



**HAL**  
open science

## Information theory, Signal Analysis and Inverse Problem

Dominique Gibert, Fernando Lopes, Vincent Courtillot, Jean-Baptiste Boulé

► **To cite this version:**

Dominique Gibert, Fernando Lopes, Vincent Courtillot, Jean-Baptiste Boulé. Information theory, Signal Analysis and Inverse Problem. arXiv, 2024, 10.48550/arXiv.2408.16361 . hal-04748228

**HAL Id: hal-04748228**

**<https://hal.science/hal-04748228v1>**

Submitted on 22 Oct 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

# Information Theory

## Signal Analysis and Processing

### Inverse Problem

Gibert Dominique, Lopes Fernando, Courtilot Vincent et Boulé Jean-Baptiste





---

# CONTENTS

<b>1</b>	<b>The Fourier transform</b>	<b>15</b>
1	Definition of the Fourier Transform . . . . .	17
2	Mathematical Physics . . . . .	17
3	Orthogonal Functions . . . . .	19
4	The Fourier Transform . . . . .	21
4.1	Theoretical Foundations and Definitions . . . . .	21
4.2	Notations . . . . .	23
4.3	Example: Extension of Potential Fields . . . . .	24
4.4	Break: The Hartley Transform . . . . .	29
5	Fourier Series . . . . .	31
5.1	Theoretical Foundations and Definitions . . . . .	31
5.2	Example: Vibrations of a Taut String . . . . .	32
6	Properties of the Fourier Transform . . . . .	35
6.1	Linearity . . . . .	36
6.2	Symmetries . . . . .	36
6.3	Similarity . . . . .	38
6.4	Translation . . . . .	41
6.5	Differentiation . . . . .	42
7	Multidimensional Fourier Transforms . . . . .	42
7.1	Example: Extension of Potential Fields . . . . .	42
7.2	General Definitions . . . . .	43

7.3	Sign Conventions in Space-Time	43
<b>2</b>	<b>Convolution and Correlation</b>	<b>45</b>
1	Convolution	46
1.1	Where Do We Encounter Convolutions?	46
1.2	Spatial Convolution	50
1.3	Convolution and Probability	51
1.4	Properties of Convolution	51
1.5	Fourier Transform of a Convolution	52
1.6	Differentiation of a Convolution	54
2	Correlation	54
<b>3</b>	<b>The Hilbert Transform</b>	<b>57</b>
1	Definition	57
2	Formulae: Hilbert Transforms	60
<b>4</b>	<b>Useful Functions in Fourier Analysis</b>	<b>61</b>
1	Catalogue of Useful Functions	62
2	Window (the "scissors")	62
3	Cardinal Sine	64
4	Triangle	65
5	Exponential Functions	67
5.1	Exponential Decaying to Infinity	67
5.2	Gaussian	67
6	Dirac Delta Function (the "photo")	68
7	Sign Function	72
8	Heaviside Distribution (the switch)	73
9	Dirac Comb (the camera)	75
10	Sine and Cosine Functions	77
11	Form: Fourier Transforms	77
<b>5</b>	<b>Sampling</b>	<b>79</b>
1	Sampling	80
1.1	Signal truncation	80
1.2	Discretization	82

1.3	Correct Discretization: Shannon Interpolation . . . . .	83
1.4	Incorrect Discretization: Spectral Aliasing . . . . .	84
1.5	Analog-to-Digital Conversion: Quantization . . . . .	86
<b>6</b>	<b>The Z-Transform</b>	<b>89</b>
1	The Utility of the Z-Transform . . . . .	90
2	Formulary: Z-Transforms . . . . .	91
<b>7</b>	<b>The Discrete Fourier Transform</b>	<b>93</b>
1	The Discrete Fourier Transform . . . . .	94
1.1	Discretization of the Fourier Transform . . . . .	94
1.2	The Fast Fourier Transform Algorithm . . . . .	95
<b>8</b>	<b>Stochastic Processes</b>	<b>103</b>
1	Definition of Stochastic Processes . . . . .	104
2	1/f Noise . . . . .	105
3	White Noise . . . . .	107
4	Brownian Noise . . . . .	109
5	Pink Noise . . . . .	110
6	Black Noise . . . . .	111
7	Stable Laws (Gauss, Cauchy, <i>etc</i> ) . . . . .	111
<b>9</b>	<b>Time-Frequency Duality</b>	<b>117</b>
1	Measuring signal duration . . . . .	119
2	The Uncertainty Principle in Signal Processing . . . . .	121
2.1	Deterministic Approach . . . . .	121
3	Causal signal duality . . . . .	123
4	Minimum Delay Signals . . . . .	124
4.1	Utility of Minimum Delay Signals . . . . .	124
5	Case of Continuous Signals . . . . .	125
6	Case of Discrete Signals . . . . .	125
7	The Cepstral Domain . . . . .	129
<b>10</b>	<b>Linear Filtering</b>	<b>131</b>
1	Filters and Z-Transforms . . . . .	132
2	Operator and Filters in Numerical Analysis . . . . .	135

3	Narrowband Filters . . . . .	136
3.1	Recursiveness and Infinite Impulse Response . . . . .	136
4	Filter Stability . . . . .	141
4.1	Back to narrow band filter . . . . .	141
4.2	The general case . . . . .	142
5	Butterworth Filters . . . . .	144
6	General Overview . . . . .	144
7	The Bilinear Transformation . . . . .	145
8	An example . . . . .	147
9	Wiener Filters . . . . .	150
10	Wiener Filtering in the Frequency Domain . . . . .	150
<b>11</b>	<b>Spectral analyses</b>	<b>155</b>
1	The spectral analysis models . . . . .	156
1.1	<i>Prony, Hildebrand, Pisarenko and Schuster</i> : Trigonometric series . . . . .	156
1.2	<i>Burg, Pisarenko,...</i> : autoregressive models . . . . .	158
2	Discrete Fourier Transform Analysis . . . . .	160
2.1	Schuster's periodogram . . . . .	160
2.2	Signal truncation effects . . . . .	162
2.3	Apodization windows . . . . .	164
2.4	Impact of a trend . . . . .	167
2.5	Statistical issues . . . . .	168
3	Autoregressive model analysis . . . . .	170
3.1	The prediction error filter . . . . .	171
3.2	Prediction error filter utility . . . . .	172
<b>12</b>	<b>Wavelet transform analysis</b>	<b>175</b>
1	Wavelets: A brief history . . . . .	177
1.1	Recent history . . . . .	177
1.2	From Joseph Fourier to Dennis Gabor . . . . .	178
1.3	From Dennis Gabor to Jean Morlet . . . . .	181
1.4	Questions addressed in this chapter . . . . .	183
2	Continuous Wavelets – Discrete Wavelets – Orthogonal Wavelets . . . . .	184
2.1	Continuous Wavelet Transform . . . . .	184

