Volume 3 of the second edition of the fully revised and updated Digital Signal and Image Processing using MATLAB®, after the first two volumes on the fundamentals and applications of Image and Signal Processing, focuses on the stochastic case. It will be of particular benefit to readers who already possess a good knowledge of MATLAB® and a command of the fundamental elements of digital signal processing, who are familiar with the fundamentals of continuous-spectrum spectral analysis and who have a certain mathematical knowledge concerning Hilbert spaces.

This volume focuses on applications but also provides a good presentation of the principles. A number of elements closer in nature to statistics than to signal processing itself are widely discussed. This choice comes from a current tendency of signal processing to use techniques from this field.

More than 200 programs and functions are provided in the MATLAB® language, with useful comments and guidance, to enable numerical experiments to be carried out, thus allowing readers to develop a deeper understanding of both the theoretical and practical aspects of this subject.

Gérard Blanchet is Professor at Telecom ParisTech, France. In addition to his research, teaching and consulting activities, he is the author of several books on automatic control systems, digital signal processing and computer architecture. He also develops tools and methodologies to improve knowledge acquisition in various fields.

Maurice Charbit is Professor at Telecom ParisTech, France. He is a teacher in probability theory, signal processing, communication theory and statistics for data processing. With regard to research, his main areas of interest are: (i) the Bayesian approach for hidden Markov models, (ii) the 3D model-based approach for face tracking, and (iii) processing for multiple sensor arrays with applications to infrasonic systems.
Digital Signal and Image Processing using MATLAB®
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreword</td>
<td></td>
<td>ix</td>
</tr>
<tr>
<td>Notations and Abbreviations</td>
<td></td>
<td>xiii</td>
</tr>
<tr>
<td><strong>1 Mathematical Concepts</strong></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1.1</td>
<td>Basic concepts on probability</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>Conditional expectation</td>
<td>9</td>
</tr>
<tr>
<td>1.3</td>
<td>Projection theorem</td>
<td>10</td>
</tr>
<tr>
<td>1.4</td>
<td>Gaussianity</td>
<td>13</td>
</tr>
<tr>
<td>1.5</td>
<td>Random variable transformation</td>
<td>18</td>
</tr>
<tr>
<td>1.5.1</td>
<td>Change of variable formula</td>
<td>18</td>
</tr>
<tr>
<td>1.5.2</td>
<td>δ-method</td>
<td>19</td>
</tr>
<tr>
<td>1.6</td>
<td>Fundamental statistical theorems</td>
<td>21</td>
</tr>
<tr>
<td>1.7</td>
<td>Other important probability distributions</td>
<td>23</td>
</tr>
<tr>
<td><strong>2 Statistical Inferences</strong></td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>2.1</td>
<td>Statistical model</td>
<td>25</td>
</tr>
<tr>
<td>2.2</td>
<td>Hypothesis tests</td>
<td>27</td>
</tr>
<tr>
<td>2.2.1</td>
<td>Simple hypotheses</td>
<td>28</td>
</tr>
<tr>
<td>2.2.2</td>
<td>Generalized Likelihood Ratio Test (GLRT)</td>
<td>33</td>
</tr>
<tr>
<td>2.2.3</td>
<td>$\chi^2$ goodness of fit test</td>
<td>39</td>
</tr>
<tr>
<td>2.3</td>
<td>Statistical estimation</td>
<td>41</td>
</tr>
<tr>
<td>2.3.1</td>
<td>General principles</td>
<td>41</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Least squares method</td>
<td>43</td>
</tr>
<tr>
<td>2.3.3</td>
<td>Least squares method: linear model</td>
<td>44</td>
</tr>
<tr>
<td>2.3.4</td>
<td>Method of moments</td>
<td>56</td>
</tr>
<tr>
<td>2.3.5</td>
<td>Maximum likelihood</td>
<td>59</td>
</tr>
<tr>
<td>2.3.6</td>
<td>Estimating a distribution</td>
<td>75</td>
</tr>
<tr>
<td>2.3.7</td>
<td>Bootstrap and others</td>
<td>79</td>
</tr>
</tbody>
</table>
3 Monte-Carlo Simulation
  3.1 Fundamental theorems ........................................ 85
  3.2 Stating the problem ........................................... 86
  3.3 Generating random variables .................................. 88
    3.3.1 The cumulative function inversion method .......... 89
    3.3.2 The variable transformation method ................. 91
    3.3.3 Acceptance-rejection method ....................... 93
    3.3.4 Sequential methods .................................... 95
  3.4 Variance reduction ........................................... 99
    3.4.1 Importance sampling .................................... 99
    3.4.2 Stratification ........................................... 103
    3.4.3 Antithetic variates .................................... 106

