

CORO-IMARO • AUVE
Autonomous Vehicles. Theoretical basis
CORO-SIP CORO-EPICO • STATES
Statistical Signal Processing and Estimation
Theory

November 25, 2020

Contents

Preface	ix
0.1 Context and objectives	ix
0.2 Requirements	ix
1 Probability theory	1
1.1 Probability space	1
1.2 Random variable (r.v.)	3
1.3 Expectation, mean, variance	6
1.4 Other features	8
1.5 Distribution models	9
1.6 Pair of r.v.: joint and marginal distributions	12
1.7 Pair of r.v.: conditional distributions	14
1.8 Triplet of r.v.	17
1.9 From probabilities to statistics	20
1.10 From the linear model to the normal distribution	22
1.11 Mixture distribution	24
1.12 Uncertainty propagation	25
2 Parametric estimation	27
2.1 Likelihood, Maximum Likelihood	27
2.2 Prior distribution, predictors	28
2.3 Posterior distribution, Bayesian estimators	28
2.4 How to obtain the posterior distribution?	29
2.5 How to propose a prior distribution?	29
2.6 Mean error analysis	30
2.7 Mean error, given the parameter	30
2.8 Mean error, Bayesian point of view	32
2.9 Linear model case	33
2.10 Probability error analysis	34
2.11 Practical remarks	35
2.12 Summary	35
3 Markov property	37
3.1 Stochastic processes: a short reminder	37
3.2 Markov process	37
3.3 Examples	38
3.4 Hidden Markov models (discrete time case)	41
3.5 Bayesian filtering	42

3.6	Linear model and Kalman filtering	43
3.7	Adapting the Kalman filter to non-linear models	45
4	Stochastic simulation	49
4.1	Random sampling	49
4.2	Pseudo-random numbers generators	49
4.3	Change of variable	50
4.4	Discrete distributions with finite support, mixture distributions	50
4.5	Rejection sampling	51
5	Monte Carlo methods	53
5.1	Direct sampling	53
5.2	Importance sampling	53
5.3	Importance sampling with auxiliary variable	54
5.4	Particle approximation	54
5.5	Application to Bayesian estimation	54
6	Particle filter	55
6.1	Principle	55
6.2	Bootstrap filter	57
6.3	Auxiliary particle filter	58
6.4	Fully adapted particle filter	59
6.5	Unscented particle filter	60
A	Mathematics	63
A.1	Matrices	63
A.2	Differentiation	64
A.3	Some functions	66
B	Stochastic processes	67
B.1	Time analysis	67
B.2	Spectral analysis	68
B.3	Autocorrelation function estimation	69
B.4	PSD estimation	70
B.5	Stationary processes and LTI systems	71
B.6	Generation process	71
C	More on probability theory	73
C.1	Bienaymé–Chebyshev inequality	73
C.2	Characteristic functions	74
C.3	Central limit theorem	75
C.4	Transformation of r.v.	75
D	More on estimation theory	79
D.1	Cramer-Rao inequality	79
D.2	If an efficient estimator exists, it is the MLE	80
D.3	Mean and variance of an i.i.d sequence	81
D.4	Orthogonality principle	81

E	Bayesian smoothing	83
----------	---------------------------	-----------

F	Matlab, Octave	85
----------	-----------------------	-----------

• Liste des algorithmes

3.1	Kalman filter	44
3.2	Stationary Kalman filter	44
3.3	Extended Kalman Filter (EKF)	46
3.4	EKF, additive noise case	46
3.5	Unscented Kalman Filter (UKF)	47
3.6	UKF, additive observation noise case	48
3.7	UKF, additive noise case	48
4.1	Rejection method for stochastic simulation	51
6.1	Bootstrap filter	57
6.2	Auxiliary particle filter	58
6.3	Fully adapted particle filter	59
6.4	Auxiliary particle filter	61

• Preface

• 0.1 Context and objectives

In the everyday life, we often see the hazard intervention:

- it does not take always the same time to go from home to work;
- a smoker can have a cancer or not;
- the fishing is not always good.

Such phenomenons are said to be **random**, or **stochastic**. To quantify them leads us to use the **probability theory**.

- Let's consider again the nicotine addiction example. Let's imagine that the doctor does not trust his patient about the daily number of cigarettes. He asks to the medical analysis laboratory to measure the nicotine blood level. The probability theory provides some tools to quantify the stochastic link between the daily number of cigarettes and the nicotine blood level.
- From this nicotine level, the doctor will be able to estimate the number of cigarettes. The **estimation theory** propose several solutions:
 - the most **likely** value;
 - the most **probable** value;
 - the **expected** value.

These notions seem similar, but have different meanings in estimation theory, in which we will distinguish the classical estimation from the **Bayesian** estimation.

- Let's go back to the fishing example. The shoal of fish path depends of numerous factors, thus depends of the hazard; it is a **random signal**, also called a **stochastic process**. The problem is to estimate this path along time, by means of the sonar onboard. The Bayesian estimation remains tractable, if we are able to write a **Markovian** representation of the path, and it becomes the **Bayesian filtering**.
- We have numerous applications:
 - spacecrafts or mobile robots localization;
 - from handwriting to text;
 - DNA sequencing...
- At the end of this course:
 1. you will understand that there is no magic algorithm to solve such problems;
 2. you will be able to question the domain specialist to elaborate a Markovian model which link the hidden quantity to the observed data;
 3. you will know how to write a Bayesian filter (or a reasonable approximation) to estimate the hidden quantity from the observed data.

• 0.2 Requirements

It is advised to review some mathematical reminders on (appendix A, page 63):

- differentiation;
- signal theory;
- (semi-)definite square matrix;
 - a symmetric matrix A is positive semi-definite if, for all vector x , $x^T A x \geq 0$;
 - every positive semi-definite matrix A has square roots R such that $A = R R^T$;

- the order between positive semi-definite matrices $A \geq B$ if $A - B$ is positive semi-definite is a partial order called the **Loewner order**.

- To avoid the distributions theory which would be necessary for a rigorous approach to probability theory, we will use in this book the intuitive concept of Dirac delta function.
- We will link it to the Kronecker delta.
- **The Kronecker delta** tests the equality of two discrete variables.
For all $(x, \bar{x}) \in \mathbb{X}^2$ where \mathbb{X} is a countable set:

$$\delta(x - \bar{x}) = \begin{cases} 1 & \text{if } x = \bar{x} \\ 0 & \text{otherwise} \end{cases} \quad \text{and then} \quad \sum_{x \in \mathbb{X}} \delta(x - \bar{x}) = 1 \quad (1)$$


- **The Dirac delta** tests the equality of two real-valued variables.

For all $(x, \bar{x}) \in \mathbb{R}^2$ (so x and \bar{x} are continuous variables):

$$\delta(x - \bar{x}) = \begin{cases} +\infty & \text{if } x = \bar{x} \\ 0 & \text{otherwise} \end{cases} \quad \text{under the condition} \quad \int_{\mathbb{R}} \delta(x - \bar{x}) \, dx = 1 \quad (2)$$

The function $x \mapsto \delta(x - \bar{x})$ is the Dirac delta located in \bar{x} .

The function $x \mapsto \alpha \delta(x - \bar{x})$ is the Dirac delta with weight α located in \bar{x} (its integral is α).

- The Dirac delta permits to extend the derivative in case of discontinuities.
In a discontinuity point, the derivative exhibits a Dirac pulse whose weight is the jump magnitude. 
- The Dirac delta and the Kronecker delta fulfill the **sifting property**:
 - for all function f from \mathbb{X} and for all $\bar{x} \in \mathbb{X}$:

$$\sum_{x \in \mathbb{X}} f(x) \delta(x - \bar{x}) = f(\bar{x}) \quad (3)$$

- for all function f from \mathbb{R} and for all $\bar{x} \in \mathbb{R}$:

$$\int_{\mathbb{R}} f(x) \delta(x - \bar{x}) \, dx = f(\bar{x}) \quad (4)$$

- We generalize the delta to functions of a variable which contains continuous and discrete components [7]:

$$\delta \left(\begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} - \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_d \end{bmatrix} \right) = \delta(x_1 - \bar{x}_1) \dots \delta(x_d - \bar{x}_d) \quad (5)$$

- In this book, δ can designate an hybrid Dirac-Kronecker delta, function of a vector variable with discrete and continuous components.

The unit integral (or sum), and the sifting property hold with this hybrid pulse:

- we integrate with respect to continuous variables;
- we sum with respect to discrete variables.

Chapter 1

Probability theory

1.1 Probability space

Let Ω be the set of the students of the university.

A student ω is drawn at random in this population.

BSc, MSc, PhD are the students in bachelor degree, master degree, doctor degree.

FR, IN, CN... are the French, Indian, Chinese... students.

With the set theory language:

- Ω is the set;
- ω is an element ($\in \Omega$);
- BSc, MSc, PhD, FR, IN, CN... are some subsets ($\subset \Omega$).

With the probability theory language:

- Ω is the universe;
- ω is an elementary event ($\in \Omega$);
- BSc, MSc, PhD, FR, IN, CN... are some events ($\subset \Omega$).

If all the students have the same chance to be drawn at random (equiprobability assumption), then the probability of the event $\Phi \subset \Omega$ is the proportion of elementary events of Ω belonging to Φ :

$$\text{Prob}(\Phi) = \frac{\text{Card } \Phi}{\text{Card } \Omega}$$

We just defined, in a natural way, in the case of a finite universe, by means of a counting interpretation, a **probability measure**:

$$\Phi \subset \Omega \longmapsto \text{Prob}(\Phi)$$

The universe, with this probability measure, is a **probability space**.

This counting interpretation (frequentist interpretation) is not necessary, but is useful to understand the calculation rules which come from the rigorous mathematical construction:¹

- $0 = \text{Prob}(\emptyset) \leq \text{Prob}(\Phi) \leq \text{Prob}(\Omega) = 1$
- $\text{Prob}(\Phi) + \text{Prob}(\bar{\Phi}) = 1$ (where $\bar{\Phi}$ is the complementary set of Φ in Ω);
- $\text{Prob}(\Phi_1 \cup \Phi_2) = \text{Prob}(\Phi_1) + \text{Prob}(\Phi_2) - \text{Prob}(\Phi_1 \cap \Phi_2)$
- $\text{Prob}(\bigcup_i \Phi_i) = \sum_i \text{Prob}(\Phi_i)$ if the Φ_i sets are disjoint;
- $\text{Prob}(\Psi) = \sum_i \text{Prob}(\Psi \cap \Phi_i)$ if the sets Φ_i form a partition of Ω (total probability law).

We will prefer to write the total probability law by means of the **conditional probabilities**.

$\text{Prob}(\Phi | \Psi)$ is the **conditional** probability to belong to Φ under the assumption to belong to Ψ .
By means of the frequentist interpretation, it is obviously:

$$\text{Prob}(\Phi | \Psi) = \frac{\text{Prob}(\Phi \cap \Psi)}{\text{Prob}(\Psi)} \quad (1.1)$$

- $\text{Prob}(\text{FR} | \text{MSc})$ is the probability that a master student is French;
- $\text{Prob}(\text{MSc} | \text{FR})$ is the probability that a French student is registered for a master degree.

1. The Kolmogorov axioms.

- ▲ Both probabilities are in general different.
- ⦿ For all event Ψ , the conditional probability measure $\Phi \mapsto \text{Prob}(\Phi | \Psi)$ fulfills the same properties than a probability measure; for example, $\text{Prob}(\bar{\Phi} | \Psi) = 1 - \text{Prob}(\Phi | \Psi)$.
- ▲ There is no general relation between $\text{Prob}(\Phi | \bar{\Psi})$ and $\text{Prob}(\bar{\Phi} | \Psi)$.

In the everyday life, there is often some ambiguity in the meaning of probabilities, since the universe, or the conditioning event, is not properly defined.

For example, a disease can be rare in the whole population, but frequent for the persons who are exposed to the agent which provokes this disease.

- T6 The events Φ and Ψ are **independent** if $\text{Prob}(\Phi \cap \Psi) = \text{Prob}(\Phi) \text{Prob}(\Psi)$, that is if $\text{Prob}(\Phi | \Psi) = \text{Prob}(\Phi)$, that is if $\text{Prob}(\Psi | \Phi) = \text{Prob}(\Psi)$.
- ⦿ If FR and MSc are independent: the proportion of French students is the same, considering either the MSc students only or the whole university; the proportion of MSc students is the same, considering either the French students only or the whole university.
- T7 The total probability law and the **Bayes law** are:

$$\text{Prob}(\Phi | \Psi) = \frac{\text{Prob}(\Psi | \Phi) \text{Prob}(\Phi)}{\text{Prob}(\Psi)} \quad (\text{Bayes law}) \quad (1.2)$$

$$\text{Prob}(\Psi) = \sum_i \text{Prob}(\Psi | \Phi_i) \text{Prob}(\Phi_i) \quad \text{if the sets } \Phi_i \text{ form a partition of } \Omega \text{ (total probability law)} \quad (1.3)$$

- T8 ▷ **Exercise 1.** In the university, 60% of the students are registered in a BSc degree, 30% of them in a MSc degree, 10% in a PhD degree. 30% of the BSc students, 40% of the MSc ones and 20% of the PhD ones are Chinese.
- a) What is the percentage of Chinese students in the university?
- b) What is the percentage of MSc students among the Chinese students?

- T9 ◁▷ **Evaluation, module 1** (screening test). The **prevalence** of a disease is the probability to be sick. The quality of a screening test is measured through the **sensibility** and the **specificity**.

Prevalence	Probability to be sick	
Sensibility	Probability that a sick person has a positive test	
Specificity	Probability that a healthy person has a negative test	

When the test is done, we can calculate the **predictive values**.

Positive predictive value	Probability to be sick if the test is positive	
Negative predictive value	Probability to be healthy if the test is negative	

- a) M is the set of sick persons (\bar{M} the healthy persons).
 T is the set of persons whose test is positive (\bar{T} the persons with negative test).
 Complete the tables above with the suitable probabilities (for example, $\text{Prob}(M | T)$, etc.)
- b) For a disease with prevalence 0.1%, and a test with sensibility 96% et specificity 98%, calculate the predictive values.

- T10 As a complement to this evaluation:

$M \cap T$ are the “true positives”,
 $\bar{M} \cap \bar{T}$ are the “true negatives”,
 $\bar{M} \cap T$ are the “false positives”,
 $M \cap \bar{T}$ are the “false negatives”.

The test accuracy is the probability that the test provides a correct result, that is $\text{Prob}((M \cap T) \cup (\bar{M} \cap \bar{T}))$.

◁▷ **Exercise 2.** Give the accuracy in function of the prevalence, the sensitivity and the specificity.

1.2 Random variable

1.2.1 Definition

A random variable (r.v.) x is a function from a probability space to a given set \mathbb{X} :

$$\begin{aligned} x : \Omega &\longrightarrow \mathbb{X} \\ \omega &\longmapsto x(\omega) \end{aligned}$$

For example:

- Ω is the university, that is the set of students;
- the student ω is drawn at random in Ω ;
- let x be the triplet made with his mean mark, his rank, the prepared diploma (we suppose that a student prepares exactly one diploma, $\{\text{BSc}, \text{MSc}, \text{PhD}\}$ is a partition of Ω).

We have defined a r.v. x from Ω to $\mathbb{X} = \mathbb{R}^+ \times \mathbb{N} \times \{\text{BSc}, \text{MSc}, \text{PhD}\}$, such that $x(\omega) = x$.

We say that x took the value x , or that x is a realization of x .

We will use upper case letters for the r.v., and lower case letters for the realizations.

1.2.2 Probability distribution

Let \mathbb{A} be a subset of \mathbb{X} .

The event $\{\omega \in \Omega \mid x(\omega) \in \mathbb{A}\}$ is noted $x \in \mathbb{A}$.

In particular, $\{\omega \in \Omega \mid x(\omega) = x\}$ is noted $x = x$.

- From the probability measure of Ω , we derive the **probability distribution** of x :

$$\mathbb{A} \subset \mathbb{X} \longmapsto \text{Prob}(x \in \mathbb{A})$$

- Exercise 3.** Pierre, Paul and Jacques left the nursery school, Pierre and Paul with 3 marbles in the pocket, Jacques with 4 ones. One of them was victim of an extortion, but his name was not revealed by the police. There are many assumptions at school about the victim identity, but, finally, people thinks that the 3 boys have the same chance of being attacked. What is the probability distribution of the loot marbles amount?

- Exercise 4.** The indicator function $\mathbf{1}_\Phi$ of a subset Φ of Ω takes the value 1 on this subset, 0 outside. Write the probability distribution of this r.v. which takes its value in $\{0, 1\}$ by means of the probability of Φ .

1.2.3 Probability mass function

If \mathbb{X} is countable, we say that x is a **discrete r.v.**.

Its distribution is described by its **probability mass function** (PMF):

$$p_x : x \in \mathbb{X} \longmapsto \text{Prob}(x = x)$$

Thus, for all $\mathbb{A} \subset \mathbb{X}$ such that the sum below is meaningful:

$$\text{Prob}(x \in \mathbb{A}) = \sum_{x \in \mathbb{A}} p_x(x) \quad (1.4)$$

- For example, let's throw a dice; the r.v. corresponds to the number of the upper side, which takes its value in $\mathbb{X} = \{1, 2, 3, 4, 5, 6\}$; for all $x \in \mathbb{X}$, $p_x(x) = \frac{1}{6}$; then $\text{Prob}(x \in \{1, 2\}) = \frac{2}{6}$.
- Let's define the r.v. DIPL which is the name of the prepared diploma (the "label" of the corresponding event):

$$\text{DIPL}(\omega) = \begin{cases} \text{"PhD"} & \text{if } \omega \in \text{PhD} \\ \text{"MSc"} & \text{if } \omega \in \text{MSc} \\ \text{"BSc"} & \text{if } \omega \in \text{BSc} \end{cases}$$

Then, $p_{\text{DIPL}}(\text{"PhD"})$ is the proportion of students who prepare a PhD.

$$p_{\text{DIPL}}(\text{"PhD"}) = \text{Prob}(\text{DIPL} = \text{"PhD"}) = \text{Prob}(\text{PhD})$$

1.2.4 Probability density function

If $\mathbb{X} = \mathbb{R}^d$, we say that X is a **continuous r.v.**

This book will reduce to the case where there exists a **probability density function** (PDF) $p_X : \mathbb{X} \rightarrow \mathbb{R}^+$ such that for all $\mathbb{A} \subset \mathbb{X}$ such that the integral below is meaningful:²

$$\text{Prob}(X \in \mathbb{A}) = \int_{\mathbb{A}} p_X(x) \, dx \quad (1.5)$$

Intuitively, $p_X(x) \, dx$ is the probability that X belongs to an hypervolume dx around x .

- The **cumulative distribution function** F_X (CDF) is defined, for all $x \in \mathbb{R}^d$, by:³

$$F_X(x) = \text{Prob}(X_1 \leq x_1 \text{ and } \dots \text{ and } X_d \leq x_d) \quad (1.6)$$

Necessarily, the PDF is the derivative of the CDF:

$$p_X = \frac{\partial^d F_X}{\partial x_1 \dots \partial x_d} \quad (1.7)$$

1.2.5 Generalization

PMF and PDF can be extended to the case where the r.v. X is hybrid, that is to say with discrete and continuous components: $X = \begin{bmatrix} X_{\text{cont}} \\ X_{\text{disc}} \end{bmatrix}$.

- Then, for all $\mathbb{A} \subset \mathbb{X}$ such that the formula below is meaningful:

$$\text{Prob}(X \in \mathbb{A}) = \sum_{\underbrace{x_{\text{disc}}}_{\mathbb{A}}} \int p_X([x_{\text{cont}}]) \, dx_{\text{cont}} \quad (1.8)$$

In this document, most of the formulae are written for continuous r.v., but are easily transposed to the discrete or hybrid cases; just remind that we integrate with respect to continuous variables, we sum with respect to discrete variables.

Remind that:

$$p_X(x) \geq 0 \quad \sum_{\underbrace{x_{\text{disc}}}_{\mathbb{X}}} \int p_X([x_{\text{cont}}]) \, dx_{\text{cont}} = 1$$

1.2.6 In practice

We rarely know the exact probability distribution.

How to obtain an approximate PMF or PDF?

- **Using situation analysis (physical law)** For a correct dice, the probability of the visible number is $\frac{1}{6}$.
- **With a model** A parameterized distribution is assumed, then we estimate the parameters.
- **With a descriptive approach (statistics, discrete r.v. case)** We observe n_T realizations of the r.v., and we measure, for each possible value, the proportion of trials which gave this value.

The PMF is the limit of these proportions when n_T tends to ∞ .⁴  

2. In the formula (1.5), if $d > 1$, the integral is a multiple integral and dx means the hypervolume $dx_1 \dots dx_d$.

3. In the formula (1.6), x_i (resp. x_i) means the i th scalar component of x (resp. x).

4. Matlab. To plot of a PMF estimated over a scalar valued population in the vector x .

`ux = unique(x); px = hist(x, ux)/length(x); stem(ux, px)`

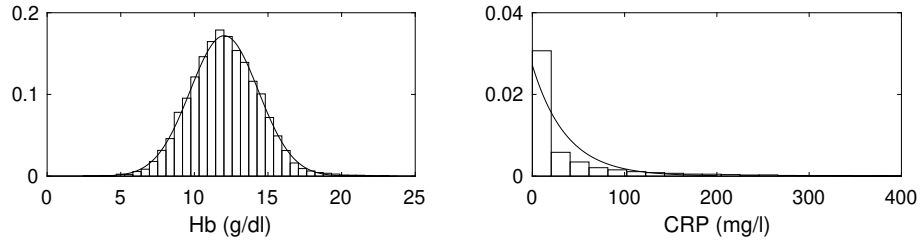


Figure 1.1: Histogram hemoglobin (Hb) and C-reactive protein (CRP). Gaussian model for Hb. Tempered exponential model for CRP


τ21 With a descriptive approach (statistics, continuous r.v. case) We assume that there is an infinite number of students in the university.

A finite number n_r of students is drawn at random, and we plot the **normalized histogram** of their average mark:

- the axis between the minimal mark and the maximal one is separated into n_b intervals with the same width;
- above each interval, a rectangle (bin) whose support is the interval and the area is equal to the proportion of students whose average mark belongs to the interval is drawn.

Thus, the histogram area is 1.

The PDF is the limit histogram when n_r , n_b and $\frac{n_r}{n_b}$ tend to ∞ .⁵

We can choose $n_b = \lfloor \sqrt{n_r} \rfloor$: the number of intervals is the square root of the population size. See [13] for other choices. 

τ22 The figure 1.1 displays the histogram of the hemoglobin (Hb) and the C-reactive protein (CRP) for patients of an hospital in Paris.

The Gauss distribution model (page 10) seems to fit in the Hb case.

The exponential distribution model (page 11), tried on the CRP, does not fit.

τ23 1.2.7 Support

The **support** $S(x)$ of the r.v. x is the set of value on which the PDF (or PMF) is strictly positive; this is the set of values that a r.v. can take:

$$S(x) = \{x \in \mathbb{X} \mid p_x(x) > 0\} \quad (1.9)$$

Let g be a function of a variable in \mathbb{X} .

For all realization x of X , we associate $g(x)$, realization of the r.v. $g \circ X$; this new r.v. is noted $g(X)$.

Let's note that we need to define g only on the support $S(x)$, since all realizations of x belong to this support; so, we can shorten the definition of a function.

Rather than writing: "let g be the function defined, for all $x \in S(x)$, by $g(x) = x^2$ ", we write:
 "let g be the function defined by $g(X) = X^2$ ".

• Let's see the r.v. $\left[\frac{\text{WEIGHT}}{\text{SIZE}} \right]$. We obtain a new r.v., the "body mass index", with $\text{BMI} = \frac{\text{WEIGHT}}{\text{SIZE}^2}$, used to quantify the stoutness of a person. We don't have to define the BMI for a 0 size, since nobody has 0 size.

τ24 **◁▷ Evaluation, module 2.** Among the functions below, defined over \mathbb{Z} or \mathbb{R} , give those which can be considered as a PMF or a PDF.

- Is this function (over \mathbb{Z}) a PMF? $p_x(x) = \begin{cases} 2 & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$
- Is this function (over \mathbb{Z}) a PMF? $p_x(x) = \begin{cases} 2 & \text{if } x = 0 \\ -1 & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$
- Is this function (over \mathbb{Z}) a PMF? $p_x(x) = \begin{cases} \frac{1}{3} & \text{if } x = -1 \text{ ou } x = 0 \text{ ou } x = 1 \\ 0 & \text{otherwise} \end{cases}$
- For which α is this function (over \mathbb{R}) a PDF? $p_x(x) = \begin{cases} \alpha x & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$

5. Matlab. To plot a normalized histogram with n_b bins of a scalar valued population in the vector x .
`[n,b] = hist(x,nb); bar(b, n/(b(2)-b(1))/sum(n), 1)`

1.3 Expectation, mean, variance

1.3.1 Mathematical expectation

The **expectation** operator, for all r.v. x and all function g of a variable in \mathbb{X} with numerical value, such that the integral below converges, gives a value noted $E(g(x))$ defined by:⁶

$$E(g(x)) = \int_{\mathbb{X}} g(x) p_x(x) dx \quad (1.10)$$

• Its purpose is to calculate a mean value.

τ26 Given the distribution of $\begin{bmatrix} \text{WEIGHT} \\ \text{SIZE} \end{bmatrix}$, we can calculate the BMI mean value.⁷

τ27 In linear transforms, the expectation operator fulfills the property below:⁸

$$E(g^T(x)) = (E(g(x)))^T \quad (1.11)$$

$$E\left(\begin{bmatrix} g_{11}(x) & g_{12}(x) \\ g_{21}(x) & g_{22}(x) \end{bmatrix}\right) = \begin{bmatrix} E(g_{11}(x)) & E(g_{12}(x)) \\ E(g_{21}(x)) & E(g_{22}(x)) \end{bmatrix} \quad (1.12)$$

$$E(g_1(x) + g_2(x)) = E(g_1(x)) + E(g_2(x)) \quad (1.13)$$

$$E(Ag(x)C + B) = A E(g(x)) C + B \quad (1.14)$$

$$E(A) = A \quad (1.15)$$

τ28 ◁▷ **Exercise 5.** Let Φ be an event. Express $\text{Prob}(\Phi)$ by means of the expectation and the indicator function of Φ .

Let x be a r.v. Express $\text{Prob}(x \in \mathbb{A})$ by means of the expectation and the indicator function of \mathbb{A} .

τ29 The **variance** gives, for all r.v. x such that the expectation below exists, a positive semi-definite matrix $\text{Var}(x)$ defined by:

$$\text{Var}(x) = E((x - E(x))(x - E(x))^T) \quad (1.16)$$

It is straightforward to show that:

$$\text{Var}(x) = E(xx^T) - E(x)E(x^T) \quad (1.17)$$

◁▷ **Exercise 6.** Prove the formula (1.17).

τ30 Under linear transforms, this operator fulfills the rule below:⁹

$$\text{Var}(Ax + B) = A \text{Var}(x) A^T \quad (1.18)$$

◁▷ **Exercise 7.** Prove the formula (1.18).

1.3.2 Mean and variance

Let x be a numerical r.v. Its mean m_x and its variance $C_{x,x}$, if they exist, are defined by:

$$m_x = E(x) \quad C_{x,x} = \text{Var}(x) \quad (1.19)$$

• The mean is also called expectation, or first **moment**, or first **cumulant**; it is a location parameter.

• The variance is also called second cumulant; it is a scale parameter; it measures how the r.v. realizations can fall away from the mean value.

For a scalar valued r.v., the variance square root is noted σ_x , and called “**standard deviation**”.¹⁰

6. In the discrete valued case:

$$E(g(x)) = \sum_{x \in \mathbb{X}} g(x) p_x(x) = \sum_{x \in \mathbb{X}} g(x) \text{Prob}(x = x)$$

7. This implies that we obtain the same result by means of the calculation below (law of the unconscious statistician):

$$E(g(x)) = \int_{\mathbb{Y}} y p_{g(x)}(y) dy$$

8. A, B, C are constant matrices such that sums and products are meaningful.

9. A and B are constant matrices are constant matrices such that sums and products are meaningful.

10. If the mean is not zero, the dimensionless ratio $\frac{\sigma_x}{m_x}$ is called “coefficient of variation”.

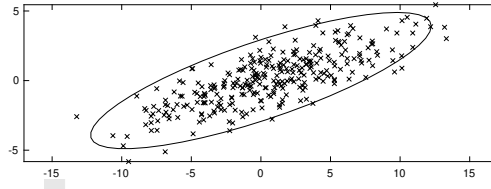


Figure 1.2: Confidence domain for a bivariate normal distribution

- If the variance is not invertible, the r.v. is **degenerate**.

The second moment is $E(\mathbf{x} \mathbf{x}^\top)$, that is $C_{\mathbf{x},\mathbf{x}} + m_{\mathbf{x}} m_{\mathbf{x}}^\top$.

To give a comparison with solid mechanics, the mean is the center of mass, the variance is the inertia around the center of mass.

1.3.3 Confidence domain

The formula (1.5) provides the probability that a r.v. belongs to a given set.

Conversely, for a given probability P_0 , there exist multiple sets such that the probability that the r.v. belongs to them is P_0 ; these sets are the P_0 confidence domains.

- In general, we look for the smallest confidence domain. For example, when searching a lost object in the sea, the probability to find the object being fixed, we look for the smallest domain in order to minimize the means to implement.
 - We can show that, if the PDF is known, with minor assumptions, the smallest confidence domain corresponds to a constant level PDF on its boundary [14]. Such a domain do not necessarily contains the mean, and can be disconnected.
- For instance, the figure 1.2 represents a 300 realizations population of a bivariate normal r.v. (page 24) with zero mean and variance $\begin{bmatrix} 25 & 8 \\ 8 & 4 \end{bmatrix}$, and the 95% confidence domain which is a contour of the PDF. This contour is elliptic in the bivariate normal distribution case.
- The knowledge of the mean and the variance provides only confidence domains at at least P_0 , by means of the Bienaymé-Tchebychev inequality, unless the variance is zero.

1.3.4 Zero-variance case

◀▶ **Exercise 8.** Let Y be a positive r.v., such that $E(Y)$ exists.

Prove that $\int_0^{+\infty} \text{Prob}(Y \geq y) dy = E(Y)$ (tip: invert integration order).

- Let Y be a positive r.v. whose mean exists.

For all $y > 0$ (**Markov inequality**):^{P1}

$$\text{Prob}(Y \geq y) \leq \frac{E(Y)}{y} \quad \text{for a positive r.v.} \quad (1.20)$$

- By means of the Markov inequality and the exercise result:

$$E(Y) = 0 \quad \text{if and only if} \quad \forall y > 0, \text{Prob}(Y < y) = 1$$

We say that the r.v. Y is null almost everywhere (or almost surely, or with probability 1); the almost sure equality is a weaker condition than the equality; but, since there is no practical consequence, this book will be simplified by saying that the r.v. Y is null, and we will note: $Y = 0$.

- The straightforward corollary is that, for all r.v. \mathbf{x} which takes its value in \mathbb{R}^d :

- \mathbf{x} is (almost surely) null if and only if its square euclidean norm has zero mean:

$$E(\mathbf{x}^\top \mathbf{x}) = 0 \quad \text{if and only if} \quad \mathbf{x} = \mathbf{0}_d$$

P1. With y a positive r.v. and $y > 0$. The r.v. inequalities $\mathbf{1}_{Y \geq y} \leq \frac{Y}{y} \mathbf{1}_{Y \geq y} \leq \frac{Y}{y}$ is preserved by expectation. Since $\text{Prob}(Y \geq \varepsilon) = E(\mathbf{1}_{Y \geq \varepsilon})$, we get the result.

- x is (almost surely) null if and only if its 2th moment is null:

$$E(x x^T) = 0_{d \times d} \text{ if and only if } x = 0_d$$

- x is (almost surely) equal to its mean value, if and only if its variance is null:

$$\text{Var}(x) = 0_{d \times d} \text{ if and only if } x = E(x)$$

Many proofs in this book will be based on these equivalences.

τ37 **◁▷ Evaluation, module 3.** A uniform distribution is constant over the support. We consider the case of a uniformly distributed r.v. over $[a, b]$, with $a < b$.

- Calculate the mean and the variance.
- How to choose a and b such that the r.v. has zero mean and unit variance?

τ38 1.4 Other features

◁▷ Exercise 9. Prove that, for a distribution with a PDF, the mean value minimizes the criterion $J(m) = E(\|x - m\|_W^2)$ with respect to m (independently from the symmetric positive definite matrix W used in the norm $\|x\|_W = \sqrt{x^T W x}$).

τ39 1.4.1 Mode, median, quartiles

There are several possibilities to define the “center” of a probability distribution of a r.v. x .

- the mean value minimizes $E(\|x - m\|_2^2)$ with respect to m ; it does not always exist;
- for a scalar valued r.v., the **median** minimizes $E(|x - m|)$ with respect to m [22]; it is a value m such that $\text{Prob}(x < m) \leq \frac{1}{2}$ and $\text{Prob}(x > m) \leq \frac{1}{2}$ ¹¹; an interval of solutions can exist, a common habit is to use the middle; the median can be extended to the vector-valued case [28];
- the **mode**, which is the value which maximizes the PDF; it is not necessarily unique; if a PDF exhibits several local maxima, the distribution is a **multimodal** one.

- ◉ In general, in the scalar valued r.v. case, for a unimodal density which spreads over the right side: mode < median < mean.

τ40 In a scalar valued r.v. case, the quartiles try to summarize the PDF shape:

- the first quartile Q_1 is a value such that $\text{Prob}(x < Q_1) \leq \frac{1}{4}$ and $\text{Prob}(x > Q_1) \leq \frac{3}{4}$;
- the second quartile is the median;
- the third quartile Q_3 is a value such that $\text{Prob}(x < Q_3) \leq \frac{3}{4}$ and $\text{Prob}(x > Q_3) \leq \frac{1}{4}$;

For Q_1 and Q_3 , an interval of solutions can exist; a common habit is to use the upper bound.

The interquartile range $Q_3 - Q_1$ is another way to measure the dispersion of a r.v.

The quartiles are a special case among **quantiles**.

τ41 1.4.2 Higher order statistics

The moments and the cumulants of order 1 and 2 (if they exist) are only partial characteristics of a r.v.. They can be completed with higher order moments and cumulants, but the characterization remains partial.

- ◉ In a nutshell:

- the moments are the coefficients of the Taylor series expansion around 0 of the first characteristic function, that is the Fourier transform of the PDF;
- the cumulants are the coefficients of the Taylor series expansion around 0 of the second characteristic function, that is the logarithm of the Fourier transform of the PDF.

11. These probabilities equal $\frac{1}{2}$ if $\text{Prob}(x = m) = 0$

For all $k \in \mathbb{N}$, there exists an invertible relation between the k first moments and the k first cumulants. With the cumulants, it is easier to obtain an interpretation which is independent of the location and the scale.

- τ42 In particular, for a scalar r.v. X , the k th moment is $E(X^k)$. Let's note $K_k(X)$ the k th cumulant; the 4 first cumulants are written in function of the 4 first moments through the formulae:

$$\begin{aligned} m_X &= K_1(X) = E(X) \\ \sigma_X^2 &= K_2(X) = E(X^2) - m_X^2 \\ K_3(X) &= E(X^3) - 3m_X E(X^2) + 2m_X^3 \\ K_4(X) &= E(X^4) - 4m_X E(X^3) + 12m_X^2 E(X^2) - 6m_X^4 \end{aligned}$$

- τ43 For a non-zero variance, we can use the normalized (and dimensionless) cumulants $\tilde{K}_3(X)$ et $\tilde{K}_4(X)$:

$$\begin{aligned} \tilde{K}_3(X) &\triangleq \frac{K_3(X)}{\sigma_X^3} = E\left(\left(\frac{X - m_X}{\sigma_X}\right)^3\right) && \text{skewness} \\ \tilde{K}_4(X) &\triangleq \frac{K_4(X)}{\sigma_X^4} = E\left(\left(\frac{X - m_X}{\sigma_X}\right)^4\right) - 3 && \text{excess Kurtosis} \end{aligned}$$

Both factors are independent of the mean and the variance, because they depend only of the standard score $\frac{X - m_X}{\sigma_X}$; they only depend of the PDF shape.

- τ44 The **skewness** measures the asymmetry of the probability distribution: if the PDF is symmetric around the mean value, it is zero (but the reciprocal is false in general).
For a unimodal density which spreads over the right side, it is positive.

τ45 1.5 Distribution models

1.5.1 Two basic distributions

“Deterministic distribution” Although we often distinguish r.v. from deterministic variables, we can present the deterministic variables as a limit case of the r.v., in which a r.v. always takes the same value x_0 . Therefore, a deterministic variable is a discrete r.v. whose PMF is:

$$p_X(x) = \begin{cases} 1 & \text{if } x = x_0 \\ 0 & \text{otherwise} \end{cases}$$

- ☛ By means of Kronecker delta, a short way to write this PMF is:

$$p_X(x) = \delta(x - x_0)$$

- ☛ But a numerical deterministic r.v. can also be considered as a continuous r.v.: in the formula above, p_X becomes the PDF, and δ becomes the Dirac delta.
- ☛ A deterministic variable has null variance.

- τ46 **Bernoulli distribution** A r.v. is Bernoulli-distributed if it only takes two distinct values x_1 and x_2 , one with the probability $\lambda \in]0, 1[$, the other with the probability $1 - \lambda$; the PMF is:

$$p_X(x) = \begin{cases} \lambda & \text{if } x = x_1 \\ 1 - \lambda & \text{if } x = x_2 \\ 0 & \text{otherwise} \end{cases}$$

- ☛ In the formula below, if p_X is a PMF, δ is the Kronecker delta; if p_X is a PDF, δ is the Dirac delta:

$$p_X(x) = \lambda \delta(x - x_1) + (1 - \lambda) \delta(x - x_2)$$

- ☛ The mean is $\lambda x_1 + (1 - \lambda) x_2$, the variance is $\lambda(1 - \lambda)(x_2 - x_1)(x_2 - x_1)^\top$.
- ☛ In the scalar case, in general, skewness and kurtosis fulfill the inequality below:

$$\tilde{K}_4(X) \geq [\tilde{K}_3(X)]^2 - 2$$

The equality holds only with the Bernoulli distribution.^{P2}

- τ47 ◁▷ **Exercise 10** (Bernoulli distribution). Complete the figure 1.3 with the PDF of the Bernoulli distribution with equiprobability, zero mean and unit variance.

