the first two values. The following theorem covers the specific case of a moment of order 4.

Theorem 1.8 (Moment of order 4) *Let X_1, X_2, X_3 and X_4 be four real or complex* centered *Gaussian random variables. Hence,*

$$\begin{aligned}\mathbb{E}\left\{X_1^{\beta_1}X_2^{\beta_2}X_3^{\beta_3}X_4^{\beta_4}\right\} &= \mathbb{E}\left\{X_1^{\beta_1}X_2^{\beta_2}\right\}\mathbb{E}\left\{X_3^{\beta_3}X_4^{\beta_4}\right\} \\ &+\mathbb{E}\left\{X_1^{\beta_1}X_3^{\beta_3}\right\}\mathbb{E}\left\{X_2^{\beta_2}X_4^{\beta_4}\right\} + \mathbb{E}\left\{X_1^{\beta_1}X_4^{\beta_4}\right\}\mathbb{E}\left\{X_2^{\beta_2}X_3^{\beta_3}\right\}\end{aligned} \tag{1.42}$$

where β_i is either "star" (conjugate variable) or "non-star" (non-conjugate variable). Hence:

$$\begin{aligned}&\operatorname{cov}\left(X_1^{\beta_1}X_2^{\beta_2}, X_3^{\beta_3}X_4^{\beta_4}\right) \\ &\quad= \mathbb{E}\left\{X_1^{\beta_1}X_2^{\beta_2}X_3^{\bar{\beta}_3}X_4^{\bar{\beta}_4}\right\} - \mathbb{E}\left\{X_1^{\beta_1}X_2^{\beta_2}\right\}\mathbb{E}\left\{X_3^{\bar{\beta}_3}X_4^{\bar{\beta}_4}\right\} \\ &\quad= \mathbb{E}\left\{X_1^{\beta_1}X_3^{\bar{\beta}_3}\right\}\mathbb{E}\left\{X_2^{\beta_2}X_4^{\bar{\beta}_4}\right\} + \mathbb{E}\left\{X_1^{\beta_1}X_4^{\bar{\beta}_4}\right\}\mathbb{E}\left\{X_2^{\beta_2}X_3^{\bar{\beta}_3}\right\}\end{aligned} \tag{1.43}$$

where $\bar{\beta}_j$ is "star" if β_j is "non-star" and conversely.

Note that, based on (1.39), for *complex* Gaussian variables, the terms $\mathbb{E}\{X_i X_j\} = 0$.

Gaussian conditional distribution

Consider two jointly Gaussian variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, taking their values from $\mathcal{X} \subset \mathbb{C}^p$ and $\mathcal{Y} \subset \mathbb{C}^q$ respectively. The respective means are noted $\mu_{\boldsymbol{X}}$ and $\mu_{\boldsymbol{Y}}$, and

$$\boldsymbol{C} = \begin{bmatrix} \text{cov}\,(\boldsymbol{X},\boldsymbol{X}) & \text{cov}\,(\boldsymbol{X},\boldsymbol{Y}) \\ \text{cov}\,(\boldsymbol{Y},\boldsymbol{X}) & \text{cov}\,(\boldsymbol{Y},\boldsymbol{Y}) \end{bmatrix} \tag{1.44}$$

is the joint covariance matrix. This produces the following results:

Property 1.10 *The conditional expectation of $\boldsymbol{X}$ given $\boldsymbol{Y}$ coincides with the orthogonal projection of $\boldsymbol{X}$ onto the affine sub-space spanned by $\mathbf{1}$ and $\boldsymbol{Y}$, written $\boldsymbol{B} + \boldsymbol{A}\boldsymbol{Y}$. Hence:*

- *the conditional expectation is expressed as:*

$$\mathbb{E}\{\boldsymbol{X}|\boldsymbol{Y}\} = \mu_{\boldsymbol{X}} + \text{cov}\,(\boldsymbol{X},\boldsymbol{Y})\,[\text{cov}\,(\boldsymbol{Y},\boldsymbol{Y})]^{-1}\,(\boldsymbol{Y} - \mu_{\boldsymbol{Y}}) \tag{1.45}$$

- *the conditional covariance is expressed as:*

$$\text{cov}\,(\boldsymbol{X}|\boldsymbol{Y}) = \text{cov}\,(\boldsymbol{X},\boldsymbol{X}) - \text{cov}\,(\boldsymbol{X},\boldsymbol{Y})\,[\text{cov}\,(\boldsymbol{Y},\boldsymbol{Y})]^{-1}\,\text{cov}\,(\boldsymbol{Y},\boldsymbol{X}) \tag{1.46}$$

- *the conditional distribution of $p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})$ is Gaussian. The mean is expressed as* (1.45) *and the covariance is given by expression* (1.46).