4 Second Order Stationary Process .................................. 107
  4.1 Statistics for empirical correlation ......................... 107
  4.2 Linear prediction of WSS processes .......................... 111
    4.2.1 Yule-Walker equations ................................ 111
    4.2.2 Levinson-Durbin algorithm ............................ 115
    4.2.3 Reflection coefficients and lattice filters .......... 120
  4.3 Non-parametric spectral estimation of WSS processes ...... 124
    4.3.1 Correlogram ............................................ 125
    4.3.2 Periodogram ............................................ 125
    4.3.3 Smoothed periodograms ................................. 128
    4.3.4 Coherence .............................................. 135

5 Inferences on HMM .................................................. 139
  5.1 Hidden Markov Models (HMM) ................................ 139
  5.2 Inferences on HMM ............................................. 142
  5.3 Gaussian linear case: the Kalman filter .................... 143
  5.4 Discrete finite Markov case ................................ 152
    5.4.1 Forward-backward formulas ............................ 153
    5.4.2 Smoothing with one instant ........................... 155
    5.4.3 Smoothing with two instants ........................... 156
    5.4.4 HMM learning using the EM algorithm ............... 156
    5.4.5 The Viterbi algorithm ................................. 158

6 Selected Topics .................................................... 163
  6.1 High resolution methods ...................................... 163
    6.1.1 Estimating the fundamental of periodic signals: MUSIC 163
    6.1.2 Introduction to array processing: MUSIC, ESPRIT .... 177
  6.2 Digital Communications ....................................... 186
    6.2.1 Introduction ............................................. 186
    6.2.2 8-phase shift keying (PSK) ............................ 189
    6.2.3 PAM modulation .......................................... 191
6.2.4 Spectrum of a digital signal .................................. 193
6.2.5 The Nyquist criterion in digital communications .......... 198
6.2.6 The eye pattern ................................................. 204
6.2.7 PAM modulation on the Nyquist channel .................... 205
6.3 Linear equalization and the Viterbi algorithm ................. 211
  6.3.1 Linear equalization ......................................... 213
  6.3.2 The soft decoding Viterbi algorithm ......................... 215
6.4 Compression ........................................................ 220
  6.4.1 Scalar Quantization .......................................... 220
  6.4.2 Vector Quantization ......................................... 222

7 Hints and Solutions .................................................. 235
H1 Mathematical concepts ............................................. 235
H2 Statistical inferences .............................................. 237
H3 Monte-Carlo simulation ............................................ 269
H4 Second order stationary process .................................. 277
H5 Inferences on HMM .................................................. 283
H6 Selected Topics ...................................................... 300

8 Appendices ............................................................. 317
A1 Miscellaneous functions .......................................... 317
A2 Statistical functions .............................................. 318
  A.2.1 Notable functions ........................................... 318
  A.2.2 Beta distribution ............................................ 318
  A.2.3 Student’s distribution ....................................... 318
  A.2.4 Chi-squared distribution ................................... 323
  A.2.5 Fisher’s distribution ......................................... 326

Bibliography ............................................................. 329

Index ................................................................. 333
Foreword

This book is the third volume in a series on digital signal processing and associated techniques. Following on from "Fundamentals" and "Advances and Applications, The Deterministic Case", it addresses the stochastic case. We shall presume throughout that readers have a good working knowledge of MATLAB® and of basic elements of digital signal processing.