τ48 1.5.2 A few probability distributions

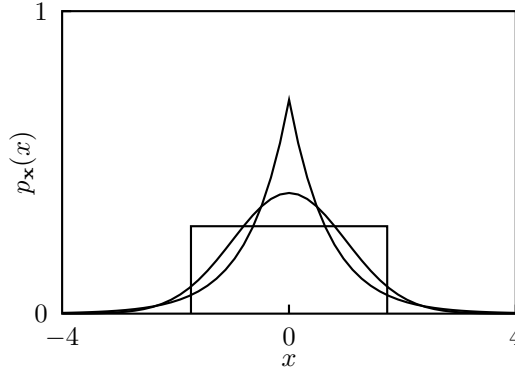


Figure 1.3: PDF of uniform, Gauss, Laplace distributions (zero mean, unit variance)

- τ49 **Uniform distribution** A real valued r.v. is uniformly distributed on the interval $[a, b]$ if its PDF is constant on this interval, and zero elsewhere:

$$p_x(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

The excess kurtosis is $-\frac{6}{5}$.

- τ50 **Laplace distribution** A real valued r.v. is Laplace-distributed with mean m_x and standard deviation σ_x if its PDF is:

$$p_x(x) = \frac{1}{\sigma_x \sqrt{2}} \exp\left(-\sqrt{2} \frac{|x - m_x|}{\sigma_x}\right)$$

The excess kurtosis is 3.

- ◉ **Univariate normal distribution** A real valued r.v. is normally distributed (or **Gauss**-distributed) with mean m_x and standard deviation σ_x if its PDF is:

$$p_x(x) = \frac{1}{\sigma_x \sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(x - m_x)^2}{\sigma_x^2}\right)$$

All higher order cumulants are zero.¹²

- τ51 Many properties hold with the normal distribution, such as the zero value of the higher order cumulants. Even if we won't define in this book the notions of "information" or "entropy" which are well known in information theory, we can say that the normal distribution is the less informative, or the most disordered among all distributions with finite variance.
- ◉ From a pragmatic point of view, this distribution often leads to simple calculations which can be easily implemented with limited numerical means. For example, on the NASA Apollo program, in 1960, a trajectory estimation filter based on Gaussian assumption was implemented.

P2. Since these factors depend only of the shape of the PDF, we can check the inequality in the case of a zero-mean and unit variance distribution. We easily check that:

$$\text{Var} \begin{pmatrix} x^2 \\ x \end{pmatrix} = \begin{bmatrix} \tilde{K}_4(x) + 2 & \tilde{K}_3(x) \\ \tilde{K}_3(x) & 1 \end{bmatrix} \quad \text{Var} (a x^2 + x - a) = (\tilde{K}_4(x) + 2) a^2 + 2 \tilde{K}_3(x) a + 1$$

The 2×2 matrix is positive semi-definite, thus its determinant is positive; this proves the inequality. The binomial in a is positive, thus its discriminant is negative; this proves the inequality. If the equality holds, the zero-mean r.v. $x^2 - \tilde{K}_3(x)x - 1$ has a null 2th moment; thus, it is a null variable; thus, x can only takes the two solutions of the 2nd order equation $x^2 - \tilde{K}_3(x)x - 1 = 0$.

12. A probability distribution is platikurtic (or sub-Gaussian) if the kurtosis is negative, mesokurtic if it is zero, leptokurtic (or super-Gaussian) if it is positive.

- Least squares optimization are often encountered (for example in regression problems), they can be explained with a probabilistic interpretation based on Gaussian distributions.
- These remarks also hold for the multivariate normal distribution, page 23, for which the PDF is, if the variance is invertible:

$$p_{\mathbf{x}}(x) = \frac{1}{\sqrt{\det(2\pi C_{\mathbf{x},\mathbf{x}})}} \exp \left[-\frac{1}{2} (x - m_{\mathbf{x}})^{\top} C_{\mathbf{x},\mathbf{x}}^{-1} (x - m_{\mathbf{x}}) \right] \quad (1.21)$$

- If \mathbf{x} is a normal r.v. which takes its value in \mathbb{R}^d , $(\mathbf{x} - m_{\mathbf{x}})^{\top} C_{\mathbf{x},\mathbf{x}}^{-1} (\mathbf{x} - m_{\mathbf{x}})$ is driven by a χ^2 distribution.

τ52 χ^2 distribution A real valued r.v. is χ^2 -distributed with d degrees of freedom if its PDF is: ¹³

$$p_{\mathbf{x}}(x) = \begin{cases} \frac{1}{2^{\frac{d}{2}} \Gamma(\frac{d}{2})} x^{\frac{d}{2}-1} \exp\left(-\frac{x}{2}\right) & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The CDF is:

$$F_{\mathbf{x}}(x) = \begin{cases} \Gamma_{\text{inc}}\left(\frac{x}{2}, \frac{d}{2}\right) & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The mean is d , the variance is $2d$.

For $d = 2$, we obtain the exponential distribution with mean 2.

τ53 Exponential distribution A real valued r.v. is exponentially distributed with mean $m_{\mathbf{x}}$ if its PDF is:

$$p_{\mathbf{x}}(x) = \begin{cases} \frac{1}{m_{\mathbf{x}}} \exp\left(-\frac{x}{m_{\mathbf{x}}}\right) & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The CDF is:

$$F_{\mathbf{x}}(x) = \begin{cases} 1 - \exp\left(-\frac{x}{m_{\mathbf{x}}}\right) & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The standard deviation is the mean value.

τ54 ◁▷ Exercise 11. Let \mathbf{x} be a real valued r.v., and \mathbf{y} is another r.v. such that $\mathbf{y} = \max(\mathbf{x}, 0)$. Write the CDF of \mathbf{y} in function of the CDF of \mathbf{x} . Deduce the PDF of \mathbf{y} in function of the PDF of \mathbf{x} and a Dirac delta function with proper weight.

τ55 ◁▷ Evaluation, module 5 (Binomial distribution). A coin is such that the probability to obtain “tail” is λ . If we throw this coin n times, the number of times for which we obtain “tail” is a binomial r.v. K with parameters n and λ , which takes its value in $\{0, \dots, n\}$; the PMF is:

$$p_K(k) = n! \frac{\lambda^k (1-\lambda)^{n-k}}{k! (n-k)!}$$

We throw this coin twice.

- a) What is the probability to never obtain “tail”?
- b) What is the probability to obtain “tail” exactly once?
- c) What is the probability to obtain “tail” twice?
- d) What is the sum of the three preceding results?
- e) Calculate the mean and the variance of the number of “tails”.

13. Γ et Γ_{inc} are the Euler gamma function and the incomplete Euler gamma function:

$$\Gamma(a) = \int_0^{+\infty} t^{a-1} e^{-t} dt \quad \Gamma_{\text{inc}}(x, a) = \frac{1}{\Gamma(a)} \int_0^x t^{a-1} e^{-t} dt$$

if $a \in \mathbb{N}$, $\Gamma(a+1) = a!$ and $\Gamma\left(a + \frac{1}{2}\right) = \frac{\prod_{k=1}^a (2k-1)}{2^a} \sqrt{\pi}$

1.6 Pair of r.v.: joint and marginal distributions

1.6.1 Definitions

Let's consider a r.v. of dimension at least 2 whose components are split under the form $\begin{bmatrix} x \\ y \end{bmatrix}$. x which takes its values in \mathbb{X} and y which takes its values in \mathbb{Y} make up the pair (x, y) . The distribution of this pair is the distribution of $\begin{bmatrix} x \\ y \end{bmatrix}$; but, this split being done, the distribution of this pair is called the **joint distribution**.

- This book will focus on the case where the joint distribution has a PDF (or PMF, or hybrid) denoted $p_{x,y}$; for all (x, y) (note that $p_{x,y}(x, y) = p_{y,x}(y, x)$):

$$p_{x,y}(x, y) = p_{\begin{bmatrix} x \\ y \end{bmatrix}}\left(\begin{bmatrix} x \\ y \end{bmatrix}\right)$$

- The probability calculation is:

$$\text{Prob}((X, Y) \in \mathbb{A}) = \int_{\mathbb{A}} p_{x,y}(x, y) \, dx \, dy \quad (1.22)$$

- Furthermore, the support is denoted $S(x, y)$.

1.6.2 The components x and y of the pair are called **marginal r.v.**

When we observe a marginal r.v. x , we ignore the component y in the pair (x, y) .

The PDF (or PMF) of the marginal r.v. are obtained by integration (or summation) of the joint PDF (or PMF) with respect to the ignored variable; for example, the PDF of x is, for all x :^{P3}

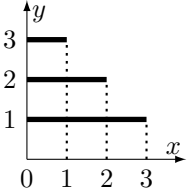
$$p_x(x) = \int_{\mathbb{Y}} p_{x,y}(x, y) \, dy \quad (1.23)$$

- The knowledge of both marginal distributions is not sufficient to know the joint distribution.

- The properties given in the section 1.2 hold:

- for the marginal r.v. x ;
- for the marginal r.v. y ;
- for the r.v. $\begin{bmatrix} x \\ y \end{bmatrix}$, that is the pair (x, y) , through the change of notation $\begin{bmatrix} x \\ y \end{bmatrix} \rightarrow (x, y)$.

Exercise 12 (Pair of r.v.: red thread). Let (x, y) be a pair of r.v. where x is a continuous valued one and y is a discrete valued one. We suppose that $p_{x,y}$ is constant over the support $([0, 3] \times \{1\}) \cup ([0, 2] \times \{2\}) \cup ([0, 1] \times \{3\})$.



- a) What is the value of $p_{x,y}$ on the support?
 Plot $x \mapsto p_{x,y}(x, y)$ for $y = 1$, for $y = 2$, and for $y = 3$
 Plot $y \mapsto p_{x,y}(x, y)$ for $0 < x < 1$, for $1 < x < 2$, and for $2 < x < 3$.
- b) Give and plot the PDF of the marginal r.v. x ; give its mean and its variance.
- c) Give and plot the PMF of the marginal r.v. y ; give its mean and its variance.

1.6.2 Mathematical expectation

With g a function from $\mathbb{X} \times \mathbb{Y}$ with numerical value, the expectation operator is:

$$\mathbb{E}(g(X, Y)) = \int_{\mathbb{X} \times \mathbb{Y}} g(x, y) p_{x,y}(x, y) \, dx \, dy \quad (1.24)$$

- In linear transforms, this operator fulfills the following rules:¹⁴

$$\mathbb{E}(g^T(X, Y)) = (\mathbb{E}(g(X, Y)))^T \quad (1.25)$$

$$\mathbb{E}\left(\begin{bmatrix} g_{11}(x, y) & g_{12}(x, y) \\ g_{21}(x, y) & g_{22}(x, y) \end{bmatrix}\right) = \begin{bmatrix} \mathbb{E}(g_{11}(x, y)) & \mathbb{E}(g_{12}(x, y)) \\ \mathbb{E}(g_{21}(x, y)) & \mathbb{E}(g_{22}(x, y)) \end{bmatrix} \quad (1.26)$$

$$\mathbb{E}(g_1(X, Y) + g_2(X, Y)) = \mathbb{E}(g_1(X, Y)) + \mathbb{E}(g_2(X, Y)) \quad (1.27)$$

$$\mathbb{E}(A g(X, Y) C + B) = A \mathbb{E}(g(X, Y)) C + B \quad (1.28)$$

P3. The proof is easy in the discrete probability case:

$p_x(x) = \text{Prob}(x = x) = \text{Prob}(\cup_y ((x, y) = (x, y))) = \sum_y \text{Prob}((x, y) = (x, y)) = \sum_y p_{x,y}(x, y)$

We will admit that the result is valid for the continuous probability case with integration with respect to continuous components.

14. A, B, C are constant matrices such that sums and products are meaningful.

τ60 The variance operator can apply to both components x and y . Furthermore, we define the **covariance** operator:

$$\text{Cov}(x, y) = E((x - E(x))(y - E(y))^T) \quad (1.29)$$

The alternate form below is straightforward to obtain:

$$\text{Cov}(x, y) = E(xy^T) - E(x)E(y^T) \quad (1.30)$$

• In linear transforms, the covariance operator fulfills the rules below:¹⁵

$$\text{Cov}(x, x) = \text{Var}(x) \quad (1.31)$$

$$\text{Cov}(y, x) = (\text{Cov}(x, y))^T \quad (1.32)$$

$$\text{Cov}(Ax + B, Cy + D) = A \text{Cov}(x, y) C^T \quad (1.33)$$

τ61 Furthermore, if the sums below are meaningful:

$$E(x + y) = E(x) + E(y) \quad (1.34)$$

$$\text{Var}(x + y) = \text{Var}(x) + \text{Var}(y) + \text{Cov}(x, y) + \text{Cov}(y, x) \quad (1.35)$$

$$\text{Cov}(x + y, x' + y') = \text{Cov}(x, x') + \text{Cov}(x, y') + \text{Cov}(y, x') + \text{Cov}(y, y') \quad (1.36)$$

τ62 1.6.3 Mean, variance, covariance

Means and variances of marginal r.v. x and y are not sufficient to retrieve the mean and the variance of the r.v. $\begin{bmatrix} x \\ y \end{bmatrix}$.

They have to be completed with the covariance $C_{x,y}$ defined by (note that $C_{y,x} = C_{x,y}^T$):

$$C_{x,y} = \text{Cov}(x, y) \quad (1.37)$$

Thus, the mean and the variance of the r.v. $\begin{bmatrix} x \\ y \end{bmatrix}$ are:

$$E\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} m_x \\ m_y \end{bmatrix} \quad \text{Var}\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} C_{x,x} & C_{x,y} \\ C_{y,x} & C_{y,y} \end{bmatrix} \quad (1.38)$$

• The **Cauchy-Schwarz inequality** is (inequality for Loewner order, see page 64):

$$C_{x,x} \geq C_{x,y} C_{y,y}^{-1} C_{y,x} \quad (1.39)$$

The equality holds if and only if there is a linear relation between x and y :^{P4}

$$C_{x,x} = C_{x,y} C_{y,y}^{-1} C_{y,x} \quad \text{if and only if} \quad x = m_x + C_{x,y} C_{y,y}^{-1} (y - m_y) \quad (1.40)$$

τ63 1.6.4 Independence, uncorrelatedness

We say that the r.v. x and y are **independent**, and we note $x \perp y$, if $p_{x,y} = p_x p_y$.

This probabilistic notion is similar to the intuitive notion; for example, the result of the blue dice does not influence the result of the red dice, the results are independent.

• We say that the numerical r.v. x and y are **uncorrelated** if their covariance $C_{x,y}$ is null.

▲ Independence implies uncorrelatedness, but the converse is false in general.

τ64 1.6.5 Pearson's correlation coefficient

If x and y are scalar valued, we define their **Pearson's correlation coefficient** $\rho_{x,y}$ by:

$$\rho_{x,y} = \frac{C_{x,y}}{\sigma_x \sigma_y} \quad (1.41)$$

• From Cauchy-Schwarz inequality (1.39), this undimensional coefficient is necessarily between -1 and 1 . Obviously, $\rho_{x,x} = 1$.

What can we deduce from its value concerning the dependence between two r.v.?

¹⁵ A, B, C, D are constant matrices such that sums and products are meaningful.

^{P4} The variance is a positive semi-definite matrix. If $C_{y,y}$ is invertible, the Schur complement of $C_{y,y}$ in $\text{Var}(\begin{bmatrix} x \\ y \end{bmatrix})$, that is $C_{x,x} - C_{x,y} C_{y,y}^{-1} C_{y,x}$ (see page 63), is positive semi-definite, which gives the **Cauchy-Schwarz inequality**. If the equality holds, just check that the r.v. $(x - m_x) - C_{x,y} C_{y,y}^{-1} (y - m_y)$ has zero mean and zero variance; then, it is almost surely zero.

- $\rho_{X,Y} = 0$: maybe independent;
- $\rho_{X,Y} \neq 0$: dependent;
- $\rho_{X,Y} = 1$: linearly dependent, the realizations of (X, Y) are on a rising line;
- $\rho_{X,Y} = -1$: linearly dependent, the realizations of (X, Y) are on a falling line.

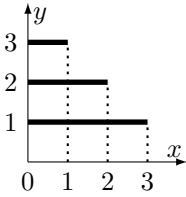
☛ In both last cases, the line contains the point (m_X, m_Y) , its slope is $\pm \frac{\sigma_Y}{\sigma_X}$.

T65 Denoting U_k the k th scalar component of a r.v. U which takes its value in \mathbb{R}^d , the variance of U is:

$$C_{U,U} = [\rho_{U_\ell, U_k} \sigma_{U_\ell} \sigma_{U_k}]_{\substack{1 \leq \ell \leq d \\ 1 \leq k \leq d}} \quad (1.42)$$

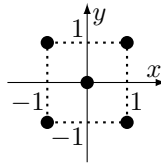
The main diagonal contains the variance of the scalar components.

T66 ◁▷ **Exercise 13** (Pair of r.v.: red thread). Let (X, Y) be a pair of r.v. where X is a continuous valued one and Y is a discrete valued one. We suppose that $p_{X,Y}$ is constant over the support $([0, 3] \times \{1\}) \cup ([0, 2] \times \{2\}) \cup ([0, 1] \times \{3\})$.



- What is the correlation coefficient between X and Y ?
- Are X and Y independent?

T67 ◁▷ **Evaluation, module 6.** (X, Y) is a pair of discrete r.v. The PMF $p_{X,Y}$ is constant over the support $\{(0, 0), (-1, -1), (1, 1), (-1, 1), (1, -1)\}$.



- Fill the table below.

x	y	$p_{X,Y}(x, y)$	$p_X(x)$	$p_Y(y)$	$p_X(x) p_Y(y)$
-1	-1				
-1	1				
0	0				
1	-1				
1	1				

- What is the correlation coefficient? Are X and Y independent?

T68 1.7 Pair of r.v.: conditional distributions

1.7.1 Définitions

Let (X, Y) be a pair of r.v.

We look at the distribution of X , when Y is fixed.

For example, what is the distribution of the mark of the math exam, in the universe restricted to the students whose mark of the physics exam is 5 over 10?

- ☛ From the conditional probability measure in the universe, we build the probability distribution of X given Y :

$$(y \in S(Y), \mathbb{A} \in \mathcal{P}(X)) \mapsto \text{Prob}(X \in \mathbb{A} \mid Y = y)$$

In the theory of continuous r.v., the event $Y = y$ has in general a null probability. We will admit that the conditional probability $\text{Prob}(X \in \mathbb{A} \mid Y = y) = \frac{\text{Prob}((X \in \mathbb{A}) \cap (Y = y))}{\text{Prob}(Y = y)}$ which is the ratio between two null probabilities, is meaningful.

- ¶69 It is often useful to calculate a conditional probability to belong to a set which depends of the condition. We will write the conditional probability distribution under the form:

$$(y \in S(Y), \mathbb{A} : \mathbb{Y} \rightarrow \mathcal{P}(\mathbb{X})) \mapsto \text{Prob}(x \in \mathbb{A}(Y) \mid Y = y)$$

- ¶70 Let's suppose that it exists a PDF (or PMF) such that, for all $y \in S(Y)$:

$$\text{Prob}(x \in \mathbb{A}(Y) \mid Y = y) = \int_{\mathbb{A}(y)} p_{X|Y}(x, y) \, dx \quad (1.43)$$

- This PDF is necessarily, for all $(x, y) \in \mathbb{X} \times S(Y)$:^{P5}

$$p_{X|Y}(x, y) = \frac{p_{X,Y}(x, y)}{\int_{\mathbb{X}} p_{X,Y}(u, y) \, du} \quad (1.44)$$

- ¶71 Remind that:

$$p_{X|Y}(x, y) \geq 0 \quad \int_{\mathbb{X}} p_{X|Y}(x, y) \, dx = 1$$

There is no general result on the integration with respect to the second variable.

- For all fixed $y \in S(Y)$, $S(X \mid Y = y)$ is the support of the conditional distribution:

$$S(X \mid Y = y) = \{x \in \mathbb{X} \mid p_{X|Y}(x, y) > 0\} \quad (1.45)$$

- ¶72 The formulae (1.23) and (1.44) permit to obtain the marginal PDF p_Y and the conditional PDF $p_{X|Y}$ in function of the joint PDF $p_{X,Y}$.

- Conversely, by combining both formulae, we obtain the join PDF given the conditional and marginal PDFs; for all (x, y) :

$$p_{X,Y}(x, y) = \begin{cases} p_{X|Y}(x, y) p_Y(y) & \text{if } y \in S(Y) \\ 0 & \text{otherwise} \end{cases} \quad (1.46)$$

- By means of this formula and the equivalent one obtained by the exchange of x and y , we easily obtain the fundamental properties below, in which only the marginal and conditional PDFs appear; for all $(x, y) \in S(X) \times S(Y)$:

$$p_{X|Y}(x, y) = \frac{p_X(x) p_{Y|X}(y, x)}{p_Y(y)} \quad (\text{Bayes law}) \quad (1.47)$$

$$p_Y(y) = \int_{S(X)} p_X(x) p_{Y|X}(y, x) \, dx \quad (\text{total probability law}) \quad (1.48)$$

- ¶73 The independence of two r.v. was defined in the previous section: X et Y are **independent** if $p_{X,Y} = p_X p_Y$.

More naturally, the independence of x and y means that the distribution of x given y does not depend of the value taken by y ; thus, the conditional distribution reduces to the marginal one:

$$x \perp\!\!\!\perp y \quad \text{if and only if} \quad \forall (x, y), p_{X|Y}(x, y) = p_X(x) \quad (1.49)$$

A deterministic variable is independent of every other r.v., since it takes always the same value.

- ¶74 ◁▷ **Exercise 14.** Two balls are consecutively drawn from an urn containing 3 white balls and 5 black balls. The 1st ball is not replaced in the urn before the 2nd sortition (drawing “without replacement”). For each drawing, we earn 1 € for a white ball, 0 € for a black one. x is the amount earned at the 1st drawing, y is the amount earned at the 2nd one.

- a) Use the assumptions to fill the tables below:

x	$p_X(x)$	x	y	$p_{Y X}(y, x)$
0		0	0	
0		0	1	
1		1	0	
1		1	1	

P5. We note that the denominator is nothing but $p_Y(y)$. The proof is easy in the discrete probability case:

$$\begin{aligned} \text{Prob}(x \in \mathbb{A}(Y) \mid Y = y) &= \frac{\text{Prob}((x \in \mathbb{A}(y)) \cap (Y = y))}{\text{Prob}(Y = y)} = \frac{\text{Prob}(\cup_{x \in \mathbb{A}(y)} [(x, Y) = (x, y)])}{\text{Prob}(Y = y)} = \frac{\sum_{x \in \mathbb{A}(y)} \text{Prob}((x, Y) = (x, y))}{\text{Prob}(Y = y)} \\ &= \frac{\sum_{x \in \mathbb{A}(y)} p_{X,Y}(x, y)}{p_Y(y)} = \sum_{x \in \mathbb{A}(y)} \frac{p_{X,Y}(x, y)}{p_Y(y)} = \sum_{x \in \mathbb{A}(y)} p_{X|Y}(x, y) \end{aligned}$$

We will assume that it remains true with integration with respect to continuous components.

b) Deduce the tables below (the last one correspond to the total earned amount):

y	$p_Y(y)$	x	y	$p_{X,Y}(x, y)$	z	$p_{X+Y}(z)$
0		0	0		0	
1		0	1		1	
		1	0		2	
		1	1			

1.7.2 Conditional expectation, mean and variance

The **conditional expectation** operator returns, for all numerical function g of a variable in $\mathbb{X} \times S(Y)$, a function which, for all $y \in S(Y)$, maps the value noted $E(g(X, Y) | Y = y)$ and defined by:

$$E(g(X, Y) | Y = y) = \int_{\mathbb{X}} g(x, y) p_{X|Y}(x, y) dx \quad (1.50)$$

This is the mean of $g(X, Y)$ given the event $Y = y$.

Let's temporarily note \bar{g} the function $y \mapsto E(g(X, Y) | Y = y)$.

As every function from $S(Y)$, it permits to transform the r.v. Y into another r.v. $\bar{g}(Y)$.

Thus, we defined:

- \bar{g} a function defined over $S(Y)$;
- $\bar{g}(y)$ the image of the realization $y \in S(Y)$ by this function, noted $E(g(X, Y) | Y = y)$;
- $\bar{g}(Y)$ the r.v. coming from the transformation of the r.v. Y by this function, noted $E(g(X, Y) | Y)$.

As mentioned above, the definition of the r.v. est equivalent to the definition of the function.

This convention can extend in many ways; for all realization y de Y , can be associated:

- the realization $\text{Prob}(X \in A(Y) | Y = y)$ of the r.v. $\text{Prob}(X \in A(Y) | Y)$;
- the realization $S(X | Y = y)$ of the random support $S(X | Y)$;
- the realization $x \mapsto p_{X|Y}(x, y)$ of the random PDF $x \mapsto p_{X|Y}(x)$.

Exercise 15. Re-write the formulae (1.43) and (1.45) by means of this convention.

For linear transforms, the conditional expectation fulfills the rules below:¹⁶

$$E(g^T(X, Y) | Y) = (E(g(X, Y) | Y))^T \quad (1.51)$$

$$E\left(\begin{bmatrix} g_{11}(X, Y) & g_{12}(X, Y) \\ g_{21}(X, Y) & g_{22}(X, Y) \end{bmatrix} | Y\right) = \begin{bmatrix} E(g_{11}(X, Y) | Y) & E(g_{12}(X, Y) | Y) \\ E(g_{21}(X, Y) | Y) & E(g_{22}(X, Y) | Y) \end{bmatrix} \quad (1.52)$$

$$E(g_1(X, Y) + g_2(X, Y) | Y) = E(g_1(X, Y) | Y) + E(g_2(X, Y) | Y) \quad (1.53)$$

$$E(A(Y)g(X, Y)C(Y) + B(Y) | Y) = A(Y)E(g(X, Y) | Y)C(Y) + B(Y) \quad (1.54)$$

$$E(A(Y) | Y) = A(Y) \quad (1.55)$$

$E(X | Y)$ is the **conditional mean**.¹⁷

The **conditional variance** is defined by:¹⁸

$$\text{Var}(X | Y) = E((X - E(X | Y))(X - E(X | Y))^T | Y) \quad (1.56)$$

The conditional variance can also be written as:

$$\text{Var}(X | Y) = E(XX^T | Y) - E(X | Y)E(X^T | Y) \quad (1.57)$$

For a linear transform, the conditional variance fulfills the rule below:¹⁹

$$\text{Var}(A(Y)X + B(Y) | Y) = A(Y)\text{Var}(X | Y)A^T(Y) \quad (1.58)$$

As an immediate corollary:

$$\text{Var}(B(Y) | Y) \text{ is null} \quad (1.59)$$

16. A , B and C are matrix functions of over $S(Y)$, such that the sums and the products below are meaningful.

17. If $E(X | Y) = E(X)$, then X and Y are uncorrelated (ex. 19, page 19).

18. Thus, for all $y \in S(Y)$, we can write: $\text{Var}(X | Y = y) = E((X - E(X | Y = y))(X - E(X | Y = y))^T | Y = y)$

19. A and B are some functions of a variable in \mathbb{Y} such that the sums and products are meaningful.

1.7.3 Laws based on the total probability law

The total probability law (1.48) can be re-written as (1.60) and permits to obtain the rules below [26]:²⁰

$$\forall x \in \mathbb{X}, p_x(x) = E(p_{x|Y}(x)) \quad (\text{total probability}) \quad (1.60)$$

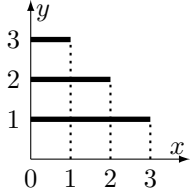
$$\text{Prob}(X \in \mathbb{A}) = E(\text{Prob}(X \in \mathbb{A} | Y)) \quad (\text{total probability}) \quad (1.61)$$

$$E(X) = E(E(X | Y)) \quad (\text{total expectation}) \quad (1.62)$$

$$\text{Var}(X) = E(\text{Var}(X | Y)) + \text{Var}(E(X | Y)) \quad (\text{total variance}) \quad (1.63)$$

It is a good training to rewrite the total expectation formula by means of the functions $y \mapsto E(X | Y = y)$ and $y \mapsto \text{Prob}(Y = y)$ for a discrete valued r.v. Y , and to explain the result.

Exercise 16 (Pair of r.v.: red thread). Let (X, Y) be a pair of r.v. where X is a continuous valued one and Y is a discrete valued one. We suppose that $p_{x,Y}$ is constant over the support $([0, 3] \times \{1\}) \cup ([0, 2] \times \{2\}) \cup ([0, 1] \times \{3\})$.



- Give and plot the PDF of $X | Y$; give its mean and its variance.
- Give the mean and the variance of X by means of the total expectation formula and the total variance one.
- Give and plot the PMF of $Y | X$; give its mean and its variance.

Evaluation, module 7. Let's flip a correct coin.

At the first throw, we earn 1 € for "tail", 0 € otherwise.

If we did not earn anything at the first throw, we do not earn anything at the second one.

If we earned 1 € at the first throw, we flip the coin again, we earn again 1 € if we obtain "tail", 0 € otherwise.

X is the earned amount at the first throw, Y is the earned amount at the second throw.

- Formalize the assumptions by filling the tables below.

x	$p_x(x)$	x	y	$p_{Y X}(y, x)$
0		0	0	
		0	1	
1		1	0	
		1	1	

- Deduce the tables below (the last one corresponds to the total earned amount).

y	$p_Y(y)$	x	y	$p_{X,Y}(x, y)$	z	$p_{X+Y}(z)$
0		0	0		0	
		0	1		1	
1		1	0		1	
		1	1		2	

1.8 Triplet of random variables

We know how to handle distributions (joint, marginal, conditional) of a pair of r.v..

For a triplet of r.v., we have also to consider, for example, the joint distribution of 2 r.v. given a third one.

- Let (X, Y, Z) be a triplet of r.v.

The PDF of the pair (X, Y) given Z is noted $p_{X,Y|Z}$.

The PDF of X given the pair (Y, Z) is noted $p_{X|Y,Z}$.

In this section, we mainly examine distributions given the r.v. Z .

²⁰ In automatic clustering or in analysis of variance (ANOVA), the first term of the decomposition of the total variance (T , "total") is called within-groups variance (W , "within"), the second term is called between-groups variance (B , "between"): $T = W + B$.

1.8.1 Joint, marginal, conditional distributions

The PDF of the marginal v.a. Y given Z is, for all $y \in \mathbb{Y}$:²¹

$$p_{Y|Z}(y) = \int_{\mathbb{X}} p_{X,Y|Z}(x, y) \, dx \quad (1.64)$$

- The PDF of X given the pair (Y, Z) is, for all $(x, y) \in \mathbb{X} \times S(Y)$:

$$p_{X|Y,Z}(x, y) = \frac{p_{X,Y|Z}(x, y)}{\int_{\mathbb{X}} p_{X,Y|Z}(u, y) \, du} \quad (1.65)$$

- From these PDFs, we can retrieve the joint distribution of the pair (X, Y) given Z , for all (x, y) , using:

$$p_{X,Y|Z}(x, y) = \begin{cases} p_{X|Y,Z}(x, y) p_{Y|Z}(y) & \text{if } y \in S(Y | Z) \\ 0 & \text{otherwise} \end{cases} \quad (1.66)$$

Exercise 17. Write the Bayes law and the total probability law which link $p_{X|Y,Z}$, $p_{X|Z}$, $p_{Y|X,Z}$ and $p_{Y|Z}$, for all $(x, y) \in S(X | Z) \times S(Y | Z)$.

We will apply the results of the previous exercise to solve the Monty Hall problem (name of the host of an American television game show).

Three opaque cups are upside down on the table. One of them hides a diamond, the other ones hide a bean. The quizmaster knows where is the diamond.

The gambler designates a cup (nobody lifts the cup).

The quizmaster lifts, among the two other cups, one which does not hide the diamond.

The gambler must lift a cup, he earns the object under it.

Exercise 18 (Monty Hall problem). The cups are numbered from 1 to 3; we note:

- D the number which hides the diamond,
- G the number pointed out by the gambler,
- Q the number lifted by the quizmaster

Fill the table. What is the probability to earn the diamond, if the gambler:

- lifts always the initially pointed out cup?
- lifts always the not initially pointed out cup?
- lifts one of them at random?

q	$p_{Q D,G}(q, d, g)$			$\sum_d p_{Q D,G}(q, d, g)$	$p_{D Q,G}(d, q, g)$			g
1								1
2								1
3								1
1								2
2								2
3								2
1								3
2								3
3								3
	1	2	3		1	2	3	
	d				d			

1.8.2 Conditional covariance

The **conditional covariance** is defined by:

$$\text{Cov}(X, Y | Z) = E((X - E(X | Z))(Y - E(Y | Z))^T | Z) \quad (1.67)$$

- It is straightforward to show that the conditional covariance can also be written as:

$$\text{Cov}(X, Y | Z) = E(X Y^T | Z) - E(X | Z) E(Y^T | Z) \quad (1.68)$$

- For linear transforms, this operator fulfills the rules below:²²

$$\text{Cov}(A(Z)X + B(Z), C(Z)Y + D(Z) | Z) = A(Z) \text{Cov}(X, Y | Z) C^T(Z)$$

21. That is, for all $(y, z) \in \mathbb{Y} \times S(Z)$, $p_{Y|Z}(y, z) = \int_{\mathbb{X}} p_{X,Y|Z}(x, y, z) \, dx$.

22. A , B , C and D are some functions, such that the sums and products below are meaningful.

- ✎ The total expectation formula permits to obtain the **total covariance formula**:

$$\text{Cov}(X, Y) = E(\text{Cov}(X, Y | Z)) + \text{Cov}(E(X | Z), E(Y | Z)) \quad (1.69)$$

τ99 ◀▷ **Exercise 19** (Laws of total expectation and total covariance). Give $\text{Cov}(E(X | Y), Y)$ in function of $\text{Cov}(X, Y)$:

- a) by means of the law of total expectation (1.62);
- b) by means of the law of total covariance (1.69) (in which we take $Y = Z$).
- c) If $E(X | Y) = E(X)$, what is the covariance $\text{Cov}(X, Y)$?

τ90 1.8.3 Conditional independence

X et Y are independent given Z , which is noted $X \perp\!\!\!\perp Y | Z$, if $p_{X,Y|Z} = p_{X|Z} p_{Y|Z}$.

- ✎ The equivalence below holds [8]:

$$X \perp\!\!\!\perp Y | Z \quad \text{if and only if} \quad p_{X|Y,Z} = p_{X|Z} \quad (1.70)$$

That means that, in the pair (Y, Z) , Z catch all the information on X (we often say that X depends only of Z).

- ▲ The conditional independence do not imply the independence. For example, we suppose that in an exam, each sheet is marked by two professors. Z is the sheet value, X and Y are the marks. Intuitively, all sheets taken together, both marks are dependent with a positive correlation. However, for a given sheet, the professors have to mark independently (the best is that they do not communicate).
- ▲ The independence does not imply the conditional independence. Let's play to heads or tails with two coins. The two results are independent.
Let's introduce the boolean variable which is true is the results are equal. Given this r.v., the two results are not independent, since the value of one of them implies the value of the other one.

τ91 The equivalence below holds:

$$X \perp\!\!\!\perp (Y, Z) \Leftrightarrow X \perp\!\!\!\perp Y | Z \text{ and } X \perp\!\!\!\perp Z \quad (1.71)$$

- ▲ If X is independent of the pair (Y, Z) , then X is independent of Y , and X is independent of Z (the independence of the whole implies the independence of each part), but the converse is false in general.

τ92 1.8.4 Recursive aspects of probability calculations

The formula (1.46) gives the joint distribution in function of the marginal distribution and the conditional distribution. By induction, we obtain (and that can be generalized to every number of r.v.):

$$p_{X,Y,Z} = p_{X|Y,Z} p_{Y|Z} p_Z \quad (1.72)$$

- ✎ The rules (1.60) to (1.63) can be interpreted as the last stage of a recursive calculation whose preceding stage is:

$$\forall x \in \mathbb{X}, p_{X|Z}(x) = E(p_{X|Y,Z}(x) | Z) \quad (\text{total probability}) \quad (1.73)$$

$$\text{Prob}(X \in \mathbb{A} | Z) = E(\text{Prob}(X \in \mathbb{A} | Y, Z) | Z) \quad (\text{total probability}) \quad (1.74)$$

$$E(X | Z) = E(E(X | Y, Z) | Z) \quad (\text{total expectation}) \quad (1.75)$$

$$\text{Var}(X | Z) = E(\text{Var}(X | Y, Z) | Z) + \text{Var}(E(X | Y, Z) | Z) \quad (\text{total variance}) \quad (1.76)$$

τ93 1.8.5 Mutual independence

We introduced only the independence between 2 r.v..

The **mutual independence** between the 3 r.v. (X, Y, Z) can be defined as follows: X is independent of the pair (Y, Z) and Y is independent of Z .

By induction, we can define the mutual independence between any number of r.v..

- ✎ We easily check that some r.v. are mutually independent if and only if the PDF of the joint distribution is the product of the marginal PDFs, that is, in the case of 3 r.v., by means of the formula (1.72):

$$p_{X,Y,Z} = p_X p_Y p_Z \quad (X, Y, Z \text{ mutually independent})$$

- ▲ The mutual independence of the triplet (X, Y, Z) implies the independance of the pair (X, Y) , of the pair (X, Z) , and of the pair (Y, Z) (the mutual independence implies the pairwise independence), but the converse does not hold in general.

T94 **Exercise 20** (Training for Markov chains). Let's play to "heads or tails" with a correct coin.

We earn 1 € if we obtain "tail", 0 € otherwise.