2.2	Orthogonal Wavelets . . . . .	185
3	How is the Orthogonal Wavelet Transform computed? . . . . .	187
3.1	The pyramid algorithm . . . . .	187
3.2	Quadrature Mirror Filters . . . . .	190
3.3	The inverse transform . . . . .	190
4	Filter, denoise and compress signals using orthogonal wavelets . . . . .	191
5	How do you filter with the continuous wavelet transform ? . . . . .	193
5.1	The Reconstruction Formula . . . . .	193
5.2	The reproducing kernel . . . . .	194
6	Asymptotic signal analysis . . . . .	195
6.1	Signaux asymptotiques . . . . .	195
6.2	Asymptotic wavelet analysis . . . . .	196
6.3	The stationary phase method . . . . .	198
6.4	The Wavelet Transform Ridge . . . . .	199
6.5	Use of non-asymptotic wavelets . . . . .	200
<b>13</b>	<b>Singular Spectrum Analysis</b>	<b>203</b>
1	Singular Spectrum Analysis (SSA) . . . . .	204
1.1	Simple algorithm presentation . . . . .	204
1.2	To see how this works in practice . . . . .	207
2	Analysis of the 4 stages of SSA . . . . .	211
2.1	Embedding . . . . .	211
2.2	Singular Value Decomposition . . . . .	212
2.3	Grouping of SVD components . . . . .	214
3	What SSA can do . . . . .	215
3.1	Trend extraction . . . . .	215
3.2	Pseudo cycle separation . . . . .	217
3.3	Nonlinear Filtering . . . . .	221
<b>14</b>	<b>Inverse Problem</b>	<b>225</b>
1	Introduction . . . . .	226
2	An inverse problem example . . . . .	226
3	General structure of inverse problems . . . . .	228
4	A little bit of history . . . . .	232



4.1	The 1960s . . . . .	232
4.2	The 1970s . . . . .	232
4.3	The 1980s . . . . .	233
4.4	The 1990s . . . . .	233
4.5	The 2000s . . . . .	234
5	Our philosophy . . . . .	234
<b>15</b>	<b>Information &amp; Inverse Problems</b>	<b>235</b>
1	The definition of information . . . . .	237
1.1	Information and Complexity . . . . .	237
1.2	Information and Probabilities . . . . .	238
1.3	Equally likely answers = maximum information . . . . .	240
1.4	About the Tunnel . . . . .	242
2	Mutual Information . . . . .	242
2.1	Coupling Information . . . . .	242
2.2	Conditional information . . . . .	245
2.3	About the Tunnel . . . . .	246
3	Case of continuous distributions . . . . .	247
3.1	There is a problem ! . . . . .	247
3.2	A new definition of Information . . . . .	248
3.3	Information conjunction . . . . .	249
4	Direct problem = information . . . . .	250
4.1	Still in the tunnel . . . . .	250
4.2	Direct problem = conditional probability . . . . .	251
5	Inverse problem = information transfer . . . . .	251
5.1	<i>a posteriori</i> conditional information . . . . .	251
5.2	The Bayes formula (discrete events) . . . . .	252
5.3	The generalised Bayes formula (continuous case) . . . . .	252
<b>16</b>	<b>Bayesian inversion</b>	<b>255</b>
1	Probabilities & Inverse Problems . . . . .	257
1.1	Probabilities, Frequencies, and Information . . . . .	257
1.2	Probability Densities . . . . .	257
1.3	Mathematical expectation value of a function . . . . .	260

1.4	Multivariate probabilities . . . . .	260
2	A few common probability distributions . . . . .	261
2.1	The normal distribution (Gauss) . . . . .	261
2.2	Generalised Gaussian distributions . . . . .	261
2.3	The <i>log-normal</i> distribution . . . . .	261
2.4	The Poisson distribution . . . . .	262
2.5	The <i>gamma</i> ( $\Gamma$ -) distribution . . . . .	262
2.6	The beta ( $\beta$ -) distribution . . . . .	262
2.7	The Pareto distribution . . . . .	263
2.8	The <i>binomial</i> distribution . . . . .	263
2.9	The Cauchy distribution . . . . .	264
2.10	The Weibull distribution . . . . .	264
3	Bayes' formula and inversion" . . . . .	264
3.1	General solution . . . . .	264
3.2	Solution for <i>a priori</i> independent data and parameters . . . . .	265
3.3	Solution for an exact physical law . . . . .	266
3.4	Solution using Bayes' formula . . . . .	266
4	The tunnel again . . . . .	267
4.1	Example 1: one data and one parameter . . . . .	267
4.2	Example 2: two data and one parameter . . . . .	269
4.3	Example 3: One data and two parameters . . . . .	272
4.4	Example 4: Two data and two parameters . . . . .	276
5	Summary of examples 1, 2, 3 and 4 . . . . .	277
<b>17</b>	<b>Monte Carlo Methods</b>	<b>279</b>
1	Introduction . . . . .	280
2	Integration by the Monte Carlo method . . . . .	281
3	Metropolis algorithm . . . . .	284
3.1	Importance sampling . . . . .	284
3.2	Markov chain . . . . .	285
3.3	The Metropolis algorithm . . . . .	286
3.4	Example . . . . .	288
3.5	Example of the tunnel . . . . .	289

<b>18</b>	<b>Simulated Annealing</b>	<b>291</b>
1	Aim of the method . . . . .	292
2	Control temperature . . . . .	292
3	Perturbing the models . . . . .	293
4	The Simulated Annealing algorithm . . . . .	294
5	Example: the traveling salesman problem . . . . .	294
5.1	Introduction . . . . .	294
5.2	Generation of models . . . . .	295
5.3	Example of how to operate . . . . .	295
<b>19</b>	<b>Methods of Least Squares</b>	<b>299</b>
1	Introduction . . . . .	300
2	Linear problem: the normal equations . . . . .	301
3	Singular Value Decomposition & Singular Vectors . . . . .	303
4	Solution provided by the spectral decomposition . . . . .	305
5	Obtained solution properties . . . . .	306
6	Example: Signal deconvolution . . . . .	307
<b>20</b>	<b>Generation of <i>a priori</i> models</b>	<b>311</b>
1	Introduction . . . . .	312
2	Convex sets: Definitions . . . . .	312
3	Projections onto convex sets . . . . .	313
3.1	Imposed Values . . . . .	313
3.2	Valeurs bornées . . . . .	313
3.3	Discontinuity . . . . .	313
3.4	Sequencing . . . . .	314
3.5	Imposed Mean . . . . .	314
3.6	Maximum Energy . . . . .	315

---

# INTRODUCTION

Signal processing is a field that is difficult to describe in a few words. However, throughout these pages, we will see that the essential goal of signal processing techniques is to [separate a message from noise](#).