Whilst our main focus is on applications, we shall also give consideration to key principles. A certain number of the elements discussed belong more to the domain of statistics than to signal processing; this responds to current trends in signal processing, which make extensive use of this type of technique.

Over 60 solved exercises allow the reader to apply the concepts and results presented in the following chapters. There will also be examples to alleviate any demonstrations that would otherwise be quite dense. These can be found in more specialist books referenced in the bibliography. 92 programs and 49 functions will be used to support these examples and corrected exercises.

Mathematical Concepts

The first chapter begins with a brief review of probability theory, focusing on the notions of conditional probability, projection theorem and random variable transformation. A number of statistical elements will also be presented, including the law of large numbers (LLN), the limit-central theorem, or the delta-method.

Statistical Inferences

The second chapter is devoted to statistical inference. Statistical inference consists of deducing interesting characteristics from a series of observations with a certain degree of reliability. A variety of techniques may be used. In this chapter, we shall discuss three broad families of techniques: hypothesis testing, parameter estimation, and the determination of confidence intervals. Key notions include Cramer-Rao bound, likelihood ratio tests, maximum likelihood approach and least square approach for linear models.
**Monte-Carlo simulation**

Monte-Carlo methods involve a set of algorithms which aim to calculate values using a pseudo-random generator. The quantities to calculate are typically integrals, and in practice, often represent the mathematical expectation of a function of interest. In cases using large dimensions, these methods can significantly reduce the calculation time required by deterministic methods. Monte-Carlo methods involve drawing a series of samples, distributed following a target distribution. The main generation methods, including importance sampling, the acceptance-rejection method, the Gibbs sampler, etc., will be presented. Another objective is to minimize the mean square error between the calculated and true values, and variance reduction methods will be studied using simulations.

**Second order stationary process**

The fourth chapter covers second order random stationary processes in the broadest sense: Wide Sense Stationary (WSS). The chapter is split into three parts, beginning with empirical second order estimators, leading to the correlogram. Then follow general and detailed results on the linear prediction which is fundamental role in the WSS time series. The third part is devoted to the non-parametric spectral estimation approaches (smooth periodograms, average periodograms, etc.). A detailed discussion on the bias-variance compromise is given.

**Inferences on HMM**

States are directly visible in simple Markov models, and the modeling process depends exclusively on transition probabilities. In hidden-state Markov models (HMM), however, states can only be seen via observed signals which are statistically linked to these states. HMMs are particularly useful as control models, using latent variables of mixtures connected to each observation.

A wide variety of problems may be encountered in relation to inference, for example seeking the sequence most likely to have produced a given series of observations; determining the a posteriori distribution of hidden states; estimating the parameters of a model; etc. Key algorithms include the Baum-Welch algorithm and the Viterbi algorithm, to cite the two best-known examples. HMM have applications in a wide range of domains, such as speech recognition (analysis and synthesis), automatic translation, handwriting analysis, activity identification, DNA analysis, etc.

**Selected Topics**

The final chapter presents applications which use many of the principles and techniques described in the preceding chapters, without falling into any of the
categories defined in these chapters. The first section is devoted to high resolution techniques (MUSIC and ESPRIT algorithms), whilst the second covers classic communication problems (coding, modulation, eye diagrams, etc.). The third section presents the Viterbi algorithm, and the fourth is given over to scalar and vectorial quantification.

Annexes

A certain number of functions are given in simplified form in the appendix. This includes a version of the boxplot function, alongside functions associated with the most common distributions (Student, $\chi^2$ and Fischer).