We throw the coin 3 times, consecutively.

Y_i is the amount of money earned at throw $i \in \{1, 2, 3\}$.

X_i is the cumulated amount of money till throw $X_i = \sum_{j=1}^i Y_j$.

a) Fill the left side table, and deduce the right side one.

x_1	x_2	$p_{X_1, X_2, X_3}(x_1, x_2, x_3)$				$p_{X_1, X_2}(x_1, x_2)$
0	0					
0	1					
0	2					
1	0					
1	1					
1	2					
		0	1	2	3	
		x_3				

x_1	x_2	$p_{X_3 X_1, X_2}(x_3, x_1, x_2)$			
0	0				
0	1				
1	1				
1	2				
		0	1	2	3
		x_3			

b) Are X_3 and X_1 independent given X_2 ?

c) Are X_3 and X_2 independent given X_1 ?

T95 **Evaluation, module 8.** X , Y et Z are three r.v. which take their value in $\{0, 1\}$; the joint PMF is:

$$p_{X,Y,Z}(x, y, z) = \frac{1}{2} \delta(y - x) \delta(z - 1 + x)$$

a) Fill the table below.

y	z	$p_{Y,Z X}(y, z, x)$		$p_{Y X}(y, x)$		$p_{Z X}(z, x)$		$p_{Y,Z}(y, z)$	$p_Y(y)$	$p_Z(z)$
0	0									
0	1									
1	0									
1	1									
		0	1	0	1	0	1			
		x		x		x				

b) Are Y and Z independent?

c) Are Y and Z independent given X ?

T96 1.9 From probabilities to statistics

1.9.1 Random sampling

Let $x : \Omega \rightarrow \mathbb{X}$ be a r.v.

It is equivalent to say:

- We obtained n_r realizations of x (from n_r draws with replacement in Ω);
- We obtained a realization of the n_r -tuple of r.v. $\vec{X}_{n_r} = (x_1, \dots, x_{n_r})$ where the r.v. x_q , $1 \leq q \leq n_r$ are mutually independent with the same distribution as x .²³
- We also say that \vec{X}_{n_r} is a **random sample** of x , with size n_r (or a n_r -**sample** of x).
- We say that \vec{X}_{n_r} is a set of **independent and identically distributed (i.i.d.)** r.v.
- We also say that \vec{X}_{n_r} is made of n_r independent copies of x .

T97 1.9.2 Empirical distribution

From a random sample \vec{X}_{n_r} of x , we obtain an **empirical** PDF (of PMF) of x with:

$$\hat{p}_X(x) = \frac{1}{n_r} \sum_{q=1}^{n_r} \delta(x - x_q) \quad (1.77)$$

23. Strictly speaking, for all q , x_q is a r.v. from the universe Ω^{n_r} to \mathbb{X} where the probability measure of Ω^{n_r} comes from the probability measure of Ω using $\text{Prob}(\Phi_1 \times \dots \times \Phi_{n_r}) = \prod_{q=1}^{n_r} \text{Prob}(\Phi_q)$, and thus, $x_q(\omega_1, \dots, \omega_{n_r}) = x(\omega_q)$.

- For a discrete r.v. (δ is the Kronecker delta), the empirical PMF is an approximation of the exact PMF. 📊 🧠
- If a component of \mathbf{x} is continuous-valued, the empirical PDF must be integrated with respect to the continuous valued component to obtain for example:
 - the empirical CDF;²⁴ 📊 🧠
 - the empirical quantiles (median...);
 - the empirical moments (mean, variance...).

🔍 **Exercise 21.** Let x be a r.v. which takes its value in \mathbb{X} , and let g be a function from \mathbb{X} which takes a numerical value. Give the approximation of $E(g(x))$ which is obtained by replacing the exact PDF with the empirical PDF in the formula (1.10).

1.9.3 Sample mean

Let X be a r.v. with a mean and a variance.

Let \vec{X}_{n_r} be a sample of x .

The empirical mean is a r.v. which is nothing but the arithmetic mean:

$$\bar{X}_{n_r} = \frac{1}{n_r} \sum_{q=1}^{n_r} X_q$$

- \bar{X}_{n_r} is a r.v. whose mean and variance are:²⁵

$$E(\bar{X}_{n_r}) = E(X) \quad \text{Var}(\bar{X}_{n_r}) = \frac{1}{n_r} \text{Var}(X)$$

- Thus, the distribution of $\sqrt{n_r}(\bar{X}_{n_r} - E(X))$ is zero-mean with variance $\text{Var}(X)$.
The **central limit theorem** says that this distribution tends to a normal one when $n_r \rightarrow +\infty$. 📊 🧠

🔍 When n_r tends to $+\infty$, the variance of \bar{X}_{n_r} tends to 0, which implies the **weak law of large numbers** (convergence in probability):^{P6}

$$\forall y > 0 \quad \lim_{n_r \rightarrow +\infty} \text{Prob}(\|\bar{X}_{n_r} - E(X)\| < y) = 1$$

But the convergence in probability does not imply that a sample mean path tends to a limit.

- We admit the **strong law of large numbers** (almost sure convergence):

$$\text{Prob}\left(\lim_{n_r \rightarrow +\infty} \bar{X}_{n_r} = E(X)\right) = 1$$

- This means that for (almost) all sequence $(x_n)_{n \in \mathbb{N}^*}$ which is a realization of $(X_n)_{n \in \mathbb{N}^*}$:

$$\lim_{n_r \rightarrow +\infty} \bar{x}_{n_r} = E(X)$$

🔍 If the calculation of a given quantity can be interpreted as the calculation of the mean of a r.v. x , a way to calculate this mean is to draw at random a realization \vec{x}_{n_r} of \vec{X}_{n_r} with a high n_r ; then, the arithmetic mean \bar{x}_{n_r} is an approximation of the mean m .

Such a method is called a **Monte-Carlo method**.

- In practice, we simulate random drawings by means of pseudo-random numbers generators. This is called **Stochastic simulation**.²⁶ 📊 🧠

24. Matlab. To plot the empirical CDF.

`stairs([min(x);sort(x(:))], (0:length(x))/length(x))`

25. The mean formula does not necessitate any assumption. The variance formula just necessitates the uncorrelatedness between the r.v. in the sample.

P6. Let's define $y_{n_r} = (\bar{X}_{n_r} - m)^T (\bar{X}_{n_r} - m)$. $E(y_{n_r}) = E(\text{trace Var}(\bar{X}_{n_r}))$ tends to 0. From Markov inequality ($\forall y >$

$0, \text{Prob}(y_{n_r} < y) \geq 1 - \frac{E(y_{n_r})}{y}$), we obtain that $\forall y, \lim_{n_r \rightarrow \infty} \text{Prob}(y_{n_r} < y) = 1$.

26. Matlab. To obtain an approximation of π (with a high n_r).

`4*mean(abs([1 j]*rand(2,nr))<1)`

1.10 From the linear model to the normal distribution

The stochastic link between two r.v. X and Y is the joint distribution (or the marginal X and the conditional $Y | X$). In many cases, a simple model can provide a structure for the distribution of Y given X , in which a modeling error W appears.

1.10.1 General model

A general enough formalization of this dependency is to suppose that there exist a function \bar{h} such that:

$$Y = \bar{h}(X, W) \text{ with } W \perp\!\!\!\perp X \quad (1.78)$$

• Thus, using the total probability law:

$$p_{Y|X}(y) = \int \delta(y - \bar{h}(x, w)) p_W(w) \, dw \quad (1.79)$$

1.10.2 Additive error model

The error is often assumed an additive one ($\bar{h}(x, w) = h(x) + w$):

$$Y = h(X) + W \text{ with } W \perp\!\!\!\perp X \quad (1.80)$$

• Thus, using the sifting property of the Dirac delta:

$$p_{Y|X}(y) = p_W(y - h(X)) \quad (1.81)$$

1.10.3 Linear model

h is assumed to be a linear function, and the independence of X et W is limited to the second order.

There exists H such that the mean and the variance given X of the r.v. $W = Y - H X$ are independent of X :

$$Y = H X + W \text{ with } \begin{cases} E(W | X) = m_W \\ \text{Var}(W | X) = C_{W,W} \end{cases} \quad (1.82)$$

• It is equivalent to assume a linear conditional mean $E(Y | X)$ and a uniform conditional variance $\text{Var}(Y | X)$:

$$\text{there exist } H, m_W, C_{W,W} \text{ such that } \begin{cases} E(Y | X) = H X + m_W \\ \text{Var}(Y | X) = C_{W,W} \end{cases} \quad (1.83)$$

• In numerous problems, X is the unknown quantity of interest, for which the observation Y brings some information. The r.v. W usually represents a measurement noise. We refer to a **linear observation model**. H is the **observation matrix**.

• If X has a mean m_X and a variance $C_{X,X}$:

- the mean and the variance of Y and the covariance are, in function of H , m_W and $C_{W,W}$:^{P7}

$$\begin{cases} m_Y = H m_X + m_W \\ C_{Y,Y} = H C_{X,X} H^T + C_{W,W} \\ C_{X,Y} = C_{X,X} H^T \end{cases} \quad (1.84)$$

- We can deduce the conditional mean and variance in function of the joint distribution parameters, if X is not degenerate:^{P8}

$$E(Y | X) = m_Y + C_{Y,X} C_{X,X}^{-1} (X - m_X) \quad \text{Var}(Y | X) = C_{Y,Y} - C_{Y,X} C_{X,X}^{-1} C_{X,Y} \quad (1.85)$$

⚠ The linear assumption (1.83) (or (1.82)), which refer to the distribution of Y given X , do not imply any property about the distribution of X given Y .

We must distinguish the two following formulations:

^{P7}. Use total expectation, total variance, total covariance formulas.

^{P8}. We invert (1.84) to obtain H , m_W and $C_{W,W}$ in function of m_Y , $C_{Y,Y}$, $C_{X,Y}$:
$$\begin{cases} H = C_{Y,X} C_{X,X}^{-1} \\ m_W = m_Y - C_{Y,X} C_{X,X}^{-1} m_X \\ C_{W,W} = C_{Y,Y} - C_{Y,X} C_{X,X}^{-1} C_{X,Y} \end{cases}$$

Then, we insert the result into (1.83).

- $Y | X$ is driven by a linear model (described above);
- $X | Y$ is driven by a linear model (obtained by exchanging X and Y in the formulas above).

Exercise 22. Let's consider d_y thermometers. Each of them provides a measure y_i of the actual temperature x . The errors are supposed zero-mean, independent, with the same variance σ^2 .

We define $Y = \begin{bmatrix} y_1 \\ \vdots \\ y_{d_y} \end{bmatrix}$

Provide a linear model which fulfills these assumptions.

Exercise 23 (Training for estimation theory and Kalman filtering).

$Y | X$ is driven by a linear model described by H , m_w and $C_{w,w}$.

The mean and the variance of X are m_x and $C_{x,x}$.

Furthermore, we assume that $X | Y$ is driven by a linear model.

Give $E(X | Y)$ and $\text{Var}(X | Y)$ in function of H , m_w , $C_{w,w}$, m_x and $C_{x,x}$.

1.10.4 Multivariate normal distribution

A r.v. X which takes its value in \mathbb{R}^d is normally distributed (or Gauss-distributed) if, for all vector $v \in \mathbb{R}^d$, $v^\top X$ is normally distributed.²⁷

Then, necessarily, X has a mean m_x and a variance $C_{x,x}$.

A normal distribution is perfectly characterized by its mean and every root Σ_x of its variance $C_{x,x} = \Sigma_x \Sigma_x^\top$.

Every linear function of a normal r.v. gives a normal r.v.

All higher order cumulants are null.

The 3 formulations below are equivalent.

1. The pair (X, Y) is normally distributed.
2. The conditions below hold:
 - X is normally distributed;
 - $Y | X$ is normally distributed;
 - $Y | X$ is driven by a linear model.²⁸
3. There exists a matrix H such that X and $Y - HX$ are normally distributed, and independent.

• In the formulations 2 and 3, we can exchange X and Y .

If the pair (X, Y) is normally distributed:

- the marginal and conditional distributions are normal;
- uncorrelatedness of X and Y implies their independence (thus, the uncorrelatedness and the independence are equivalent).

• If the r.v. X and Y are normally distributed (each of both), and not independent:

- the joint and marginal distributions are not necessarily normal;
- the uncorrelatedness does not imply the independence.

1.10.5 Non degenerate case

If the variance $C_{x,x}$ is invertible, we can show that the PDF is, for all x :²⁹

$$p_x(x) = \frac{1}{\sqrt{\det(2\pi C_{x,x})}} \exp \left[-\frac{1}{2} (x - m_x)^\top C_{x,x}^{-1} (x - m_x) \right] \quad (1.86)$$

27. For all $v \in \mathbb{R}^d$, $v^\top X$ is driven by an univariate normal distribution.

28. The means, variances and covariance can be calculated thanks to formulas (1.84) and (1.85). We can deduce that if $g(\cdot, C)$ is the PDF of the centered normal distribution with variance C , for all (x, y) :

$$g(x - m_x, C_{x,x}) g(y - (Hx + m_w), C_{w,w}) = g \left(\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} m_x \\ Hm_x + m_w \end{bmatrix}, \begin{bmatrix} C_{x,x} & C_{x,x} H^\top \\ H C_{x,x} & H C_{x,x} H^\top + C_{w,w} \end{bmatrix} \right)$$

Conversely:

$$g \left(\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} m_x \\ m_y \end{bmatrix}, \begin{bmatrix} C_{x,x} & C_{x,y} \\ C_{y,x} & C_{y,y} \end{bmatrix} \right) = g(x - m_x, C_{x,x}) g(y - (m_y + C_{y,x} C_{x,x}^{-1} (x - m_x)), C_{y,y} - C_{y,x} C_{x,x}^{-1} C_{x,y})$$

29. See the degenerate case in [30].

- If X takes its value in \mathbb{R}^d , the r.v. $K^2 = (X - m_X)^T C_{X,X}^{-1} (X - m_X)$ is χ^2 -distributed with d degrees of freedom (page 11).

The smallest confidence domain at level $P_0 \in [0, 1]$ is the ellipsoid:

$$\{x \in \mathbb{R}^d \mid (x - m_X)^T C_{X,X}^{-1} (x - m_X) \leq F_K^{-1}(P_0)\}$$

1.10.6 Non degenerate bivariate case

This is a the $d = 2$ case. $K^2 = (X - m_X)^T C_{X,X}^{-1} (X - m_X)$ is χ^2 -distributed with 2 degrees of freedom (2 mean exponential distribution).

- The confidence domain at level P_0 is the ellipse:^{30 31}

$$\{x \in \mathbb{R}^2 \mid (x - m_X)^T C_{X,X}^{-1} (x - m_X) \leq -2 \log(1 - P_0)\}$$

◁▷ **Exercise 24.** X is a zero mean unit variance r.v., and $Y = \Sigma X + m$, where m and Σ have correct dimensions. Give the mean and the variance of Y .

- From the result of this exercise, we have a way to simulate a normal r.v. with known mean and variance by means of the random generators available in scientific calculation softwares.³²

◁▷ **Exercise 25.** With Matlab or Octave, simulate 100 realizations of a bivariate normal r.v. $\begin{bmatrix} X \\ Y \end{bmatrix}$ with zero mean, correlation coefficient 0.97, the first component with variance 4, the second one with variance 1.

Plot the confidence ellipse at 95%.

Plot the line $x \mapsto E(Y \mid X = x)$ (formula (1.85)).

Plot the line $y \mapsto E(X \mid Y = y)$.

1.11 Mixture distribution

A continuous r.v. is driven by a mixture distribution if its PDF is written as:

$$p_X(x) = \sum_{z \in \{\zeta_1, \dots, \zeta_{n_c}\}} \lambda_z f_z(x) \quad (1.87)$$

with:

- $\{\zeta_1, \dots, \zeta_{n_c}\}$ a labels set;
 - for all z , f_z a positive valued function with unit integral;
 - for all z , $\lambda_z \geq 0$, and $\sum_z \lambda_z = 1$.
- A mixture distribution can be used when the r.v. X show phenomenons from different clusters:
- f_z is the PDF of the r.v. in the cluster labelled z ;
 - λ_z is the probability of the cluster labelled z .

For example, if X represent the mark for an exam in which there were good and bad students, but no average one: $\{\zeta_1, \zeta_2\} = \{\text{"good student"}, \text{"bad student"}\}$.

- A mixture distribution can be used in problems not related to clustering ones; thus the cluster label is meaningless, we will use $\{\zeta_1, \dots, \zeta_{n_c}\} = \{1, \dots, n_c\}$.

Let's introduce the discrete r.v. Z which takes its value in $\{\zeta_1, \dots, \zeta_{n_c}\}$.

By means of the total probability law:

$$p_X(x) = \sum_{z \in \{\zeta_1, \dots, \zeta_{n_c}\}} \underbrace{\text{Prob}(Z = z)}_{\lambda_z} \underbrace{p_{X|Z}(x, z)}_{f_z(x)} \quad (1.88)$$

30. The boundary of this domain is written in polar coordinates, by means of every square root Σ_X of $C_{X,X}$:

$$\left\{ \sqrt{-2 \log(1 - P_0)} \Sigma_X \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix} + m_X \mid \theta \in [0, 2\pi] \right\}$$

31. Matlab. To plot the P_0 confidence ellipse (normal distribution with mean m and variance C).

```
t = linspace(0, 2*pi, 100);
X = sqrt(-2*log(1-P0))*chol(C, 'lower')*[cos(t); sin(t)] + m*ones(1, length(t));
plot(X(1,:), X(2,:))
```

32. Matlab. To generate a Gaussian n_t -sample, with mean m and variance C .

```
x = chol(C, 'lower')*randn(length(C), nr) + m*ones(1, nr);
```

A mixture distribution can be seen as the marginal distribution of the continuous-valued r.v. in a pair composed of a discrete-valued r.v. and a continuous-valued one; the joint distribution PDF is $p_{x,z}(x, z) = \lambda_z f_z(x)$.

T115 On the figure 1.4 the PDF of a Gaussian mixture is represented with full line together with $p_{x,z}(x, \zeta_1)$ and $p_{x,z}(x, \zeta_2)$ in dotted line.

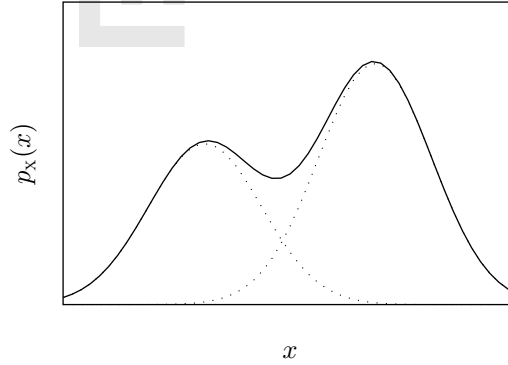


Figure 1.4: PDF of a Gaussian mixture

T116 1.12 Uncertainty propagation

Let x be a r.v. with known probability distribution.

Let $h : x \mapsto h(x)$ be a function defined on the support of x . A few ideas to determine the probability distribution of $h(x)$ are available in the appendix.

In this section, x has known mean and variance, and h is a non-linear function.

We want to approximate the mean and the variance of $h(x)$, and the covariance between $h(x)$ and x .

We want to define an **uncertainty propagation**, that is a transform UP such that:

$$\text{if } y = h(x) \text{ then } (m_y, C_{y,y}, C_{x,y}) \simeq \text{UP} (h, m_x, C_{x,x}) \quad (1.89)$$

T117 If h is differentiable, with Jacobian matrix $\frac{\partial h}{\partial x^\top}$, a natural solution consists in a linearization around the mean value of x :

$$h(x) \simeq h(m_x) + \frac{\partial h}{\partial x^\top}(m_x) (x - m_x)$$

We obtain, if $y = h(x)$:

$$\begin{aligned} m_y &\simeq h(m_x) \\ C_{y,y} &\simeq \left[\frac{\partial h}{\partial x^\top}(m_x) \right] C_{x,x} \left[\frac{\partial h}{\partial x^\top}(m_x) \right]^\top \\ C_{x,y} &\simeq C_{x,x} \left[\frac{\partial h}{\partial x^\top}(m_x) \right]^\top \end{aligned} \quad (1.90)$$

T118 Another solution, the **unscented transform** (UT), consists in the determination of a set of n_σ vectors ξ_q (the so-called σ -points) and weights ω_q , $1 \leq q \leq n_\sigma$ such that:

$$\sum_{q=1}^{n_\sigma} \omega_q = 1 \quad m_x = \sum_{q=1}^{n_\sigma} \omega_q \xi_q \quad C_{x,x} = \sum_{q=1}^{n_\sigma} \omega_q (\xi_q - m_x) (\xi_q - m_x)^\top \quad (1.91)$$

• Thus, if $y = h(x)$, $(m_y, C_{y,y}, C_{x,y}) \simeq \text{UT} (h, m_x, C_{x,x})$, that is:

$$\begin{aligned} m_y &\simeq \sum_{q=1}^{n_\sigma} \omega_q h(\xi_q) \text{ (noted } \bar{y} \text{ below)} \\ C_{y,y} &\simeq \sum_{q=1}^{n_\sigma} \omega_q (h(\xi_q) - \bar{y}) (h(\xi_q) - \bar{y})^\top \\ C_{x,y} &\simeq \sum_{q=1}^{n_\sigma} \omega_q (\xi_q - m_x) (h(\xi_q) - \bar{y})^\top \end{aligned} \quad (1.92)$$

We can easily check that, if h is linear, there is no approximation.

- †119 How to obtain the σ -points? Let Σ_x be a square root of $C_{x,x}$ ($C_{x,x} = \Sigma_x \Sigma_x^T$), and $\Sigma_x^{(q)}$ be the q th column of Σ_x , d_x be the size of x , $\lambda > -d_x$ be a tuning scale parameter; we obtain $n_\sigma = 2d_x + 1$ σ -points, with:

$$\begin{aligned} \omega_q &= \frac{1}{2(d_x + \lambda)} & \xi_q &= m_x + \sqrt{d_x + \lambda} \Sigma_x^{(q)} & 1 \leq q \leq d_x \\ \omega_{q+d_x} &= \frac{1}{2(d_x + \lambda)} & \xi_{q+d_x} &= m_x - \sqrt{d_x + \lambda} \Sigma_x^{(q)} & 1 \leq q \leq d_x \\ \omega_{2d_x+1} &= \frac{\lambda}{d_x + \lambda} & \xi_{2d_x+1} &= m_x \end{aligned} \quad (1.93)$$

The σ -points (but the central one) are distributed over the ellipsoid $\{x \mid (x - m_x)^T C_{x,x}^{-1} (x - m_x) = d_x + \lambda\}$.

- There is no convincing policy about the choice of the λ parameter [29]. The simplest solution is $\lambda = 0$, so that the central σ -point is not used; this is a cubature [2, 15]. If $\lambda < 0$, the weight of the central σ -point is negative, so that we can obtain a non positive definite variance $C_{y,y}$; a solution is to cancel the effect of the central σ -point in this variance calculation [21].

T120 Chapter 2

Parametric estimation

T121 The objective is to evaluate an unknown quantity $x^* \in \mathbb{X}$, the **parameter**, from an **observation** (or **data**) $y \in \mathbb{Y}$ which is linked to this parameter.

For example:

- x^* is the actual temperature, y is the measures given by several thermometers;
- x^* is the ship position (North, West), y is the azimuth of several stars.
- x^* is the state of the patient (has he a heart disease?), y is the biological signal (ECG);

• An **estimator** is a function \hat{x} which, for all observation, returns an estimate:

$$\begin{aligned} \hat{x}: \mathbb{Y} &\longrightarrow \mathbb{X} \\ y &\longmapsto \hat{x}(y) \end{aligned} \quad (2.1)$$

This definition is rather fuzzy, since an estimator is nothing but a statistic¹ which takes its value in \mathbb{X} . Intuitively, an estimator should give an estimate “close” to the actual value.

T122 A stochastic context is assumed: the observation y is the realization of a r.v. Y .

The process which produced this observation depends of a parameter X , random or not, which, in the experiment, took the unknown value x^* .

Thus, we have to make explicit the expression of the r.v. $\hat{x}(Y)$.

• The objective is now:

- to represent this stochastic link between the parameter X and the observation Y ,
- to define an estimator $\hat{x}(Y)$,
- to measure its performances (we have to make explicit the notion of “closeness” mentioned above).

T123 2.1 Likelihood, Maximum Likelihood

The problem formalization needs to define the conditional distribution of $Y | X$, obtained by physical insight, or by intuition.²

- With fixed x , the function $y \mapsto p_{Y|X}(y, x)$ is the observation PDF, assuming that X is x .
- With fixed y , the function $x \mapsto p_{Y|X}(y, x)$ is called the **likelihood**.

A value of the parameter is likely if the data is probable with this value.

The likelihood measures the adequacy of the parameter to the data.

• In the classical estimation theory,³ it is the unique formalization.

It permits to define the **Maximum Likelihood estimator (MLE)** estimator:

$$\hat{x}_{MLE}(Y) = \arg \max_{x \in \mathbb{X}} p_{Y|X}(Y, x) \quad (2.2)$$


• Other estimators which rely on a likelihood structure are possible, for exemple if $Y | X$ is driven by a linear model.

1. That is a function of the observation

2. The conditional distribution with PDF $y \mapsto p_{Y|X}(y, x^*)$ is the data generating process, but the actual value x^* is unknown.

3. R.A. Fisher, 1922.

◁▷ **Exercise 26.** The observation Y is uniformly distributed between 0 and x . We want to estimate x .

- Plot the PDF $y \mapsto p_{Y|X}(y, x)$ for a fixed parameter x , the likelihood $x \mapsto p_{Y|X}(y, x)$, for an observation y , and the function $(x, y) \mapsto p_{Y|X}(y, x)$ (perspective drawing). 
- Give $\hat{x}_{MLE}(Y)$.

2.2 Prior distribution, predictors

In the Bayesian estimation theory (from Thomas Bayes's name, 1702-1761, but this theory developed in the fifties), the complementary assumption is that, if the experiment is repeated, the parameter itself is modified, and is driven by a probability distribution, the **prior distribution**.

This distribution characterizes the prior information about the parameter, before any observation.

⦿ A **Bayesian estimator** makes a compromise between:

- the information brought by the data (by means of the likelihood),
- and this prior information.

The doctor diagnosis uses the biological examinations, but also the family medical history.

⦿ In the extreme, the medical history could lead the doctor not to ask for further examination.

Thus, it is natural to propose an evaluation of the parameter with no data, that is a **predictor**, which returns a prediction only from the prior distribution.

⦿ The **MMSE** predictor returns⁴ the mean value:

$$\tilde{x}_{MMSE} = E(X) \quad \text{prior mean} \quad (2.3)$$

MMSE stands for “**Minimum Mean Square Error**”; the significance will appear further.

⦿ The maximum *a priori* predictor (**MAP**) returns the most probable value:

$$\tilde{x}_{MAP} = \arg \max_{x \in \mathbb{X}} p_X(x) \quad \text{maximum } a \text{ priori} \quad (2.4)$$

⦿ But there are other policies. For example, if the parameter can be split into two components, $X = (X_C, X_D)$:⁵

$$\begin{aligned} \tilde{x}_D &= \arg \max_{x_D} p_{X_D}(x_D) && \text{marginal maximum } a \text{ priori} \\ \tilde{x}_C &= E(X_C | X_D = \tilde{x}_D) && \text{conditional prior mean} \end{aligned} \quad (2.5)$$

⦿ The estimator (or predictor) choice can be pragmatic (in some situations, it can be easier to compute a maximum rather than an expectation).

⦿ It can depend of the interpretation we want to give to the result:

- the birth rate is 1.9 children per woman (that is the mean value);
- if we encounter an unknown woman, we can expect that she has 2 children (that is the most probable value).

2.3 Posterior distribution, Bayesian estimators

We now observe the r.v. Y .

The **posterior distribution** is the distribution of the parameter X given the observation Y .

To build some estimators, we convert the predictors defined above in the prior context to the posterior context.

⦿ The **MMSE** estimator is defined by:

$$\hat{x}_{MMSE}(Y) = E(X | Y) \quad \text{posterior mean} \quad (2.6)$$

4. For numerical parameters.

5. Let's consider an urn which contains balls, some of them in balsa and the others in lead. The lead balls have all the same weight. We get ready to draw a ball from the urn. What can we expect on the category, which is a discrete r.v. x_d (balsa or lead), and on the weight, which is a continuous variable x_c , with the knowledge of the joint distribution of the pair (x_c, x_d) ? A “rational” predictor consists in choosing the maximum *a priori* for the category (that is the most probable category), then the prior mean given this category for the weight (that is the mean weight in the most probable category).

- The **Maximum a Posteriori** estimator (**MAP**) is defined by:

$$\hat{x}_{\text{MAP}}(Y) = \arg \max_{x \in \mathbb{X}} p_{X|Y}(x) \quad \text{maximum a posteriori} \quad (2.7)$$

- If the r.v. X can be split as $X = \begin{bmatrix} x_c \\ x_d \end{bmatrix}$, we can use the estimator below:

$$\begin{aligned} \hat{x}_d(Y) &= \arg \max_{x_d} p_{x_d|Y}(x_d) && \text{marginal maximum a posteriori} \\ \hat{x}_c(Y) &= E(x_c | x_d = \hat{x}_d(Y), Y) && \text{conditional posterior mean} \end{aligned} \quad (2.8)$$

- †129 The **LMMSE** (“Linear MMSE”) estimator relies on the second order characteristics of the joint distribution of (X, Y) :

$$\hat{x}_{\text{LMMSE}}(Y) = m_X + C_{X,Y} C_{Y,Y}^{-1} (Y - m_Y) \quad (2.9)$$

It will be rewritten page 33 when $Y | X$ is driven by a linear model.⁶

◁▷ **Exercise 27.** In this exercise, the problem is inverted, we look for an estimator $\hat{y}(x)$, for the simplified model $Y = HX + W$ with $E(W | X) = m_W$ (no condition on $\text{Var}(W | X)$).

- Write $\hat{y}_{\text{MMSE}}(x)$ and $\hat{y}_{\text{LMMSE}}(x)$.
- Prove that $\hat{y}_{\text{MMSE}}(x) = \hat{y}_{\text{LMMSE}}(x)$ if and only if this simplified model is fulfilled.

†130 2.4 How to obtain the posterior distribution?

In the previous exercise, we saw that a constrained posterior distribution can lead to a Bayesian estimator.

But, in general, we have:

- the likelihood $p_{Y|X}$, then the distribution of $Y | X$,
- and the prior distribution, that is the distribution of X .

It is equivalent to say that we have the distribution of the pair (X, Y) .

- By means of the Bayes law, we obtain the posterior distribution; for all (x, y) :

$$p_{X|Y}(x, y) = \frac{p_X(x) p_{Y|X}(y, x)}{p_Y(y)} \quad (2.10)$$

Since the denominator does not depend of x , we often write:

$$p_{X|Y}(x, y) \propto p_X(x) p_{Y|X}(y, x) \quad (2.11)$$

This formula is the basis one in Bayesian inference.

†131 2.5 How to propose a prior distribution?

This information depends of the interpretation we give to probabilities.

- In the **Frequentist interpretation**, a probability measures the frequency of appearance of an event in the assumption that we are able to repeat endlessly the experiment.
- In the **Bayesian interpretation**, probability measures the subjective belief that an event can occur.

- ▲ A prior distribution can be frequentist; for example, the prevalence of a disease is based on statistics. Do not make a confusion between “Bayesian estimation” (in which the parameter is equipped with a prior distribution) and “Bayesian interpretation” (which refers to the subjective signification of the probabilities).

- ▲ Is it possible that a prior distribution does not bring any information? This is the question of **Noninformative** distributions.

With a uniform prior, the MAP estimator gives back the MLE. Thus, a uniform distribution would be noninformative (Laplace proposal 200 years ago).

This is controversial [4] and leads to technical problems when the domain \mathbb{X} is not bounded (**improper** distributions). See [25] for more information.

- †132 From a pragmatic point of view, a Bayesian prior distribution allows more flexibility in Bayesian estimation.

The prior distribution is chosen in a parametric family:

6. $Y | X$ is driven by a linear model if $Y = HX + W$ with $\begin{cases} E(W | X) = m_W \\ \text{Var}(W | X) = C_{W,W} \end{cases}$

- which permits to intuitively set a confidence level,⁷
 - which is **conjugate** for the likelihood [11], this means that the posterior distribution belongs to the same family.⁸
- ☉ Thus, if the observation is a sample (Y_1, Y_2, \dots) of a r.v. Y , i.i.d. given the parameter, sequentially processed, for all x :

$$\begin{array}{ll}
 \text{prior} & p_x(x) \\
 \text{taking } Y_1 \text{ into account} & p_{x|Y_1}(x) \propto p_x(x) p_{Y|X}(Y_1, x) \\
 \text{taking } Y_2 \text{ into account} & p_{x|Y_1, Y_2}(x) \propto p_{x|Y_1}(x) p_{Y|X}(Y_2, x)
 \end{array} \quad (2.12)$$

The posterior distribution of one stage becomes the prior distribution of the next stage.

With a conjugate prior distribution for the likelihood, all stages are mathematically identical.

T133 2.6 Mean error analysis

We evaluate a parameter x where all the d_x components are numeric.

A predictor provides a prediction \tilde{x} .

From an observation Y , an estimator provides an estimate $\hat{x}(Y)$.

- ☉ The prediction is corrupted by the **prediction error** $\tilde{x} - x$.

The estimate is corrupted by the **estimation error** $\hat{x}(Y) - x$.

If a component of the error is positive, it is an overestimation.

T134 The performances of a predictor or an estimator can be measured by means of:⁹

- the **bias** (Bias, vector valued), that is the mean error, and the **error variance** (Errvar, positive definite matrix valued),
- or the **Mean Square Error** (MSE, positive real valued).

We define a norm $\|x\|_W = \sqrt{x^T W x}$, where W is a symmetric positive definite matrix W .

The MSE is the mean of the square norm of the error.

Necessarily, $MSE = \|\text{Bias}\|_W^2 + \text{trace}(W \text{Errvar})$

The bias, the error variance, the MSE are expressed:

- given the parameter lp in classical estimation;
- by letting (x, Y) jointly varying, in Bayesian estimation.

T135 2.7 Mean error, given the parameter

The **bias** of the estimator is the function from \mathbb{X} to \mathbb{R}^{d_x} which returns the error mean value:

$$\begin{aligned}
 \text{Bias}_x^{lp}(x) &= E(\hat{x}(Y) - x \mid x) \\
 &= E(\hat{x}(Y) \mid x) - x
 \end{aligned} \quad (2.13)$$

If the bias is uniformly null, the estimator is **unbiased**.

A positive bias component means a tendency to overestimate.

- ☉ The **variance** of the estimator is the function from \mathbb{X} to $\mathbb{R}^{d_x \times d_x}$ which returns the estimation error variance, that is a symmetric positive semi-definite matrix:

$$\begin{aligned}
 \text{Errvar}_x^{lp}(x) &= \text{Var}(\hat{x}(Y) - x \mid x) \\
 &= \text{Var}(\hat{x}(Y) \mid x)
 \end{aligned} \quad (2.14)$$

▲ Given the parameter, the estimation variance and the estimation error variance coincide.

T136 The bias measures in what extent the estimated value may vary from the actual value, on average. The variance measures the scattering of a large number of realizations.

For a given bias, it is natural to look for an estimator with low variance.

- ☉ But, in an estimation problem such that:

7. For a normal distribution, the mean is the assumed parameter, and higher is the variance, lower is the confidence in this assumed value.
 8. For example, if the prior distribution is normal, parameterized by its mean and its variance, the posterior distribution is normal, with a mean and a variance updated in function of the data.
 9. If the quantities exist.

- the likelihood is differentiable with respect to the parameter, with continuous differential,
- the support $S(Y | x)$ is independent of the parameter,

every estimator should respect the **Cramer-Rao inequality**, that is, for an unbiased estimator:¹⁰

$$\text{Errvar}_x^{\text{lp}}(x) \geq [\text{FI}(x)]^{-1} \quad (\text{for an unbiased estimator}) \quad (2.15)$$

where FI is the **Fisher information**, function which returns a positive semi-definite matrix defined by:¹¹

$$\text{FI}(x) = \text{Var} \left(\frac{\partial \ln p_{Y|x}}{\partial x}(Y, x) \mid x \right) \quad (2.16)$$

A high Fisher information means that a parameter variation implies a strong observation variation.

T137 An **efficient estimator** is an unbiased estimator which reaches this bound.

If an efficient estimator exists, it is unique, and it is the MLE (proof in appendix D.2).

A Unbiasedness and efficiency are not always reachable:

- an unbiased estimator does not necessarily exist;
- if some exist, there is not necessarily an efficient one among them (it is even a school case [18]);
- the MLE is not necessarily efficient.

A Unbiasedness and efficiency are arbitrary goals:

- they are not preserved by a non linear change of parameterization;¹²
- their existence depends of the parameterization;
- but the MLE is coherent with respect to a re-parameterization; this is the **invariance principle** [6].¹³

o The **MVUE** (“Minimum Variance Unbiased estimator”) may exist without being efficient (the variance does not reach the CRLB).

The **BLUE** (“Best Linear Unbiased estimator”) is a minimum variance unbiased estimator among the linear estimators.

T138 The **Mean square error (MSE)** is a positive scalar valued function which quantifies the estimator quality:¹⁴

$$\begin{aligned} \text{MSE}_x^{\text{lp}}(x) &= E \left(\|\hat{x}(Y) - x\|_W^2 \mid x \right) \\ &= \left\| \text{Bias}_x^{\text{lp}}(x) \right\|_W^2 + \text{trace} (W \text{Errvar}_x^{\text{lp}}(x)) \end{aligned} \quad (2.17)$$

If the goal is to minimize the MSE, there are usually neither unbiasedness nor minimal variance, but a compromise.

T1A We have postulated the existence and the unicity of the MLE.

But we can face some problems.

- The likelihood has no maximum for some observations.
- There is a countable set of global maxima.
- There is a non countable set of global maxima, due to an over-parameterization.
- The criterion has to be optimized: with a local optimization method, this can lead to a local maximum of the likelihood.