This definition assumes that we know what the desired message is or what noise needs to be eliminated, which necessitates relying on *a priori* considerations borrowed from the physics or chemistry of the problems being addressed. Without a fine understanding of the issues under study, the most sophisticated signal processing techniques in the world risk "spinning their wheels." It is important to understand the relationships between signal processing and chemistry and physics, as this is where the models used in signal processing find their justification. In our quest to isolate the message, we will see that it is often wise to transport the information carried by the signal into another "world," dual to the initial "world," where the information becomes more readable. The most well-known example is probably the transition from the time domain to the frequency domain *via* the Fourier transform . The choice of the "host" world depends heavily on our *a priori* knowledge of the problem at hand, and signal processing methods are like "glasses" through which we view a "landscape" of information where entities need to be recognized. Therefore, we will work on crafting glasses suited to our vision that provide the clearest possible images of the landscape. But that is not enough: you can process beautiful images of California obtained by the [SPOT satellite\\*](#) and recognize very nice "roads" but not the slightest fault if you do not have this concept in mind. It is not the role of signal processing to conceptualize the entities to be recognized in the landscape of information,

---

\*Satellite for Earth Observation. A family of French Earth observation satellites, initially launched between 1985 and 2002, and later between 2012 and 2014. URL: <http://www.intelligence-airbusds.com/en/99-spotmaps-high-resolution-colour-satellite-images>

but that of other disciplines such as geology, chemistry, and physics. A first piece of advice: [learn a bit of signal processing and a lot of other things!](#)

The choice of the dual space in which to transport the information contained in the signal critically depends on the models adopted to represent the signals. This modeling, explicit or implicit, allows signal processing techniques to be integrated into the theory of inverse problems. This approach is very beneficial for understanding the importance of choosing signal models and for clarifying the notion of resolution. Some classic signal processing problems, such as deconvolution, directly fall under the theory of inverse problems and are better understood in this context.

In practice, the array of available signal processing techniques allows for a progressive approach and gradually clarifies the understanding of a particular signal. In all cases, the physics of the phenomena causing the signal provides valuable insights into the nature of the message to be extracted. For example, in satellite altimetry, a geophysicist aiming to study the geoid will seek to correct the undulations of the sea surface for their temporal variability, which is precisely the signal of interest to the oceanographer studying ocean currents. This antagonism of objectives can be illustrated in all branches of global physics: [one's signal may be another's noise](#). We touch here on a very general human principle. The signals studied in geophysics are extremely varied and require a vast array of processing methods. As a result, many techniques are employed by geophysicists, and sometimes, when the standard array is no longer sufficient, some of them develop new methodologies that prove to be very broad in scope. This is the case with methods based on the criterion of entropy maximization, for example, or the case of wavelets. There are also rediscoveries such as the "Sompi" method ([Kumazawa et al., 1990](#)), which closely resembles the method invented by Baron de [Prony](#)—itself close to Fourier analysis ([Hauer et al., 1990](#))—in 1795!

This course should be considered an introduction aimed at raising awareness among geophysicists dealing with signals. I have followed a classic approach based on the Fourier transform, from which I develop a number of "selected pieces" chosen either for their universal character (sampling, the uncertainty principle, *etc*) or for their great practical utility (linear filtering, spectral analysis, *etc*). The choice of the Fourier transform is both simple and in line with what is generally done in the literature on signal processing. Nevertheless, it remains debatable as sine and cosine functions, which have an unbounded support, do not always have a physical meaning. However, these functions have the immense advantage of being the [eigenfunctions of most of the major partial differential equations in mathematical physics](#) expressed in Cartesian coordinates. This is what makes them successful, along with plane harmonic waves in seismology. But the Earth is round, we drill cylindrical wells, and Cartesian coordinates are not always the best suited. We then have to abandon them along with

sine and cosine functions, which give way to spherical harmonics, Bessel functions , *etc* . Many geophysical signals must therefore be processed using models other than the Fourier transform (wavelets, spherical harmonics, *etc* ), but many points covered in this course (sampling, aliasing, duality, *etc* ) remain valid and adaptable to these function bases. Some readers will undoubtedly find this course scandalously incomplete. This is the result of a simple principle to which I have adhered unflinchingly: I only discuss techniques that I have personally used. It seemed indispensable to me, in a course with a practical aim, to adopt such a principle because merely reading the specialized literature generally does not provide a precise idea of the operational character of the theories developed there. This concern to help the reader form a personal opinion is concretized by the fact that they can recreate all the figures in this book using the [Matlab®](#) functions accompanying the book. It is, of course, possible to change the initial parameter values to test the limits of the presented techniques. These functions can also be used to carry out a number of additional practical exercises.