Remarques

The notation used in this book is intended to conform to current usage; in cases where this is not the case, every care has been taken to remove any ambiguity as to the precise meaning of terms. On a number of occasions, we refer to nitnslseries instead of nitnsltime series or nitnslsequences to avoid confusion.
Notations and Abbreviations

\[ \emptyset \quad \text{empty set} \]
\[ \sum_{k,n} = \sum_k \sum_n \]
\[ \text{rect}_T(t) = \begin{cases} 1 & \text{when } |t| < T/2 \\ 0 & \text{otherwise} \end{cases} \]
\[ \text{sinc}(x) = \frac{\sin(\pi x)}{\pi x} \]
\[ 1(x \in A) = \begin{cases} 1 & \text{when } x \in A \\ 0 & \text{otherwise} \end{cases} \quad \text{(indicator function of } A) \]
\[ (a,b] = \{x : a < x \leq b\} \]
\[ \delta(t) = \{ \begin{array}{ll} \text{Dirac distribution when } t \in \mathbb{R} \\ \text{Kronecker symbol when } t \in \mathbb{Z} \end{array} \] \]
\[ \text{Re}(z) \quad \text{real part of } z \]
\[ \text{Im}(z) \quad \text{imaginary part of } z \]
\[ i \text{ or } j = \sqrt{-1} \]
\[ x(t) \Rightarrow X(f) \quad \text{Fourier transform} \]
\[ (x \ast y)(t) \quad \text{continuous time convolution} \]
\[ = \int_{\mathbb{R}} x(u)y(t-u)du \]
\[ (x \ast y)(t) \quad \text{discrete time convolution} \]
\[ = \sum_{u \in \mathbb{Z}} x(u)y(t-u) \]
\[ x \text{ or } x \quad \text{vector } x \]
\[ I_N \quad (N \times N)\text{-dimension identity matrix} \]
\[ A^* \quad \text{complex conjugate of } A \]
\[ A^T \quad \text{transpose of } A \]
\[ A^{\dagger} \quad \text{transpose-conjugate of } A \]
\[ A^{-1} \quad \text{inverse matrix of } A \]
\[ A^\# \quad \text{pseudo-inverse matrix of } A \]

- \( \mathbb{P} \): probability measure
- \( \mathbb{P}_\theta \): probability measure indexed by \( \theta \)
- \( \mathbb{E}\{X\} \): expectation of \( X \)
- \( \mathbb{E}_\theta\{X\} \): expectation of \( X \) under the distribution \( \mathbb{P}_\theta \)
- \( X_c = X - \mathbb{E}\{X\} \): zero-mean random variable
- \( \text{var}(X) = \mathbb{E}\{|X_c|^2\} \): variance of \( X \)
- \( \text{cov}(X, Y) = \mathbb{E}\{X_cY^*_c\} \): covariance of \( (X, Y) \)
- \( \text{cov}(X) = \text{cov}(X, X) = \text{var}(X) \): variance of \( X \)
- \( \mathbb{E}\{X|Y\} \): conditional expectation of \( X \) given \( Y \)

- \( a \xrightarrow{\mathcal{L}} b \) or \( a \xrightarrow{\mathcal{D}} b \): \( a \) converges in law to \( b \)
- \( a \xrightarrow{\mathbb{P}} b \): \( a \) converges in distribution to \( b \)
- \( a \xrightarrow{a.s.} b \): \( a \) converges almost surely to \( b \)