T140 **<▷ Exercise 28.** The observation Y is uniformly distributed between 0 and x . We estimate x (see exercise 26).

- Give the bias, the variance, the MSE of the estimator $\hat{x}(Y) = \alpha Y + \beta$.
- Write the bias, the variance, the MSE of:
 - the MLE,
 - the BLUE,
 - the estimator $\hat{x}(Y) = \alpha Y$ which minimizes the MSE.

<▷ Exercise 29. The observation Y is uniformly distributed between 0 and $1/x$. We estimate x .

- Give the MLE.
- Does it exist an unbiased estimator?

10. The Cramer-Rao inequality can be extended to the biased case (proof in appendix D.1).

11. The gradient $\frac{\partial}{\partial x}$ is the gradient with respect to the 2nd variable only, that is the parameter.

12. E.g., the square root of an unbiased estimation of the variance is not an unbiased estimation of the standard deviation.

13. E.g., the MLE of a variance is the square of the MLE of a standard deviation.

14. Reminder: W is a positive definite matrix and $\|x\|_W = \sqrt{x^T W x}$.

2.8 Mean error, Bayesian point of view

A priori

The bias, the error variance, the mean square error of a predictor \tilde{x} are defined below; they are rewritten by means of the MMSE predictor:¹⁵

$$\begin{aligned} \text{Bias}_{\tilde{x}} &= E(\tilde{x} - x) = \tilde{x} - \tilde{x}_{\text{MMSE}} \\ \text{Errvar}_{\tilde{x}} &= \text{Var}(\tilde{x} - x) = \text{Var}(x) \\ \text{MSE}_{\tilde{x}} &= E(\|\tilde{x} - x\|_W^2) = \|\tilde{x} - \tilde{x}_{\text{MMSE}}\|_W^2 + \text{trace}(W \text{Var}(x)) \end{aligned} \quad (2.18)$$

- The error variance does not depend of the predictor.
Only the MMSE predictor nullifies the bias and minimizes the MSE.

$$\tilde{x}_{\text{MMSE}} = E(x) \quad \begin{cases} \text{Bias}_{\tilde{x}_{\text{MMSE}}} = 0_{d_x} \\ \text{Errvar}_{\tilde{x}_{\text{MMSE}}} = \text{Var}(x) \end{cases} \quad (2.19)$$

A posteriori

The bias, the error variance, the MSE of a Bayesian estimator are defined below; they are rewritten by means of the MMSE estimator:^{P1}

$$\begin{aligned} \text{Bias}_{\hat{x}} &= E(\hat{x}(Y) - x) = E(\hat{x}(Y) - \hat{x}_{\text{MMSE}}(Y)) \\ \text{Errvar}_{\hat{x}} &= \text{Var}(\hat{x}(Y) - x) = \text{Var}(\hat{x}(Y) - \hat{x}_{\text{MMSE}}(Y)) + E(\text{Var}(x | Y)) \\ \text{MSE}_{\hat{x}} &= E(\|\hat{x}(Y) - x\|_W^2) = E(\|\hat{x}(Y) - \hat{x}_{\text{MMSE}}(Y)\|_W^2) + \text{trace}(W E(\text{Var}(x | Y))) \end{aligned} \quad (2.20)$$

- Only the MMSE estimator nullify the bias, and minimizes the error variance, minimizes the MSE.^{16 P2}

$$\hat{x}_{\text{MMSE}}(Y) = E(x | Y) \quad \begin{cases} \text{Bias}_{\hat{x}_{\text{MMSE}}} = 0_{d_x} \\ \text{Errvar}_{\hat{x}_{\text{MMSE}}} = E(\text{Var}(x | Y)) \end{cases} \quad (2.21)$$

⚠ Do not muddle the estimation variance and the estimation error variance!

¶ If the observation is numeric with d_y components, we can decide to look for:

- an unbiased estimator with minimal error variance,
- or an estimator with minimal MSE,

among the linear estimators.

- Both objectives lead to the LMMSE estimator: ^{P3}

$$\hat{x}_{\text{LMMSE}}(Y) = m_x + C_{x,Y} C_{Y,Y}^{-1} (Y - m_Y) \quad \begin{cases} \text{Bias}_{\hat{x}_{\text{LMMSE}}} = 0_{d_x} \\ \text{Errvar}_{\hat{x}_{\text{LMMSE}}} = C_{x,x} - C_{x,Y} C_{Y,Y}^{-1} C_{Y,x} \end{cases} \quad (2.22)$$

If $x | Y$ is driven by a linear model, the LMMSE estimator is the MMSE one.

15. Reminder: W is a positive definite matrix, $\|x\|_W = \sqrt{x^T W x}$, and $\text{MSE} = \|\text{Bias}\|_W^2 + \text{trace}(W \text{Errvar})$.

P1. We express the bias, the error variance, the MSE given the observation of the MMSE estimator:

$$\begin{aligned} E(\hat{x}(Y) - x | Y) &= \hat{x}(Y) - \hat{x}_{\text{MMSE}}(Y) \\ \text{Var}(\hat{x}(Y) - x | Y) &= \text{Var}(x | Y) \\ E(\|\hat{x}(Y) - x\|_W^2 | Y) &= \|\hat{x}(Y) - \hat{x}_{\text{MMSE}}(Y)\|_W^2 + \text{trace}(W \text{Var}(x | Y)) \end{aligned}$$

The we use total expectation formula and the total variance formula.

16. Whatever the weighting matrix W is.

P2. The variance is minimized if $\text{Var}(\hat{x}(Y) - \hat{x}_{\text{MMSE}}(Y))$ is null, that is if the r.v. $\hat{x}(Y) - \hat{x}_{\text{MMSE}}(Y)$ is (almost surely) constant; among minimal variance estimators, none but the MMSE estimator nullify the bias.

Likewise, the MSE is minimized if $E(\|\hat{x}(Y) - \hat{x}_{\text{MMSE}}(Y)\|_W^2)$ is null, that is if the r.v. $\hat{x}(Y) - \hat{x}_{\text{MMSE}}(Y)$ is (almost surely) null. Thus, none but the MMSE estimator minimizes the MSE.

P3. Let's consider a linear estimator: $\hat{x}(Y) = m_x + B + A(Y - m_Y)$

By means of standard calculations, we obtain:

$$\begin{aligned} \text{Bias}_{\hat{x}} &= B \\ \text{Errvar}_{\hat{x}} &= C_{x,x} - C_{x,Y} C_{Y,Y}^{-1} C_{Y,x} + (A - C_{x,Y} C_{Y,Y}^{-1}) C_{Y,Y} (A - C_{x,Y} C_{Y,Y}^{-1})^T \\ \text{MSE}_{\hat{x}} &= \text{trace}[C_{x,x} - C_{x,Y} C_{Y,Y}^{-1} C_{Y,x}] + B^T B + \text{trace}[(A - C_{x,Y} C_{Y,Y}^{-1}) C_{Y,Y} (A - C_{x,Y} C_{Y,Y}^{-1})^T] \end{aligned}$$

In the variance formula, the last term is necessarily positive semi-definite, and is null only if $A = C_{x,Y} C_{Y,Y}^{-1}$. To cancel the bias, one takes $B = 0_d$. We obtain the same result by searching the linear estimator which minimizes the MSE.

T144 From prior to posterior

The MMSE predictor, the LMMSE estimator and the MMSE estimator are all unbiased. The error variances are sorted in decreasing order below:¹⁷

$$\begin{array}{ll} \text{MMSE predictor} & \text{Errvar}_{\hat{x}_{\text{MMSE}}} = C_{x,x} \\ \text{LMMSE estimator} & \text{Errvar}_{\hat{x}_{\text{LMMSE}}} = C_{x,x} - C_{x,y} C_{y,y}^{-1} C_{y,x} \\ \text{MMSE estimator} & \text{Errvar}_{\hat{x}_{\text{MMSE}}} = C_{x,x} - \text{Var}(E(x | y)) \end{array} \quad (2.23)$$

⊕ Thus, with respect to the MMSE predictor:

- the LMMSE estimator decreases the variance only if x and y are correlated,
- the MMSE estimator decreases the variance only if $E(x | y)$ is not uniform.

T145 2.9 Linear model case

We assume that $y | x$ is driven by a linear model (page 22); there exist a known matrix H and a r.v. w with known mean m_w and variance $C_{w,w}$ such that:

$$y = Hx + w \text{ with } \begin{cases} E(w | x) = m_w \\ \text{Var}(w | x) = C_{w,w} \end{cases} \quad (2.24)$$

We can derive the optimal linear estimators.

T146 The BLUE is written as:^{P4}

$$\hat{x}_{\text{BLUE}}(y) = (H^T C_{w,w}^{-1} H)^{-1} H^T C_{w,w}^{-1} (y - m_w) \quad \begin{cases} \text{Bias}_{\hat{x}_{\text{BLUE}}}^p(x) = 0_{d_x} \\ \text{Errvar}_{\hat{x}_{\text{BLUE}}}^p(x) = (H^T C_{w,w}^{-1} H)^{-1} \end{cases} \quad (2.25)$$

It minimizes the criterion $(y - Hx)^T C_{w,w}^{-1} (y - Hx)$ with respect to x .¹⁸

⊕ If x has a mean m_x and a variance $C_{x,x}$, the LMMSE estimator is written as¹⁹

$$\begin{aligned} \hat{x}_{\text{LMMSE}}(y) &= m_x + C_{x,x} H^T (H C_{x,x} H^T + C_{w,w})^{-1} (y - H m_x - m_w) \\ &\begin{cases} \text{Bias}_{\hat{x}_{\text{LMMSE}}} = 0_{d_x} \\ \text{Errvar}_{\hat{x}_{\text{LMMSE}}} = C_{x,x} - C_{x,x} H^T (H C_{x,x} H^T + C_{w,w})^{-1} H C_{x,x} \end{cases} \end{aligned} \quad (2.26)$$

T147 Can we make a link between the numerous estimators?

- If w is normally distributed and independent of x , the BLUE is the MLE, and is efficient.^{P5}

17. The MMSE estimator is, by construction, the estimator with the lowest error variance. We can confirm that the LMMSE estimator error variance is indeed greater, since the Schur complement of $\text{Var}(E(x | y))$ in $\text{Var}\left(\begin{bmatrix} x \\ y \end{bmatrix}\right)$ is positive definite (we use also the result of the exercise 19, page 19).

P4. Let's define: $A_{\text{opt}} = (H^T C_{w,w}^{-1} H)^{-1} H^T C_{w,w}^{-1}$. Let's use a linear estimator: $\hat{x}(y) = b + A(y - m_w)$.
Remarking that $A_{\text{opt}} H = I_{d_x}$, standard calculations lead to the bias and the variance given the parameter:

$$\text{Bias}_{\hat{x}}^p(x) = b + (A H - I_{d_x}) x$$

$$\text{Errvar}_{\hat{x}}^p(x) = A C_{w,w} A^T = (A - A_{\text{opt}}) C_{w,w} (A - A_{\text{opt}})^T - A_{\text{opt}} C_{w,w} A_{\text{opt}}^T + A C_{w,w} A_{\text{opt}}^T + A_{\text{opt}} C_{w,w} A^T$$

To obtain the unbiasedness, b must be null, and A must fulfill the constraint $I_{d_x} = A H$; we easily check that, for such A , $A C_{w,w} A_{\text{opt}}^T = (H^T C_{w,w}^{-1} H)^{-1}$; thus, the variance estimator is $(A - A_{\text{opt}}) C_{w,w} (A - A_{\text{opt}})^T + A_{\text{opt}} C_{w,w} A_{\text{opt}}^T$, necessarily minimal for $A = A_{\text{opt}}$.

18. This is the **Generalized Least Squares** estimator (GLS). The **Ordinary Least Squares** estimator (OLS) is $\hat{x}_{\text{OLS}}(y) = (H^T H)^{-1} H^T (y - m_w)$

19. This is a re-writing of (2.22), by means of the transform (1.84):

$$\begin{cases} m_y = H m_x + m_w \\ C_{y,y} = H C_{x,x} H^T + C_{w,w} \\ C_{x,y} = C_{x,x} H^T \end{cases} \quad \text{and} \quad \begin{cases} \hat{x}_{\text{LMMSE}}(y) = m_x + C_{x,x} C_{y,y}^{-1} (y - m_y) \\ \text{Errvar}_{\hat{x}_{\text{LMMSE}}} = C_{x,x} - C_{x,x} C_{y,y}^{-1} C_{y,x} \end{cases}$$

P5. Let's write the likelihood; for all (x, y) :

$$p_{y|x}(y, x) = p_w(y - Hx) = \frac{1}{\sqrt{\det(2\pi C_{w,w})}} \exp\left[-\frac{1}{2} (y - Hx - m_w)^T C_{w,w}^{-1} (y - Hx - m_w)\right]$$

We obtain the solution by cancelling the log-likelihood gradient which is, for all (x, y) :

$$\frac{\partial \ln p_{y|x}}{\partial x}(y, x) = H^T C_{w,w}^{-1} (y - Hx - m_w)$$

- If x and w are normally distributed and independent, the LMMSE estimator is the MMSE one.²⁰
- If $C_{w,w}$ is invertible, by means of the Woodbury matrix inversion lemma, the LMMSE estimator is written as:

$$\hat{x}_{\text{LMMSE}}(Y) = (C_{x,x}^{-1} + H^T C_{w,w}^{-1} H)^{-1} [C_{x,x}^{-1} m_x + H^T C_{w,w}^{-1} (Y - m_w)]$$

$$\begin{cases} \text{Bias}_{\hat{x}_{\text{LMMSE}}} = 0_{d_x} \\ \text{Errvar}_{\hat{x}_{\text{LMMSE}}} = (C_{x,x}^{-1} + H^T C_{w,w}^{-1} H)^{-1} \end{cases} \quad (2.27)$$

With a weak prior information, that is $C_{x,x}^{-1}$ almost null, we get back the BLUE.²¹

If the observation is highly noisy, that is $C_{w,w}^{-1}$ almost null, we get back the MMSE predictor.

- ✎ **Exercise 30.** Let's consider d_y thermometers. Each of them provides a measure Y_i of the actual temperature x . Given x , the errors are supposed zero-mean, independent, with the same variance σ^2 .
- write the BLUE and its variance.
 - What is the criterion which is minimized by this solution?

2.10 Probability error analysis

We are interested in the estimation of a discrete valued parameter.

In an From an observation Y , an estimator provides an estimate $\hat{x}(Y)$.

- If $\hat{x}(Y) \neq x$, we make an error.

The performances of an estimator can be measured by means of the error probability.

- ✎ **Given the parameter**, the error probability is the function from \mathbb{X} to $[0, 1]$ defined as:

$$\begin{aligned} \text{Errprob}_x^p(x) &= \text{Prob}(\hat{x}(Y) \neq x \mid x) \\ &= 1 - \text{Prob}(\hat{x}(Y) = x \mid x) \\ &= 1 - \int_{\{y \in \mathbb{Y} \mid \hat{x}(y) = x\}} p_{Y|x}(y) \, dy \end{aligned} \quad (2.28)$$

For one value of the parameter, this probability is null (so minimal), if the integration domain is the support of the conditional distribution.

- Thus, it is possible to nullify this error probability for all possible values of the parameter only if the supports $S(Y \mid x = x)$, $x \in \mathbb{X}$ form a partition of $S(Y)$.

The MLE provides the unique x such that $p_{Y|x}(y, x)$ is not zero, and nullify the error probability.

Nevertheless, in the general case, a compromise is necessary.

- **From a Bayesian point of view**, the error probability is a number in $[0, 1]$ defined as:

$$\text{Errprob}_{\hat{x}} = \text{Prob}(\hat{x}(Y) \neq x) \quad (2.29)$$

The MAP estimator minimizes the error probability.^{P622}

20. More generally, it is sufficient that $x \mid Y$ is driven by a linear model.

21. Since the BLUE is unbiased with uniform variance given the parameter, we obtain, by means of the total expectation and total variance formulae, without assumption about the prior distribution, that, for this estimator: $\text{Bias}_{\hat{x}_{\text{BLUE}}} = 0$ and $\text{Errvar}_{\hat{x}_{\text{BLUE}}} = (H^T C_{w,w}^{-1} H)^{-1}$

P6. We express the error probability given the observation:

$$\begin{aligned} \text{Errprob}_x^o(Y) &= \text{Prob}(\hat{x}(Y) \neq x \mid Y) \\ &= 1 - \text{Prob}(\hat{x}(Y) = x \mid Y) \end{aligned} \quad (2.30)$$

Remind that $\hat{x}_{\text{MAP}}(Y) = \arg \max_{x \in \mathbb{X}} \text{Prob}(x = x \mid Y)$.

Thus, the MAP estimator minimizes this error probability. From the total probability law, $\text{Errprob}_{\hat{x}} = E(\text{Errprob}_{\hat{x}}^o(Y))$.

Thus, the MAP estimator minimizes this probability too.

22. In classification problems, this definition of the error probability is generally used.

2.11 Practical remarks

In numerous actual cases, the observation is a sample $\vec{Y}_{n_r} = (Y_1, \dots, Y_{n_r})$ of a r.v. Y such that:

- the distribution of Y the parameter to estimate X has a known form;
- The sample is assumed i.i.d. given the unknown parameter.

The estimator corresponds to the definition of $\hat{x}(\vec{Y}_{n_r})$.

- The likelihood is, for all $x \in S(X)$:

$$p_{\vec{Y}_{n_r}|X}(\vec{Y}_{n_r}, x) = \prod_{k=1}^{n_r} p_{Y|X}(Y_k, x) \quad (2.31)$$

- Since the log function is increasing, the likelihood maximization is equivalent to the log-likelihood maximization:

$$\hat{x}_{MLE}(\vec{Y}_{n_r}) = \arg \max_x \ell_{\vec{Y}_{n_r}}(x) \quad \text{with} \quad \ell_{\vec{Y}_{n_r}}(x) = \sum_{k=1}^{n_r} \ln p_{Y|X}(Y_k, x) \quad (2.32)$$

This trick simplifies the calculation in general.

⌊> **Exercise 31.** We want to check if a coin is correctly balanced. X is the probability to obtain “tail”. We throw n_r times the coin. For n from 1 to n_r , Y_n takes the value 1 if we obtain “tail”, and takes the value 0 if we obtain “head”.

Write the MLE of X from a sample of Y with size n_r .

⌊> **Exercise 32.** The r.v. Y takes the value $c \in \{1, \dots, n_c\}$ with the probability x_c . Write the MLE of $X = (X_1, \dots, X_{n_c})$ from a sample of Y with size n_r .

Beware! The problem has the constraint $\sum_{c=1}^{n_c} x_c = 1$, and can be easily solved using Lagrange multipliers technique.

2.12 Summary

The estimation of an unknown parameter X from an observation Y is formalized by:

- the distribution of $Y | X$ in the classical estimation.
- the distribution of $X | Y$ in the Bayesian estimation (in practice, the distribution of (X, Y)).

This problem reduces to an optimization one (MLE, MAP) or an integration one (MMSE).

- The LMMSE estimator can be written by means of the second order properties $m_Y, C_{Y,Y}, m_X, C_{X,X}, C_{X,Y}$.
- A linear model is an assumption on the conditional mean (assumed linear) and the conditional variance (assumed uniform).

if $X | Y$ is driven by a linear model, the LMMSE estimator is the MMSE one.

If $Y | X$ is driven by a linear model, we can write the BLUE and re-write the LMMSE estimator.

- The linear assumption is a school case; even if the formula $Y = HX + w$ is fulfilled, the variance of w usually must be estimated (see page 81). But, in general, there is no explicit solution. The optimization problem or the integration one has to be solved numerically.

T155 Chapter 3

Markov property

T156 3.1 Stochastic processes: a short reminder

A general introduction to stochastic processes is available page 67.

A **stochastic process** (or **random signal**) x is a function of time and chance.

- In the discrete time case, for all $n \in \mathbb{Z}$, $x[n]$ is a r.v., a stochastic process is also called a **time series**
- In the continuous time case, for all $t \in \mathbb{R}$, $x(t)$ is a r.v.

Every realization of a stochastic process is called a **trajectory**, or a **sample path**.

- ⊕ The distribution of the random process corresponds to the distribution of all n_t -tuple $(x(t_1), \dots, x(t_{n_t}))$, for all number of times n_t and for all distinct times (t_1, \dots, t_{n_t}) .

- ⊕ A stochastic process is **independent** if the r.v. of all n_t -tuple are mutually independent.¹

Such a process is completely unpredictable, since the knowledge of the trajectory in some times does not bring any information on the signal value at another time.

- ⊕ A process is **independent and identically distributed** (i.i.d.) if it is independent and if the distribution of $x(t)$ does not depend of t .

T157 A process is **white** if it exists a positive semi-definite matrix Q such that, for all (t_1, t_2) :

$$\text{Cov}(x(t_2), x(t_1)) = Q \delta(t_2 - t_1)$$

where δ is the Dirac delta function (continuous time case) or the Kronecker delta (discrete time case).

Q is the **power spectral density**, or **power spectrum**.²

- ▲ If x is a discrete time signal, its variance is Q .

If x is a continuous time signal, its variance is infinite.

- ⊕ An i.i.d. process is white.
- ⊕ An independent stochastic process is a particular (and degenerate) Markov chain.

T158 3.2 Markov process

A stochastic process x is a **Markov process** (or a **Markov chain** in the discrete time case) if, given the process at the current time, past and future of the process are independent.

For all increasing sequence of times $(t_{-n_{\text{past}}}, \dots, t_{-1}, t_0, t_1, \dots, t_{n_{\text{future}}})$:

$$\underbrace{(x(t_{-n_{\text{past}}}), \dots, x(t_{-1}))}_{\text{past}} \perp \underbrace{(x(t_1), \dots, x(t_{n_{\text{future}}}))}_{\text{future}} \mid \underbrace{x(t_0)}_{\text{current}} \quad (3.1)$$

- ⊕ The r.v. $x(t)$ (or the value it takes) is called the **state** of the process at time t .
The current state trapped all the information about the process future contained in the process past.
The path who led to this state does not matter.

1. That is $p_{x(t_1), \dots, x(t_{n_t})} = \prod_{n=1}^{n_t} p_{x(t_n)}$.

2. In general, the power spectrum is a function of frequency, constant for a white process.

- It is sufficient that the property (3.1) holds for $n_{\text{future}} = 1$ to hold for any n_{future} ; thus, x is Markovian if and only if, for all increasing sequence $(t_{-n_{\text{past}}}, \dots, t_{-1}, t_0, t_1)$:^{P1}

$$\underbrace{(x(t_{-n_{\text{past}}}), \dots, x(t_{-1}))}_{\text{past}} \perp\!\!\!\perp \underbrace{x(t_1)}_{\text{future}} \mid \underbrace{x(t_0)}_{\text{current}} \quad (3.2)$$

Intuitively, a Markov process is completely characterized by:

- the distribution of $x(t_{\min})$, where t_{\min} is the initial time, from which we start to observe the process;
 - the distribution of $x(t_1) \mid x(t_0)$ for all t_0 , where t_1 is the close future of t_0 .
- In the discrete time case, we choose $t_{\min} = 1$. The process is described by:
- the **initial distribution**, that is the distribution of $x[1]$;
 - the **transition distribution**, that is the distribution of $x[n+1] \mid x[n]$, for all $n \geq 1$.
- In the continuous time case,³ we choose $t_{\min} = 0$. The process is described by:
- the **initial distribution**, that is the distribution of $x(0)$;
 - the **transition distribution**, that is the distribution of $x(t+h) \mid x(t)$ when $h \rightarrow 0$, for all $t \geq 0$.
- For a continuous valued state, this can be interpreted as the expression of the PDF of the derivative $\dot{x}(t)$. This interpretation will be used in this book, but we should have in mind that a strong extension of the notions of derivative and integration (for example, the Itô calculus) is needed for a satisfactory mathematical theory.
- If the transition distribution does not depend of n or t , the process is **homogeneous**.

3.3 Examples

Random walk (discrete time, continuous or discrete state)

This model can be used, for example, for the search of lost bodies in the sea.

There exist an i.i.d. random sequence $v = (v[n])_{n \geq 1}$, independent of $x[1]$, such that, for all $n \geq 1$:

$$x[n+1] = x[n] + v[n] \quad (3.3)$$

v is called the **increments** sequence.

- x is a Markov chain for which the transition distribution is, for all $n \geq 1$, and for all (x, x^+) :^{P2}

$$p_{x[n+1] \mid x[n]}(x^+, x) = p_{v[n]}(x^+ - x) \quad (3.4)$$

A random walk corresponds to the cumulative sum of an independent sequence; for all $n \geq 1$:

$$x[n] = x[1] + \sum_{k=1}^{n-1} v[k]$$

P1. By induction, by means of the formula (1.66):

$$p_{x(t_{n_{\text{future}}}), \dots, x(t_1) \mid x(t_0), x(t_{-1}), \dots, x(t_{-n_{\text{past}}})} = \prod_{n=0}^{n_{\text{future}}-1} p_{x(t_{n+1}) \mid x(t_n), \dots, x(t_0), \dots, x(t_{-n_{\text{past}}})}$$

3. If the state sample paths are continuous (in the usual mathematical meaning), the process is a **diffusion process**; if the state sample paths are piecewise constant, the process is a **jump process**.

P2. The event $\{x[n] = x[n], \dots, x[2] = x[2], x[1] = x[1]\}$ is the event $\{v[n-1] = x[n] - x[n-1], \dots, v[1] = x[2] - x[1], x[1] = v[1]\}$; thus: $p_{x[n+1] \mid x[n], \dots, x[1]}(x[n+1], x[n], \dots, x[1]) = p_{x[n+1] - x[n] \mid x[n], \dots, x[1]}(x[n+1] - x[n], x[n], \dots, x[2], x[1]) = p_{v[n] \mid v[n-1], \dots, v[1], x[1]}(x[n+1] - x[n], x[n] - x[n-1], \dots, x[2] - x[1], x[1]) = p_{v[n]}(x[n+1] - x[n])$

Conversely, if x is a Markov chain such that $p_{x[n+1] \mid x[n]}(x^+, x)$ depends only of $x^+ - x$. We define $v[n] = x[n+1] - x[n]$. The event $\{v[n-1] = v[n-1], \dots, v[1] = v[1], x[1] = x[1]\}$ is $\{x[n] = x[1] + \sum_{k=1}^{n-1} v[k], \dots, x[2] = x[1] + v[2], x[1] = x[1]\}$; thus: $p_{v[n] \mid v[n-1], \dots, v[1], x[1]}(v[n], v[n-1], \dots, v[1], x[1]) = p_{x[n+1] - x[n] \mid x[n], \dots, x[2], x[1]}(v[n], x[1] + \sum_{k=1}^{n-1} v[k], \dots, x[1] + v[2], x[1]) = p_{x[n+1] \mid x[n], \dots, x[2], x[1]}(x[1] + \sum_{k=1}^n v[k], x[1] + \sum_{k=1}^{n-1} v[k], \dots, x[1] + v[2], x[1]) = p_{x[n+1] \mid x[n]}(x[1] + \sum_{k=1}^n v[k], x[1] + \sum_{k=1}^{n-1} v[k])$ depends only of $v[n]$. The sequence v is independent (along time), is independent of $x[1]$.

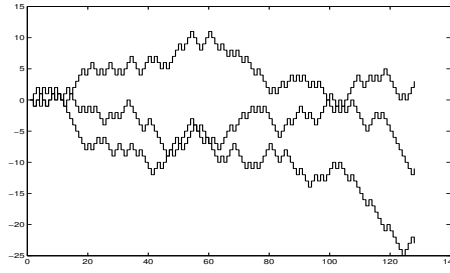


Figure 3.1: Three sample paths of a random walk

✎ With the complementary assumptions that:

- the initial state $x[1]$ is deterministic and zero-valued,
- v is zero-mean, with variance Q ,

then, for all (n, n') :

$$\begin{aligned} E(x[n]) &= 0 \\ \text{Var}(x[n]) &= Q(n-1) \\ \text{Cov}(x[n], x[n']) &= Q \min(n-1, n'-1) \end{aligned}$$

¶162 The figure 3.1 represents three trajectories of a random walk whose jumps take the value ± 1 with equiprobability. We can extend to the case where the initial state is not zero and where the jumps are not zero-mean, as in the **gambler's ruin** problem.

¶163 **Wiener process (continuous time, continuous state)**

It is the continuous time version of the random walk.

It is also called **Brownian motion**, since the botanist Robert Brown described such phenomenons in the motion of particles in water.

It is used in mathematical finance.

✎ There exist an i.i.d. random signal $v : t \geq 0 \mapsto v(t)$, independent of $x(0)$, such that, for all $t \geq 0$:

$$\dot{x}(t) = v(t) \quad (3.5)$$

Thus, for all $t \geq 0$:

$$x(t) = x(0) + \int_0^t v(\tau) d\tau$$

¶164 With the complementary assumptions that:

- the initial state $x(0)$ is deterministic and zero-valued,
- v is zero-mean, with power spectral density Q ,

then, for all (t, t') :⁴

$$\begin{aligned} E(x(t)) &= 0 \\ \text{Var}(x(t)) &= Qt \\ \text{Cov}(x(t), x(t')) &= Q \min(t, t') \end{aligned}$$

Strictly speaking, this formula, together with the Gaussian assumption, should be used to define the Wiener process. Thus, we can show that the sample paths are continuous, but they are not differentiable, at every time. The definition $\dot{x} = v$ is easy to use, but implies a strong extension of the notion of derivative.

¶165 **Poisson process (continuous time, discrete state)**

It is a counting model:

- number of persons arriving in a queue (in queuing theory),
- number of failures of an apparatus since its first putting into service (in reliability theory).

4. Let's define $\rho(x, t) = p_{x(t)}(x)$. assumed Gaussian, $\rho(x, t) = \frac{1}{\sqrt{2\pi t Q}} e^{-\frac{1}{2t} x^T Q^{-1} x}$, which fulfills $\frac{\partial \rho}{\partial t} = \frac{1}{2} \text{trace } Q \frac{\partial^2 \rho}{\partial x \partial x^T}$; if $Q = q I_{d_x}$ we obtain the diffusion equation $\frac{\partial \rho}{\partial t} = \frac{q}{2} \text{trace} \frac{\partial^2 \rho}{\partial x \partial x^T}$ (where $\text{trace} \frac{\partial^2 \rho}{\partial x \partial x^T}$) this trace is the Laplacian of ρ .

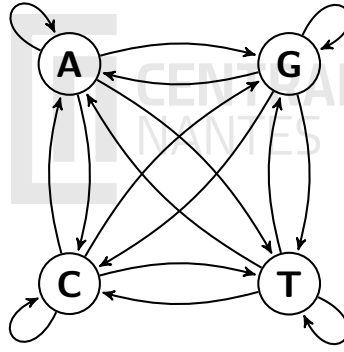


Figure 3.2: State graph of a finite state space Markov chain [10]

- The process takes its value in \mathbb{N} . The sample paths are initially zero valued, and increasing. The transition PMF is, for $h > 0$ small enough, for all $t \geq 0$, for all (x^+, x) :

$$\text{Prob}\left(x(t+h) = x^+ \mid x(t) = x\right) = \begin{cases} 0 & \text{if } x^+ < x & (\text{the process is increasing}) \\ 0 & \text{if } x^+ > x+1 & (\text{no simultaneous arrivals}) \\ \lambda h & \text{if } x^+ = x+1 & (\lambda \text{ is the process intensity}) \\ 1 - \lambda h & \text{if } x^+ = x \end{cases} \quad (3.6)$$

- With these natural assumptions, $x(t)$ is Poisson-distributed; for all $x \in \mathbb{N}$:

$$\text{Prob}(x(t) = x) = \exp(-\lambda t) \frac{(\lambda t)^x}{x!} \quad (3.7)$$

T166 ◀▷ **Exercise 33** (The assumption (3.6) implies (3.7)). Let x be a Poisson process.

Reminder: the solution of the differential equation $\dot{g}(t) = -\lambda g(t) + f(t)$ is $g(t) = e^{-\lambda t} g(0) + \int_0^t e^{-\lambda(t-\tau)} f(\tau) d\tau$.

- By means of the total probability law, write $p_{x(t+h)}$ in function of $p_{x(t)}$.
- Let's note $g_x(t) = p_{x(t)}(x)$. Write the differential equation which drives g_x in which g_{x-1} appears (take care of the particular case $x = 0$, by comparison with the general case $x > 0$).
- Use a recursion to prove the formula (3.7).

T167 **DNA (discrete “time”, discrete state)** A DNA strand is a sequence of 4 types of nucleotides: adenine, cytosine, guanine, thymine (A, C, G, T).

The sequential examination of a DNA strand can be considered as a trajectory of a Markov chain for which the state takes its value in the the **finite state space** $\{A, C, G, T\}$.

- The transition distribution corresponds to $\text{Prob}\left(x[n+1] = x^+ \mid x[n] = x\right)$ for all x and x^+ in the state space, represented, by the **transition matrix** (4×4 in this DNA model).

T168 It can be represented by a state graph (figure 3.2) in which each arrow should be labelled with a probability (or suppressed if the probability is 0) [10].

T169 Nevertheless, the homogeneity assumption of this Markov chain modeling the DNA strand is erroneous, since there are some pieces of this brand in which the dinucleotide CG is over-represented: the CpG islands. Another model with an homogeneous Markov chain uses the state space $\{A_+, C_+, G_+, T_+, A_-, C_-, G_-, T_-\}$ in which the $+$ index indicates that the nucleotide is inside a CpG island, and the $-$ index indicates that the nucleotide is not in such an island.

- The Markov chain takes its value in a 8 elements state space. Looking at a nucleotide along a DNA strand, we do not know if it belongs to a CpG island or not. The Markov chain is said to be “hidden”.

3.4 Hidden Markov models (discrete time case)

An **Hidden Markov Model** (HMM) corresponds to 2 stochastic processes:

- the state process $(x[n])_{n \geq 1}$ which, in general, cannot be observed,
- the observation process $(y[n])_{n \geq 1}$, in practice the output of some sensors,

➤ Given the present state, past and future are independent with the following convention:

$$\underbrace{\left(\begin{bmatrix} x[1] \\ y[1] \end{bmatrix}, \dots, \begin{bmatrix} x[n-1] \\ y[n-1] \end{bmatrix} \right)}_{\text{past}} \perp\!\!\!\perp \underbrace{\left(y[n], \begin{bmatrix} x[n+1] \\ y[n+1] \end{bmatrix}, \dots \right)}_{\text{future}} \mid x[n] \quad (3.8)$$

➤ It is necessary and sufficient to have, for all $n \geq 1$:

$$\underbrace{\left(\begin{bmatrix} x[1] \\ y[1] \end{bmatrix}, \dots, \begin{bmatrix} x[n-1] \\ y[n-1] \end{bmatrix} \right)}_{\text{past}} \perp\!\!\!\perp (y[n], x[n+1]) \mid x[n] \quad (3.9)$$

1171 We can show that the state process is Markovian.⁵

➤ In general, the future state and the present observation are assumed independent; for all $n \geq 1$:⁶

$$x[n+1] \perp\!\!\!\perp y[n] \mid x[n] \quad (3.10)$$

➤ Thus, the HMM is characterized by the following distributions:

- the **initial distribution**, that is the distribution of the initial state $x[1]$;
- the **transition distribution**, that is the distribution of $x[n+1] \mid x[n]$, for all $n \geq 1$;
- the **emission distribution**, that is the distribution of $y[n] \mid x[n]$, for all $n \geq 1$ (also known as the **local likelihood**).

1172 A path of the state process and of the observation process is generated with the induction on the figure 3.3.

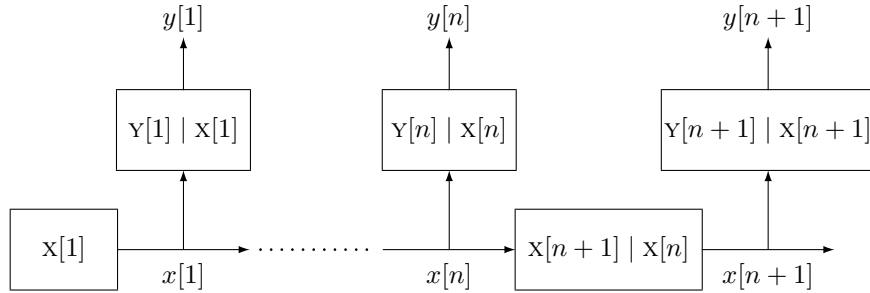


Figure 3.3: Hidden Markov Model

1173 The trajectography problems, that is the location of a moving target, are some typical examples of HMM. The state contains the coordinates of the target. The observation contains some goniometric measures. The transition distribution corresponds to the assumption on the target trajectory (for example, it almost follows a straight line). The emission distribution corresponds to the noise measurement distribution).

➤ The goal is to estimate the state path from the observation path.
This can be:

5. The condition (3.9) is equivalent to the 3 hypotheses below [12]:

- the state process is a Markov chain;
- given the state sequence, the observation sequence is independent;
- the observation $y[n]$ at time n , given the state sequence, depend only of $(x[n], x[n+1])$.

The last two hypotheses form the **memoryless channel** assumption; for all $n \geq 1$:

$$\begin{aligned} p_{x[n+1] \mid x[1:n]} &= p_{x[n+1] \mid x[n]} && \text{Markovian state} \\ p_{y[1:n] \mid x[1:n+1]} &= \prod_{k=1}^n p_{y[k] \mid x[k], x[k+1]} && \text{Memoryless channel} \end{aligned}$$

6. The condition (3.10) can also write: $p_{y[n] \mid x[n], x[n+1]} = p_{y[n] \mid x[n]}$. Thus, (3.9) and (3.10) are equivalent to the 2 conditions below:

$$\begin{aligned} p_{x[n+1] \mid x[1:n]} &= p_{x[n+1] \mid x[n]} && \text{Markovian state} \\ p_{y[1:n] \mid x[1:n]} &= \prod_{k=1}^n p_{y[k] \mid x[k]} && \text{Memoryless channel} \end{aligned}$$

A smoothing the future observation values are available to estimate the state at a given time;

A filtering: the current state has to be estimated “on the flight”, from past data only, for an **online** implementation.