Let $g(\boldsymbol{Y})$ be the second member of (1.45), and let us demonstrate that $g(\boldsymbol{Y})$ is the conditional expectation of $\boldsymbol{X}$ given $\boldsymbol{Y}$. A rapid calculation shows that $\mathbb{E}\left\{(\boldsymbol{X} - g(\boldsymbol{Y}))\boldsymbol{Y}^H\right\} = 0$. Consequently, the random vectors $\boldsymbol{Z} = (\boldsymbol{X} - g(\boldsymbol{Y}))$ and $\boldsymbol{Y}$ are uncorrelated. As the vectors are jointly Gaussian, following property (1.9), they are independent and hence $\mathbb{E}\{\boldsymbol{Z}|\boldsymbol{Y}\} = \mathbb{E}\{\boldsymbol{Z}\}$. Using the second member of (1.45), we obtain $\mathbb{E}\{\boldsymbol{Z}\} = 0$. On the other hand:

$$\mathbb{E}\{\boldsymbol{X} - g(\boldsymbol{Y})|\boldsymbol{Y}\} \;=\; \mathbb{E}\{\boldsymbol{X}|\boldsymbol{Y}\} - g(\boldsymbol{Y})$$

It follows that $\mathbb{E}\{\boldsymbol{X}|\boldsymbol{Y}\} = g(\boldsymbol{Y})$. To demonstrate expression (1.46), let us denote $\boldsymbol{X}^c = \boldsymbol{X} - \mu_{\boldsymbol{X}}$ and $\boldsymbol{X}^c_{\boldsymbol{Y}} = \mathbb{E}\{\boldsymbol{X}|\boldsymbol{Y}\} - \mu_{\boldsymbol{X}}$. Hence, successively:

$$\begin{aligned} \mathbb{E}\left\{(\boldsymbol{X}^c - \boldsymbol{X}^c_{\boldsymbol{Y}})((\boldsymbol{X}^c - \boldsymbol{X}^c_{\boldsymbol{Y}})^H|\boldsymbol{Y}\right\} &= \mathbb{E}\left\{(\boldsymbol{X}^c - \boldsymbol{X}^c_{\boldsymbol{Y}})(\boldsymbol{X}^c - \boldsymbol{X}^c_{\boldsymbol{Y}})^H\right\} \\ &= \mathbb{E}\left\{(\boldsymbol{X}^c - \boldsymbol{X}^c_{\boldsymbol{Y}})\boldsymbol{X}^{c,H}\right\} \\ &= \text{cov}\,(\boldsymbol{X},\boldsymbol{X}) - \text{cov}\,(\boldsymbol{X},\boldsymbol{Y})\,[\text{cov}\,(\boldsymbol{Y}\boldsymbol{Y})]^{-1}\text{cov}\,(\boldsymbol{Y},\boldsymbol{X}) \end{aligned}$$

where, in the first equality, we use the fact that $(\boldsymbol{X}^c - \boldsymbol{X}^c_{\boldsymbol{Y}})$ is independent of $\boldsymbol{Y}$. In conclusion, the conditional distribution of $\boldsymbol{X}$, given $\boldsymbol{Y}$, is written:

$$\begin{aligned} p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}) \;=\; &\mathcal{N}\left(\mu_X + \text{cov}\,(\boldsymbol{X},\boldsymbol{Y})\,[\text{cov}\,(\boldsymbol{Y}\boldsymbol{Y})]^{-1}(\boldsymbol{Y} - \mu_{\boldsymbol{Y}}),\right. \\ &\left.\text{cov}\,(\boldsymbol{X},\boldsymbol{X}) - \text{cov}\,(\boldsymbol{X},\boldsymbol{Y})\,[\text{cov}\,(\boldsymbol{Y}\boldsymbol{Y})]^{-1}\text{cov}\,(\boldsymbol{Y},\boldsymbol{X})\right) \end{aligned} \tag{1.47}$$