- ADC: Analog to Digital Converter
- ADPCM: Adaptive Differential PCM
- AMI: Alternate Mark Inversion
- AR: Autoregressive
- ARMA: AR and MA
- BER: Bit Error Rate
- bps: bits per second
- cdf: cumulative distribution function
- CF: Clipping Factor
- CZT: Causal z-Transform
- DAC: Digital to Analog Converter
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCT</td>
<td>Discrete Cosine Transform</td>
</tr>
<tr>
<td>d.o.f.</td>
<td>degree of freedom</td>
</tr>
<tr>
<td>DFT</td>
<td>Discrete Fourier Transform</td>
</tr>
<tr>
<td>DTFT</td>
<td>Discrete Time Fourier Transform</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation Maximization</td>
</tr>
<tr>
<td>ESPRIT</td>
<td>Estimation of Signal Parameter via Rotational Invariance Techniques</td>
</tr>
<tr>
<td>FIR</td>
<td>Finite Impulse Response</td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier Transform</td>
</tr>
<tr>
<td>FT</td>
<td>Continuous Time Fourier Transform</td>
</tr>
<tr>
<td>GLRT</td>
<td>Generalized Likelihood Ratio Test</td>
</tr>
<tr>
<td>GEM</td>
<td>Generalized Expectation Maximization</td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian Mixture Model</td>
</tr>
<tr>
<td>HDB</td>
<td>High Density Bipolar</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Model</td>
</tr>
<tr>
<td>IDFT</td>
<td>Inverse Discrete Fourier Transform</td>
</tr>
<tr>
<td>i.i.d./iid</td>
<td>independent and identically distributed</td>
</tr>
<tr>
<td>IIR</td>
<td>Infinite Impulse Response</td>
</tr>
<tr>
<td>ISI</td>
<td>InterSymbol Interference</td>
</tr>
<tr>
<td>KKT</td>
<td>Karush-Kuhn-Tucker</td>
</tr>
<tr>
<td>LDA</td>
<td>Linear Discriminant Analysis</td>
</tr>
<tr>
<td>LBG</td>
<td>Linde, Buzzo, Gray (algorithm)</td>
</tr>
<tr>
<td>LMS</td>
<td>Least Mean Squares</td>
</tr>
<tr>
<td>MA</td>
<td>Moving Average</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Square Error</td>
</tr>
<tr>
<td>MUSIC</td>
<td>MUltiple SIgнал Charaterization</td>
</tr>
<tr>
<td>PAM</td>
<td>Pulse Amplitude Modulation</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>PSK</td>
<td>Phase Shift Keying</td>
</tr>
<tr>
<td>QAM</td>
<td>Quadrature Amplitude Modulation</td>
</tr>
<tr>
<td>rls</td>
<td>recursive least squares</td>
</tr>
<tr>
<td>rms</td>
<td>root mean square</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operational Characteristic</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal to Noise Ratio</td>
</tr>
<tr>
<td>r.v./rv</td>
<td>random variable</td>
</tr>
<tr>
<td>STFT</td>
<td>Short Term Fourier Transform</td>
</tr>
<tr>
<td>TF</td>
<td>Transfer Function</td>
</tr>
<tr>
<td>WSS</td>
<td>Wide (Weak) Sense Stationary (Second Order) Process</td>
</tr>
</tbody>
</table>
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:

$$p_X(n) = \Pr(X = a_n) \quad (1.1)$$

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:

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

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 $F_X(+\infty) = 1$. 

Digital Signal and Image Processing using MATLAB®: Advances and Applications: The Stochastic Case,
© ISTE Ltd 2015. Published by ISTE Ltd and John Wiley & Sons, Inc.
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:

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

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

$\Pr(X = a_n, Y = b_k)$ 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:

$$p_X(n) = \Pr(X = a_n) = \sum_{k=0}^{+\infty} p_{XY}(n,k)$$

$$p_Y(k) = \Pr(Y = b_k) = \sum_{n=0}^{+\infty} p_{XY}(n,k)$$

$p_X(n)$ 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:

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

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:

\[
F_X(x) = \Pr(X \leq x) = \int_{-\infty}^{x} p_X(u)du
\]

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 \( \Delta \) of \( \mathbb{R}^2 \), the probability that the pair \( (X, Y) \) belongs to \( \Delta \) is given by:

\[
\Pr((X, Y) \in \Delta) = \int_{\Delta} P_{XY}(x, y)dx\,dy
\]

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

\[
\int_{\mathbb{R}^2} P_{XY}(x, y)dx\,dy = 1
\]

\( P_{XY}(x, y) \) 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:

\[
p_X(x) = \int_{-\infty}^{+\infty} P_{XY}(x, y)dy
\]

\[
p_Y(y) = \int_{-\infty}^{+\infty} P_{XY}(x, y)dx
\]