3.5 Bayesian filtering

We observe $Y[1:n]$, with growing n .

We observe $Y[1:n]$, with growing n . The goal is to write the recursion on the probability distributions which will permit to estimate the current state, even to predict the future state and the future observation.

- For the sake of simplicity, we introduce the exponent $^{|n}$ which means “given the data $Y[1:n]$ ”. The Bayesian filtering is a recursion on the PDFs below; for all (y, x, x^+) :

$$\begin{aligned} p_{Y[n]|Y[1:n-1]}(y) & \text{ noted } p_{Y[n]}^{n-1}(y) \\ p_{X[n]|Y[1:n]}(x) & \text{ noted } p_{X[n]}^n(x) \\ p_{X[n+1]|Y[1:n]}(x^+) & \text{ noted } p_{X[n+1]}^n(x^+) \end{aligned}$$

- The recursion is started with the PDF $p_{X[1]}^0$, simply the prior PDF of the initial state; for all x^+ :⁷

$$\text{prior PDF of the first state} \quad p_{X[1]}^0(x^+) = p_{X[1]}(x^+) \quad (3.11)$$

- Then, for $n \geq 1$, we propagate the PDFs by means of the total probability law and the Bayes law; for all (y, x, x^+) :

$$\text{prior PDF of the } n\text{th observation} \quad p_{Y[n]}^{n-1}(y) = \int p_{Y[n]|X[n]}(y, x) p_{X[n]}^{n-1}(x) dx \quad (3.12)$$

$$n\text{th observation} \quad Y[n] \quad \leftarrow \text{sensors}$$

$$\text{posterior PDF of the } n\text{th state} \quad p_{X[n]}^n(x) = \frac{p_{Y[n]|X[n]}(Y[n], x) p_{X[n]}^{n-1}(x)}{p_{Y[n]}^{n-1}(Y[n])} \quad (3.13)$$

$$\text{prior PDF of the } (n+1)\text{th state} \quad p_{X[n+1]}^n(x^+) = \int p_{X[n+1]|X[n]}(x^+, x) p_{X[n]}^n(x) dx \quad (3.14)$$

- The formula (3.13) was obtained thanks to the notational trick $p_{X[n]|Y[n]}^{n-1}(x, Y[n]) = p_{X[n]}^n(x)$.

- From these PDFs, we can deduce Bayesian estimates, for example the MMSE estimations below, together with their (co)variances given the data:⁸

$$\begin{aligned} n\text{th observation prediction} \quad \hat{Y}^{n-1}[n] &= E^{n-1}(Y[n]) \\ C_{Y,Y}[n] &= \text{Var}^{n-1}(Y[n]) \\ C_{X,Y}[n] &= \text{Cov}^{n-1}(X[n], Y[n]) \\ n\text{th state estimation} \quad \hat{X}^n[n] &= E^n(X[n]) \\ P^n[n] &= \text{Var}^n(X[n]) \\ (n+1)\text{th state prediction} \quad \hat{X}^n[n+1] &= E^n(X[n+1]) \\ P^n[n+1] &= \text{Var}^n(X[n+1]) \end{aligned}$$

- In the linear Gaussian case, we obtain the **Kalman filter**, which propagates directly the quantities above, these quantities being necessary and sufficient to represent the Gaussian distributions.

7. In this book, the observation starts at time $n = 1$; thus, the exponent $^{|0}$ actually means “given nothing”.

8. For example, $E^n(X[n]) = \int x p_{X[n]}^n(x) dx$.

3.6 Linear model and Kalman filtering

The linear model is written, for all $n \geq 1$, as:⁹

$$\begin{cases} Y[n] = H_n X[n] + h_n + w[n] \\ X[n+1] = F_n X[n] + f_n + v[n] \end{cases} \quad (3.15)$$

under the assumptions:

- the matrix and vector series $(F_n)_{n \geq 1}$, $(f_n)_{n \geq 1}$, $(H_n)_{n \geq 1}$, $(h_n)_{n \geq 1}$ are known;¹⁰
- the series $(w[n])_{n \geq 1}$ is zero-mean, independent, with known variance Q_n ;
- the series $(v[n])_{n \geq 1}$ is zero-mean, independent, with known variance R_n ;
- the mean and the variance of the initial state $X[1]$ are known;
- the series $(w[n])_{n \geq 1}$, $(v[n])_{n \geq 1}$, and the initial state $X[1]$ are mutually independent.

T178 If the distributions of $X[1]$, $v[n]$, $w[n]$ are Gaussian, then, the distributions involved in the Bayesian filtering (3.12), (3.13) and (3.14) remain Gaussian, the recursion is written in function of their mean and their variance. The means provide the MMSE estimates.

- The Bayes filter becomes the Kalman filter, which is a recursive implementation of the MMSE applied to the Gaussian linear model.
- If Gaussian assumption on the sequences v , w or the initial state is removed, we can show that the Kalman filter is nothing but an implementation of the LMMSE estimator.
- Thus, the variances below (which can be calculated in advance, since they do not depend of the observations) have the meaning of estimation error variance in the Bayesian point of view.

$$C_{Y,Y}[n] = \text{Var}(\hat{Y}^{[n-1]}[n] - Y[n]) \quad P^{[n]}[n] = \text{Var}(\hat{X}^{[n]}[n] - X[n]) \quad P^{[n]}[n+1] = \text{Var}(\hat{X}^{[n]}[n+1] - X[n+1])$$

T179 Thus, we recognize in the Kalman filter (algo. 3.1) the formulas of the LMMSE estimator applied to the linear model.

T180 The fundamental equation is the state estimation one, which corrects the prediction by means:

- of the **Kalman gain** $K[n] = C_{X,Y}[n] C_{Y,Y}^{-1}[n]$;
- of the **innovation** process $(Y[n] - \hat{Y}^{[n-1]}[n])_{n \geq 1}$.¹¹

$$\hat{X}^{[n]}[n] = \hat{X}^{[n-1]}[n] + K[n] (Y[n] - \hat{Y}^{[n-1]}[n])$$

together with the error variance $P^{[n]}[n]$ for which numerous formulations exist.¹²

T181 If the matrices F_n , H_n , Q_n , R_n are time-independent, the matrix $P^{[n-1]}[n]$ tends to the limit $P[\infty]$ which fulfills the **discrete Riccati equation** obtained by grouping the formulas of variances update:

$$P[\infty] = F P[\infty] F^T - F P[\infty] H^T (H P[\infty] H^T + R)^{-1} H P[\infty] F^T + Q$$

- We obtains a highly simplified filter, by replacing the time-varying Kalman gain $K[n]$ by its limit $K[\infty]$:

$$K[\infty] = P[\infty] H^T (H P[\infty] H^T + R)^{-1}$$

T182 This filter asymptotically behaves like the Kalman filter, but is slower in the transience after the initialization (algo. 3.2).

9. This is an HMM whose PDFs are:

$$\begin{aligned} p_{X[1]} & \text{with mean } m_{X[1]} \text{ and variance } C_{X[1],X[1]} \\ p_{Y[n]|X[n]}(y, x) &= p_{w[n]}(y - H_n x - h_n) \quad \text{where } w[n] \text{ is zero-mean with variance } R_n \\ p_{X[n+1]|X[n]}(x^+, x) &= p_{v[n]}(x^+ - F_n x - f_n) \quad \text{where } v[n] \text{ is zero-mean with variance } Q_n \end{aligned}$$

10. They can be independent of time n . The index n is introduced for the sake of the generality.

11. Furthermore, the innovation sequence is uncorrelated.

12. With the Kalman gain $K[n]$, the variance update can also be written as:

$$\begin{aligned} P^{[n]}[n] &= (I - K[n] H_n) P^{[n-1]}[n] & (\text{the simplest}) \\ P^{[n]}[n] &= P^{[n-1]}[n] - K[n] C_{Y,Y}[n] K^T[n] \\ P^{[n]}[n] &= (I - K[n] H_n) P^{[n-1]}[n] (I - K[n] H_n)^T + K[n] R_n K^T[n] & (\text{Joseph form, better conditioned [31]}) \\ (P^{[n]}[n])^{-1} &= (P^{[n-1]}[n])^{-1} + H_n^T R_n^{-1} H_n & (\text{then } K[n] = P^{[n]}[n] H_n^T R_n^{-1}) \end{aligned}$$

Algorithm 3.1 Kalman filter**Works on the model** $H_n, h_n, F_n, f_n, R_n, Q_n, m_{x[1]}, C_{x[1],x[1]}$

$$\text{with } \begin{cases} Y[n] = H_n X[n] + h_n + w[n] \\ X[n+1] = F_n X[n] + f_n + v[n] \end{cases} \quad \text{and} \quad \begin{cases} R_n = C_{w[n],w[n]} \\ Q_n = C_{v[n],v[n]} \end{cases}$$

Initialization

$$\begin{array}{lll} \text{prediction of } x[1] \downarrow & \hat{x} \leftarrow m_{x[1]} & \text{provides } \hat{x}^{[0]}[1] \\ & P \leftarrow C_{x[1],x[1]} & P^{[0]}[1] \end{array}$$

Loop ($n \geq 1$)

$$\begin{array}{lll} \text{prediction of } Y[n] & \hat{Y} \leftarrow H_n \hat{x} + h_n & \hat{Y}^{[n-1]}[n] \\ & C_{x,Y} \leftarrow P H_n^T & C_{x,Y}[n] \\ & C_{Y,Y} \leftarrow H_n P H_n^T + R_n & C_{Y,Y}[n] \\ \rightarrow \text{observation of } Y[n] & Y \leftarrow \text{sensors} & Y[n] \\ \text{estimation of } x[n] & \hat{x} \leftarrow \hat{x} + C_{x,Y} C_{Y,Y}^{-1} (Y - \hat{Y}) & \hat{x}^{[n]}[n] \rightarrow \\ & P \leftarrow P - C_{x,Y} C_{Y,Y}^{-1} C_{x,Y}^T & P^{[n]}[n] \\ \text{prediction of } x[n+1] \downarrow & \hat{x} \leftarrow F_n \hat{x} + f_n & \hat{x}^{[n]}[n+1] \\ & P \leftarrow F_n P F_n^T + Q_n & P^{[n]}[n+1] \end{array}$$

Algorithm 3.2 Stationary Kalman filter**Works on the model** $H, h_n, F, f_n, R, Q, m_{x[1]}$

$$\text{with } \begin{cases} Y[n] = H X[n] + h_n + w[n] \\ X[n+1] = F X[n] + f_n + v[n] \end{cases} \quad \text{and} \quad \begin{cases} R = C_{w[n],w[n]} \\ Q = C_{v[n],v[n]} \end{cases}$$

Preliminaries

$$\begin{array}{lll} \text{Solve}/P & P = F P F^T - F P H^T (H P H^T + R)^{-1} H P F^T + Q & \text{provides } P[\infty] \\ \text{Calculate} & K \leftarrow P H^T (H P H^T + R)^{-1} & K[\infty] \end{array}$$

Initialization

$$\text{prediction of } x[1] \downarrow \quad \hat{x} \leftarrow m_{x[1]} \quad \hat{x}^{[0]}[1]$$

Loop ($n \geq 1$)

$$\begin{array}{lll} \text{prediction of } Y[n] & \hat{Y} \leftarrow H \hat{x} + h_n & \hat{Y}^{[n-1]}[n] \\ \rightarrow \text{observation of } Y[n] & Y \leftarrow \text{sensors} & Y[n] \\ \text{estimation of } x[n] & \hat{x} \leftarrow \hat{x} + K (Y - \hat{Y}) & \hat{x}^{[n]}[n] \rightarrow \\ \text{prediction of } x[n+1] \downarrow & \hat{x} \leftarrow F \hat{x} + f_n & \hat{x}^{[n]}[n+1] \end{array}$$

3.7 Adapting the Kalman filter to non-linear models

We suppose that the data generating process is, for all $n \geq 1$:

$$\begin{cases} Y[n] = h_n(X[n], W[n]) \\ X[n+1] = f_n(X[n], V[n]) \end{cases} \quad (3.16)$$

under the following assumptions:

- the function series $(f_n)_{n \geq 1}$ and $(h_n)_{n \geq 1}$ are known;
- the series $(V[n])_{n \geq 1}$ is zero-mean, independent, with known variance Q_n ;
- the series $(W[n])_{n \geq 1}$ is zero-mean, independent, with known variance R_n ;
- the mean and the variance of the initial state $X[1]$ are known;
- the series $(V[n])_{n \geq 1}$, $(W[n])_{n \geq 1}$, and the initial state $X[1]$ are mutually independent.

☛ We do not derive the distributions involved in this HMM.¹³

We draw inspiration from the Kalman filter to derive an approximation of the Bayesian filter which propagates only some means and (co)variances.

†184 We maintain the LMMSE state update:

$$\text{nth observation prediction} \quad \hat{Y}^{[n-1]}[n] = E^{[n-1]}(h_n(X[n], W[n])) \quad (3.17)$$

$$C_{Y,Y}[n] = \text{Var}^{[n-1]}(h_n(X[n], W[n]))$$

$$C_{X,Y}[n] = \text{Cov}^{[n-1]}(X[n], h_n(X[n], W[n]))$$

$$\text{nth observation} \quad Y[n] \leftarrow \text{sensors}$$

$$\text{nth state estimation} \quad \hat{X}^{[n]}[n] \simeq \hat{X}^{[n-1]}[n] + C_{X,Y}[n] C_{Y,Y}^{-1}[n] (Y[n] - \hat{Y}^{[n-1]}[n]) \quad (3.18)$$

$$P^{[n]}[n] \simeq P^{[n-1]}[n] - C_{X,Y}[n] C_{Y,Y}^{-1}[n] C_{X,Y}^T[n]$$

$$(n+1)\text{th state prediction} \quad \hat{X}^{[n]}[n+1] = E^{[n]}(f_n(X[n], V[n])) \quad (3.19)$$

$$P^{[n]}[n+1] = \text{Var}^{[n]}(f_n(X[n], V[n]))$$

☛ The formulas (3.17) and (3.19) are approximated through uncertainty propagation (page 25).

†185 If we propagate the uncertainty through linearization, we obtain the **Extended Kalman filter** (EKF, algo. 3.3).

†186 The case of additive noises is provided in the algo. 3.4.

13. The PDFs are $p_{Y[n]|X[n]}(y, x) = \int \delta(y - h_n(x, w)) p_{W[n]}(w) \, dw$ et $p_{X[n+1]|X[n]}(x^+, x) = \int \delta(x^+ - f_n(x, v)) p_{V[n]}(v) \, dv$.

Algorithm 3.3 Extended Kalman Filter (EKF)**Works on the model** $h_n, f_n, R_n, Q_n, m_{x[1]}, C_{x[1],x[1]}$

$$\text{with } \begin{cases} Y[n] = h_n(x[n], w[n]) \\ x[n+1] = f_n(x[n], v[n]) \end{cases} \quad \text{and } \begin{cases} R_n = C_{w[n],w[n]} \\ Q_n = C_{v[n],v[n]} \end{cases}$$

Initialization

$$\begin{array}{lll} \text{prediction of } x[1] \downarrow & \hat{x} \leftarrow m_{x[1]} & \text{provides } \hat{x}^{[0]}[1] \\ & \downarrow & \\ & P \leftarrow C_{x[1],x[1]} & P^{[0]}[1] \end{array}$$

Loop ($n \geq 1$)

$$\begin{array}{lll} \text{Jacobian matrix of } h_n & H_x \leftarrow \frac{\partial h_n}{\partial x^T}(\hat{x}, 0) \text{ and } H_w \leftarrow \frac{\partial h_n}{\partial w^T}(\hat{x}, 0) & \\ \text{prediction of } Y[n] & \hat{Y} \leftarrow h_n(\hat{x}, 0) & \hat{Y}^{[n-1]}[n] \\ & C_{Y,Y} \leftarrow H_x P H_x^T + H_w R_n H_w^T & C_{Y,Y}[n] \\ & C_{X,Y} \leftarrow P H_x^T & C_{X,Y}[n] \\ \rightarrow \text{observation of } Y[n] & Y \leftarrow \text{sensors} & Y[n] \\ \text{estimation of } x[n] & \hat{x} \leftarrow \hat{x} + C_{X,Y} C_{Y,Y}^{-1} (Y - \hat{Y}) & \hat{x}^{[n]}[n] \rightarrow \\ & P \leftarrow P - C_{X,Y} C_{Y,Y}^{-1} C_{X,Y}^T & P^{[n]}[n] \\ \text{Jacobian matrix of } f_n & F_x \leftarrow \frac{\partial f_n}{\partial x^T}(\hat{x}, 0) \text{ and } F_v \leftarrow \frac{\partial f_n}{\partial v^T}(\hat{x}, 0) & \\ \text{prediction of } x[n+1] \downarrow & \hat{x} \leftarrow f_n(\hat{x}, 0) & \hat{x}^{[n]}[n+1] \\ & \downarrow & \\ & P \leftarrow F_x P F_x^T + F_v Q_n F_v^T & P^{[n]}[n+1] \end{array}$$

Algorithm 3.4 EKF, additive noise case**Works on the model** $h_n, f_n, R_n, Q_n, m_{x[1]}, C_{x[1],x[1]}$

$$\text{with } \begin{cases} Y[n] = h_n(x[n]) + w[n] \\ x[n+1] = f_n(x[n]) + v[n] \end{cases} \quad \text{and } \begin{cases} R_n = C_{w[n],w[n]} \\ Q_n = C_{v[n],v[n]} \end{cases}$$

Initialization

$$\begin{array}{lll} \text{prediction of } x[1] \downarrow & \hat{x} \leftarrow m_{x[1]} & \text{provides } \hat{x}^{[0]}[1] \\ & \downarrow & \\ & P \leftarrow C_{x[1],x[1]} & P^{[0]}[1] \end{array}$$

Loop ($n \geq 1$)

$$\begin{array}{lll} \text{Jacobian matrix of } h_n & H \leftarrow \frac{\partial h_n}{\partial x^T}(\hat{x}) & \\ \text{prediction of } Y[n] & \hat{Y} \leftarrow h_n(\hat{x}) & \hat{Y}^{[n-1]}[n] \\ & C_{X,Y} \leftarrow P H^T & C_{X,Y}[n] \\ & C_{Y,Y} \leftarrow H P H^T + R_n & C_{Y,Y}[n] \\ \rightarrow \text{observation of } Y[n] & Y \leftarrow \text{sensors} & Y[n] \\ \text{estimation of } x[n] & \hat{x} \leftarrow \hat{x} + C_{X,Y} C_{Y,Y}^{-1} (Y - \hat{Y}) & \hat{x}^{[n]}[n] \rightarrow \\ & P \leftarrow P - C_{X,Y} C_{Y,Y}^{-1} C_{X,Y}^T & P^{[n]}[n] \\ \text{Jacobian matrix of } f_n & F \leftarrow \frac{\partial f_n}{\partial x^T}(\hat{x}) & \\ \text{prediction of } x[n+1] \downarrow & \hat{x} \leftarrow f_n(\hat{x}) & \hat{x}^{[n]}[n+1] \\ & \downarrow & \\ & P \leftarrow F P F^T + Q_n & P^{[n]}[n+1] \end{array}$$

- T187** If we propagate the uncertainty through the unscented transform UT page 25,¹⁴ we obtain the UKF (algo. 3.5), or the cubature Kalman filter (CKF) if the cubature is used.
- T188** The UKF provides naturally some σ -points for $(\hat{x}^n[n+1], P^n[n+1])$. In the case of an additive observation noise, we save one σ -points calculation as in the original algorithm [20] (algo. 3.6):

$$\text{if } y[n] = h_n(x[n]) + w[n] \quad \text{then} \quad \text{Var}^{n-1}(h_n(x[n]) + w[n]) = \text{Var}^{n-1}(h_n(x[n])) + R_n$$

- T189** If the state noise is also additive:

$$x[n+1] = f_n(x[n]) + v[n] \quad \text{then} \quad \text{Var}^n(f_n(x[n]) + v[n]) = \text{Var}^n(f_n(x[n])) + Q_n$$

we can either use the algorithm 3.6, with one σ -points calculation in dimension $2d_x$, or take into account the formula above; this leads to two σ -points calculation in dimension d_x (algo. 3.7).

- T190** Both EKF and UKF are some adaptations of the standard Kalman filter to a non-linear model. If they are applied to a linear model, we retrieve exactly the Kalman filter.

In general, the matrices $C_{y,y}[n]$, $P^n[n]$, $P^n[n+1]$ depend of the data and cannot be considered as estimation error variance in a bayesian meaning.

- ⊕ It is hard to give general theoretical results on the performances of such approximation-based algorithms. In practice, the UKF seems to provide less estimation error than the EKF. It does not need to compute the Jacobian matrices. We have to compare the numerical costs of the unscented transform and the Jacobian matrix computation.
- ⊕ If the EKF or the UKF do not provide suitable results, we can use **sequential Monte Carlo** methods such as the particle filter to approximate the Bayes filter.

14. if $y = h(x)$ then $(m_y, C_{y,y}, C_{x,y}) \simeq \text{UT}(h, m_x, C_{x,x})$

Algorithm 3.5 Unscented Kalman Filter (UKF)

Works on the model $h_n, f_n, R_n, Q_n, m_{x[1]}, C_{x[1],x[1]}$

$$\text{with } \begin{cases} y[n] = h_n(x[n], w[n]) \\ x[n+1] = f_n(x[n], v[n]) \end{cases} \quad \text{and} \quad \begin{cases} R_n = C_{w[n],w[n]} \\ Q_n = C_{v[n],v[n]} \end{cases}$$

Initialization

$$\begin{array}{lll} \text{prediction of } x[1] \downarrow & \hat{x} \leftarrow m_{x[1]} & \text{provides } \hat{x}^{[0]}[1] \\ \downarrow & P \leftarrow C_{x[1],x[1]} & P^{[0]}[1] \end{array}$$

Loop ($n \geq 1$)

$$\begin{array}{lll} \text{prediction of } y[n] & (\hat{y}, C_{y,y}, C_{x,y}) \leftarrow \text{UT} \left(h_n, \begin{bmatrix} \hat{x} \\ 0 \end{bmatrix}, \begin{bmatrix} P & 0 \\ 0 & R_n \end{bmatrix} \right) & \hat{y}^{[n-1]}[n], C_{y,y}[n], C_{x,y}[n] \\ \rightarrow \text{observation of } y[n] & y \leftarrow \text{sensors} & y[n] \\ \text{estimation of } x[n] & \hat{x} \leftarrow \hat{x} + C_{x,y} C_{y,y}^{-1} (y - \hat{y}) & \hat{x}^{[n]}[n] \rightarrow \\ & P \leftarrow P - C_{x,y} C_{y,y}^{-1} C_{x,y}^\top & P^{[n]}[n] \\ \text{prediction of } x[n+1] \downarrow & (\hat{x}, P) \leftarrow \text{UT} \left(f_n, \begin{bmatrix} \hat{x} \\ 0 \end{bmatrix}, \begin{bmatrix} P & 0 \\ 0 & Q_n \end{bmatrix} \right) & \hat{x}^{[n]}[n+1], P^{[n]}[n+1] \end{array}$$

Algorithm 3.6 UKF, additive observation noise case**Works on the model** $h_n, f_n, R_n, Q_n, m_{x[1]}, C_{x[1],x[1]}$

$$\text{with } \begin{cases} Y[n] = h_n(X[n]) + W[n] \\ X[n+1] = f_n(X[n], V[n]) \end{cases} \quad \text{and} \quad \begin{cases} R_n = C_{w[n],w[n]} \\ Q_n = C_{v[n],v[n]} \end{cases}$$

Initialization

$$\begin{array}{lll} \text{prediction of } X[1] \downarrow & \hat{X} \leftarrow m_{x[1]} & \text{provides } \hat{X}^{[0]}[1] \\ \downarrow & P \leftarrow C_{x[1],x[1]} & P^{[0]}[1] \\ \text{prediction of } Y[1] \downarrow & (\hat{Y}, C_{Y,Y}, C_{X,Y}) \leftarrow \text{UT}(h_1, \hat{X}, P) & \hat{Y}^{[0]}[1], C_{X,Y}[1] \\ \downarrow & C_{Y,Y} \leftarrow C_{Y,Y} + R_1 & C_{Y,Y}[1] \end{array}$$

Loop ($n \geq 1$)

$$\begin{array}{lll} \rightarrow \text{observation of } Y[n] & Y \leftarrow \text{sensors} & Y[n] \\ \text{estimation of } X[n] & \hat{X} \leftarrow \hat{X} + C_{X,Y} C_{Y,Y}^{-1} (Y - \hat{Y}) & \hat{X}^{[n]}[n] \rightarrow \\ & P \leftarrow P - C_{X,Y} C_{Y,Y}^{-1} C_{X,Y}^T & P^{[n]}[n] \\ \text{pred. } Y[n+1] \text{ and } X[n+1] \downarrow & \left(\begin{bmatrix} \hat{X} \\ \hat{Y} \end{bmatrix}, \begin{bmatrix} P & C_{X,Y} \\ C_{X,Y}^T & C_{Y,Y} \end{bmatrix} \right) \leftarrow \text{UT} \left(\begin{bmatrix} f_n \\ h_{n+1} \circ f_n \end{bmatrix}, \begin{bmatrix} \hat{X} \\ 0 \end{bmatrix}, \begin{bmatrix} P & 0 \\ 0 & Q_n \end{bmatrix} \right) & \\ & \hat{X}^{[n]}[n+1], \hat{Y}^{[n]}[n+1], P^{[n]}[n+1], C_{X,Y}[n+1] & \\ \downarrow & C_{Y,Y} \leftarrow C_{Y,Y} + R_{n+1} & C_{Y,Y}[n+1] \end{array}$$

Algorithm 3.7 UKF, additive noise case**Works on the model** $h_n, f_n, R_n, Q_n, m_{x[1]}, C_{x[1],x[1]}$

$$\text{with } \begin{cases} Y[n] = h_n(X[n]) + W[n] \\ X[n+1] = f_n(X[n]) + V[n] \end{cases} \quad \text{and} \quad \begin{cases} R_n = C_{w[n],w[n]} \\ Q_n = C_{v[n],v[n]} \end{cases}$$

Initialization

$$\begin{array}{lll} \text{prediction of } X[1] \downarrow & \hat{X} \leftarrow m_{x[1]} & \text{provides } \hat{X}^{[0]}[1] \\ \downarrow & P \leftarrow C_{x[1],x[1]} & P^{[0]}[1] \end{array}$$

Loop ($n \geq 1$)

$$\begin{array}{lll} \text{prediction of } Y[n] & (\hat{Y}, C_{Y,Y}, C_{X,Y}) \leftarrow \text{UT}(h_n, \hat{X}, P) & \hat{Y}^{[n-1]}[n], C_{X,Y}[n] \\ & C_{Y,Y} \leftarrow C_{Y,Y} + R_n & C_{Y,Y}[n] \\ \rightarrow \text{observation of } Y[n] & Y \leftarrow \text{sensors} & Y[n] \\ \text{estimation of } X[n] & \hat{X} \leftarrow \hat{X} + C_{X,Y} C_{Y,Y}^{-1} (Y - \hat{Y}) & \hat{X}^{[n]}[n] \rightarrow \\ & P \leftarrow P - C_{X,Y} C_{Y,Y}^{-1} C_{X,Y}^T & P^{[n]}[n] \\ \text{prediction of } X[n+1] \downarrow & (\hat{X}, P) \leftarrow \text{UT}(f_n, \hat{X}, P) & \hat{X}^{[n]}[n+1] \\ \downarrow & P \leftarrow P + Q_n & P^{[n]}[n+1] \end{array}$$

T191 Chapter 4

Stochastic simulation

T192 4.1 Random sampling

We consider the universe of the students of the university.

The function which returns for each student his average mark is a r.v. X .

The r.v. (X_1, \dots, X_{n_r}) obtained in a repetition of this operation n_r times in the same experimental conditions is called a n_r -**sample** of the r.v. X .

This n_r -sample is an i.i.d. sequence, which is constituted from n_r independent copies of the r.v. X .

We performed a random **sampling**.

- ⊕ The **stochastic simulation** consists to generate, by means of available calculation means, a sequence of numbers which can be considered as a realization of a n_r -sample of a r.v.

T193 For example, if we want to predict the weather conditions, we derive an algorithm based on a probabilistic model. If, on actual data, we obtain strange results, is this due to the model, to the prediction algorithm, or to a programming error?

We must test the algorithm on simulated data! If the prediction is correct on simulated data, but not on actual data, the model validity becomes questionable.

- ⊕ Furthermore, with **Monte Carlo methods**, we perform a computation thanks to random sampling.

For example, let's try to estimate the value of π ; we sample according to a uniform distribution over the square $[-1, 1] \times [-1, 1]$; the number of realizations in the unit circle is an approximation of $\frac{\pi}{4}$.¹

T194 4.2 Pseudo-random numbers generators



Figure 4.1: White noise generator

It exists some analog devices which simulate the hazard in continuous time, the **generators** of **pseudo-random** numbers in this document are algorithms which create a sequence of numbers such that a bunch of statistical tests assesses that this sequence is a sample realization with a given probability distribution.

- ⊕ An actual generator of random numbers does not exist (the prefix “pseudo” is important) since these algorithms are deterministic: we can always generate twice the same sequence, and the obtained sequences are periodic (with a very great period).

T195 For example, a **linear congruential generator** provide a sequence $(u_q)_{q \geq 1}$ by means of the recursion below, with

1. Matlab. To obtain an approximation of π (with a high n_r).
`4*mean(abs([1 j]*rand(2,nr))<1)`

s_q integer-valued:

$$\begin{aligned} u_q &= \frac{s_q}{T} \\ s_{q+1} &= (a s_q + b) \bmod T \end{aligned}$$

It must be initialized by s_1 (the **seed**).

The sequence period is at most T , some values of a and b provide a period T .

If $n_r \ll T$, (u_1, \dots, u_{n_r}) may be viewed as a realization of a n_r -sample of a r.v. U uniformly distributed between 0 and 1.²

- ⊛ This algorithm behavior highly depends of the choice of a , b and T . An implementation which was widely used in the seventies was later proven to be a poor choice, so that some scientific results became doubtful [19].

⊛ Thereafter:

- we will assume that the available calculation resources (computer, operating system, programming language with its libraries) provide a reliable uniformly distributed generator.
- U will designate a r.v. we can simulate, but it will often be a uniformly distributed r.v.

T196 4.3 Change of variable

If the r.v. X to be simulated can be obtained from the r.v. U with the transform $X = h(U)$, we just have to simulate U and to apply the transform h to the result.

- ⊛ In particular, in the dimension 1 case, in the CDF method, we use $h = F_X^{-1}$ and U uniformly distributed between 0 and 1.³
- ⊛ The Box-Muller transform provides a sample driven by a standardized normal distribution (page 75).⁴
- ⊛ To get a sample with a given mean and a given variance from a zero mean unit variance sample, we left multiply by a variance square root (the Cholesky decomposition for example), then we add the mean value.⁵

T197 4.4 Discrete distributions with finite support, mixture distributions

To simulate a r.v. Z which takes its value in $\{\zeta_1, \dots, \zeta_{n_c}\}$, with $\text{Prob}(Z = z) = \lambda_z$:

- we simulate a r.v. uniformly distributed between 0 and 1;
- then we select ζ_c if the result belongs to the interval $[\sum_{\ell=1}^{c-1} \lambda_{\zeta_\ell}, \sum_{\ell=1}^c \lambda_{\zeta_\ell}]$.⁶
- ⊛ To simulate a r.v. X driven by a mixture distribution $p_X(x) = \sum_{z \in \{\zeta_1, \dots, \zeta_{n_c}\}} \lambda_z f_z(x)$ (page 24), we simulate a pair (X, Z) in two steps:
 - we simulate Z according to the procedure above; let ζ_c be the result;
 - we simulate X with the distribution of PDF f_{ζ_c} .

T198 We simulate a n_r -sample $(Z_q)_{1 \leq q \leq n_r}$ of Z .

Let K_c the number of elements which take the value ζ_c in this sample:

$$K_c = \text{Card}\{q \in \{1, \dots, n_r\} \mid Z_q = \zeta_c\}$$

$K = (\underbrace{K_1, \dots, K_{n_c}}_{\text{sum}=n_r})$ is driven by a **multinomial distribution** with parameters $(n_r, \underbrace{\lambda_{\zeta_1}, \dots, \lambda_{\zeta_{n_c}}}_{\text{sum}=1})$. Its PMF is:

$$p_K(k_1, \dots, k_{n_c}) = n_r! \prod_{c=1}^{n_c} \frac{\lambda_{\zeta_c}^{k_c}}{k_c!}$$

2. Matlab. To generate a n_r -sample uniform distributed between 0 and 1 (the seed is processed by the function **rng**).

$\mathbf{x} = \text{rand}(1, n_r);$

3. If F_X^{-1} is available in the used programming language.

4. Matlab. To generate a n_r -sample driven by a zero mean and unit variance Gaussian distribution in \mathbb{R}^d .

$\mathbf{x} = \text{randn}(d, n_r);$

5. Matlab. To generate a Gaussian n_r -sample, with mean m and variance C .

$\mathbf{x} = \text{chol}(C, 'lower') * \text{randn}(\text{length}(C), n_r) + m * \text{ones}(1, n_r);$

6. We choose ζ_{n_c} after $n_c - 1$ tests, and, for all $c < n_c$, we choose ζ_c after d_c tests. The mean number of tests is $\sum_{c=1}^{n_c} c \lambda_c - \lambda_{n_c}$. We should sort the λ_c in decreasing order, in order to minimize this number.

The proposed algorithm permits to perform a **multinomial sampling**, that is the simulation of a r.v. driven by a multinomial distribution.^{7 8}

T199 4.5 Rejection sampling

We want to simulate a r.v. X (without simple generator), with PDF p_X .

To do this, we simulate a pair of independent r.v. (\tilde{X}, U) :

- \tilde{X} is an instrumental r.v. (for which we have a generator), with PDF $p_{\tilde{X}}$, such that the coefficient ν below (obviously in the interval $[0, 1]$ is strictly positive:

$$\nu = \inf_{x \in S(X)} \frac{p_{\tilde{X}}(x)}{p_X(x)}$$

and such that we can determine a value $a \in]0, \nu]$ (thus, for all x , $a p_X(x) \leq p_{\tilde{X}}(x)$);

- U is uniformly distributed between 0 and 1.

T200 Let T be the indicator function of the event $a p_X(\tilde{X}) \geq U p_{\tilde{X}}(\tilde{X})$; this test r.v. takes its value in $\{0, 1\}$.

Thus, the distribution of \tilde{X} given the event $T = 1$ is the distribution of X , the acceptance ratio is a .^{P1}

$$\begin{aligned} p_{\tilde{X}|T}(x, 1) &= p_X(x) \\ p_T(1) &= a \end{aligned}$$

In practice, we repeat the simulation of \tilde{X} and U until the test is true.⁹

• For a fast algorithm, we must select an instrumental distribution such that:

- the acceptance ratio a is close to 1, to avoid a high number of lost trials;
- the instrumental r.v. is easily simulated;
- the test is easily calculated.

T201 The rejection method (algo. 4.1) is typically used to simulate a gamma distributed r.v. [23].

7. Matlab. To generate a n_r -sample x of a discrete r.v. z which takes its value in $\{1, \dots, n_c\}$ such that $\text{Prob}(z = c) = \lambda_c$, together with one realization k with n_c components ($\lambda = (\lambda_1, \dots, \lambda_{n_c})$, the c th component of k is the number of occurrences of the value c in the array z).

`[k,z] = histc(rand(1,nr), [-Inf cumsum(lambda(1:end-1)) Inf]); k(end) = [];`

8. There exist some approximation of the multinomial sampling, with less tests and less random trials (stratified sampling, Kitagawa sampling...) [17]. For example, the systematic Kitagawa sampling needs only one sortition, whatever the length n_r is.

P1. Let's define $g = a \frac{p_X}{p_{\tilde{X}}}$. The joint distribution of (\tilde{X}, T) is written as:

$$p_{\tilde{X},T}(x, 1) = \int_0^1 p_{\tilde{X},T|U}(x, 1, u) \, du = \int_0^1 \underbrace{p_{T|\tilde{X},U}(1, x, u)}_{\begin{cases} 1 & \text{if } u \leq g(x) \\ 0 & \text{otherwise} \end{cases}} \underbrace{p_{\tilde{X}|U}(x, u)}_{p_{\tilde{X}}(x)} \, du = \int_0^{g(x)} p_{\tilde{X}}(x) \, du = p_{\tilde{X}}(x) g(x) = a p_X(x)$$

The acceptance ratio is: $\text{Prob}(a p_X(\tilde{X}) \geq U p_{\tilde{X}}(\tilde{X})) = p_T(1) = \int p_{\tilde{X},T}(x, 1) \, dx = \int a p_X(x) \, dx = a$

The conditional distribution becomes: $p_{\tilde{X}|T}(x, 1) = \frac{p_{\tilde{X},T}(x, 1)}{p_T(1)} = p_X(x)$

9. More generally, the target PDF p_X and the instrumental $p_{\tilde{X}}$ must be known up to a multiplicative constant, that is $p_X = \alpha \tilde{p}_X$ and $p_{\tilde{X}} = \beta \tilde{\tilde{p}}_X$ with unknown α or β . We must determine $\tilde{a} > 0$ (beware! \tilde{a} can be greater than 1) such that for all \tilde{x} , $\tilde{a} \tilde{p}_X(\tilde{x}) \leq \tilde{\tilde{p}}_X(\tilde{x})$, and the test becomes $\tilde{a} \tilde{p}_X(\tilde{x}) \geq U \tilde{\tilde{p}}_X(\tilde{x})$. Thus, $a = \frac{\beta}{\alpha} \tilde{a}$.

Algorithm 4.1 Rejection method for stochastic simulation

Inputs

- Target PDF \tilde{p}_X (up to a multiplicative constant).
- Instrumental PDF $\tilde{\tilde{p}}_X$ (up to a multiplicative constant) and associated generator.
- $\tilde{a} > 0$ such that for all $x \in S(X)$, $\tilde{p}_X(x) \geq \tilde{a} \tilde{\tilde{p}}_X(x)$.