Note that the distribution for random vector $\mathbb{E}\{\boldsymbol{X}|\boldsymbol{Y}\}$ should not be confused with the conditional distribution $p_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})$ of $\boldsymbol{X}$ given $\boldsymbol{Y}$. We shall restrict ourselves to the real, scalar case, taking μ_X and μ_Y as the respective means of X and Y, and

$$\boldsymbol{C} = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$$

with $-1 \leq \rho \leq 1$ as the covariance matrix. The conditional distribution of X given Y has a probability density $p_{X|Y}(x;y) = \mathcal{N}(\mu_X + \rho\sigma_X(y-\mu_Y)/\sigma_Y, \sigma_X^2(1-\rho^2))$. On the other hand the random variable distribution $\mathbb{E}\{X|Y\}$ has a probability density of $\mathcal{N}(\mu_X, \rho^2\sigma_X^2)$. Indeed based on equation (1.45), $\mathbb{E}\{\mathbb{E}\{X|Y\}\} = \mu_X$ and $\mathbb{E}\left\{(\mathbb{E}\{X|Y\} - \mu_X)^2\right\} = \rho^2\sigma_X^2\sigma_Y^2/\sigma_Y^2 = \rho^2\sigma_X^2$.

1.5 Random variable transformation

1.5.1 Change of variable formula

In many cases, it is necessary to determine the distribution of $Y = g(X)$ from the distribution of X. We shall consider this question in the context of continuous random vectors of dimension 2; the generalization to higher dimensions is straightforward.

Take two random variables X_1 and X_2 with a joint probability density $p_{X_1X_2}(x_1,x_2)$ and two measurable functions $g_1(x_1,x_2)$ and $g_2(x_1,x_2)$. We shall consider the two random variables:

$$\begin{cases} Y_1 &= g_1(X_1,X_2) \\ Y_2 &= g_2(X_1,X_2) \end{cases}$$

and assume that the transformation defined in this way is bijective. For any pair (y_1,y_2), there is a single solution (x_1,x_2). We may therefore write:

$$\begin{cases} X_1 &= h_1(Y_1,Y_2) \\ X_2 &= h_2(Y_1,Y_2) \end{cases}$$

In this case, the probability distribution of the random variables (Y_1,Y_2) has a density of:

$$p_{Y_1Y_2}(y_1,y_2) = p_{X_1X_2}\left(h_1(y_1,y_2), h_2(y_1,y_2)\right)\left|\det\left\{\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}}\right\}\right| \tag{1.48}$$

where $\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}}$ denotes the *Jacobian* of $\boldsymbol{h} : \boldsymbol{y} \to \boldsymbol{x}$ which is expressed as:

$$\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}} = \begin{bmatrix} \dfrac{\partial h_1(y_1,y_2)}{\partial y_1} & \dfrac{\partial h_2(y_1,y_2)}{\partial y_1} \\ \dfrac{\partial h_1(y_1,y_2)}{\partial y_2} & \dfrac{\partial h_2(y_1,y_2)}{\partial y_2} \end{bmatrix}$$

In cases where the transformation is not bijective, it is necessary to sum all of the solutions giving the pair (x_1, x_2) as a function of the pair (y_1, y_2).

Note that the Jacobian of a bijective function has one particularly useful property. Taking a bijective function $\boldsymbol{x} \in \mathbb{R}^d \leftrightarrow \boldsymbol{y} \in \mathbb{R}^d$, we have:

$$\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}} \times \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \boldsymbol{I}_d$$

This property allows us to calculate the Jacobian using the expression which is easiest to calculate, and, if necessary, to take the inverse.