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:

$$\Pr(X = a_n, Y \in [a, b]) = \int_a^b p_{XY}(n, y)dy$$  \hspace{1cm} (1.9)

where the function $p_{XY}(n, y)$, with $n \in \{0, \ldots, k, \ldots\}$ and $y \in \mathbb{R}$, is $\geq 0$ 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:

$$p_{X_1X_2\ldots X_n}(x_1, x_2, \ldots, x_n) = p_{X_1}(x_1)p_{X_2}(x_2)\ldots p_{X_n}(x_n)$$  \hspace{1cm} (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:

$$p_{X_1}(x_1) = \int \cdots \int p_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) \, dx_2 \ldots dx_n$$

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

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

(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:

\[
\phi_{X_1 \ldots X_n}(u_1, \ldots, u_n) = \mathbb{E}\left\{ e^{j u_1 X_1 + \cdots + j u_n X_n} \right\} = \mathbb{E}\left\{ \prod_{k=1}^{n} e^{j u_k X_k} \right\}
\]  

(1.12)

Because \( |e^{j u X}| = 1 \), the characteristic function exists and is continuous even if the moments \( \mathbb{E}\{X^k\} \) do not exist. The Cauchy probability distribution, for example, the probability density function of which is \( p_X(x) = \frac{1}{\pi(1 + x^2)} \), 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 \( \phi_{X_k}(u_k) = \mathbb{E}\{e^{j u_k X_k}\} = \phi_{X_1 \ldots X_n}(0, \ldots, 0, u_k, 0, \ldots, 0) \).

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:

\[
\text{var}(X) = \mathbb{E}\{(X - \mathbb{E}\{X\})^2\} = \mathbb{E}\{X^2\} - (\mathbb{E}\{X\})^2
\]  

(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 \( b \):

\[
\mathbb{E}\{aX + b\} = a\mathbb{E}\{X\} + b
\]  

(1.14)

\[
\text{var}(aX + b) = a^2 \text{var}(X)
\]  

(1.15)

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

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. The covariance of \(X\) and \(Y\) is the quantity defined by:

\[
\text{cov}(X, Y) = \mathbb{E}\{(X - \mathbb{E}\{X\})(Y^* - \mathbb{E}\{Y^*\})\} = \mathbb{E}\{XY^*\} - \mathbb{E}\{X\}\mathbb{E}\{Y^*\}
\]

(1.16)

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

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

\[
\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\text{var}(Y)}}
\]

(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 \(C\) is the matrix with the generating element \(C_{ij} = \text{cov}(X_i, X_j)\) for \(1 \leq i \leq n\) and \(1 \leq j \leq n\).

Matrix notation: if we write

\[
X = [X_1 \ldots X_n]^T
\]

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

\[
\mathbb{E}\{X\} = [\mathbb{E}\{X_1\} \ldots \mathbb{E}\{X_n\}]^T
\]

the covariance matrix:

\[
C = \mathbb{E}\{(X - \mathbb{E}\{X\})(X - \mathbb{E}\{X\})^H\} = \mathbb{E}\{XX^H\} - \mathbb{E}\{X\}\mathbb{E}\{X\}^H
\]

(1.18)

and the correlation matrix

\[
R = DCD
\]

(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.
8  Digital Signal and Image Processing using MATLAB®

with

\[
D = \begin{bmatrix}
C_{11}^{-1/2} & 0 & \ldots & 0 \\
0 & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & 0 \\
0 & \ldots & 0 & C_{nn}^{-1/2}
\end{bmatrix}
\]  \hspace{1cm} (1.20)

\(R\) is obtained by dividing each element \(C_{ij}\) of \(C\) 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 \(a \in \mathbb{C}^n\), we have \(a^H Ca \geq 0\).