Returns a realization x of X .

Algorithm

- Repeat
 - Draw u uniformly distributed between 0 and 1
 - Draw \tilde{x} according to the distribution with PDF $\propto \tilde{\tilde{p}}_X$
 - until $\tilde{a} \tilde{p}_X(\tilde{x}) \geq u \tilde{\tilde{p}}_X(\tilde{x})$
 - $x = \tilde{x}$
-

T202 Chapter 5

Monte Carlo methods

We want to evaluate $E(\phi(x)) = \int \phi(x) p_x(x) dx$ for a r.v. x and a function ϕ defined on the support of x such that this expectation exists.

T203 5.1 Direct sampling

We can simulate x .

It is not necessary to express p_x .

- Let $(x_q)_{1 \leq q \leq n_r}$ be a n_r -sample of x . The arithmetic mean $\sum_{q=1}^{n_r} \frac{1}{n_r} \phi(x_q)$ is a r.v. with mean $E(\phi(x))$ (that is the quantity to evaluate), and variance $\frac{1}{n_r} \text{Var}(\phi(x))$. Thus, we can write:

$$E(\phi(x)) \simeq \sum_{q=1}^{n_r} \frac{1}{n_r} \phi(x_q) \quad (5.1)$$

For every realization $(x_q)_{1 \leq q \leq n_r}$ of $(x_q)_{1 \leq q \leq n_r}$, the empirical mean $\sum_{q=1}^{n_r} \frac{1}{n_r} \phi(x_q)$ is an approximation of $E(\phi(x))$.

T204 5.2 Importance sampling

We can simulate an instrumental r.v. \tilde{x} (the support of \tilde{x} must contain the support of x).

We can calculate the ratio $\frac{p_x}{p_{\tilde{x}}}$.

We easily check that the expectation $E(\phi(x))$ can be written as:

$$E(\phi(x)) = E\left(\frac{p_x(\tilde{x})}{p_{\tilde{x}}(\tilde{x})} \phi(\tilde{x})\right)$$

- Let $(\tilde{x}_q)_{1 \leq q \leq n_r}$ be a n_r -sample of \tilde{x} . We can write:

$$E(\phi(x)) \simeq \sum_{q=1}^{n_r} \underbrace{\frac{1}{n_r} \frac{p_x(\tilde{x}_q)}{p_{\tilde{x}}(\tilde{x}_q)}}_{\omega_q(\tilde{x}_q)} \phi(\tilde{x}_q) \quad (5.2)$$

For every realization $(\tilde{x}_q)_{1 \leq q \leq n_r}$ of $(\tilde{x}_q)_{1 \leq q \leq n_r}$, the weighted mean $\sum_{q=1}^{n_r} \omega_q(\tilde{x}_q) \phi(\tilde{x}_q)$ is an approximation of $E(\phi(x))$. The Monte-Carlo method relies on a **weighted sampling**, also called **importance sampling**.¹

1. If $\text{Var}\left(\frac{p_x(\tilde{x})}{p_{\tilde{x}}(\tilde{x})} \phi(\tilde{x})\right) < \text{Var}(\phi(x))$, we obtain a better approximation than with direct sampling; the variance reduction through importance sampling is beyond the scope of this document.

T205 5.3 Importance sampling with auxiliary variable

Typically, we use this method if x is driven by a mixture distribution; x is a marginal of the pair (x, z) , with z the mixture component.

We can simulate the instrumental pair (\hat{x}, \hat{z}) (the support of (\hat{x}, \hat{z}) must contain the support of (x, z)) [24].

We can calculate $\frac{p_{x,z}}{p_{\hat{x},\hat{z}}}$.

We easily check the equality below, where \hat{z} is the auxiliary variable:

$$E(\phi(x)) = E\left(\frac{p_{x,z}(\hat{x}, \hat{z})}{p_{\hat{x},\hat{z}}(\hat{x}, \hat{z})} \phi(\hat{x})\right)$$

- Let $(\hat{x}_q, \hat{z}_q)_{1 \leq q \leq n_r}$ be a n_r -sample of (\hat{x}, \hat{z}) . We can write:

$$E(\phi(x)) \simeq \sum_{q=1}^{n_r} \frac{1}{n_r} \underbrace{\frac{p_{x,z}(\hat{x}_q, \hat{z}_q)}{p_{\hat{x},\hat{z}}(\hat{x}_q, \hat{z}_q)}}_{\omega(\hat{x}_q, \hat{z}_q)} \phi(\hat{x}_q) \quad (5.3)$$

For every realization $(\hat{x}_q, \hat{z}_q)_{1 \leq q \leq n_r}$ of $(\hat{x}, \hat{z})_{1 \leq q \leq n_r}$, the weighted mean $\sum_{q=1}^{n_r} \omega(\hat{x}_q, \hat{z}_q) \phi(\hat{x}_q)$ is an approximation of $E(\phi(x))$.

T206 5.4 Particle approximation

The formulas (5.1), (5.2) ou (5.3) provide an approximation of $E(\phi(x))$ which can be re-written as:

$$E(\phi(x)) \simeq \int \phi(x) \hat{p}_x(x) dx \quad \text{with} \quad \hat{p}_x(x) = \sum_{q=1}^{n_r} \omega_q \delta(x - \hat{x}_q) \quad \text{and} \quad \omega_q = \begin{cases} \frac{1}{n_r} & \text{direct} \\ \omega(\hat{x}_q) & \text{importance} \\ \omega(\hat{x}_q, \hat{z}_q) & \text{aux. var.} \end{cases}$$

The pulses sum \hat{p}_x is a **particle** approximation of the actual PDF of x . It is random since it depends of the **particles** $(\hat{x}_q)_{1 \leq q \leq n_r}$ and of the auxiliary variables $(\hat{z}_q)_{1 \leq q \leq n_r}$.

- ▲ In the importance sampling case, the weights are random, with mean $1/n_r$; their sum is random, with mean 1. A realization of \hat{p}_x is not a PDF. To keep a PDF interpretation, we must divide the weights by their sum.

T207 The exercise below shows that the particle approximation by itself is an approximation of the PMF of a discrete r.v., but cannot be used as the PDF of a continuous r.v.

◁▷ **Exercise 34.** Show that for all x , $E(\hat{p}_x(x)) = p_x(x)$, and that $\text{Var}(\hat{p}_x(x))$ is finite if x is discrete, infinite if x is continuous.

T208 5.5 Application to Bayesian estimation

Let x be the parameter to estimate. ϕ is the identity function, \hat{p}_x a particle approximation of the *a priori* PDF for a prediction, of the *a posteriori* PDF for an estimation.

- If x is numerically valued, the MMSE predictor approximation is the particles weighted mean:

$$\check{x}_{\text{MMSE}} = \sum_{q=1}^{n_r} \omega_q \hat{x}_q$$

- If x is discrete, the MAP predictor approximation is the value for which the sum of associated weights is maximal (if the weights are equal, it is the majority vote):²

$$\check{x}_{\text{MAP}} = \arg \max_x \sum_{q|\hat{x}_q=x} \omega_q$$

- The transposition to *a posteriori* estimators is straightforward.

2. If x has continuous and discrete components, that is $x = (x_c, x_d)$, we can use the marginal MAP for the discrete components, and the conditional MMSE for the continuous ones: $\check{x}_d = \arg \max_{x_d} \sum_{q|\hat{x}_{q,d}=x_d} \omega_q$ and $\check{x}_c = \frac{1}{\sum_{q|\hat{x}_{q,d}=\check{x}_d} \omega_q} \sum_{q|\hat{x}_{q,d}=\check{x}_d} \omega_q \hat{x}_{c,q}$

T209 Chapter 6

Particle filter

We try to recursively estimate the state of an HMM. We implement a Monte Carlo based approximation of the Bayesian filter, that is a **Sequential Monte Carlo (SMC)** method.

Unlike the Kalman filter, we do not only transmit the mean and the variance, but a particle approximation of some probability distributions.

T210 6.1 Principle

Lets's consider a Markov model with state $(x[n])_{n \geq 1}$ and observation $(y[n])_{n \geq 1}$. To simplify the writing:

the initial PDF $p_{x[1]}(x)$ is noted $\rho(x)$

the transition PDF $p_{x[n+1]|x[n]}(x^+, x)$ is noted $\kappa_{n+1}(x^+, x)$

- ⊕ We remind the Bayes filtering (page 42) with the notations above ((3.12) and (3.13) are merged):

$$\text{initialization} \quad p_{x[1]}^0(x^+) = \rho(x^+) \quad (6.1)$$

$$\text{recursion } (n \geq 1) \quad p_{x[n]}^n(x) = \frac{p_{y[n]|x[n]}(y[n], x) p_{x[n]}^{n-1}(x)}{\int p_{y[n]|x[n]}(y[n], u) p_{x[n]}^{n-1}(u) \, du} \quad (6.2)$$

$$p_{x[n+1]}^n(x^+) = \int \kappa_{n+1}(x^+, x) p_{x[n]}^n(x) \, dx \quad (6.3)$$

- ⊕ The particle filter [5] consists:

- in a preliminary particle approximation of the initial distribution, through sampling;
- in a propagation of this approximation with the Bayes filter;
- a resampling step is then necessary.

T211 This **condensation** (“**conditional density propagation**”) reminds the evolution theory (figure 6.1), in which the highly adapted individuals are selected.

T212 We use an instrumental stochastic process $\hat{x} = (\hat{x}[n])_{n \geq 1}$ which must be a Markov chain given the observations $Y = (y[n])_{n \geq 1}$. To simplify the writing:

instrumental initial PDF $p_{\hat{x}[1]|Y}(x)$ is noted $\hat{\rho}(x)$

instrumental transition PDF $p_{\hat{x}[n+1]|\hat{x}[n], Y}(x^+, x)$ is noted $\hat{\kappa}_{n+1}(x^+, x)$

T213 **To start up**, we need a particle approximation of the initial distribution with PDF ρ . This is the initial sampling, based on the instrumental distribution with PDF $\hat{\rho}$.

For all $q \in \{1, \dots, n_r\}$:

- an particle is sampled according to this distribution;
- its weight is calculated thanks to the importance sampling (5.2).

We obtain the particle approximation below:

$$\hat{p}_{x[1]}^0(x) = \sum_{q=1}^{n_r} \tilde{\omega}_q[1] \delta(x - \hat{x}_q[1]) \quad \text{with} \quad \tilde{\omega}_q[1] = \frac{1}{n_r} \frac{\rho(\hat{x}_q[1])}{\hat{\rho}(\hat{x}_q[1])}$$

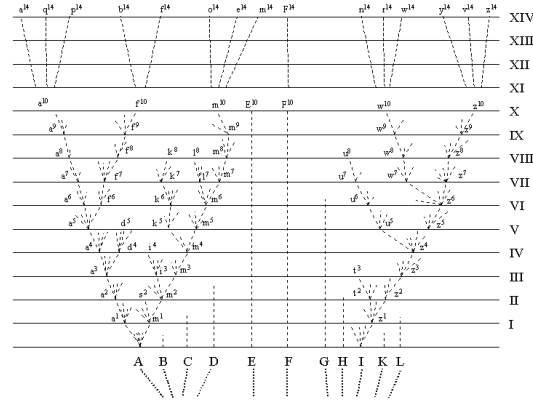


Figure 6.1: unique illustration in “Origin of Species” (Ch. DARWIN, 1859)

τ214 The recursion of the Bayesian filtering is fed with a particle approximation:

$$\hat{p}_{\mathbf{x}[n]}^{n-1}(x) = \sum_{q=1}^{n_r} \tilde{\omega}_q[n] \delta(x - \hat{\mathbf{x}}_q[n]) \quad (6.4)$$

We apply the formula (6.2); the particles are unchanged, their weight is multiplied by their local likelihood:

$$\hat{p}_{\mathbf{x}[n]}^n(x) = \sum_{q=1}^{n_r} \omega_q[n] \delta(x - \hat{\mathbf{x}}_q[n]) \quad \text{with } \omega_q[n] = \frac{\tilde{\omega}_q[n] p_{Y[n]|\mathbf{x}[n]}(Y[n], \hat{\mathbf{x}}_q[n])}{\sum_{z=1}^{n_r} \tilde{\omega}_z[n] p_{Y[n]|\mathbf{x}[n]}(Y[n], \hat{\mathbf{x}}_z[n])} \quad (6.5)$$

We apply the formula (6.3); we obtain a mixture distribution:

$$\hat{p}_{\mathbf{x}[n+1]}^n(x^+) = \sum_{q=1}^{n_r} \omega_q[n] \kappa_{n+1}(x^+, \hat{\mathbf{x}}_q[n]) \quad (6.6)$$

τ215 To go ahead, we need a particle approximation of the distribution with PDF $x^+ \mapsto \sum_{q=1}^{n_r} \omega_q[n] \kappa_{n+1}(x^+, \hat{\mathbf{x}}_q[n])$. This is the resampling.

We choose (we will see how to do this later) an instrumental PDF $x^+ \mapsto \sum_{q=1}^{n_r} \dot{\omega}_q[n] \dot{\kappa}_{n+1}(x^+, \hat{\mathbf{x}}_q[n])$.

• For all $q \in \{1, \dots, n_r\}$:

- the parent particle index $z_q[n+1]$ is selected according to the distribution $\text{Prob}(Z = z) = \dot{\omega}_z[n]$;¹
- the child particle $\hat{\mathbf{x}}_q[n+1]$ mutates according to the distribution $x^+ \mapsto \dot{\kappa}_{n+1}(x^+, \hat{\mathbf{x}}_{z_q[n+1]}[n])$.
- its weight is calculated thanks to the importance sampling with auxiliary variable (5.3).²

Thus, the formula (6.6) is replaced by:

$$\hat{p}_{\mathbf{x}[n+1]}^n(x) = \sum_{q=1}^{n_r} \tilde{\omega}_q[n+1] \delta(x - \hat{\mathbf{x}}_q[n+1]) \quad \text{with } \tilde{\omega}_q[n+1] = \frac{1}{n_r} \frac{\kappa_{n+1}(\hat{\mathbf{x}}_q[n+1], \hat{\mathbf{x}}_z[n]) \omega_z[n]}{\dot{\kappa}_{n+1}(\hat{\mathbf{x}}_q[n+1], \hat{\mathbf{x}}_z[n]) \dot{\omega}_z[n]} \Big|_{z=z_q[n+1]} \quad (6.7)$$

τ216 For online processing, the instrumental chain, which is sampled and resampled, can rely on the current observation.

If $\dot{\kappa}_{n+1}$ rely on $(Y[k])_{1 \leq k \leq n}$ (and $\dot{\rho}$ on no observation), we can include a prediction step in the filter. This is the **bootstrap filter** case, in which the instrumental chain do not rely on any observation.

Sif $\dot{\kappa}_{n+1}$ rely on $(Y[k])_{1 \leq k \leq n+1}$ (and $\dot{\rho}$ on $Y[1]$), this is meaningless, prediction and estimation steps must be merged.

1. $z \in \{1, \dots, n_r\}$. See multinomial sampling, page 51. $z_q[n+1]$ is the **auxiliary** variable.

2. We could use the standard importance sampling (5.2), at the cost of $2n_r$ PDF calculation per particle: $\tilde{\omega}_q[n+1] =$

$$\frac{1}{n_r} \frac{\sum_{z=1}^{n_r} \kappa_{n+1}(\hat{\mathbf{x}}_q[n+1], \hat{\mathbf{x}}_z[n]) \omega_z[n]}{\sum_{z=1}^{n_r} \dot{\kappa}_{n+1}(\hat{\mathbf{x}}_q[n+1], \hat{\mathbf{x}}_z[n]) \dot{\omega}_z[n]}$$

T217 6.2 Bootstrap filter

The particles are sampled according to the initial distribution and the transition one of the model of the data, that is: $\hat{\rho} = \rho$, $\hat{\kappa}_{n+1} = \kappa_{n+1}$, and $\hat{\omega}_q[n] = \omega_q[n]$ for all q .

The sampling of the distributions is then direct (no importance sampling).

T218 We get the “bootstrap” filter (algo. 6.1). The weights are proportional to the local likelihood.³⁴

3. We can store the re-distributed particles in the same variable, as in the next note

4. Matlab. Resampling in the particle filter.

`[~,z] = histc(rand(1,nr),[-Inf cumsum(omega(1:end-1)) Inf]); x = x(:,z);`

Algorithm 6.1 Bootstrap filter

Can be applied to a model with state $(x[n])_{n \geq 1}$, observation $(y[n])_{n \geq 1}$, **if we are able to**

- simulate the HMM state sequence,
- express $p_{Y[n]|X[n]}$.

Initialization

sampling $\downarrow \forall q$ $x_q \leftarrow$ sampling of $x[1]$ provides $(x_q[1])_{1 \leq q \leq n_r}$

prediction of $x[1]$ $\hat{x} \leftarrow \sum_{q=1}^{n_r} \frac{1}{n_r} x_q$ $\hat{x}^{[0]}[1]$

Loop ($n \geq 1$)

\rightarrow observation of $Y[n]$ $Y \leftarrow$ sensors $Y[n]$

weights update $\forall q$ $\omega_q \leftarrow p_{Y[n]|X[n]}(Y, x_q)$ (then normalization) $(\omega_q[n])_{1 \leq q \leq n_r}$

estimation of $x[n]$ $\hat{x} \leftarrow \sum_{q=1}^{n_r} \omega_q x_q$ $\hat{x}^{[n]}[n] \rightarrow$

resampling $\forall q$ $z_q \leftarrow$ sampling with the discrete distr. $(\omega_z)_{1 \leq z \leq n_r}$ $(z_q[n+1])_{1 \leq q \leq n_r}$

$\forall q$ $x_q^- \leftarrow x_{z_q}$ $(x_{z_q[n+1]}[n])_{1 \leq q \leq n_r}$

$\downarrow \forall q$ $x_q \leftarrow$ sampling of $x[n+1] \mid x[n] = x_q^-$ $(x_q[n+1])_{1 \leq q \leq n_r}$

prediction of $x[n+1]$ $\hat{x} \leftarrow \sum_{q=1}^{n_r} \frac{1}{n_r} x_q$ $\hat{x}^{[n]}[n+1]$

T219 6.3 Auxiliary particle filter

The bootstrap filter has few calculation per particle, but we need many particles because many of them are not selected.

It can be defective if the local likelihood of all the particles is 0. A solution is to sample the particle with a distribution based on the current observation (this is **adaptation** [24]).

T220 The observation must precede the random sampling, a prediction is meaningless. We re-organize the general algorithm by merging the estimation stage and the prediction one (algo. 6.2).

We must now propose some instrumental distributions based on the current observation.

Algorithm 6.2 Auxiliary particle filter

Can be applied to a model with state $(X[n])_{n \geq 1}$, observation $(Y[n])_{n \geq 1}$, initial distribution ρ , transition κ_{n+1} , **if we are able to**

- simulate a chain $(\hat{X}[n])_{n \geq 1}$ with initial distribution $\hat{\rho}$ and transition $\hat{\kappa}_{n+1}$,
- express $\frac{\rho p_{Y[1]|X[1]}}{\hat{\rho}}$ and $\frac{\kappa_n p_{Y[n]|X[n]}}{\hat{\kappa}_n}$.

Initial time ($n = 1$)

→ observation of $Y[1]$	$Y \leftarrow \text{sensors}$	provides $Y[1]$
sampling $\downarrow \forall q$	$X_q \leftarrow \text{sampling of } \hat{X}[1]$	$(\hat{X}_q[1])_{1 \leq q \leq n_r}$
$\downarrow \forall q$	$\omega_q \leftarrow \frac{\rho(X_q)}{\hat{\rho}(X_q)} p_{Y[1] X[1]}(Y, X_q)$ (then normalization)	$(\omega_q[1])_{1 \leq q \leq n_r}$
estimation of $X[1]$	$\hat{X} \leftarrow \sum_{q=1}^{n_r} \omega_q X_q$	$\hat{X}^1[1] \rightarrow$

Loop ($n \geq 2$)

→ observation of $Y[n]$	$Y \leftarrow \text{sensors}$	$Y[n]$
resampling $\forall z$	$\hat{\omega}_z \leftarrow \text{strategy to be defined}$	$(\hat{\omega}_z[n-1])_{1 \leq z \leq n_r}$
$\forall q$	$Z_q \leftarrow \text{sampling with the discrete distr. } (\hat{\omega}_z)_{1 \leq z \leq n_r}$	$(Z_q[n])_{1 \leq q \leq n_r}$
$\forall q$	$X_q^- \leftarrow X_{Z_q}$	$(\hat{X}_{Z_q}[n-1])_{1 \leq q \leq n_r}$
$\forall q$	$\omega_q^- \leftarrow \omega_{Z_q}$	$(\omega_{Z_q}[n-1])_{1 \leq q \leq n_r}$
$\forall q$	$\hat{\omega}_q^- \leftarrow \hat{\omega}_{Z_q}$	$(\hat{\omega}_{Z_q}[n-1])_{1 \leq q \leq n_r}$
$\downarrow \forall q$	$X_q \leftarrow \text{sampling of } \hat{X}[n] \mid \hat{X}[n-1] = X_q^-$	$(\hat{X}_q[n])_{1 \leq q \leq n_r}$
$\downarrow \forall q$	$\omega_q \leftarrow \frac{\omega_q^- \kappa_n(X_q, X_q^-)}{\hat{\omega}_q^- \hat{\kappa}_n(X_q, X_q^-)} p_{Y[n] X[n]}(Y, X_q)$ (then normalization)	$(\omega_q[n])_{1 \leq q \leq n_r}$
estimation of $X[n]$	$\hat{X} \leftarrow \sum_{q=1}^{n_r} \omega_q X_q$	$\hat{X}^n[n] \rightarrow$

6.4 Fully adapted particle filter

For all time n , for all q , the weights $\omega_q[n]$ are some functions of $\hat{X}_{1:n_r}[1:n]$, $Z_{1:n_r}[2:n]$, $Y[1:n]$.

But they become deterministic and equal if we choose:^{P1}

$$\hat{\rho} = p_{X[1]|Y[1]} \quad (6.8)$$

$$\forall n \geq 2, \hat{\omega}_z[n-1] \propto \omega_z[n-1] p_{Y[n]|X[n-1]}(Y[n], \hat{X}_z[n-1]) \quad (6.9)$$

$$\forall n \geq 2, \hat{\kappa}_n = p_{X[n]|X[n-1], Y[n]} \quad (6.10)$$

T222 There are very few particles without child, the diversity is preserved during the selection (algo. 6.3).

T223 But it is rarely possible to sample according to the distributions (6.10). We sampled according to an “analogous” distribution, for example a normal distribution with suitable mean and variance. And it is rarely possible to express the conditional distribution of the formula (6.9). An approximate calculation is made, for example with a normal PDF with suitable mean and variance.

P1. Given $\mathcal{C}[n] = (\hat{X}_{1:n_r}[1:n-1], Z_{1:n_r}[2:n], Y[1:n])$, $\omega_q[n]$ remains random as a function of $\hat{X}_{1:n_r}[n]$. We easily check that the weights mean value is:

$$\begin{aligned} E(\omega_q[1] | \mathcal{C}[1]) &= \frac{1}{n_r} p_{Y[1]}(Y[1]) \\ \forall n \geq 2 \quad E(\omega_q[n] | \mathcal{C}[n]) &= \frac{\omega_z[n-1] p_{Y[n]|X[n-1]}(Y[n], \hat{X}_z[n-1])}{n_r \hat{\omega}_z[n-1]} \Big|_{z=\hat{X}_q[n]} \end{aligned}$$

Thus, all the weights have the same mean with (6.9). Then we check that $\text{Var}(\omega_q[n] | \mathcal{C}[n]) = 0$, if we choose (6.8) et (6.10).

Algorithm 6.3 Fully adapted particle filter

Can be applied to a model with state $(X[n])_{n \geq 1}$, observation $(Y[n])_{n \geq 1}$, **if we are able to**

- simulate $X[1]$ given $Y[1]$, and $X[n]$ given $(X[n-1], Y[n])$, $n \geq 2$,
- express $p_{Y[n]|X[n-1]}$, $n \geq 2$.

Initial time ($n = 1$)

→ observation of $Y[1]$	$Y \leftarrow \text{sensors}$	provides $Y[1]$
sampling $\downarrow \forall q$	$X_q \leftarrow \text{sampling of } X[1] Y[1] = Y$	$(\hat{X}_q[1])_{1 \leq q \leq n_r}$
estimation of $X[1]$	$\hat{X} \leftarrow \sum_{q=1}^{n_r} \frac{1}{n_r} X_q$	$\hat{X}^{[1]}[1] \rightarrow$

Loop ($n \geq 2$)

→ observation of $Y[n]$	$Y \leftarrow \text{sensors}$	$Y[n]$
selection distribution $\forall z$	$\hat{\omega}_z \leftarrow p_{Y[n] X[n-1]}(Y, X_z)$ (then normalization)	$(\hat{\omega}_z[n-1])_{1 \leq z \leq n_r}$
resampling $\forall q$	$Z_q \leftarrow \text{sampling with the discrete distr. } (\hat{\omega}_z)_{1 \leq z \leq n_r}$	$(Z_q[n])_{1 \leq q \leq n_r}$
	$\forall q \quad X_q^- \leftarrow X_{Z_q}$	$(\hat{X}_{Z_q[n]}[n-1])_{1 \leq z \leq n_r}$
$\downarrow \forall q$	$X_q \leftarrow \text{sampling of } X[n] X[n-1] = \hat{X}_q^-, Y[n] = Y$	$(\hat{X}_q[n])_{1 \leq q \leq n_r}$
estimation of $X[n]$	$\hat{X} \leftarrow \sum_{q=1}^{n_r} \frac{1}{n_r} X_q$	$\hat{X}^{[n]}[n] \rightarrow$

6.5 Unscented particle filter

We draw at random the n_r initial particles $\hat{x}_q[1]$ with the normal distribution with mean and variance $E(x[1] | Y[1])$ and $\text{Var}(x[1] | Y[1])$.

We draw at random the n_r particles $\hat{x}_q[n]$, $n \geq 2$ with the normal distribution with mean and variance $E(x[n] | x[n-1] = \hat{x}_{z_q[n]}[n-1], Y[n])$ and $\text{Var}(x[n] | x[n-1] = \hat{x}_{z_q[n]}[n-1], Y[n])$.

But these means and variances are difficult to calculate. We approximate them by means of an uncertainty transform (linearization or unscented transform).

T225 We have the non linear model (3.16) with the assumptions page 45:

$$\begin{cases} Y[n] &= h_n(x[n], w[n]) \\ x[n+1] &= f_n(x[n], v[n]) \end{cases}$$

We have 2 stages, which are analogous to the prediction step and the estimation step of a Kalman filter, starting from the parent particle $\hat{x}_q[n-1]$.

We first calculate, for each particle, an approximation $\check{x}_q^{[n-1]}[n]$ of the mean $E(x[n] | x[n-1] = \hat{x}_q[n-1])$, that is:

$$\check{x}_q^{[n-1]}[n] \simeq E(f_{n-1}(x[n-1], v[n-1]) | x[n-1] = \hat{x}_q[n-1])$$

and an approximation $P_q^{[n-1]}[n]$ of the associated variance, through uncertainty transformation due to the function $v \mapsto f_{n-1}(\hat{x}_q[n-1], v)$, for $v[n-1]$ zero-mean with variance Q_{n-1} (we note that the state is fixed, the cloud of particles takes its variation into account).

T226 Then, we correct with $Y[n]$ to obtain an approximation $\check{x}_q^{[n]}[n]$ of $E(x[n] | x[n-1] = \hat{x}_q[n-1], Y[n])$ and an approximation $P_q^{[n]}[n]$ of the associated variance, through LMMSE.

The approximations below

$$\begin{aligned} \check{y}_q[n] &\simeq E(h_n(x[n], w[n]) | x[n-1] = \hat{x}_q[n-1]) \\ C_{x,y;q}[n] &\simeq \text{Cov}(x[n], h_n(x[n], w[n]) | x[n-1] = \hat{x}_q[n-1]) \\ C_{y,y;q}[n] &\simeq \text{Var}(h_n(x[n], w[n]) | x[n-1] = \hat{x}_q[n-1]) \end{aligned}$$

are obtained by uncertainty transformation due to the function h_n , for $w[n]$ zero-mean with variance R_n , and $x[n]$ with mean $\check{x}_q^{[n-1]}[n]$ and variance $P_q^{[n-1]}[n]$ calculated at the first stage.

The particle redistribution is made with the discrete distribution below ($g(., P)$ is the normal zero-mean PDF with variance P), obtained by an approximation of the formula (6.9):

$$\hat{\omega}_q[n-1] \propto \omega_q[n-1] g(Y[n] - \check{y}_q[n], C_{y,y;q}[n])$$

T227 We can use the unscented transform to propagate the uncertainty (algo. 6.4), to obtain the algorithm in [1]), but we can also use a linearization.



Algorithm 6.4 Unscented particle filter

Works on the model $h_n, f_n, R_n, Q_n, m_{x[1]}, C_{x[1],x[1]}$

$$\text{with } \begin{cases} Y[n] = h_n(X[n], W[n]) \\ X[n+1] = f_n(X[n], V[n]) \end{cases} \quad \text{and} \quad \begin{cases} R_n = C_{w[n],w[n]} \\ Q_n = C_{v[n],v[n]} \end{cases}$$

if we are able to express the initial distribution ρ and the transition distribution κ_{n+1}

Notation $g(\cdot, P)$ is the PDF of the zero-mean normal PDF with variance P . $\mathcal{N}(x, P)$ is the normal distribution with mean x and variance P .

Initial time ($n = 1$)

→ observation	$Y \leftarrow \text{sensors}$	provides $Y[1]$
	$\check{X}_1 \leftarrow m_{x[1]} + C_{x,y;1} C_{y,y;1}^{-1} (Y - \check{Y}_1)$	$\check{X}_1^n[n]$
	$P_1 \leftarrow C_{x[1],x[1]} - C_{x,y;1} C_{y,y;1}^{-1} C_{x,y;1}^\top$	$P_1^n[n]$
sampling $\downarrow \forall q$	$X_q \leftarrow \text{sampling driven by } \mathcal{N}(\check{X}_1, P_1)$	$(\check{X}_q[1])_{1 \leq q \leq n_r}$
$\downarrow \forall q$	$\omega_q \leftarrow \frac{\rho(X_q) p_{Y[1] X[1]}(Y, X_q)}{g(X_q - \check{X}_1, P_1)} \text{ (then normalization)}$	$(\omega_q[1])_{1 \leq q \leq n_r}$
estimation	$\hat{X} \leftarrow \sum_{q=1}^{n_r} \omega_q X_q$	$\hat{X}^1[1] \rightarrow$

Loop ($n \geq 2$)

→ observation	$Y \leftarrow \text{sensors}$	$Y[n]$	
$\forall q$	$(\check{X}_q, P_q) \leftarrow \text{UT}(v \mapsto f_{n-1}(X_q, v), 0, Q_{n-1})$	$(\check{X}_q^{n-1}[n], P_q^{n-1}[n])_{1 \leq q \leq n_r}$	
$\forall q$	$(\check{Y}_q, C_{y,y;q}, C_{x,y;q}) \leftarrow \text{UT}\left(h_n, \begin{bmatrix} \check{X}_q \\ 0 \end{bmatrix}, \begin{bmatrix} P_q & 0 \\ 0 & R_n \end{bmatrix}\right)$	$(\check{Y}_q[n], C_{y,y;q}[n], C_{x,y;q}[n])_{1 \leq q \leq n_r}$	
	$\forall q$	$\check{X}_q \leftarrow \check{X}_q + C_{x,y;q} C_{y,y;q}^{-1} (Y[n] - \check{Y}_q)$	$(\check{X}_q^n[n])_{1 \leq q \leq n_r}$
	$\forall q$	$P_q \leftarrow P_q - C_{x,y;q} C_{y,y;q}^{-1} C_{x,y;q}^\top$	$(P_q^n[n])_{1 \leq q \leq n_r}$
	$\forall z$	$\hat{\omega}_z \leftarrow \omega_z g(Y - \check{Y}_z, C_{y,y;z}) \text{ (then normalization)}$	$(\hat{\omega}_z[n-1])_{1 \leq z \leq n_r}$
resampling	$\forall q$	$Z_q \leftarrow \text{sampling with the discrete distr. } (\hat{\omega}_z)_{1 \leq z \leq n_r}$	$(Z_q[n])_{1 \leq q \leq n_r}$
	$\forall q$	$\check{X}_q^- \leftarrow \check{X}_{Z_q}$	$(\check{X}_{Z_q}^-[n])_{1 \leq q \leq n_r}$
	$\forall q$	$P_q^- \leftarrow P_{Z_q}$	$(P_{Z_q}^-[n])_{1 \leq q \leq n_r}$
	$\forall q$	$X_q^- \leftarrow X_{Z_q}$	$(\check{X}_{Z_q}^-[n-1])_{1 \leq q \leq n_r}$
	$\forall q$	$\omega_q^- \leftarrow \omega_{Z_q}$	$(\omega_{Z_q}^-[n-1])_{1 \leq q \leq n_r}$
	$\forall q$	$\hat{\omega}_q^- \leftarrow \hat{\omega}_{Z_q}$	$(\hat{\omega}_{Z_q}^-[n-1])_{1 \leq q \leq n_r}$
$\downarrow \forall q$	$X_q \leftarrow \text{sampling driven by } \mathcal{N}(\check{X}_q^-, P_q^-)$	$(\check{X}_q^n[n])_{1 \leq q \leq n_r}$	
$\downarrow \forall q$	$\omega_q \leftarrow \frac{\omega_q^- \kappa_n(X_q, X_q^-) p_{Y[n] X[n]}(Y, X_q)}{\hat{\omega}_q^- g(X_q - \check{X}_q^-, P_q^-)} \text{ (then norm.)}$	$(\omega_q^n[n])_{1 \leq q \leq n_r}$	
estimation	$\hat{X} \leftarrow \sum_{q=1}^{n_r} \omega_q X_q$	$\hat{X}^n[n] \rightarrow$	

Appendix A

Mathematics

A.1 Matrices

0_d is the $d \times 1$ null vector, $0_{d \times n}$ is the $d \times n$ null matrix, I_d is the $d \times d$ identity matrix.

A Toeplitz matrix has constant diagonals: $M_{\ell,k}$ depends only of $\ell - k$.

A square matrix has the same number of rows and columns.

The concepts of “determinant”, “invertibility” and “diagonalization” of a square matrix are assumed to be known.

The **trace** of a square matrix M , denoted $\text{trace } M$, is the sum of the diagonal elements.

Let M and M' be two square matrices, and λ be a scalar number:

$$\det(\lambda M) = \lambda^d \det M \quad (\text{A.1})$$

$$\text{trace}(\lambda M) = \lambda \text{trace } M \quad (\text{A.2})$$

$$\det(M M') = \det M \det M' \quad (\text{A.3})$$

$$\text{trace}(M + M') = \text{trace } M + \text{trace } M' \quad (\text{A.4})$$

If M is a diagonalizable matrix, its determinant is the eigenvalues product, its trace is the eigenvalues sum.

Let A and B be two matrices such that the products AB and BA are meaningful (therefore, these products are square); then:

$$\text{trace}(AB) = \text{trace}(BA) \quad (\text{A.5})$$

Let us consider the block matrix $\begin{bmatrix} A & B \\ D & C \end{bmatrix}$ with square A and C . If C is invertible, then $\begin{bmatrix} A & B \\ D & C \end{bmatrix}$ is invertible if and only if the **Schur complement** of C in $\begin{bmatrix} A & B \\ D & C \end{bmatrix}$, that is $A - B C^{-1} D$, is invertible.¹

Symmetrically, if A is invertible, then $\begin{bmatrix} A & B \\ D & C \end{bmatrix}$ is invertible if and only if the Schur complement of A in $\begin{bmatrix} A & B \\ D & C \end{bmatrix}$, that is $C - D A^{-1} B$, is invertible.²

If A and C are invertible, then $A - B C^{-1} D$ is invertible if and only if $C - D A^{-1} B$ is invertible, and we obtain the **Woodbury inversion lemma**:

$$(A - B C^{-1} D)^{-1} = A^{-1} + A^{-1} B (C - D A^{-1} B)^{-1} D A^{-1} \quad (\text{A.6})$$

We obtain a symmetrical relation by exchanging A and C , B and D ; furthermore:

$$C^{-1} D (A - B C^{-1} D)^{-1} = (C - D A^{-1} B)^{-1} D A^{-1} \quad (\text{A.7})$$

A matrix M is **symmetric** if it is equal to its transpose matrix: $M = M^T$.

A matrix M is **diagonal** all the terms but the main diagonal ones are zero: $M_{\ell,k} = 0$ if $\ell \neq k$.

A matrix M is **orthogonal** if it is invertible and its inverse is equal to its transpose: $M^{-1} = M^T$.

Every symmetric matrix is diagonalizable with real eigenvalues and orthogonal change of basis matrix.

1. ... and we obtain:

$$\begin{bmatrix} A & B \\ D & C \end{bmatrix}^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & C^{-1} \end{bmatrix} + \begin{bmatrix} I & \\ -C^{-1} D & \end{bmatrix} (A - B C^{-1} D)^{-1} \begin{bmatrix} I & -B C^{-1} \end{bmatrix}$$

2. ... and we obtain:

$$\begin{bmatrix} A & B \\ D & C \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -A^{-1} B & \\ I & \end{bmatrix} (C - D A^{-1} B)^{-1} \begin{bmatrix} -D A^{-1} & I \end{bmatrix}$$

A symmetric matrix M is **positive semi-definite** (resp. **positive definite**) if, for all non-zero vector x :

$$x^\top M x \geq 0 \quad (\text{resp. } x^\top M x > 0)$$

A positive definite matrix is an invertible positive semi-definite matrix. The inverse of a positive definite matrix is positive definite, the eigenvalues are inverted, the change of basis matrix remains unchanged.