Example 1.1 (Law of the sum of two random variables)
As an example, let us consider two random variables X_1 and X_2 with a joint probability density $p_{X_1X_2}(x_1, x_2)$. We wish to determine the joint distribution of the pair (Y_1, Y_2), defined by the following transformation:

$$\begin{cases} Y_1 = X_1 \\ Y_2 = X_1 + X_2 \end{cases} \Leftrightarrow \begin{cases} X_1 = Y_1 \\ X_2 = Y_2 - Y_1 \end{cases}$$

where the determinant of the Jacobian has a value of 1. Applying (1.48), we obtain the following probability density for the pair (Y_1, Y_2):

$$p_{Y_1Y_2}(y_1, y_2) = p_{X_1X_2}(y_1, y_2 - y_1)$$

From this, the probability density of $Y_2 = X_1 + X_2$ may be deduced by identifying the marginal distribution of Y_2. We obtain:

$$p_{Y_2}(y_2) = \int_{\mathbb{R}} p_{X_1X_2}(y_1, y_2 - y_1) dy_1$$

In cases where X_1 and X_2 are independent:

$$p_{X_1X_2}(x_1, x_2) = p_{X_1}(x_1) p_{X_2}(x_2)$$

hence:

$$p_{Y_2}(y_2) = \int_{\mathbb{R}} p_{X_1}(y_1) p_{X_2}(y_2 - y_1) dy_1$$

which is the expression of the convolution product $(p_{X_1} \star p_{X_2})(y_2)$.

1.5.2 δ-method

In cases where only the first two moments are considered, under very general conditions, the δ-method allows us to obtain approximate formulas for the mean and the covariance of:

$$\boldsymbol{Y} = \boldsymbol{g}(\boldsymbol{X}) \tag{1.49}$$

for any function $\boldsymbol{g} : \mathbb{R}^m \mapsto \mathbb{R}^q$. Let $\mu_{\boldsymbol{X}} = \mathbb{E}\{\boldsymbol{X}\} \in \mathbb{R}^m$. Assuming that $\boldsymbol{g}$ is differentiable at point $\mu_{\boldsymbol{X}}$ and using the first order Taylor expansion of $\boldsymbol{g}$ in the neighborhood of $\mu_{\boldsymbol{X}}$, we write

$$\boldsymbol{Y} = \boldsymbol{g}(\boldsymbol{X}) \approx \boldsymbol{g}(\mu_{\boldsymbol{X}}) + \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}=\mu_{\boldsymbol{X}}} (\boldsymbol{X} - \mu_{\boldsymbol{X}}) \tag{1.50}$$

where

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}=\mu_{\boldsymbol{X}}} = \begin{bmatrix} \dfrac{\partial g_1}{\partial x_1}(\mu_{\boldsymbol{X}}) & \cdots & \dfrac{\partial g_1}{\partial x_m}(\mu_{\boldsymbol{X}}) \\ \vdots & & \vdots \\ \dfrac{\partial g_q}{\partial x_1}(\mu_{\boldsymbol{X}}) & \cdots & \dfrac{\partial g_q}{\partial x_m}(\mu_{\boldsymbol{X}}) \end{bmatrix}$$

is the $q \times m$ Jacobian matrix of $\boldsymbol{g}$ performed at point $\mu_{\boldsymbol{X}}$. For the sake of simplicity, this is noted $\boldsymbol{J}(\mu_{\boldsymbol{X}})$ below. Therefore, taking the expectation of (1.50), we get at first order

$$\mathbb{E}\{\boldsymbol{Y}\} \approx \boldsymbol{g}(\mu_{\boldsymbol{X}}) + \boldsymbol{J}(\mu_{\boldsymbol{X}}) \times \mathbb{E}\{\boldsymbol{X} - \mu_{\boldsymbol{X}}\} = \boldsymbol{g}(\mu_{\boldsymbol{X}}) + 0$$

then

$$\boldsymbol{Y} - \mathbb{E}\{\boldsymbol{Y}\} \approx \boldsymbol{J}(\mu_{\boldsymbol{X}})(\boldsymbol{X} - \mu_{\boldsymbol{X}})$$

Therefore, according to the definition (1.18) of $\mathrm{cov}\,(\boldsymbol{Y})$, we have

$$\mathrm{cov}\,(\boldsymbol{g}(\boldsymbol{X})) \approx \boldsymbol{J}(\mu_{\boldsymbol{X}})\,\mathrm{cov}\,(\boldsymbol{X})\;\boldsymbol{J}^H(\mu_{\boldsymbol{X}})$$