---

---

# CHAPTER 1

---

## THE FOURIER TRANSFORM

<b>1</b>	<b>Definition of the Fourier Transform</b> . . . . .	<b>17</b>
<b>2</b>	<b>Mathematical Physics</b> . . . . .	<b>17</b>
<b>3</b>	<b>Orthogonal Functions</b> . . . . .	<b>19</b>
<b>4</b>	<b>The Fourier Transform</b> . . . . .	<b>21</b>
4.1	Theoretical Foundations and Definitions . . . . .	21
4.2	Notations . . . . .	23
4.3	Example: Extension of Potential Fields . . . . .	24
4.4	Break: The Hartley Transform . . . . .	29
<b>5</b>	<b>Fourier Series</b> . . . . .	<b>31</b>
5.1	Theoretical Foundations and Definitions . . . . .	31
5.2	Example: Vibrations of a Taut String . . . . .	32
<b>6</b>	<b>Properties of the Fourier Transform</b> . . . . .	<b>35</b>
6.1	Linearity . . . . .	36
6.2	Symmetries . . . . .	36
6.3	Similarity . . . . .	38
6.4	Translation . . . . .	41



6.5	Differentiation . . . . .	42
<b>7</b>	<b>Multidimensional Fourier Transforms . . . . .</b>	<b>42</b>
7.1	Example: Extension of Potential Fields . . . . .	42
7.2	General Definitions . . . . .	43
7.3	Sign Conventions in Space-Time . . . . .	43

### 1 Definition of the Fourier Transform

Almost all works on signal processing are based on the Fourier transform, which associates with a function  $f(t)$  its Fourier transform  $F(u)$ . The expressions we have adopted for the direct and inverse Fourier transforms are those used by Bracewell in his book (Bracewell et Bracewell, 1986). They have the advantage of being symmetric and easy to remember.

$$F(u) = \int_{-\infty}^{+\infty} f(t) \exp(-2i\pi ut) dt \tag{1.1}$$

where  $u \in \mathbb{R}$  is referred to as the frequency. For a wide class of functions,  $f(t)$ , the above integral equation is invertible and the original function can be reconstructed using the inverse Fourier transform,

$$f(t) = \int_{-\infty}^{+\infty} F(u) \exp(+2i\pi ut) du \tag{1.2}$$

Many signal processing operations involve computing the Fourier transform of the signal, inspecting it, applying a series of simple operations to it, and finally reconstructing the processed signal by computing an inverse Fourier transform. Faced with this approach, a novice\* often wonders: "Why this Fourier transform? Why not my Zébulon-Klack transform", defined by the following relation,

$$ZK(\zeta, \chi) = \int_{-\infty}^{+\infty} \arctan[f^2(t)] (\cosh \zeta + \sinh \chi) dt, \tag{1.3}$$

of which I am very proud!". I have never had to use the Zébulon-Klack transform, but there may be a domain in mathematical physics where it is quite useful. Why not, since it is precisely in mathematical physics that the Fourier transform finds its justification. Ultimately, things are not as definitive as they might first appear, and it is important to explore the domain where the Fourier transform proves to be useful.

### 2 Mathematical Physics

The language of physics is constructed using mathematics. The laws of physics are expressed in the form of equations, which physicists spend considerable time solving within various contexts

---

\*That is, someone who dares to ask the right questions!

of complexity. The same law can be presented in very different mathematical forms. For instance, Newton's law of universal gravitation can be written as

$$\|\vec{f}\| = G \frac{m_1 m_2}{r^2} \quad (2.1)$$

where  $G$  is the universal gravitational constant, and  $\vec{f}$  is the mutual attraction force between the two masses  $m_1$  and  $m_2$  separated by the distance  $r$ . However, one can also describe the law of universal gravitation using Poisson's equation ,

$$\nabla^2 \Phi = -4\pi G \rho \quad (2.2)$$

where  $\rho$  represents the mass density of the material in the considered region, and  $\Phi$  is a potential whose gradient provides the gravitational attraction. A similar approach can be applied to the laws of electromagnetism, *etc* Poisson's equation (2.2) is a partial differential equation that allows for a local formulation of gravitation within the framework of field theory. This local expression of physical laws is generally more satisfying to the mind as it removes the "magical" notion of action at a distance. I do not intend to delve further into this fascinating subject; interested readers may profitably consult the works of Feynman (1980)\* or Thom et Noël (1991)†. This topic is often present in non-local formulations which are extensively used for practical reasons (e.g., geometric optics and ray theory in seismology). Partial differential equations are ubiquitous in physics, and it is remarkable that a few of these equations cover a vast range of mathematical physics, as illustrated by the chapter "The Same Equations Have the Same Solutions" in the physics course by Feynman *et al.* (2013). We shall mention only, the Laplace equation,

$$\nabla^2 \psi = 0 \quad (2.3)$$

the wave equation,

$$\nabla^2 \psi - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \psi = 0 \quad (2.4)$$

---

\*"The Nature of Physics"

†"Predicting is not Explaining"

and the diffusion equation,

$$\nabla^2 \psi - \frac{1}{\kappa} \frac{\partial}{\partial t} \psi = 0. \quad (2.5)$$

The solution of these equations, that is, finding the field  $\psi$  while considering boundary conditions, initial conditions, etc., can only be achieved numerically in complex cases. Simple cases can be handled analytically through methods such as separation of variables and Green's functions (see, for instance, the books by [Morse et Feshbach \(1953\)](#)). In this chapter and the one on convolution, we will see that these two techniques bestow a particular status upon the Fourier transform, though not upon the Zébulon-Klack transform!

## 3 Orthogonal Functions

The method of separation of variables, pioneered by Bernoulli in the mid-18th century ([Bernoulli, 1753](#)), involves selecting a coordinate system (Cartesian, spherical, cylindrical, etc.) in which the unknown field is expressed as the product of functions, each depending on only one coordinate,

$$\psi(x, y, z, t) = f(x) \cdot g(y) \cdot h(z) \cdot s(t) \quad (3.1)$$

When this solution form is substituted into the partial differential equation to be solved, it results in a system of differential equations coupled by arbitrary constants, referred to as "separation constants." The analytical form of the partial differential equation and the resulting differential equations depends on the choice of coordinate system. For example, in Cartesian coordinates  $(x, y, z)$ , the wave equation is

$$\frac{\partial^2}{\partial x^2} \psi + \frac{\partial^2}{\partial y^2} \psi + \frac{\partial^2}{\partial z^2} \psi - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \psi = 0 \quad (3.2)$$

whereas in spherical coordinates  $(r, \theta, \phi)$ , it is written as

$$\frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial \psi}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial \psi}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 \psi}{\partial \phi^2} - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \psi = 0 \quad (3.3)$$