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

\[
\text{cov} \left( \sum_{i=1}^{m} v_i^* X_i, \sum_{j=1}^{n} w_j^* Y_j \right) = \sum_{i=1}^{m} \sum_{j=1}^{n} v_i^* w_j \text{cov}(X_i, Y_j)
\]  \hspace{1cm} (1.21)

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

\[
\text{cov}(A, B) = \mathbb{E} \left\{ V^H (X - \mathbb{E}\{X\}) (Y - \mathbb{E}\{Y\})^H W \right\} = \sum_{i=1}^{m} \sum_{j=1}^{n} v_i^* w_j \text{cov}(X_i, Y_j)
\]

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

\[
\text{cov} \left( V^H X, W^H Y \right) = V^H C W
\]  \hspace{1cm} (1.22)

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

\[
\begin{bmatrix}
Y_1 \\
\vdots \\
Y_q
\end{bmatrix} = A \begin{bmatrix}
X_1 \\
\vdots \\
X_n
\end{bmatrix} + b
\]

where \( A \) is a matrix and \( b \) is a non-random vector with the adequate sizes. We then have:

\[
\mathbb{E}\{Y\} = A\mathbb{E}\{X\} + b \\
C_Y = AC_XA^H
\]

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 \( \text{var}(X_i) = \sigma^2 \) and if \( \text{cov}(X_i, X_j) = 0 \) for \( i \neq j \). Hence their covariance matrix can be expressed:

\[
C = \sigma^2 I_n
\]

where \( I_n \) 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 \( Y \) taking values respectively in \( X \subset \mathbb{R} \) and \( Y \subset \mathbb{R}^q \) with joint probability density \( p_{X,Y}(x, y) \). The conditional expectation of \( X \) given \( Y \), is a (measurable) real valued function \( g(Y) \) such that for any other real valued function \( h(Y) \) we have:

\[
\mathbb{E}\{|X - g(Y)|^2\} \leq \mathbb{E}\{|X - h(Y)|^2\}
\]

(1.23)

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

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

\[
g(y) = \int_X x p_{X|Y}(x, y)dx
\]
with
\[ p_{X|Y}(x, y) = \frac{p_{XY}(x, y)}{p_Y(y)} \quad \text{and} \quad p_Y(y) = \int_x p_{XY}(x, y) \, dx \] (1.24)

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

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

1. **linearity:** \( \mathbb{E} \{a_1 X_1 + a_2 X_2 | Y \} = a_1 \mathbb{E} \{X_1 | Y\} + a_2 \mathbb{E} \{X_2 | Y\} \);

2. **orthogonality:** \( \mathbb{E} \{(X - \mathbb{E} \{X | Y\}) h(Y)\} = 0 \) for any function \( h : Y \mapsto \mathbb{R} \);

3. \( \mathbb{E} \{h(Y) f(X) | Y\} = h(Y) \mathbb{E} \{f(X) | Y\} \), for all functions \( f : X \mapsto \mathbb{R} \) and \( h : Y \mapsto \mathbb{R} \);

4. \( \mathbb{E} \{\mathbb{E} \{f(X, Y) | Y\}\} = \mathbb{E} \{f(X, Y)\} \) for any function \( f : X \times Y \mapsto \mathbb{R} \); specifically
   \[ \mathbb{E} \{\mathbb{E} \{X | Y\}\} = \mathbb{E} \{X\} \]

5. **refinement by conditioning:** it can be shown (see page 13) that
   \[ \text{cov} \{\mathbb{E} \{X | Y\}\} \leq \text{cov} \{X\} \] (1.25)
   The variance is therefore reduced by conditioning;

6. if \( X \) and \( Y \) are independent, then \( \mathbb{E} \{f(X) | Y\} = \mathbb{E} \{f(X)\} \). Specifically, \( \mathbb{E} \{X | Y\} = \mathbb{E} \{X\} \). The reverse is not true;

7. \( \mathbb{E} \{X | Y\} = X \), if and only if \( X \) is a function of \( Y \).

### 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 \( \lambda \) 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:

\[
\|X - X_0\|^2 = (X, X - X_0) - (X_0, X - X_0) = \|X\|^2 - (X, X_0) \tag{1.29}
\]

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 $\varepsilon \in \mathcal{C}$:

$$(X|\varepsilon) = \frac{\langle X, \varepsilon \rangle}{\langle \varepsilon, \varepsilon \rangle} \varepsilon$$

(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 $(X|\mathcal{C'}) = (X|\mathcal{C})$;
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:

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

(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 $\varepsilon$ 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.