Let M be a positive definite matrix $d \times d$, and r a positive real number. The set $\{x \in \mathbb{R}^d \mid x^\top M x = r\}$ is centered on 0_d ; it is an interval for $d = 1$, an ellipse for $d = 2$, an ellipsoid for $d = 3$. In this book, we will call this set an ellipsoid, for all d .

The relation between some positive semi-definite matrices denoted $M \geq M'$ if $M - M'$ is positive semi-definite is a partial order, called **Loewner** order (or Löwner) [3, 9, 16]. If $M \geq M'$, the ellipsoid $\{x \in \mathbb{R}^d \mid x^\top M x = r\}$ is inside the ellipsoid $\{x \in \mathbb{R}^d \mid x^\top M' x = r\}$. If M and M' are invertible:

$$M \geq M' \text{ if and only if } M'^{-1} \geq M^{-1}$$

The eigenvalues of a positive definite matrix are strictly positive, the eigenvalues of a positive semi-definite matrix are non-negative.

Let us consider the block matrix $M = \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix}$ such that C is positive definite. Then M is positive definite (respectively semi-definite) if and only if the Schur complement of C in M is positive definite (respectively semi-definite).

If M is positive semi-definite, there exist some matrices L such that $M = L L^\top$ (we say that L is a square root of M). Among these matrices, some are lower triangular with positive elements on the main diagonal. If M is positive definite, such a solution is unique, and the elements on the main diagonal are strictly positive (**Cholesky decomposition**).³

If M is positive semi-definite:^{P1}

$$(\det M)^{\frac{1}{d}} \leq \frac{\text{trace } M}{d} \quad (\text{A.8})$$

Let M_0 be a positive definite matrix $d \times d$; we have to solve the minimization below:

$$\arg \min_M \left(\log \det M + \text{trace}(M^{-1} M_0) \right) \quad (\text{A.9})$$

The solution, among all positive definite matrices, is M_0 ;

the solution, among all Toeplitz diagonal positive definite matrices is $\frac{\text{trace } M_0}{d} I_d$.^{P2}

A.2 Differentiation

Let's consider the differentiable (once or twice) vector valued functions of a vector valued variable.

Let $f : x \mapsto f(x)$ be a differentiable function:

- $\frac{\partial f}{\partial x^\top}$ is the **Jacobian matrix**;
- If f is scalar valued:
 - $\frac{\partial f}{\partial x}$ is the **gradient** (transposed Jacobian matrix);

3. Matlab. `L = chol(M, 'lower')`

P1. The left term is the eigenvalues geometric mean, the right term is the eigenvalues arithmetic mean. The inequality of arithmetic and geometric means Specifies that the geometric mean of positive numbers is always lower than the arithmetic mean. More generally, this property holds for every matrix with positive eigenvalues (not necessarily symmetric).

P2. At first, let's show that the eigenvalues of a product of positive definite matrices are strictly positive: let A and B two positive definite matrices; there exist an invertible matrix L such that $A = L L^\top$; let λ be an eigenvalue of AB ; if v is an associated eigenvector, then:

$$\lambda v = A B v = L L^\top B v = L L^\top B L L^{-1} v$$

Pre-multiplying by L^{-1} , we obtain:

$$\lambda (L^{-1} v) = L^\top B L (L^{-1} v)$$

λ is then an eigenvalue of the matrix $L^\top B L$, which is positive definite (with associated eigenvector $L^{-1} v$). Then, $\lambda \in \mathbb{R}^{+*}$.

The eigenvalues $(\lambda_1, \dots, \lambda_d)$ of $M^{-1} M_0$ are strictly positive, then:

$$J(M) - J(M_0) = -\log \det(M^{-1} M_0) + \text{trace } M^{-1} M_0 - \text{trace } I_d = -\log \prod_{k=1}^d \lambda_k + \sum_{k=1}^d \lambda_k - d = \sum_{k=1}^d (\lambda_k - 1 - \log \lambda_k)$$

$(\lambda_k - 1 - \log \lambda_k)$ is minimal for $\lambda_k = 1$, then $J(M) - J(M_0)$ is minimal if the eigenvalues of $M^{-1} M_0$ are all 1, that is if $M = M_0$. The solution in the Toeplitz diagonal case is straightforward.

$-\frac{\partial^2 f}{\partial x \partial x^\top}$ is the **Hessian** matrix (Jacobian matrix of the gradient, symmetric).

Under matrix form (ℓ is the row index, k is the column index):

$$\frac{\partial f}{\partial x^\top}(x) = \left[\frac{\partial f_\ell}{\partial x_k}(x) \right]_{\ell,k} \quad \frac{\partial f}{\partial x}(x) = \left[\frac{\partial f}{\partial x_\ell}(x) \right]_\ell \quad \frac{\partial^2 f}{\partial x \partial x^\top}(x) = \left[\frac{\partial^2 f}{\partial x_\ell \partial x_k}(x) \right]_{\ell,k}$$

Reciprocal function If $f : x \mapsto f(x)$ is invertible, with reciprocal function $f^{-1} : y \mapsto f^{-1}(y)$:

$$\frac{\partial f^{-1}}{\partial y^\top}(y) = \left[\frac{\partial f}{\partial x^\top}(f^{-1}(y)) \right]^{-1} \quad \frac{\partial f^{-1}}{\partial y^\top}(f(x)) = \left[\frac{\partial f}{\partial x^\top}(x) \right]^{-1} \quad (\text{A.10})$$

Integration By means of the differentiable and invertible change of variable $y = f(x)$, we obtain:

$$\int_A g(y) \, dy = \int_{f^{-1}(A)} g(f(x)) \left| \det \frac{\partial f}{\partial x^\top}(x) \right| \, dx \quad (\text{A.11})$$

Truncated series At order 1, around x_0 :

$$f(x) \approx f(x_0) + \frac{\partial f}{\partial x^\top}(x_0) (x - x_0)$$

At order 2, if f is scalar valued:

$$f(x) \approx f(x_0) + (x - x_0)^\top \frac{\partial f}{\partial x}(x_0) + \frac{1}{2} (x - x_0)^\top \frac{\partial^2 f}{\partial x \partial x^\top}(x_0) (x - x_0)$$

Composition Let's consider $g \circ f : x \mapsto g(f(x))$ the composition of $f : x \mapsto f(x)$ and $g : y \mapsto g(y)$:

$$\frac{\partial(g \circ f)}{\partial x^\top}(x) = \frac{\partial g}{\partial y^\top}(f(x)) \frac{\partial f}{\partial x^\top}(x)$$

If f and g are scalar valued (thus, g is a function from \mathbb{R} to \mathbb{R}):

$$\begin{aligned} \frac{\partial(g \circ f)}{\partial x}(x) &= \frac{\partial f}{\partial x}(x) g'(f(x)) \\ \frac{\partial^2(g \circ f)}{\partial x \partial x^\top}(x) &= \frac{\partial^2 f}{\partial x \partial x^\top}(x) g'(f(x)) + \frac{\partial f}{\partial x}(x) g''(f(x)) \left[\frac{\partial f}{\partial x}(x) \right]^\top \end{aligned}$$

Product If g is scalar valued:

$$\frac{\partial(fg)}{\partial x^\top}(x) = \frac{\partial f}{\partial x^\top}(x) g(x) + f(x) \left[\frac{\partial g}{\partial x}(x) \right]^\top$$

If f and g take their value in the same set:

$$\frac{\partial(f^\top g)}{\partial x}(x) = \left[\frac{\partial f}{\partial x^\top}(x) \right]^\top g(x) + \left[\frac{\partial g}{\partial x^\top}(x) \right]^\top f(x)$$

Linear function

$$\text{if } f(x) = Ax + b \text{ then } \frac{\partial f}{\partial x^\top}(x) = A$$

Quadratic function

$$\text{if } f(x) = x^\top Ax + b^\top x + c \text{ then } \frac{\partial f}{\partial x}(x) = (A + A^\top)x + b \text{ and } \frac{\partial^2 f}{\partial x \partial x^\top}(x) = A + A^\top$$

In this book, we use the result below, where C is positive definite:

$$\begin{aligned} &\text{if } f(x) = \frac{1}{2} (b - Hx)^\top C^{-1} (b - Hx) \text{ with } C \text{ positive definite} \\ &\text{then } \frac{\partial f}{\partial x}(x) = -H^\top C^{-1} (b - Hx) \end{aligned} \quad (\text{A.12})$$

Dirac pulse and change of variable Let g be a differentiable and invertible function, with Jacobian matrix $\frac{\partial g}{\partial x^T}$; for all (x, y) :^{P3}

$$\delta(x - g^{-1}(y)) = \delta(y - g(x)) \left| \det \frac{\partial g}{\partial x^T}(x) \right| \quad (\text{A.13})$$

A.3 Some functions

Error function It is the odd function defined, for all $x \in \mathbb{R}$, by:

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) \, dt$$

It increases from -1 to 1 .

The CDF of the univariate normal distribution with zero mean and unit variance is $x \mapsto \frac{1}{2} + \frac{1}{2} \text{erf}\left(\frac{x}{\sqrt{2}}\right)$.⁴

Γ function It is defined, for all $x > 0$, by

$$\Gamma(x) = \int_0^{+\infty} t^{x-1} \exp(-t) \, dt$$

It is unimodal and positive, it tends to $+\infty$ in 0 and $+\infty$, it is minimal for $x \simeq 1,4616$ where it takes the value $\Gamma(x) \simeq 0.8856$. For all $x > 0$:

$$\Gamma(x+1) = x \Gamma(x)$$

It can be seen as an interpolation of the factorial, since, for all $n \in \mathbb{N}$:

$$\Gamma(n+1) = n!$$

For all $n \in \mathbb{N}$ (a special case is $\Gamma(\frac{1}{2}) = \sqrt{\pi}$)

$$\Gamma\left(n + \frac{1}{2}\right) = \frac{\prod_{k=1}^n (2k-1)}{2^n} \sqrt{\pi}$$

For all $\alpha > 0$ and all $\beta > 0$:

$$\int_0^1 x^{\alpha-1} (1-x)^{\beta-1} \, dx = \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

P3. Let's fix y . If $x \neq g^{-1}(y)$, both sides are zero. If we integrate with respect to x , we obtain 1 for both sides: through extraction property on the left side, through property (A.11) on the right side. Thus, we obtain the Dirac pulse with weight 1 in $g^{-1}(y)$ on both sides.

4. Matlab. To define the PDF and the CDF on an univariate normal distribution with mean m and standard deviation σ .

```
pdfGauss = @(x, m, sigma) 1/sqrt(2*pi)/sigma*exp(-0.5*((x-m)/sigma).^2);
cdfGauss = @(x, m, sigma) 0.5 + 0.5*erf((x-m)/sigma/sqrt(2));
```

Appendix B

Stochastic processes

A **stochastic process** (or **random signal**) is a function of time and chance.

In the discrete-time case, a stochastic process is also called a **time series**.

Every realization of a stochastic process is called a **trajectory**, or a **sample path**.

For example, if we consider the rolling angle of two identical boats at the same place under the same weather conditions, we do not obtain the same signals; nevertheless, the trajectories have a similar “look”.

If Y is a stochastic process, for all $t \in \mathbb{R}$, $Y(t)$ (or, for all $n \in \mathbb{Z}$, $Y[n]$) is a r.v.

▲ A process can take discrete values, or continuous values, or hybrid values. The time can be continuous or discrete. Do not make any confusion between these two aspects.

B.1 Time analysis

The **probability distribution** of the random process corresponds to the distribution of $Y(t_1), \dots, Y(t_{n_t})$, for all number of times n_t and for all distinct times (t_1, \dots, t_{n_t}) .

A stochastic process is **independent** if $Y(t_1), \dots, Y(t_{n_t})$ are mutually independent, for all n_t and for all (t_1, \dots, t_{n_t}) . Such a process is completely unpredictable: the knowledge of the trajectory in some times does not bring any information on the signal value at another time.

A process is **independent and identically distributed** (i.i.d.) if it is independent and if the distribution of $Y(t)$ does not depend of t .

Usually, we observe only one trajectory. We have to make simplifying assumptions such that a statistical analysis can be made with only one trajectory: the stationarity and the ergodicity.

Stationarity (strict-sense) The probabilistic properties don't depend of the time origin; for all number of times n_t , for all initial time t , for all $n_t - 1$ lags $(\tau_1, \dots, \tau_{n_t-1})$ with respect to the initial time:

the distribution of $(Y(t + \tau_1), \dots, Y(t + \tau_{n_t-1}), \dots, Y(t))$ do not depend of t .

An i.i.d. process is stationary.

But, for the sake of simplicity, we usually reduce to second order probabilistic properties, to define a weaker stationarity assumption, the “wide-sense” stationarity.

Stationarity (wide-sense) For all initial time t and all lag τ :

the mean $E(Y(t))$ does not depend of t ,

the variance $\text{Var}(Y(t))$ does not depend of t ,

the covariance $\text{Cov}(Y(t + \tau), Y(t))$ only depends on τ .

Thus, the process distribution is only characterized by:

- the mean $m_Y = E(Y(t))$;
- the autocorrelation function $c_{Y,Y} : \tau \mapsto \text{Cov}(Y(t + \tau), Y(t))$.¹

The mean and the autocorrelation function are the second-order statistics.

1. It would be better to call it “autocovariance function”, because it corresponds to nothing but the covariance between the signal at a time and the same signal at another time, but “autocorrelation function” is widely used.

Obviously, the strict-sense stationarity implies the wide-sense stationarity. The converse is generally false. But, if the process is normally distributed (that is, for all number of times n_t and for all times (t_1, \dots, t_{n_t}) , $(Y(t_1), \dots, Y(t_{n_t}))$ is normally distributed), the strict-sense stationarity and the wide-sense stationarity are equivalent).

Ergodicity (in the mean and the autocorrelation) For all realization y of the wide-sense stationary process Y :

$$\begin{aligned} m_Y &= \langle y(t) \rangle_t && \text{ergodicity in the mean} \\ c_{Y,Y}(\tau) &= \langle (y(t+\tau) - m_Y)(y(t) - m_Y)^\top \rangle_t && \text{ergodicity in the autocorrelation} \end{aligned}$$

We used above the time average operator:

$$\begin{aligned} \langle \psi(t) \rangle_t &= \lim_{t_{\max} \rightarrow +\infty} \frac{1}{2t_{\max}} \int_{-t_{\max}}^{+t_{\max}} \psi(t) dt && \text{continuous time} \\ \langle \psi[n] \rangle_n &= \lim_{n_{\max} \rightarrow +\infty} \frac{1}{2n_{\max} + 1} \sum_{n=-n_{\max}}^{n_{\max}} \psi[n] && \text{discrete time} \end{aligned}$$

In practice, the ergodicity is nothing but an assumption.²

Every function (or sequence) cannot represent an autocorrelation function. In particular, $c_{Y,Y}(0)$ is the variance of the signal and must be positive semi-definite. The autocorrelation has the symmetry below; for all τ :

$$c_{Y,Y}(-\tau) = c_{Y,Y}^\top(\tau)$$

Furthermore, the Cauchy-Schwarz inequality (1.39) gives:

$$c_{Y,Y}(0) \geq c_{Y,Y}(-\tau) [c_{Y,Y}(0)]^{-1} c_{Y,Y}(\tau)$$

For a scalar-valued process, we obtain $c_{Y,Y}(0) \geq |c_{Y,Y}(\tau)|$; in general, a finite scalar-valued sequence $(c[0], \dots, c[\tau])$ can represent the $\tau + 1$ first terms of an autocorrelation function of a discrete time scalar-valued stochastic process if and only if the Toeplitz matrix $[c[\ell - k]]_{\substack{0 \leq \ell \leq \tau \\ 0 \leq k \leq \tau}}$ is a positive definite matrix; this matrix is the **autocorrelation matrix**.

The autocorrelation function measures in what extent the signal at time $t + \tau$ is similar to the signal at time t . For a scalar-valued signal, it is in general positive and decreasing (also this is not mathematically necessary); if it decreases slowly when τ moves away from 0, this indicates that the realizations of the stochastic process are smooth, with slow variations; conversely, if it decreases rapidly to 0, the realizations are “nervous”, with fast variations.

In the extreme, for the **white noise**, there is full uncorrelation between 2 different times. The autocorrelation is null everywhere but in $\tau = 0$; there exists a positive semi-definite matrix Q such that:

$$c_{Y,Y} = Q \delta$$

where δ is the Dirac delta (continuous time case) or the Kronecker delta (discrete time case). Obviously, an i.i.d. process is a white noise.

▲ If Y is discrete time, its variance is Q (in the scalar case, Q has the same dimensionality than the square of the signal). If Y is continuous time, its variance is infinite (in the scalar case, Q has the same dimensionality than the square of the signal, multiplied by a time). In both cases, the power spectrum (defined below) is Q .

The whiteness concept comes from an analogy with the white colour in which all frequencies are present, and can be interpreted in the spectral domain.

B.2 Spectral analysis

The **power spectral density** (PSD, or **power spectrum**) $S_{Y,Y}$ of a wide-sense stationary stochastic process Y is the Fourier transform of the autocorrelation function; for all frequency f in the continuous time case (or all normalized frequency λ in the discrete time case):

$$S_{Y,Y}(f) = \mathcal{F}_{cc} c_{Y,Y}(f) = \int_{-\infty}^{+\infty} c_{Y,Y}(\tau) \exp(-j 2\pi f \tau) d\tau \quad \text{continuous time} \quad (\text{B.1})$$

$$S_{Y,Y}(\lambda) = \mathcal{F}_{dc} c_{Y,Y}(\lambda) = \sum_{\tau=-\infty}^{+\infty} c_{Y,Y}[\tau] \exp(-j 2\pi \lambda \tau) \quad \text{discrete time} \quad (\text{B.2})$$

2. A necessary and sufficient condition for the ergodicity in the mean is that $\langle y(t) \rangle_t$ is independent of the trajectory, that is to say deterministic.

A necessary and sufficient condition for the ergodicity in the autocorrelation is that for all τ , $\langle (y(t+\tau) - m_Y)(y(t) - m_Y)^\top \rangle_t$ is deterministic.

Obviously, the PSD of a white noise is uniform.

A sine wave has a PSD located in its frequency.

◁▷ **Exercise 35** (Sine wave PSD). Let's consider the scalar-valued continuous-time stochastic process defined, for all $t \in \mathbb{R}$, by $Y(t) = a \cos(2\pi f_0 t + \Phi)$, where Φ is a r.v. uniformly distributed between 0 and 2π .

a) Show that this process is wide-sense stationary, and that its autocorrelation function is written as, for all τ ,

$$c_{Y,Y}(\tau) = \frac{a^2}{2} \cos(2\pi f_0 \tau).$$

b) What is its PSD.

We can see that the PSD permits to identify some periodic components in a stationary stochastic process, just like the Energy spectral density (that is the square modulus of the Fourier transform) permits to identify some periodic components in a deterministic signal.

The **Wiener-Khinchine theorem** connects the notions of PSD and ESD: for a zero-mean process Y .

We note $w_{t_{\max}}$ is the rectangular window over the interval $\{t \mid |t| \leq t_{\max}\}$.

Under the assumption that $\int_{-\infty}^{\infty} |\tau c_{Y,Y}(\tau)| d\tau$ (or $\sum_{\tau=-\infty}^{+\infty} |\tau c_{Y,Y}(\tau)|$) is convergent, then the PSD is written as, for a scalar-valued process:³

$$S_{Y,Y}(f) = \lim_{t_{\max} \rightarrow +\infty} \frac{1}{2t_{\max}} \mathbb{E} \left(\left| \mathcal{F}_{cc}(Y w_{t_{\max}})(f) \right|^2 \right) \quad \text{continuous time} \quad (\text{B.3})$$

$$S_{Y,Y}(\lambda) = \lim_{n_{\max} \rightarrow +\infty} \frac{1}{2n_{\max} + 1} \mathbb{E} \left(\left| \mathcal{F}_{dc}(Y w_{n_{\max}})(\lambda) \right|^2 \right) \quad \text{discrete time} \quad (\text{B.4})$$

B.3 Autocorrelation function estimation

Let's suppose that the stochastic process Y is discrete time, wide-sense stationary, with zero mean, ergodic in the autocorrelation. How to estimate its autocorrelation function, that is to define a function \hat{c} which returns, for all lag τ and all windowed sample path $y[1:n_t] = (y[1], \dots, y[n_t])$, an approximation $\hat{c}(\tau, y[1:n_t])$ of $c_{Y,Y}(\tau)$?

It is equivalent to define the r.v. $\hat{c}(\tau, Y[1:n_t])$, more simply noted $\hat{C}_{Y,Y}^{[n_t]}(\tau)$, where the exponent $^{[n_t]}$ means "given $Y[1:n_t]$ ".

A natural estimation is:

$$\hat{C}_{Y,Y}^{[n_t]}(\tau) = \begin{cases} \frac{1}{n_t - |\tau|} \sum_{n=1-\min(0,\tau)}^{n_t - \max(0,\tau)} Y[n + \tau] Y^T[n] & \text{if } |\tau| < n_t \\ 0 & \text{otherwise} \end{cases} \quad (\text{B.5})$$

The difference $\hat{C}_{Y,Y}^{[n_t]}(\tau) - c_{Y,Y}(\tau)$ is the **estimation error**, its mean $\mathbb{E} \left(\hat{C}_{Y,Y}^{[n_t]}(\tau) \right) - c_{Y,Y}(\tau)$ is the **bias**. We easily check that the estimation (B.5) is unbiased (that is with null bias). Using the ergodicity assumption, the asymptotic variance is null (when $n_t \rightarrow +\infty$): this estimation is **consistent**. But nothing implies that this estimation is mathematically an autocorrelation function.

It is often better to use the **biased** estimation below:

$$\hat{C}_{Y,Y}^{[n_t]}(\tau) = \begin{cases} \frac{1}{n_t} \sum_{n=1-\min(0,\tau)}^{n_t - \max(0,\tau)} Y[n + \tau] Y^T[n] & \text{if } |\tau| < n_t \\ 0 & \text{otherwise} \end{cases} \quad (\text{B.6})$$

This estimation is biased, but asymptotically unbiased and consistent. It provides a valid autocorrelation sequence.

3. For a vector-valued process:

$$S_{Y,Y}(f) = \lim_{t_{\max} \rightarrow +\infty} \frac{1}{2t_{\max}} \mathbb{E} \left(\left[\mathcal{F}_{cc}(Y w_{t_{\max}})(f) \right] \left[\mathcal{F}_{cc}(Y w_{t_{\max}})(f) \right]^H \right) \quad \text{continuous time}$$

$$S_{Y,Y}(\lambda) = \lim_{n_{\max} \rightarrow +\infty} \frac{1}{2n_{\max} + 1} \mathbb{E} \left(\left[\mathcal{F}_{dc}(Y w_{n_{\max}})(\lambda) \right] \left[\mathcal{F}_{dc}(Y w_{n_{\max}})(\lambda) \right]^H \right) \quad \text{discrete time}$$

where the H exponent indicates the Hermitian transpose (conjugate transpose).

B.4 PSD estimation

Let's suppose that the stochastic process Y is discrete time, wide-sense stationary, with zero mean, ergodic in autocorrelation. How to estimate its PSD, that is to define a function \hat{S} which returns, for all frequency λ and all windowed sample path $y[1:n_t] = (y[1], \dots, y[n_t])$, an approximation $\hat{S}(\lambda, y[1:n_t])$ of $S_{Y,Y}(\lambda)$?

It is equivalent to define the r.v. $\hat{S}(\lambda, Y[1:n_t])$, more simply noted $\hat{S}_{Y,Y}^{[n_t]}(\lambda)$, where the exponent $^{[n_t]}$ means “given $Y[1:n_t]$ ”.

Starting from the formula (B.2), it is natural to propose, as an estimation of the PSD, the Fourier transform of an estimation of the autocorrelation function:

$$\hat{S}_{Y,Y}^{[n_t]}(\lambda) = \mathcal{F}_{\text{dc}} \hat{C}_{Y,Y}^{[n_t]}(\lambda) \quad (\text{B.7})$$

$$= \sum_{\tau=-n_t+1}^{n_t-1} \hat{C}_{Y,Y}^{[n_t]}(\tau) \exp(-j 2\pi \lambda \tau) \quad (\text{B.8})$$

Starting from the formula (B.4), in which we “forget” the expectation (since only one sample path is available) and the time limit (since the observed sample path is only finite length), it is natural to propose, as an estimation of the PSD, the ESD of the sample path divided by its length, that is, for a scalar valued signal:⁴

$$\hat{S}_{Y,Y}^{[n_t]}(\lambda) = \frac{1}{n_t} |\mathcal{F}_{\text{dc}}(y w_{n_t})(\lambda)|^2 \quad (\text{B.9})$$

$$= \frac{1}{n_t} \left| \sum_{n=1}^{n_t} y[n] \exp(-j 2\pi \lambda n) \right|^2 \quad (\text{B.10})$$

By means of tedious calculations, we can show that both estimations (B.8) and (B.10) are equal, if the biased estimation (B.6) of the autocorrelation function is used. This is the **periodogram**.

The **periodogram**. It exhibits spectral leakage due to the time windowing. We can show that the mean of the periodogram is the actual PSD convolved with the square Dirichlet kernel, which implies a low resolution (due to the width of the main lobe of the Dirichlet kernel) and secondary lobes appearance: if there are two periodic components with close frequencies, we obtain in general only one lobe around these frequencies (low resolution); furthermore, the secondary lobes from a high power periodic component can mask the spectral contribution of weak components (low sensitivity).

Furthermore, we can show that, despite the ergodicity in autocorrelation, this estimation is not consistent: the variance of the estimation does not tend to zero when $n_t \rightarrow +\infty$. For example, for a scalar valued i.i.d. normally distributed sequence, the standard deviation of the estimation error has the level of the actual DSP.

Some modified versions of the periodogram, deduced from formulation (B.10), try to correct these problems. The use of a smoothed window (with respect to the rectangular window, such as the Hamming, Hann, Bartlett windows...) with lower secondary lobes, improve the sensitivity, but decreases the resolution. Beware! To keep the asymptotic unbiasedness, we must normalize by the window energy; for a window w_{n_t} , for a scalar valued signal:⁵

$$\hat{S}_{Y,Y}^{[n_t]}(\lambda) = \frac{1}{\sum_{n=1}^{n_t} w_{n_t}^2[n]} |\mathcal{F}_{\text{dc}}(y w_{n_t})(\lambda)|^2 \quad (\text{B.11})$$

$$= \frac{1}{\sum_{n=1}^{n_t} w_{n_t}^2[n]} \left| \sum_{n=1}^{n_t} y[n] w_{n_t}[n] \exp(-j 2\pi \lambda n) \right|^2 \quad (\text{B.12})$$

To decrease the variance of the estimation error, we can divide the sample path into several segments. The estimated DSP is the arithmetic mean of the DSPs estimated on all the segments. These segments can overlap. Unfortunately, the resolution decreases.

4. For a vector valued signal

$$\begin{aligned} \hat{S}_{Y,Y}^{[n_t]}(\lambda) &= \frac{1}{n_t} \left[\mathcal{F}_{\text{dc}}(y w_{n_t})(\lambda) \right] \left[\mathcal{F}_{\text{dc}}(y w_{n_t})(\lambda) \right]^H \\ &= \frac{1}{n_t} \left[\sum_{n=1}^{n_t} y[n] \exp(-j 2\pi \lambda n) \right] \left[\sum_{n=1}^{n_t} y[n] \exp(-j 2\pi \lambda n) \right]^H \end{aligned}$$

5. For a vector valued signal

$$\hat{S}_{Y,Y}^{[n_t]}(\lambda) = \frac{1}{\sum_{n=1}^{n_t} w_{n_t}^2[n]} \left[\mathcal{F}_{\text{dc}}(y w_{n_t})(\lambda) \right] \left[\mathcal{F}_{\text{dc}}(y w_{n_t})(\lambda) \right]^H$$

The **Welch** periodogram combine all these variants; it is parameterized by the length of the segments, the overlap ration, the window shape. The choice of these parameters depends of the phenomenons we try to enlighten.

All these modifications decreases the resolution. A possible “high resolution” approach consists in modeling the random signal by means of a **generation process** based on linear time invariant systems (LTI).

B.5 Stationary processes and LTI systems

Let us consider a LTI system with impulse response h , whose input is a stochastic process w , whose output is $Y = h * w$ (figure B.1).

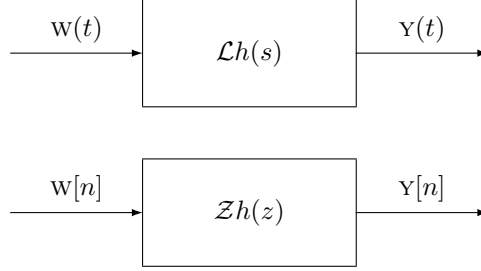


Figure B.1: LTI system

We easily show that LTI filtering preserves the stationarity, and that the output autocorrelation function is written as:

$$c_{Y,Y} = h * c_{w,w} * \underline{h}^T \quad (\text{B.13})$$

where \underline{h} is the time-reversed sequence h ($\underline{h}[n] = h[-n]$).

Using Fourier transform, the PSD of Y is written as, in the discrete time case:⁶

$$S_{Y,Y}(\lambda) = \mathcal{F}_{dc} h(\lambda) S_{w,w}(\lambda) [\mathcal{F}_{dc} h(\lambda)]^H \quad (\text{B.14})$$

In particular, if w is a white noise with spectral density Q :

$$S_{Y,Y}(\lambda) = \mathcal{F}_{dc} h(\lambda) Q [\mathcal{F}_{dc} h(\lambda)]^H \quad (\text{B.15})$$

This is the fundamental property leading to the modeling by means of a generation process.

B.6 Generation process

In this section, this book reduces to a discrete time scalar-valued stationary stochastic process. A generation process consists in the assumption that the signal Y is the response of a LTI system whose input is a white noise with DSP Q .

$$S_{Y,Y}(\lambda) = Q |\mathcal{F}_{dc} h(\lambda)|^2 \quad (\text{B.16})$$

The frequency response of this generation process permits to shape the DSP. Although there exist some mathematical conditions to the existence of such a model (we speak of **spectral factorization**), in practice, we assume that such a representation is possible.

We choose a generation process such that:

- The poles and the zeros are inside the unit circle;
- the system has direct feedthrough ($h[0] \neq 0$) and monic ($h[0] = 1$); indeed, since the input white noise is fictitious, it can be time-shifted, and its variance is a scale parameter;
- the transfer function is rational; there exists a set of coefficients $(a_1, \dots, a_{n_a}, c_1, \dots, c_{n_c})$ such that it is written as:

$$\mathcal{Z}h(z) = \frac{1 + c_1 z^{-1} + \dots + c_{n_c} z^{-n_c}}{1 + a_1 z^{-1} + \dots + a_{n_a} z^{-n_a}}$$

6. For a continuous time signal, just replace the Fourier transform \mathcal{F}_{dc} by the Fourier transform \mathcal{F}_{cc}

This structure can be considered as sufficiently general, especially with high orders n_a and n_c . Nevertheless, the aim of simplicity leads to choose small orders.

Si $n_c = 0$, we have an **all-pole**, or **autoregressive (AR)** model. It is good at representing stochastic processes whose DSP exhibits spikes (the spikes frequency are close to the argument of the poles next to the unit circle) (figure B.2). It is used in speech modeling.

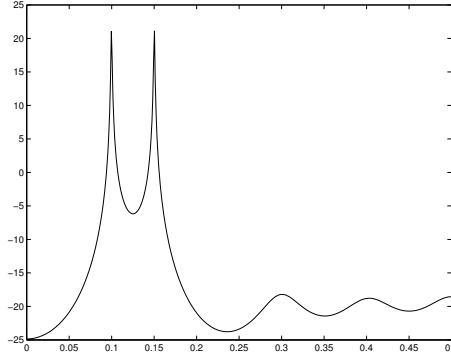


Figure B.2: Peaky nature of AR spectrum

If $n_a = 0$, we have a **Moving Average (MA)** model. It is good at representing stochastic processes whose DSP exhibits strong attenuations. It is used in economy.

The general case corresponds to ARMA models.

Equivalence rules between the transfer function and the state space representation permit to write the representation below, for an ARMA model (with $N = \max(n_a, n_c)$):

$$Y[n] = C X[n] + W[n] \quad (\text{B.17})$$

$$X[n+1] = A X[n] + B W[n] \quad (\text{B.18})$$

avec:

$$A = \begin{bmatrix} -a_1 & 1 & 0 & \cdots & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & & \ddots & \ddots & 0 \\ -a_{N-1} & 0 & \cdots & \cdots & 0 & 1 \\ -a_N & 0 & \cdots & \cdots & \cdots & 0 \end{bmatrix} \quad B = \begin{bmatrix} c_1 - a_1 \\ \vdots \\ \vdots \\ \vdots \\ c_N - a_N \end{bmatrix} \quad (\text{B.19})$$

$$C = [1 \quad 0 \quad \cdots \quad \cdots \quad \cdots \quad 0]$$

In this representation, the state vector x fulfills the **Markov** property: at time n , $x[n]$ caught all the information about the past $\{x[m] \mid m \leq n\}$. The Markov processes will be studied in chapter 3.

Once a model structure has been chosen, we have to estimate, from one sample path $y[1:n_t]$, the value of the parameters $(Q, a_1, \dots, a_{n_a}, c_1, \dots, c_{n_c})$. Then, the estimated DSP will be calculated by means of formula (B.16). Furthermore, the sample path has been encoded by means of $n_a + n_c + 1$ parameter; this can be useful for clustering and automatic diagnosis purposes. The estimated model can also be used for prediction. The problem of parameter estimation will be investigated in chapter 2.

Appendix C

More on probability theory

C.1 Bienaymé–Chebyshev inequality

Let x be a r.v. with value in \mathbb{R}^d , which has a mean and a variance;
for all $\varepsilon > 0$, and all positive definite matrix C :^{P1}

$$\text{Prob} \left((x - m_x)^\top C^{-1} (x - m_x) \geq \varepsilon \right) \leq \frac{\text{trace } C^{-1} C_{x,x}}{\varepsilon} \quad (\text{C.1})$$

With a probability P_0 , by choosing $\varepsilon = \frac{\text{trace } C^{-1} C_{x,x}}{1-P_0}$ in (C.1), we obtain, in the scalar case ($d = 1$):

$$\text{Prob} \left(x \in \left[m_x \mp \frac{\sigma_x}{\sqrt{1-P_0}} \right] \right) \geq P_0 \quad (\text{C.2})$$

The interval $\left[m_x \mp \frac{\sigma_x}{\sqrt{1-P_0}} \right]$, centered over m_x , is a confidence interval at at least P_0 .

In the general case, we obtain:

$$\text{Prob} \left((x - m_x)^\top C^{-1} (x - m_x) < \frac{\text{trace } C^{-1} C_{x,x}}{1-P_0} \right) \geq P_0 \quad (\text{C.3})$$

The ellipsoid $\{x \in \mathbb{R}^d \mid (x - m_x)^\top C^{-1} (x - m_x) < \frac{\text{trace } C^{-1} C_{x,x}}{1-P_0}\}$, centered over m_x , is a confidence domain at at least P_0 .

If $C_{x,x}$ is invertible, choosing $C = C_{x,x}$ provides the smallest confidence domain at at least P_0 which can be obtained by the Bienaymé–Chebyshev inequality; greater the variance is (Loewner order), bigger this smallest domain is.^{P2}

The variance is just an indication of a r.v. scattering, which permits to obtain, thanks to the Bienaymé–Chebyshev inequality, some too big confidence domains.

For example, the figure C.1 represents a population of 300 realizations of a zero-mean normal r.v. with variance $\begin{bmatrix} 4 & 8 \\ 8 & 25 \end{bmatrix}$, the confidence domain at 95% build over an isodensity contour (full line), and the confidence domains at at least 95% provided by the Bienaymé–Chebyshev inequality (dotted line) for $C = I_2$ and $C = C_{x,x}$.

P1. We use (A.5) and apply the Markov inequality page 7 to the r.v. $(x - m_x)^\top C^{-1} (x - m_x)$:

$$\begin{aligned} \mathbb{E}((x - m_x)^\top C^{-1} (x - m_x)) &= \mathbb{E}(\text{trace}(C^{-1} (x - m_x)(x - m_x)^\top)) = \text{trace}(\mathbb{E}(C^{-1} (x - m_x)(x - m_x)^\top)) \\ &= \text{trace}(C^{-1} \underbrace{\mathbb{E}((x - m_x)(x - m_x)^\top)}_{C_{x,x}}) \end{aligned}$$

P2. Let V be the volume of the ellipsoid defined in (C.3), and V_0 be the volume of a sphere with volume $\sqrt{\frac{d}{1-P_0}}$ in dimension d . By the change of variable $y = \sqrt{\frac{d}{\text{trace } C^{-1} C_{x,x}}} R^{-1} (x - m_x)$ where R is a square root of C ($C = R R^\top$):

$$V = \int_{\{x \in \mathbb{R}^d \mid (x - m_x)^\top C^{-1} (x - m_x) \leq \frac{\text{trace } C^{-1} C_{x,x}}{1-P_0}\}} dx = V_0 \sqrt{\left(\frac{\text{trace } C^{-1} C_{x,x}}{d} \right)^d \det C}$$

Thanks to (A.8) applied to the matrix $C^{-1} C_{x,x}$, we obtain that the volume is minimal for $C = C_{x,x}$.

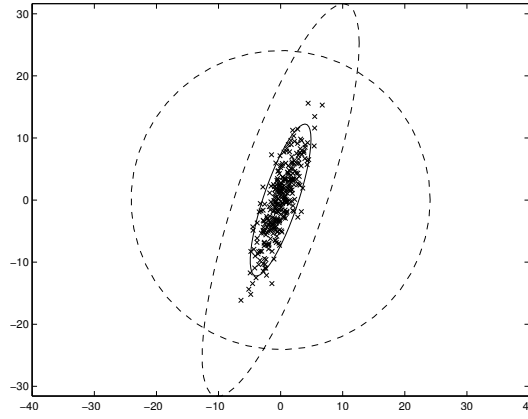


Figure C.1: Confidence domain for a bivariate normal r.v.