In all cases, the coupled differential equations can be expressed in the form of a Sturm-Liouville

equation,

$$\frac{d}{dt} \left[ l(t) \frac{\partial P}{\partial t} \right] + [m(t) + u \cdot n(t)] P(t) = 0 \quad (3.4)$$

where  $u$  is the separation constant and the functions  $l(t)$ ,  $m(t)$ , and  $n(t) > 0$  are determined by the chosen coordinate system. The solutions,  $P$ , are as numerous as the allowed values for the separation constant, denoted as  $P(t | u)$ . These solutions are known as the eigenfunctions of the differential equation. They have the important property of being mutually orthogonal, meaning they satisfy

$$\int n(t) P(t | u_1) P^*(t | u_2) dt = 0 : \text{if } u_1 \neq u_2 \quad (3.5)$$

where  $*$  denotes the complex conjugate and a weighted inner product with the function  $n(t)$  is used. The bounds of the integral above depend on the range of the solutions being sought. The solution,  $s(t)$ , is a linear combination of all the particular solutions, which are the eigenfunctions,

$$s(t) = \int_{\Omega} S(u) P(t, u) du \quad (3.6)$$

where  $\Omega$  is the set of permissible values for  $u$ . The coefficients,  $S(u)$ , in this linear combination are adjusted according to the boundary conditions and initial conditions that the field  $\psi$  must satisfy. These coefficients indicate "how much" of each eigenfunction  $P(t | u)$  is involved in the "composition" of the function  $s(t)$ . To better understand their role, one can compare this situation to the more classical context of vector analysis, where a vector is decomposed into a basis. In this case,  $s(t)$  plays the role of the vector to be decomposed, the eigenfunctions  $P(t | u)$  are analogous to the basis vectors<sup>\*</sup>, and  $S(u)$  can be considered as the function providing the components of  $s(t)$  in the basis  $\{P(t | u); u \in \Omega\}$ . For the problem of constructing the solution  $\psi$  to be well-posed, the partial differential equation to be solved must be accompanied by boundary and/or initial conditions that uniquely determine the components  $S(u)$  by forming the inner product between the field expression at the boundaries and the basis functions<sup>†</sup>. This point is discussed very clearly in the books by [Morse et Feshbach \(1953\)](#). Each function,  $f$ ,  $g$ ,  $h$ , and  $s$  in the expression of the field  $\psi$  can thus be written as a linear combination of the eigenfunctions of the corresponding Sturm-Liouville equation.

---

<sup>\*</sup>which are generally infinite in number

<sup>†</sup>the example of potential field extension provided later illustrates this computation.

## 4 The Fourier Transform

### 4.1 Theoretical Foundations and Definitions

The prominent role of the Fourier transform in signal processing is justified by the fact that many partial differential equations in physics lead to Sturm-Liouville differential equations where  $n(t) = 1$ , and whose eigenfunctions are the cos and sin functions. The solutions then take the form

$$s(t) = \int_0^{+\infty} S_{\cos}(u) \cos(2\pi ut) du + \int_0^{+\infty} S_{\sin}(u) \sin(2\pi ut) du \quad (4.1)$$

where the coefficients  $S_{\cos}$  and  $S_{\sin}$  are known as the Fourier coefficients. [Joseph Fourier \(1768–1830\)](#), born in Auxerre, submitted his first paper on polynomial root approximations to the Académie des Sciences in 1789. After spending several years in Egypt, he was appointed Prefect in Grenoble in 1802. In 1807, he presented a paper on heat propagation to the Académie des Sciences. His major work ([Fourier, 1822](#)), "Théorie analytique de la chaleur" (Analytical Theory of Heat), was published in 1822, and a few months later, he was appointed perpetual secretary of the Académie des Sciences. It is worth noting that Fourier became interested in statistics as early as 1798 and was recognized by the Académie des Sciences from 1816 as a specialist in insurance, statistics, and probability. Although some sums of trigonometric series had been calculated by Euler (1707–1783), the history of trigonometric series can be traced back to the solution of the vibrating strings problem ([Bernoulli, 1753](#)). The question of representing an arbitrary function, possibly discontinuous, by a trigonometric series quickly arose—a representation that the leading mathematicians of the time (1750) deemed impossible. It was not until fifty years later that Fourier addressed this issue while working on his analytical theory of heat. His initial work (1807) concerned only trigonometric series, and it was in 1812 that he introduced the Fourier integral. [He is credited with the notation  \$\int\_a^b\$ .](#)

The variable  $u \geq 0$  is called the frequency. This result is highly significant, indicating that [in many physical problems, the solutions can be expressed as a linear combination of cosine and sine functions](#). It is possible to modify the expression of the solution above to match the form of the Fourier transform we encountered at the beginning of this chapter, eq. (1.1). The calculation is straightforward and uses the following Euler identities,

$$\cos(2\pi ut) = \frac{\exp(2i\pi ut) + \exp(-2i\pi ut)}{2} \quad (4.2)$$

---

and

$$\sin(2\pi ut) = \frac{\exp(2i\pi ut) - \exp(-2i\pi ut)}{2i} \quad (4.3)$$

Some algebraic manipulations then yield

$$s(t) = \int_{-\infty}^{+\infty} S(u) \exp(2i\pi ut) du, \quad (4.4)$$

where, this time, the frequency  $u$  can take negative values. The function  $S(u)$  is called the Fourier transform of  $s(t)$  and is given by

$$\left\{ \begin{array}{ll} S(u) = [S_{\cos}(u) - iS_{\sin}(u)]/2 & u \geq 0 \\ S(u) = [S_{\cos}(-u) + iS_{\sin}(-u)]/2 & u \leq 0 \end{array} \right\} \quad (4.5)$$