It is worth noting that $\mathrm{cov}\,(\boldsymbol{g}(\boldsymbol{X}))$ is a $q \times q$ matrix and $\mathrm{cov}\,(\boldsymbol{X})$ is a $m \times m$ matrix. In summary we have:

$$\begin{cases} \mathbb{E}\{\boldsymbol{g}(\boldsymbol{X})\} & \approx & \boldsymbol{g}(\mu_{\boldsymbol{X}}) \\ \mathrm{cov}\,(\boldsymbol{g}(\boldsymbol{X})) & \approx & \boldsymbol{J}(\mu_{\boldsymbol{X}})\,\mathrm{cov}\,(\boldsymbol{X})\;\boldsymbol{J}^H(\mu_{\boldsymbol{X}}) \end{cases} \tag{1.51}$$

The δ-method is commonly used when calculating the mean and the covariance of $\boldsymbol{g}(\boldsymbol{X})$ is either intractable or the probability distribution of $\boldsymbol{X}$ is not fully specified.

Exercise 1.1 (δ-method) (see page 235) Consider two random variables (X_1, X_2), Gaussian and independent, with means of μ_1 and μ_2 respectively, and with the same variance σ^2. Using the pair (X_1, X_2), we determine the pair (R, θ) by bijective transformation:

$$(X_1, X_2) = \boldsymbol{h}(R, \theta) : \begin{cases} X_1 &=& R\cos(\theta) \in \mathbb{R} \\ X_2 &=& R\sin(\theta) \in \mathbb{R} \end{cases} \Leftrightarrow$$

$$(R, \theta) = \boldsymbol{g}(X_1, X_2) : \begin{cases} R &=& |X_1 + jX_2| = \sqrt{X_1^2 + X_2^2} \in \mathbb{R}^+ \\ \theta &=& \arg(X_1 + jX_2) \in (0, 2\pi) \end{cases}$$

Use the δ-method to determine the covariance of the pair (R, θ). Use this result to deduce the variance of R. This may be compared with the theoretical value given by:

$$\operatorname{var}(R) = 2\sigma^2 + (\mu_1^2 + \mu_2^2) - \frac{\pi\sigma^2}{2} L_{1/2}^2 \left(\frac{-(\mu_1^2 + \mu_2^2)}{2\sigma^2} \right)$$

where $L_{1/2}(x) = {}_1F_1\left(-\frac{1}{2}; 1; x\right)$ is the hypergeometric function. We see that, when $(\mu_1^2+\mu_2^2)/\sigma^2$ tends toward infinity, $\operatorname{var}(R)$ tends toward σ^2. Additionally, when $\mu_1 = \mu_2 = 0$, we have $\operatorname{var}(R) = (4-\pi)\sigma^2/2 \approx 0.43\sigma^2$.

1.6 Fundamental statistical theorems

The following two theorems form the basis of statistical methods, and are essential to the validity of Monte-Carlo methods, which are presented in brief in Chapter 3. In very general conditions, these theorems imply that the empirical average of a series of r.v.s will converge toward the mean. The first theorem, often (erroneously) referred to as a *law*, sets out this convergence; the second theorem states that this convergence is "distributed in a Gaussian manner".

Theorem 1.9 (Law of large numbers) *Let X_n be a series of random vectors of dimension d, independent and identically distributed, with a mean vector $m = \mathbb{E}\{X_1\} \in \mathbb{R}^d$ and finite covariance. In this case,*

$$\frac{1}{N} \sum_{n=1}^{N} X_n \xrightarrow{a.s.}_{N \to +\infty} \mathbb{E}\{X_1\} = m$$

and convergence is almost sure.