C.2 Characteristic functions

Let \mathbf{x} be a r.v. with value in \mathbb{R}^d , with PDF $p_{\mathbf{x}}$.

Its **first characteristic function** is defined, for all $\mathbf{v} \in \mathbb{R}^d$ such that the integral below converges, by:¹

$$\phi_{\mathbf{x}}(\mathbf{v}) = \mathbb{E}(\exp(j \mathbf{v}^T \mathbf{x})) = \int_{\mathbb{R}^d} p_{\mathbf{x}}(\mathbf{x}) \exp(j \mathbf{v}^T \mathbf{x}) \, d\mathbf{x}$$

The k th **moment** corresponds to the enumeration below:

$$\underbrace{(k_1, \dots, k_d)}_{\sum_{i=1}^d k_i = k} \in \mathbb{Z}^d \mapsto \frac{\partial^k \phi_{\mathbf{x}}}{\partial(jv_1)^{k_1} \dots \partial(jv_d)^{k_d}}(0_d) = \mathbb{E}(\mathbf{x}_1^{k_1} \dots \mathbf{x}_d^{k_d})$$

The first moment corresponds to the mean $m_{\mathbf{x}}$, the second moment corresponds to $\mathbb{E}(\mathbf{x} \mathbf{x}^T)$.

For a scalar r.v., the k th moment corresponds to $\mathbb{E}(x^k)$.

The Taylor series of $\phi_{\mathbf{x}}$ at 0_d is written as:²

$$\begin{aligned} \phi_{\mathbf{x}}(\mathbf{v}) &= 1 + \sum_{k=1}^{+\infty} \frac{1}{k!} \left[\sum_{i=1}^d (jv_i) \frac{\partial \phi_{\mathbf{x}}}{\partial(jv_i)}(0_d) \right]^{[k]} \\ &= 1 + (j\mathbf{v})^T m_{\mathbf{x}} + \frac{1}{2} (j\mathbf{v})^T \mathbb{E}(\mathbf{x} \mathbf{x}^T) (j\mathbf{v}) + (\text{order} > 2 \text{ terms}) \end{aligned}$$

The **second characteristic function** is the logarithm of the first one:

$$\psi_{\mathbf{x}} = \log \phi_{\mathbf{x}}$$

The k th **cumulant** corresponds to the enumeration below:

$$\underbrace{(k_1, \dots, k_d)}_{\sum_{i=1}^d k_i = k} \in \mathbb{Z}^d \mapsto \frac{\partial^k \psi_{\mathbf{x}}}{\partial(jv_1)^{k_1} \dots \partial(jv_d)^{k_d}}(0_d)$$

From the Taylor series of $\log(1 + \varepsilon)$ at 0, we obtain the relation between the moments and the cumulants.

The first cumulant corresponds to the mean $m_{\mathbf{x}}$, the second cumulant corresponds to the variance $C_{\mathbf{x}, \mathbf{x}}$.

The Taylor series of $\psi_{\mathbf{x}}$ at 0_d is written as:

$$\begin{aligned} \psi_{\mathbf{x}}(\mathbf{v}) &= \sum_{k=1}^{+\infty} \frac{1}{k!} \left[\sum_{i=1}^d (jv_i) \frac{\partial \psi_{\mathbf{x}}}{\partial(jv_i)}(0_d) \right]^{[k]} \\ &= (j\mathbf{v})^T m_{\mathbf{x}} + \frac{1}{2} (j\mathbf{v})^T C_{\mathbf{x}, \mathbf{x}} (j\mathbf{v}) + (\text{order} > 2 \text{ terms}) \end{aligned}$$

1. This is the Fourier transform, the inverse transform is: $p_{\mathbf{x}}(\mathbf{x}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \phi_{\mathbf{x}}(\mathbf{v}) \exp(-j \mathbf{v}^T \mathbf{x}) \, d\mathbf{v}$

2. We use the symbolic power notation;

e.g., the symbolic square is written as $\left[\sum_{i=1}^d v_i \frac{\partial \phi_{\mathbf{x}}}{\partial v_i}(\mathbf{v}) \right]^{[2]} = \sum_{i=1}^d \sum_{\ell=1}^d v_i v_{\ell} \frac{\partial^2 \phi_{\mathbf{x}}}{\partial v_i \partial v_{\ell}}(\mathbf{v})$.

A linear transform leads to:

$$\phi_{H(x-b)}(v) = \exp(-j v^T H b) \phi_x(H^T v) \quad \psi_{H(x-b)}(v) = -j v^T H b + \psi_x(H^T v)$$

In particular, from a standardized r.v. $\tilde{x} = \Sigma_x^{-1}(x - m_x)$ with $C_{x,x} = \Sigma_x \Sigma_x^T$:

$$\phi_x(v) = \exp(j v^T m_x) \phi_{\tilde{x}}(\Sigma_x^T v) \quad \psi_x(v) = j v^T m_x + \psi_{\tilde{x}}(\Sigma_x^T v) \quad (C.4)$$

If x and y are independent:

$$\phi_{x+y} = \phi_x \phi_y \quad \psi_{x+y} = \psi_x + \psi_y \quad (C.5)$$

This can be generalized to any number of independent r.v.

For a standardized univariate normal r.v., we have:^{P3}

$$\phi_x(v) = \exp\left(-\frac{v^2}{2}\right) \quad \psi_x(v) = -\frac{v^2}{2} \quad (C.6)$$

For a normal r.v., we have:^{P4}

$$\psi_x(v) = (jv)^T m_x + \frac{1}{2}(jv)^T C_{x,x} (jv)$$

Thus, for a normal r.v., all the cumulants of order > 2 are null.

C.3 Central limit theorem

Let $(x[n])_{n \in \mathbb{N}^*}$ a series of independent r.v with value in \mathbb{R}^d , with the same mean m and the same variance C .

When $n_t \rightarrow +\infty$, the r.v. $y = \frac{1}{\sqrt{n_t}} \sum_{n=1}^{n_t} (x[n] - m)$ is driven by the normal distribution with zero mean and variance C .^{P5}

C.4 Transformation of random variables

Let w be a r.v., with PDF (or PMF) p_w .

Let $\phi: w \mapsto \phi(w)$ be a function defined on the support of w .

What is the distribution of $\phi(w)$?

By means of the law of total probability, noting that $p_{\phi(w)|w}(y, w) = \delta(y - \phi(w))$, we obtain:

$$p_{\phi(w)}(y) = \sum_{w \in S(w)} \delta(y - \phi(w)) p_w(w) \quad \text{discrete case}$$

$$p_{\phi(w)}(y) = \int_{S(w)} \delta(y - \phi(w)) p_w(w) dw \quad \text{continuous case}$$

In the discrete case, the solution is simple and natural; for all y :

$$p_{\phi(w)}(y) = \sum_{w|\phi(w)=y} p_w(w)$$

In the continuous case, the problem is more complex, this book will give the solution in some special cases.

P3. The first characteristic function is $\phi_x(v) = \int_{\mathbb{R}} p_x(x) \exp(j v x) dx = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp(-\frac{x^2}{2}) \cos(v x) dx$

By derivating ϕ_x , and using integration by part, we obtain the differential equation $\phi'_x(v) + v \phi_x(v) = 0$.

With the initial condition $\phi_x(0) = 1$, we obtain (C.6).

P4. Thanks to the formulas C.4 and C.5.

P5. Under matrix form, $y = \underbrace{\left[\frac{1}{\sqrt{n_t}} I_d \quad \cdots \quad \frac{1}{\sqrt{n_t}} I_d \right]}_H \underbrace{\begin{bmatrix} x[1]-m \\ \vdots \\ x[n_t]-m \end{bmatrix}}_z$.

Thus, thanks to C.4 and C.5 (the components of z are independent), then by Taylor expansion for great n_t :

$$\phi_y(v) = \phi_z(H^T v) = \prod_{n=1}^{n_t} \phi_{x[n]} \left(\frac{1}{\sqrt{n_t}} v \right) = \phi_{x[1]}^{n_t} \left(\frac{1}{\sqrt{n_t}} v \right) = \left(1 - \frac{1}{2n_t} v^T C v + o(n_t^{-1}) \right)^{n_t}$$

where $o(n_t^{-1})$ is a function such that $n_t o(n_t^{-1})$ tends to 0 when n_t tends to $+\infty$.

Composing with an expansion of $\log(1 + \varepsilon)$ at 0:

$$\psi_y(v) = n_t \log \left(1 - \frac{1}{2n_t} v^T C v + o(n_t^{-1}) \right) = -\frac{1}{2} v^T C v + o(1) = \frac{1}{2} (jv)^T C (jv) + o(1)$$

where $o(1)$ is the function which tends to 0 when n_t tends to $+\infty$.

Thus, the second characteristic function ψ_y tends to the characteristic function of a zero mean normal distribution with variance C .

C.4.1 Differentiable and invertible transformation

If the function $\phi : w \mapsto \phi(w)$ is invertible and differentiable, then, for all $y \in S(w)$:

$$p_{\phi(w)}(\phi(w)) = \frac{1}{\left| \det \frac{\partial \phi}{\partial w^T}(w) \right|} p_w(w) \quad (\text{C.7})$$

By means of the change of variable $y = \phi(w)$, we obtain, for all $y \in S(\phi(w))$:^{P6}

$$p_{\phi(w)}(y) = \left| \det \frac{\partial \phi^{-1}}{\partial y^T}(y) \right| p_w(\phi^{-1}(y)) \quad (\text{C.8})$$

In the scalar case, a particular case corresponds to the case where $\phi = F_Y^{-1}$ (inverse of the cumulative distribution function of Y) and where W is uniformly distributed on $(0, 1)$. Then, $p_{F_Y^{-1}(W)} = p_Y$.

◁▷ **Exercise 36** (Cauchy distribution). Let U be a r.v. uniformly distributed between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$.

- Give the PDF of $\tan U$.
- What about the mean of $\tan U$?

◁▷ **Exercise 37** (Rayleigh distribution).

- Let U be a scalar-valued r.v. uniformly distributed between 0 and 1.
Give the PDF of $\sqrt{-2 \log U}$.
What is the name of this distribution?
- Give the PDF of $\sqrt{-2 \log U}$ (Rayleigh distribution with mean value $\sqrt{\frac{\pi}{2}}$).

◁▷ **Exercise 38** (Box-Muller transformation).

Let R and Θ be two independent r.v. R is Rayleigh distributed with mean value $\sqrt{\frac{\pi}{2}}$.
 Θ is uniformly distributed between 0 and 2π .
Give the PDF of the pair $(R \cos \Theta, R \sin \Theta)$.
What is the name of this distribution?

C.4.2 Linear invertible transformation

If ϕ is linear, that is $\phi(w) = Aw + b$ (then $\frac{\partial \phi}{\partial w^T}(w) = A$) with A invertible, the transformation becomes:

$$p_{Aw+b}(y) = \frac{1}{|\det A|} p_w(A^{-1}(y - b)) \quad (\text{C.9})$$

In particular, if the r.v. has a mean and a variance, by taking $b = -\Sigma_w^{-1} m_w$ and $A = \Sigma_w^{-1}$ where Σ_w is a square root of $C_{w,w}$ ($C_{w,w} = \Sigma_w \Sigma_w^T$, for example, the Cholesky decomposition), we obtain the standard score, that is a r.v. with zero mean and unit variance (the variance is the identity matrix). This operation is often useful in the scalar case, since it is often easier to perform calculation on the standard score, and to correct afterwards by means of the formulae below:

$$\text{with } \tilde{w} = \frac{w - m_w}{\sigma_w}, \quad \begin{cases} p_w(w) &= \frac{1}{\sigma_w} p_{\tilde{w}}\left(\frac{w - m_w}{\sigma_w}\right) \\ F_w(w) &= F_{\tilde{w}}\left(\frac{w - m_w}{\sigma_w}\right) \end{cases}$$

▲ In the multivariate case, the distribution of the standard score usually depends of the chosen square root of the variance.

◁▷ **Exercise 39** (Uniform distribution). In the exercise 9, page 8, we calculated the mean and the variance of a r.v. uniformly distributed between a and b . Let's find again the same result by another way.

- Give the mean and the variance of the uniform distribution between 0 and 1.
- Give a linear change of variable to transform a r.v. which is uniformly distributed between 0 and 1 into a r.v. which is uniformly distributed between a and b .
- Deduce the mean and the variance of a r.v. which is uniformly distributed between a and b .

P6. By means of the change of variable $w = \phi^{-1}(u)$ and the differential calculus rule (A.11):

$$p_{\phi(w)}(y) = \int_{\phi(S(w))} \delta(y - u) p_w(\phi^{-1}(u)) \left| \det \frac{\partial \phi^{-1}}{\partial y^T}(u) \right| du$$

We immediately obtain the formula (C.8), and, by means of the formula (A.10), the formula (C.7).

C.4.3 Non invertible transformation

If ϕ is not invertible, because the output set has a lower dimension than the input set, a solution can consist in completing this function to obtain an invertible function for which we can apply the formula (C.8), then to calculate the marginal distribution.

◁ **Exercise 40** (Distribution of a sum of random variables). Let (w_1, w_2) be a pair of real r.v.. The aim is to express the PDF of $w_1 + w_2$ by means of the joint PDF p_{w_1, w_2} .

- a) Let pose $Y_1 = w_1 + w_2$, $Y_2 = w_1 - w_2$. By means of the formula (C.9), give $p_{Y_1, Y_2}(y_1, y_2)$ in function of p_{w_1, w_2} .
- b) Deduce the marginal PDF $p_{Y_1}(y_1)$ (use the change of variable $w_1 = \frac{y_1 + y_2}{2}$).
- c) If w_1 et w_2 are independent, what is the operator to use between the marginal PDF p_{w_1} are p_{w_2} to obtain p_{Y_1} ?

Appendix D

More on estimation theory

D.1 Cramer-Rao inequality

We want to estimate a parameter x from an observation Y , thanks to the likelihood $p_{Y|X}$.

Let's see the regularity conditions below:

- R1** for all $x \in \mathbb{X}$, for all $y \in S(Y | X = x)$, the likelihood $p_{Y|X}(y, \cdot)$ is differentiable in x ;
- R2** for all $x \in \mathbb{X}$, for all $y \in S(Y | X = x)$, the gradient $\frac{\partial p_{Y|X}}{\partial x}(y, \cdot)$ is continuous in x ;
- R2+** as R2, but furthermore, the gradient $\frac{\partial p_{Y|X}}{\partial x}(y, \cdot)$ is differentiable and the hessian is continuous in x ;
- R3** the support $S(Y | X = x)$ is uniform (independent of x).

If the condition R1 is fulfilled, we call **score function** the function $\frac{\partial \ln p_{Y|X}}{\partial x}$, and we call **Fisher information** the function FI from \mathbb{X} in $\mathbb{R}^{d_x \times d_x}$ which returns a positive semi-definite matrix, defined by:

$$FI(x) = \text{Var} \left(\frac{\partial \ln p_{Y|X}}{\partial x}(Y, x) \mid x \right)$$

The Fisher information can also be written as below:

$$\begin{aligned} \text{If R2 and R3 are fulfilled} \quad 0_d &= E \left(\frac{\partial \ln p_{Y|X}}{\partial x}(Y, x) \mid x \right) \quad \text{donc} \\ FI(x) &= E \left(\left[\frac{\partial \ln p_{Y|X}}{\partial x}(Y, x) \right] \left[\frac{\partial \ln p_{Y|X}}{\partial x}(Y, x) \right]^T \mid x \right) \\ \text{If R2+ and R3 are fulfilled} \quad FI(x) &= -E \left(\frac{\partial^2 \ln p_{Y|X}}{\partial x \partial x^T}(Y, x) \mid x \right) \end{aligned}$$

Let's consider an estimator with a differentiable bias Bias^{lp} . If R1, R2 and R3 are fulfilled, thus the estimator variance is bounded by the Cramer-Rao bound, that is, for all x :

$$\text{Errvar}^{lp}(x) \geq \left[I_{d_x} + \frac{\partial \text{Bias}^{lp}}{\partial x^T}(x) \right] FI^{-1}(x) \left[I_{d_x} + \frac{\partial \text{Bias}^{lp}}{\partial x^T}(x) \right]^T \quad (D.1)$$

Furthermore, if the bound is reached, thus, for all x :

$$\hat{x}(Y) = x + \text{Bias}^{lp}(x) + \left[I_{d_x} + \frac{\partial \text{Bias}^{lp}}{\partial x^T}(x) \right] FI^{-1}(x) \frac{\partial \ln p_{Y|X}}{\partial x}(Y, x) \quad (D.2)$$

Thus, if an estimator reaches the Cramer-Rao bound, it is unique.

Proof We note $p = p_{Y|X}$, $\ell = \ln p_{Y|X}$, $\text{Bias}^{lp} = B$. We will often use the identity $\frac{\partial p}{\partial x} = \frac{\partial \ell}{\partial x} p$. The formulas are for all $x \in \mathbb{X}$. The integrals are over the uniform support $S(Y | X = x)$.

The integral of a PDF is 1: $1 = \int p(y, x) dy$. By differentiation:

$$0_{d_x} = \int \frac{\partial \ell}{\partial x}(y, x) p(y, x) dy = E \left(\frac{\partial \ell}{\partial x}(Y, x) \mid x = x \right) \quad (D.3)$$

With a new differentiation, we show the equivalence of the Fisher information formulas:

$$\begin{aligned} 0_{d_x \times d_x} &= \int \frac{\partial^2 \ell}{\partial x \partial x^\top}(y, x) p(y, x) \, dy + \int \left[\frac{\partial \ell}{\partial x}(y, x) \right] \left[\frac{\partial \ell}{\partial x}(y, x) \right]^\top p(y, x) \, dy \\ &= \mathbb{E} \left(\frac{\partial^2 \ell}{\partial x \partial x^\top}(Y, X) \mid X = x \right) + \mathbb{E} \left(\left[\frac{\partial \ell}{\partial x}(Y, X) \right] \left[\frac{\partial \ell}{\partial x}(Y, X) \right]^\top \mid X = x \right) \end{aligned}$$

The Jacobian matrix of the bias $B(x) = \int (\hat{x}(y) - x) p(y, x) \, dy$ is written as:

$$\begin{aligned} \frac{\partial B}{\partial x^\top}(x) &= \int (\hat{x}(y) - x) \left[\frac{\partial \ell}{\partial x}(y, x) \right]^\top p(y, x) \, dy - \int I_{d_x} p(y, x) \, dy \\ &= \mathbb{E} \left((\hat{x}(Y) - X) \left[\frac{\partial \ell}{\partial x}(Y, X) \right]^\top \mid X = x \right) - I_{d_x} \end{aligned} \quad (\text{D.4})$$

By combining (D.3) and (D.4):

$$\mathbb{E} \left([\hat{x}(Y) - X - B(X)] \left[\frac{\partial \ell}{\partial x}(Y, X) \right]^\top \mid X = x \right) = I_{d_x} + \frac{\partial B}{\partial x^\top}(x)$$

Let $A(x)$ be the positive semi-definite matrix defined by:

$$A(x) = \mathbb{E} \left(\begin{bmatrix} \hat{x}(Y) - X - B(X) \\ \frac{\partial \ell}{\partial x}(Y, X) \end{bmatrix} \begin{bmatrix} \hat{x}(Y) - X - B(X) \\ \frac{\partial \ell}{\partial x}(Y, X) \end{bmatrix}^\top \mid X = x \right) = \begin{bmatrix} \text{Errvar}^{\text{lp}}(x) & I_{d_x} + \frac{\partial B}{\partial x^\top}(x) \\ \left[I_{d_x} + \frac{\partial B}{\partial x^\top}(x) \right]^\top & \text{FI}(x) \end{bmatrix}$$

The Schur complement of $\text{Errvar}^{\text{lp}}(x)$ in $A(x)$ is positive semi-definite, this proves the Cramer-Rao inequality.

Let's define, for the sake of simplicity, $\beta(x) = I_{d_x} + \frac{\partial B}{\partial x^\top}(x)$. If the Cramer-Rao bound is reached, $A(x)$ is written as:

$$A(x) = \begin{bmatrix} \beta(x) & \text{FI}^{-1}(x) \beta^\top(x) \\ \beta^\top(x) & \text{FI}(x) \end{bmatrix}$$

We easily check:

$$\begin{bmatrix} I_{d_x} \\ -\text{FI}^{-1}(x) \beta^\top(x) \end{bmatrix}^\top A(x) \begin{bmatrix} I_{d_x} \\ -\text{FI}^{-1}(x) \beta^\top(x) \end{bmatrix} = 0_{d_x \times d_x}$$

Thus:

$$\mathbb{E} \left(\begin{bmatrix} I_{d_x} \\ -\text{FI}^{-1}(X) \beta^\top(X) \end{bmatrix}^\top \begin{bmatrix} \hat{x}(Y) - X - B(X) \\ \frac{\partial \ell}{\partial x}(Y, X) \end{bmatrix} \begin{bmatrix} \hat{x}(Y) - X - B(X) \\ \frac{\partial \ell}{\partial x}(Y, X) \end{bmatrix}^\top \begin{bmatrix} I_{d_x} \\ -\text{FI}^{-1}(X) \beta^\top(X) \end{bmatrix} \mid X = x \right) = 0_{d_x \times d_x}$$

The second order moment of a r.v. being null, this r.v. is (almost surely) null:

$$\begin{bmatrix} I_{d_x} \\ -\text{FI}^{-1}(x) \beta^\top(x) \end{bmatrix}^\top \begin{bmatrix} \hat{x}(Y) - x - B(x) \\ \frac{\partial \ell}{\partial x}(Y, x) \end{bmatrix} = 0_{d_x}$$

We obtain the formula (D.2).

D.2 If an efficient estimator exists, it is the MLE

If the estimator \hat{x} is efficient, its bias is uniformly null, the formula (D.2) is rewritten as:

$$\hat{x}(Y) = x + [\text{FI}(x)]^{-1} \frac{\partial \ln p_{Y|X}}{\partial x}(Y, x) \quad (\text{for an efficient estimator}) \quad (\text{D.5})$$

If this formula holds for all x , it is true for $\hat{x}_{\text{MLE}}(Y)$ which nullifies $\frac{\partial \ln p_{Y|X}}{\partial x}(Y, \hat{x}_{\text{MLE}}(Y))$.

Thus, if an efficient estimator exists, it is unique, and it is the MLE.

D.3 Estimation of the mean and the variance of an i.i.d sequence

Let's consider a sequence of r.v. $Y = (Y[n])_{1 \leq n \leq n_t}$, all of them with size d_y . The goal is to estimate the values of m^* and C^* taken by the parameters M et C , from Y , assuming that:

- given M and C , for all n , $Y[n]$ has the mean M and the variance C ;
- given M and C , the sequence $(Y[n])_{1 \leq n \leq n_t}$ is independent.

The empirical estimate of the mean is:

$$\hat{m}(Y) = \frac{1}{n_t} \sum_{n=1}^{n_t} Y[n]$$

The empirical estimate of the variance, if the mean is known and takes the value m , is:

$$\hat{C}_{\text{emp}}(Y, m) = \frac{1}{n_t} \sum_{n=1}^{n_t} (Y[n] - m) (Y[n] - m)^T$$

Thus, it is natural to use the estimator of the variance below:

$$\hat{C}(Y) = \begin{cases} \hat{C}_{\text{emp}}(Y, m^*) & \text{if } m^* \text{ is known} \\ \hat{C}_{\text{emp}}(Y, \hat{m}(Y)) & \text{otherwise} \end{cases}$$

These estimators are the ML ones under the assumption that $Y[n]$ is normally distributed;^{P1} furthermore, it is easy to check that the mean estimator is unbiased with variance $\frac{1}{n} C$ (even if $Y[n]$ is not normally distributed):

$$E(\hat{m}(Y) | M, C) = M \quad \text{Var}(\hat{m}(Y) - M | M, C) = \frac{1}{n_t} C$$

The variance estimator is unbiased only in the known mean case:

$$E(\hat{C}(Y) | M, C) = \begin{cases} C & \text{if } m^* \text{ is known} \\ \frac{n_t-1}{n_t} C & \text{otherwise} \end{cases}$$

In the unknown mean case, the estimator $\frac{n_t}{n_t-1} \hat{C}(Y)$ of the variance is unbiased.

In many applications, the parameter of interest is the mean. $\frac{1}{n_t} \hat{C}(Y)$ is an approximation of the estimator variance, and can be used to measure the confidence in the mean estimate.

D.4 Orthogonality principle

We consider the space of r.v. with value in \mathbb{X} , which x and all estimator $\hat{x}(Y)$ belong to. We use the inner product defined, for all x and x' , by $E(x^T x')$.

The MMSE estimator $\hat{x}_{\text{MMSE}}(Y)$ is the orthogonal projection of x on the estimators subspace.^{P2}

$$\forall \hat{x} : Y \rightarrow \mathbb{X} \quad E((\hat{x}_{\text{MMSE}}(Y) - x)^T \hat{x}(Y)) = 0 \quad (\text{D.6})$$

The LMMSE estimator $\hat{x}_{\text{LMMSE}}(Y)$ is the orthogonal projection of x on the linear estimators subspace.

$$\forall \hat{x}_{\text{LIN}} : Y \rightarrow \mathbb{X} \text{ linear} \quad E((\hat{x}_{\text{LMMSE}}(Y) - x)^T \hat{x}_{\text{LIN}}(Y)) = 0 \quad (\text{D.7})$$

Thus, only the LMMSE estimator fulfills:

$$\begin{cases} E(\hat{x}_{\text{LMMSE}}(Y) - x) = 0_{d_x} \\ E((\hat{x}_{\text{LMMSE}}(Y) - x) Y^T) = 0_{d_x \times d_y} \end{cases} \quad (\text{D.8})$$

P1. The log-likelihood is:

$$\begin{aligned} \log p_{Y|M,C}(y, m, C) &= \log \prod_{n=1}^{n_t} p_{Y[n]|M,C}(y[n], m, C) = \log \prod_{n=1}^{n_t} \frac{1}{\sqrt{\det(2\pi C)}} \exp \left[-\frac{1}{2} (y[n] - m)^T C^{-1} (y[n] - m) \right] \\ &= \log \left((\det(2\pi C))^{-\frac{n_t}{2}} \exp \left[-\frac{1}{2} \sum_{n=1}^{n_t} (y[n] - m)^T C^{-1} (y[n] - m) \right] \right) \\ &= -\frac{n_t}{2} \log \det(2\pi C) - \frac{1}{2} \sum_{n=1}^{n_t} (y[n] - m)^T C^{-1} (y[n] - m) = -\frac{n_t}{2} [\log \det(2\pi C) + \text{trace } C^{-1} \hat{C}_{\text{emp}}(y, m)] \end{aligned}$$

By cancelling the gradient of the penultimate form with respect to m , we obtain the empirical mean. In the last form, we retrieve the criterion (A.9), page 64, which is maximal for $C = \hat{C}_{\text{emp}}(y, m)$.

P2. Let $\phi(Y)$ be an estimator. For all estimator $\hat{x}(Y)$: $E((\phi(Y) - x)^T \hat{x}(Y) | Y) = (\phi(Y) - \hat{x}_{\text{MMSE}}(Y))^T \hat{x}(Y)$. By means of the total expectation formula: $E((\phi(Y) - x)^T \hat{x}(Y)) = E((\phi(Y) - \hat{x}_{\text{MMSE}}(Y))^T \hat{x}(Y))$. Thus, this quantity is 0 for $\phi = \hat{x}_{\text{MMSE}}$. Conversely, if it is 0 for all \hat{x} , it is 0 for $\hat{x} = \phi - \hat{x}_{\text{MMSE}}$. Thus, the norm of $\phi(Y) - \hat{x}_{\text{MMSE}}(Y)$ is 0, so $\phi(Y) = \hat{x}_{\text{MMSE}}(Y)$ almost surely.

Appendix E

Bayesian smoothing

The sequences $(x[n])_{n \geq 1}$ and $(Y[n])_{n \geq 1}$ are driven by a HMM with initial distribution $p_{x[1]}$, transition distribution $p_{x[n+1]|x[n]}$ and emission distribution $p_{Y[n]|x[n]}$.

We observed $Y[1:n_t]$. We express $p_{x[n]}^{n_t}$ to estimate $x[n]$ from past and future observations.

The algorithm contains a forward recursion and a backward one.

The forward recursion (n increases from 1 to n_t) is nothing but the Bayesian filtering, and returns $p_{x[n]}^n$, $1 \leq n \leq n_t$.

The backward recursion below (n decreases from $n_t - 1$ to 1), with the initialization $p_{x[n_t]}^{n_t}$, provides the solution:

$$p_{x[n+1]}^n(x^+) = \int p_{x[n+1]|x[n]}(x^+, x) p_{x[n]}^n(x) dx \quad (\text{E.1})$$

$$p_{x[n]}^{n_t}(x) = p_{x[n]}^n(x) \int \frac{p_{x[n+1]|x[n]}(x^+, x) p_{x[n+1]}^{n_t}(x^+)}{p_{x[n+1]}^n(x^+)} dx^+ \quad (\text{E.2})$$

The computation (E.1) is included in the Bayesian filtering and is not necessary if we stored $p_{x[n+1]}^n$, $1 \leq n \leq n_t - 1$.

For the linear model described by (3.15) page 43, the Bayes smoother is called the **Kalman smoother**, or the **Rauch–Tung–Striebel smoother** [27]. The forward recursion is the Kalman filter which provides $\hat{x}^n[n]$ and $P^n[n]$. The backward recursion is:

$$\hat{x}^{n_t}[n+1] = F_n \hat{x}^n[n] + f_n \quad (\text{E.3})$$

$$P^{n_t}[n+1] = F_n P^n[n] F_n^T + Q_n \quad (\text{E.4})$$

$$L[n] = P^n[n] F_n^T \left[P^{n_t}[n+1] \right]^{-1} \quad (\text{E.5})$$

$$\hat{x}^{n_t}[n] = \hat{x}^n[n] + L[n] (\hat{x}^{n_t}[n+1] - \hat{x}^n[n+1]) \quad (\text{E.6})$$

$$P^{n_t}[n] = P^n[n] + L[n] (P^{n_t}[n+1] - P^n[n+1]) L[n]^T \quad (\text{E.7})$$

The computations (E.3) et (E.4) are included in the Kalman filtering and are not necessary if we stored $\hat{x}^n[n+1]$ and $P^n[n+1]$, $1 \leq n \leq n_t - 1$.

(E.7) is also optional.

Appendix F

Matlab, Octave

- To plot of a PMF estimated over a scalar valued population in the vector x 4
`ux = unique(x); px = hist(x, ux)/length(x); stem(ux, px)`
- To plot a normalized histogram with n_b bins of a scalar valued population in the vector x 5
`[n,b] = hist(x,nb); bar(b, n/(b(2)-b(1))/sum(n), 1)`
- To plot the empirical CDF 21
`stairs([min(x);sort(x(:))], (0:length(x))/length(x))`
- To obtain an approximation of π (with a high n_r) 21
`4*mean(abs([1 j]*rand(2,nr))<1)`
- To plot the P_0 confidence ellipse (normal distribution with mean m and variance C) 24
`t = linspace(0,2*pi,100);
X = sqrt(-2*log(1-P0))*chol(C,'lower')*[cos(t); sin(t)] + m*ones(1,length(t));
plot(X(1,:),X(2,:))`
- To generate a Gaussian n_r -sample, with mean m and variance C 24
`x = chol(C,'lower')*randn(length(C),nr) + m*ones(1,nr);`
- To obtain an approximation of π (with a high n_r) 49
`4*mean(abs([1 j]*rand(2,nr))<1)`
- To generate a n_r -sample uniformy distributed between 0 and 1 (the seed is processed by the function `rng`) ... 50
`x = rand(1,nr);`
- To generate a n_r -sample driven by a zero mean and unit variance Gaussian distribution in \mathbb{R}^d 50
`x = randn(d,nr);`
- To generate a Gaussian n_r -sample, with mean m and variance C 50
`x = chol(C,'lower')*randn(length(C),nr) + m*ones(1,nr);`
- To generate a n_r -sample x of a discrete r.v. z which takes its value in $\{1, \dots, n_c\}$ such that $\text{Prob}(z = c) = \lambda_c$, together with one realization k with n_c components ($\lambda = (\lambda_1, \dots, \lambda_{n_c})$, the c th component of k is the number of occurrences of the value c in the array z) 51
`[k,z] = histc(rand(1,nr),[-Inf cumsum(lambda(1:end-1)) Inf]); k(end) = [];`
- Resampling in the particle filter 57
`[~,z] = histc(rand(1,nr),[-Inf cumsum(omega(1:end-1)) Inf]); x = x(:,z);`
- To define the PDF and the CDF on an univariate normal distribution with mean m and standard deviation σ 66
`pdfGauss = @(x, m, sigma) 1/sqrt(2*pi)/sigma*exp(-0.5*((x-m)/sigma).^2);
cdfGauss = @(x, m, sigma) 0.5 + 0.5*erf((x-m)/sigma/sqrt(2));`

Bibliography

- [1] C. Andrieu, M. Davy, and A. Doucet. Improved auxiliary particle filtering: applications to time-varying spectral analysis. In *Proceedings of the 11th IEEE Signal Processing Workshop on Statistical Signal Processing*, pages 309–312, Singapore, Aug. 2001. doi: 10.1109/SSP.2001.955284.
- [2] I. Arasaratnam and S. Haykin. Cubature Kalman filters. *IEEE Trans. Autom. Control*, 54:1254–1269, June 2009.
- [3] J. Baksalary. On the Lowner, Minus, and Star Partial Orderings of Nonnegative Definite Matrices and Their Squares. *Linear algebra and its applications*, 151:135–141, 1991.
- [4] J. M. Bernardo. Noninformative priors do not exist: A discussion. *J. Statistics Planning and Inference*, 65: 159–189, 1997.
- [5] F. Campillo. Filtrage particulaire & modèles de markov cachés. Cours de master, 2006. URL <ftp://ftp.irisa.fr/local/sigma2/campillo/cours/2006-master2-toulon.pdf>.
- [6] G. Casella and R. Berger. *Statistical Inference*. Brooks/Cole, 2nd edition, 2001.
- [7] S. Chakraborty. Some Applications of Dirac’s Delta Function in Statistics for More Than One Random Variable. *Applications and Applied Mathematics*, 3(1):42–54, 2008.
- [8] A. Dawid. Conditional independance in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31, 1979.
- [9] P. Druilhet. Support du cours de statistique inférentielle. ENSAI, Rennes, 2004.
- [10] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. Biological sequence analysis: probabilistic models of proteins and nucleic acids, 1998.
- [11] D. Fink. A compendium of conjugate priors, 1997.
- [12] G. D. Forney. The Viterbi algorithm. *Proc. of the IEEE*, 61:268 – 278, Mar. 1973.
- [13] D. Freedman and P. Diaconis. On the histogram as a density estimator: L2 theory. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 57:453–476, 1981.
- [14] J. Garcia, Z. Kutalik, K. Cho, and O. Wolkenhauer. Level sets and minimum volume sets of probability density functions. *International Journal of Approximate Reasoning*, 34(1):25 – 47, 2003. ISSN 0888-613X. doi: DOI:10.1016/S0888-613X(03)00052-5. URL <http://www.sciencedirect.com/science/article/B6V07-48WPVM7-1/2/0f5d45074298bea8cc4e5fd4e7655ceb>.
- [15] F. Gustafsson and G. Hendeby. Some relations between extended and unscented kalman filters. *IEEE Transactions on Signal Processing*, 60(2):545–555, Feb. 2012.
- [16] J. Hauke and A. Markiewicz. On Orderings Induced by the Loewner Partial Ordering. *Applicationes Mathematicae*, 22(2):145–154, 1994.
- [17] J. D. Hol, T. B. Schon, and F. Gustafsson. On resampling algorithms for particle filters. In *Nonlinear Statistical Signal Processing Workshop, 2006 IEEE*, pages 79–82, Sept. 2006. doi: 10.1109/NSSPW.2006.4378824.
- [18] A. Høst-Madsen. On the existence of efficient estimators. *IEEE Transactions on Signal Processing*, 48:3028–3031, Jan. 2000.
- [19] A. Johansen and L. Evers. Monte Carlo Methods. Lecture Notes, Department of Mathematics, University of Bristol, Nov. 2007.

- [20] S. Julier and J. Uhlmann. A new extension of the Kalman filter to nonlinear systems. In *Int. Symp. Aerospace/Defense Sensing, Simul. and Controls, Orlando, FL*, 1997. URL <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.5.2891>.
- [21] S. Julier, J. Uhlmann, and F. Durrant-Whyte. A New Method for the Nonlinear Transformation of Means and Covariances in Filters and Estimators. *IEEE Transactions on Automatic Control*, 45(3):477–482, Mar. 2000.
- [22] Y. Lee. Graphical Demonstration of an Optimality Property of the Median. *The American Statistician*, 49(4):369–372, Nov. 1995.
- [23] G. Marsaglia and W. Tsang. A Simple Method for Generating Gamma Variables. *ACM Transactions on Mathematical Software*, 26(3):363–372, Sept. 2000.
- [24] M. Pitt and N. Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446):590–599, 1999.
- [25] C. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer, 2007.
- [26] G. Saporta. *Probabilités, analyse des données et statistique*. Technip, 2006.
- [27] S. Särkkä. *Bayesian Filtering and Smoothing*. Cambridge University Press, 2013.
- [28] C. G. Small. A survey of multidimensional medians. *International Statistical Review / Revue Internationale de Statistique*, 58(3):263–277, 1990. ISSN 03067734, 17515823. URL <http://www.jstor.org/stable/1403809>.
- [29] O. Straka, J. Dunik, and M. Simandl. Scaling parameter in unscented transform: Analysis and specification. In *American Control Conference (ACC)*, pages 5550–5555, Montreal, QC, June 2012.
- [30] F. van Perlo-ten Kleij. *Contributions to multivariate analysis with applications in marketing*. PhD thesis, University of Groningen, 2004. URL <http://irs.ub.rug.nl/ppn/270867074>.
- [31] M. Verhaegen and P. Van Dooren. Numerical aspects of different Kalman filter implementations. *IEEE Transactions on Automatic Control*, 31(10):907–917, 1986. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1104128.