$S(u)$  can be computed using the orthogonality property of the eigenfunctions,

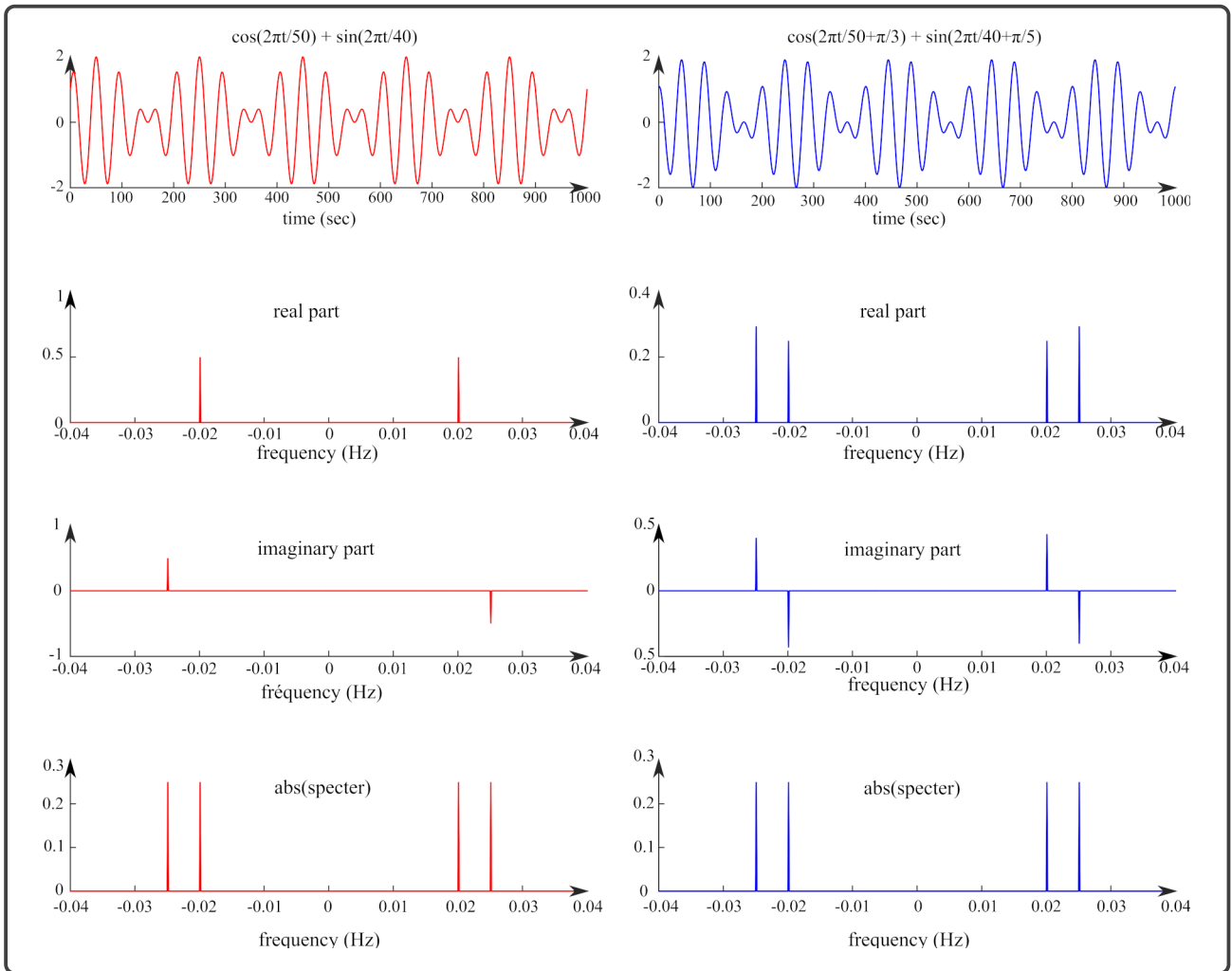
$$\int_{-\infty}^{+\infty} \exp(2i\pi ut) \exp(-2i\pi vt) dt = \delta(u - v) \quad (4.6)$$

where  $\delta(t \neq 0) = 0$ . We find

$$S(u) = \int_{-\infty}^{+\infty} s(t) \exp(-2i\pi ut) dt \quad (4.7)$$

which is the expression for the direct Fourier transform. The function [fourier\\_01.m](#) computes the Fourier transforms of simple signals and illustrates the role of the real and imaginary parts (Figure 1.1).

## 4. THE FOURIER TRANSFORM



**Figure 1.1:** Examples of Fourier transforms of simple signals. On the left, the real part of the transform corresponds to the frequency carried by the cosine component of the signal, and the imaginary part corresponds to the sine component. On the right, phase shifts cause the two frequencies present in the signal to decompose into the real and imaginary parts of the Fourier transform.

### 4.2 Notations

We will employ two notations. The first, implicitly adopted up to this point, represents functions using lowercase letters in their original physical space (eg  $f(t)$ ) and their images in the Fourier dual space using uppercase letters (eg  $F(u)$ ). The second notation will be used only when the first is not applicable, representing the direct and inverse transformation operations by  $\mathcal{F}$  and  $\mathcal{F}^{-1}$ , respectively. Therefore, we have:

$$\mathcal{F} f(t) = F(u) \tag{4.8}$$



---

and

$$\mathcal{F}^{-1}F(u) = f(t) \quad (4.9)$$

### 4.3 Example: Extension of Potential Fields

This example will introduce an initial application of the Fourier transform in geophysics: the extension, either upwards or downwards, of potential fields such as the Earth's magnetic field or the gravitational field. In the atmosphere, which we will approximate as a vacuum, these two fields satisfy Laplace's equation:

$$\nabla^2 v = 0 \quad (4.10)$$

where  $v$  is the field to be extended from a surface on which it is assumed to be perfectly known. The field in question may be a potential or a component of a geophysical field such as gravity or the magnetic field. Let us restrict ourselves to a two-dimensional Cartesian geometry where Laplace's equation is written as:

$$\frac{\partial^2}{\partial x^2} v(x, z) + \frac{\partial^2}{\partial z^2} v(x, z) = 0 \quad (4.11)$$

By separating variables, we seek a solution of the form:

$$v(x, z) = l(x)m(z), \quad (4.12)$$

which, when substituted into equation (4.10), yields:

$$\frac{1}{l(x)} \frac{d^2}{dx^2} l(x) = -\frac{1}{m(z)} \frac{d^2}{dz^2} m(z) \quad (4.13)$$

This equation must be satisfied for all pairs  $(x, z)$ , which is only possible if each term is equal to a real constant, **the famous separation constant  $\alpha$** ,

$$\frac{d^2}{dx^2} l(x) - \alpha l(x) = 0 \quad (4.14)$$

#### 4. THE FOURIER TRANSFORM

---

and

$$\frac{d^2}{dz^2}m(z) + \alpha m(z) = 0 \quad (4.15)$$

The separation of variables has transformed the initial partial differential equation into a system of two coupled differential equations. If  $\alpha > 0$ , we find:

$$l(x, \alpha > 0) = L_+(\alpha) \exp(+\sqrt{\alpha}x) + L_-(\alpha) \exp(-\sqrt{\alpha}x), \quad (4.16)$$

and if  $\alpha \leq 0$ ,

$$l(x, \alpha \leq 0) = L_{\cos}(\alpha) \cos(\sqrt{|\alpha|x}) + L_{\sin}(\alpha) \sin(\sqrt{|\alpha|x}). \quad (4.17)$$