One fundamental example is that of empirical frequency, which converges toward the probability. Let X_n be a series of N random variables with values in $a_1, a_2, \ldots, a_J$ and let f_j be the empirical frequency, defined as the relationship between the number of values equal to a_j and the total number N. In this case:

$$f_j = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}(X_n = a_j) \xrightarrow{a.s.} \mathbb{E}\{\mathbb{1}(X_1 = a_j)\} = \mathbb{P}\{X_1 = a_j\}$$

Theorem 1.10 (Central limit theorem) *Let X_n be a series of random vectors of dimension d, independent and identically distributed, of mean vector $m = \mathbb{E}\{X_1\}$ and covariance matrix $C = \operatorname{cov}(X_1, X_1)$. In this case:*

$$\sqrt{N} \left(\frac{1}{N} \sum_{n=1}^{N} X_n - m \right) \xrightarrow{d}_{N \to +\infty} \mathcal{N}(0, C)$$

with convergence in distribution.

Convergence in distribution is defined as follows:

Definition 1.18 (Convergence in distribution) *A set of r.v. U_N is said to converge in distribution toward an r.v. U if, for any bounded continuous function f, when N tends toward infinity, we have:*

$$\mathbb{E}\{f(U_N)\} \to_{N\to\infty} \mathbb{E}\{f(U)\} \tag{1.52}$$

Theorem 1.10 is the basis for calculations of confidence intervals (see definition 2.6), and is used as follows: we approximate the probability distribution of the random vector $\sqrt{N}\left(N^{-1}\sum_{n=1}^{N} X_n - m\right)$, for which the expression is often impossible to calculate, by the Gaussian distribution. For illustrative purposes, consider the case where $d = 1$, taking $\widehat{m}_N = N^{-1}\sum_{n=1}^{N} X_n$. Hence, for any $c > 0$:

$$\mathbb{P}\left\{\sqrt{N}\left(\widehat{m}_N - m\right) \in (-\varepsilon, +\varepsilon)\right\} \approx 2\int_0^{\varepsilon} \frac{1}{\sigma\sqrt{2\pi}} e^{-u^2/2\sigma^2} du$$

Letting $\epsilon = c\sigma$, we have:

$$\mathbb{P}\left\{\widehat{m}_N - \frac{c\sigma}{\sqrt{N}} < m \leq \widehat{m}_N + \frac{c\sigma}{\sqrt{N}}\right\} \approx 2\int_0^{c} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

Aim for a probability equal typically to 0.05, $c = 1.96$.

As expected, the smaller σ and/or the higher N, the narrower, i.e. "better", the interval will be.

Exercise 1.2 (Asymptotic confidence interval) (see page 236) Consider a sequence of N independent random Bernoulli variables X_k such that $\mathbb{P}\{X_k = 1\} = p$. To estimate the proportion p, we consider $\widehat{p} = \frac{1}{N}\sum_{k=1}^{N} X_k$.

1. Using the central limit theorem 1.10, determine the asymptotic distribution of $\widehat{p}$.

2. Use the previous result to deduce the approximate expression of the probability that p will lie within the interval between $\widehat{p} - \epsilon/\sqrt{N}$ and $\widehat{p} + \epsilon/\sqrt{N}$.

3. Use this result to deduce an interval which ensures that this probability will be higher than $100\,\alpha\%$, expressed as a function of N and α: typically, $\alpha = 0.95$.

4. Write a program which verifies this asymptotic behavior.

The following theorem, known as the *continuity theorem*, allows us to extend the central limit theorem to more complicated functions:

Theorem 1.11 (Continuity) *Let $\boldsymbol{U}_N$ be a series of random vectors of dimension d such that*

$$\sqrt{N}(\boldsymbol{U}_N - \boldsymbol{m}) \xrightarrow{d}_{N\to+\infty} \mathcal{N}(\boldsymbol{0}_d, \boldsymbol{C})$$

and let $\boldsymbol{g}$ be a function $\mathbb{R}^d \mapsto \mathbb{R}^q$ supposed to be twice continuously differentiable. Thus,

$$\sqrt{N}(\boldsymbol{g}(\boldsymbol{U}_N) - \boldsymbol{g}(\boldsymbol{m})) \xrightarrow{d}_{N\to+\infty} \mathcal{N}(\boldsymbol{0}_q, \boldsymbol{\Gamma})$$

where $\boldsymbol{\Gamma} = \partial\boldsymbol{g}(\boldsymbol{m})\, \boldsymbol{C}\, \partial^T \boldsymbol{g}(m)$ and where

$$\partial\boldsymbol{g} = \begin{bmatrix} \dfrac{\partial g_1(u_1,\ldots,u_d)}{\partial u_1} & \cdots & \dfrac{\partial g_1(u_1,\ldots,u_d)}{\partial u_d} \\ \vdots & & \vdots \\ \dfrac{\partial g_q(u_1,\ldots,u_d)}{\partial u_1} & \cdots & \dfrac{\partial g_q(u_1,\ldots,u_d)}{\partial u_d} \end{bmatrix}$$

is the Jacobian of $\boldsymbol{g}$ and $\partial\boldsymbol{g}(\boldsymbol{m})$ the Jacobian calculated at point $\boldsymbol{m}$.

Applying theorem (1.11), consider the function associating vector $\boldsymbol{U}_N$ with its ℓ-th component, which is written:

$$\boldsymbol{U}_N \mapsto U_{N,\ell} = \boldsymbol{E}_\ell^T \boldsymbol{U}_N$$

where $\boldsymbol{E}_\ell$ is the vector of dimension d of which all components are equal to 0, with the exception of the ℓ-th, equal to 1. Direct application of the theorem gives:

$$\sqrt{N}(U_{N,\ell} - m_\ell) \xrightarrow{d} \mathcal{N}(0, C_{\ell\ell})$$

where m_ℓ is the ℓ-th component of m and $C_{\ell\ell}$ the ℓ-th diagonal element of $\boldsymbol{C}$.

1.7 Other important probability distributions

This section presents a non-exhaustive list of certain other important probability distributions. Some of the associated functions, which are not available in the basic version of MATLAB®, are given in simplified form in the Appendix.

Uniform distribution over (a, b) : noted $\mathcal{U}(a, b)$ of density

$$p_X(x; a, b) = \frac{1}{b-a} \mathbb{1}(x \in (a, b)) \tag{1.53}$$

where $a < b$. The mean is equal to $(b+a)/2$ and the variance to $(b-a)^2/12$.

Exponential distribution : noted $E(\theta)$, of density

$$p_X(x;\theta) = \theta^{-1} e^{-x/\theta} \mathbb{1}(x \geq 0) \tag{1.54}$$

with $\theta > 0$. The mean is equal to θ and the variance to θ^2. We can easily demonstrate that $E(\theta) = \theta E(1)$.

Gamma distribution : noted $G(k, \theta)$, of density

$$p_X(x;(k,\theta)) = \frac{1}{\Gamma(k)\theta^k} e^{-x/\theta} x^{k-1} \mathbb{1}(x > 0) \tag{1.55}$$

where $\theta \in \mathbb{R}^+$ and $k \in \mathbb{R}^+$. The mean is equal to $k\theta$ and the variance to $k\theta^2$. Note that $E(\theta) = G(1, \theta)$.

χ^2 distribution with k d.o.f. : noted χ^2_k. The r.v. $Y = \sum_{i=1}^{k} X_i^2$ where X_i are k Gaussian, independent, centered r.v.s of variance 1 follows a χ^2 distribution with k degrees of freedom (d.o.f.). The mean is equal to k and the variance to $2k$.

Fisher distribution with (k_1, k_2) d.o.f. : noted $F(k_1, k_2)$. Let X and Y be two real, centered Gaussian vectors of respective dimensions k_1 and k_2, with respective covariance matrices I_{k_1} and I_{k_2}, and independent of each other, then the r.v.

$$F_{k_1,k_2} = \frac{k_1^{-1} X^T X}{k_2^{-1} Y^T Y} \tag{1.56}$$

follows a Fisher distribution with (k_1, k_2) d.o.f.

Student distribution with k d.o.f. : noted T_k. Let X be a real, centered Gaussian vector, with a covariance matrix I_k, and Y a real, centered Gaussian vector, of variance 1 and independent of X. The r.v.

$$T_k = \frac{Y}{\sqrt{k^{-1} \sum_{i=1}^{k} X_i^2}} \tag{1.57}$$

follows a Student distribution with k d.o.f.

We can show that if Z follows a Student distribution with k degrees of freedom, then Z^2 follows a Fisher distribution with $(1, k)$ degrees of freedom.