Identical solutions are obtained for  $m(z)$ , though the sign of the constant  $\alpha$  should be reversed:

$$m(z, \alpha \leq 0) = M_+(\alpha) \exp(+\sqrt{|\alpha|z}) + M_-(\alpha) \exp(-\sqrt{|\alpha|z}) \quad (4.18)$$

and,

$$m(z, \alpha > 0) = M_{\cos}(\alpha) \cos(\sqrt{\alpha}z) + M_{\sin}(\alpha) \sin(\sqrt{\alpha}z). \quad (4.19)$$

In the most general case, the solution  $v(x, z)$  is a linear combination of the solutions above for all possible values of the separation constant  $\alpha$ . However, not all obtained solutions are necessarily physically acceptable. For instance, consider the specific case of calculating a field in the half-space  $z \geq 0$  with sources located entirely in the half-space  $z < 0$ . In such a configuration, physical considerations indicate that  $v \rightarrow 0$  as  $z \rightarrow +\infty$ , which eliminates the solutions  $M_+(\alpha) \exp(+\sqrt{|\alpha|z})$ ,  $M_{\cos}(\alpha) \cos(\sqrt{\alpha}z)$ , and  $M_{\sin}(\alpha) \sin(\sqrt{\alpha}z)$ . Ultimately, the acceptable solutions are:

$$l(x, \alpha \leq 0) = L_{\cos}(\alpha) \cos(\sqrt{|\alpha|x}) + L_{\sin}(\alpha) \sin(\sqrt{|\alpha|x}), \quad (4.20)$$

and

$$m(z, \alpha \leq 0) = M_-(\alpha) \exp(-\sqrt{|\alpha|}z). \quad (4.21)$$

The most general solution that can be constructed is therefore:

$$v(x, z) = \int_{-\infty}^0 \left[ V_{\cos}(\alpha) \cos(\sqrt{|\alpha|x}) + V_{\sin}(\alpha) \sin(\sqrt{|\alpha|x}) \right] \exp(-\sqrt{|\alpha|}z), d\alpha \quad (4.22)$$

This expression resembles the Fourier transform discussed at the beginning of this chapter. The resemblance becomes clearer by performing the variable change  $\sqrt{|\alpha|} \rightarrow 2\pi u$  and using Euler's identities to switch to complex notation:

$$v(x, z) = \int_{-\infty}^{+\infty} V(u) \exp(2i\pi ux) \exp(-2\pi|u|z), du \quad (4.23)$$

At  $z = 0$ , the expression is exactly the same as the inverse Fourier transform:

$$v(x, 0) = \int_{-\infty}^{+\infty} V(u) \exp(2i\pi ux), du \quad (4.24)$$

Thus, by direct Fourier transform, we have:

$$V(u) = \int_{-\infty}^{+\infty} v(x, 0) \exp(-2i\pi ux), dx \quad (4.25)$$

We are now able to write the complete chain of calculations for extending, upwards, a known potential field at  $z = 0$ :

$$v(x, 0) \xrightarrow{\mathcal{F}} V(u) \mapsto V(u) \exp(-2\pi|u|z) \xrightarrow{\mathcal{F}^{-1}} v(x, z) \quad (4.26)$$

To illustrate this, the program [prolonDemo01.m](#) calculates the upward extension of a gravity anomaly obtained using the [talwani.m](#) function and produces Figure (1.2). The method of [Talwani et al. \(1959\)](#) allows, as in magnetism, for the calculation of the theoretical gravity anomaly of any body, such as a polygon, as shown in Figure (1.3). This anomaly is the sum of the horizontal ( $X_i$ ) and vertical ( $Z_i$ ) contributions – using the notation from Talwani's paper – from each of the  $n$  sides of the polygon ABCDEF:

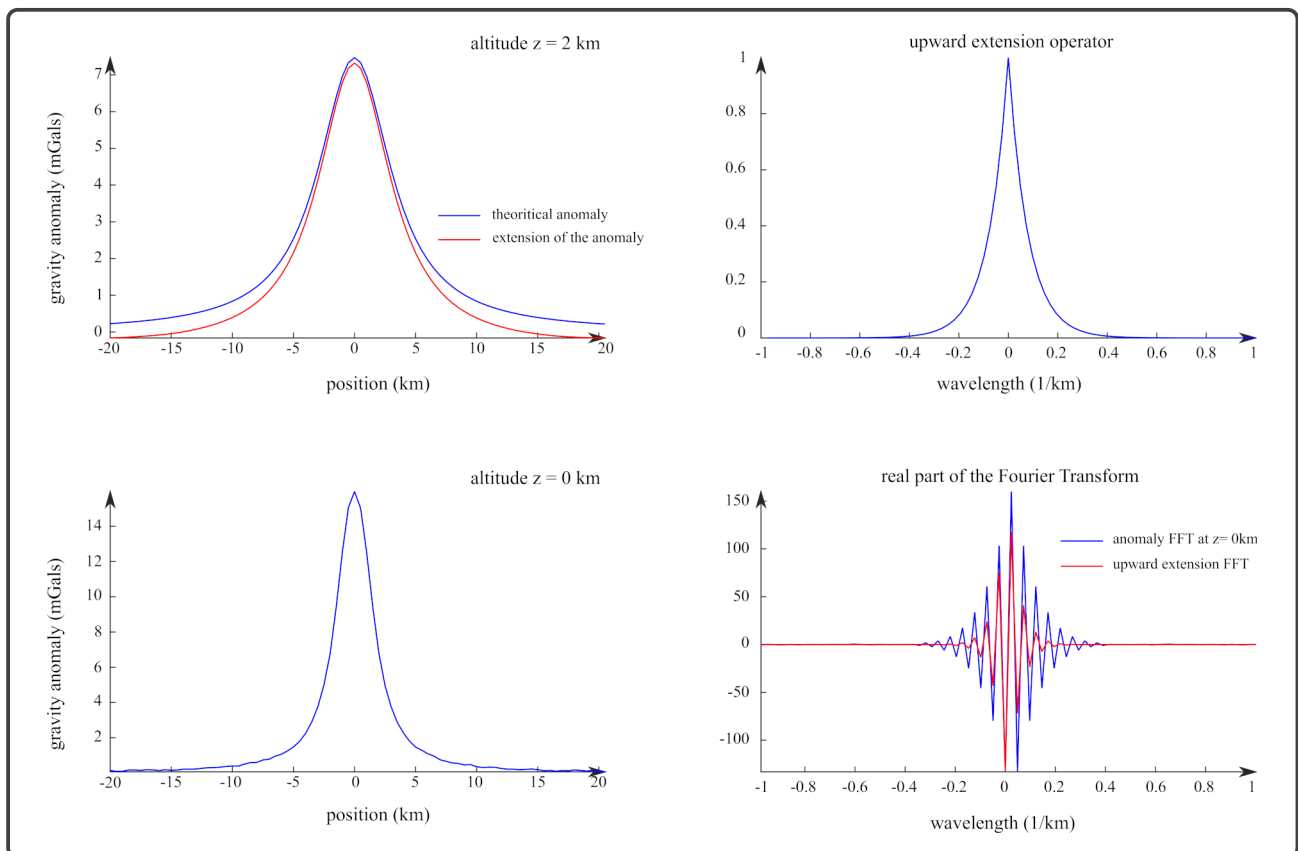
## 4. THE FOURIER TRANSFORM

$$V = 2G\rho \sum_{i=1}^n Z_i \quad (4.27)$$

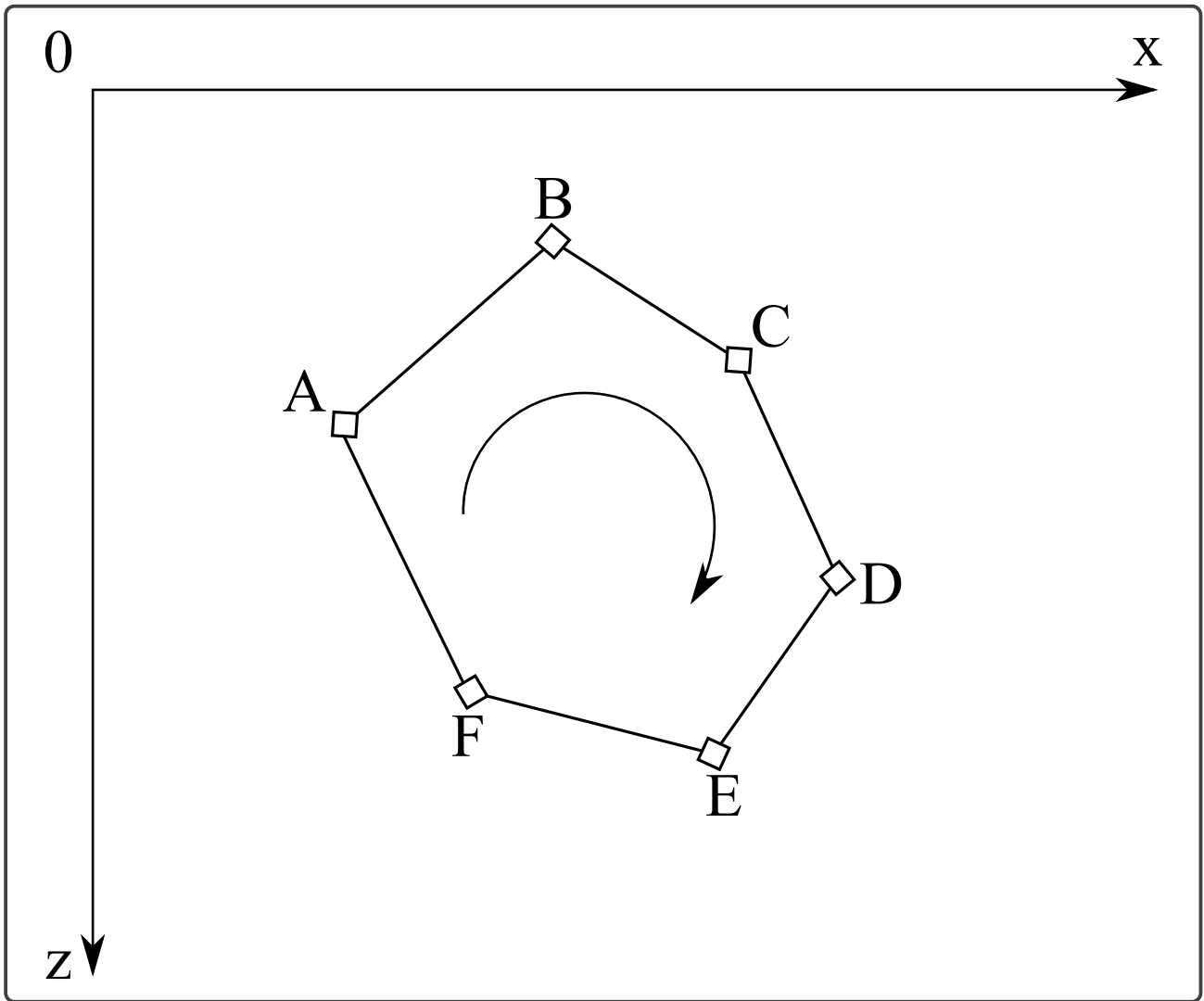
and

$$H = 2G\rho \sum_{i=1}^n H_i \quad (4.28)$$

where  $G$  is the universal gravitational constant and  $\rho$  is the volumetric density of the object.



**Figure 1.2: Illustration of the upward extension of a gravity anomaly. The discrepancy between the theoretical anomaly and the extended anomaly is due to numerical inaccuracies resulting from sampling.**



**Figure 1.3: Polygon ABCDEF, of infinite dimension along Oy, used to calculate the theoretical magnetic or gravity anomaly produced by the perturbing body.**

We will now describe the processing chain of the code [prolonDemo01.m](#). This chain consists of three stages. The first stage involves calculating the Fourier transform (line 27),  $V(u)$ , of the field measured at  $z = 0$ . The second stage is the application of the extension operator, which entails computing the product of  $V(u)$  and the function  $\exp(-2\pi|u|z)$  (line 34). The final stage involves calculating the inverse Fourier transform of this product to obtain the field at the desired altitude  $z$  (line 35). Similar calculations can be performed for other potential field transformations, such as computing horizontal or vertical derivatives, downward extension, pole reduction of magnetic anomalies, etc. In all cases, the first and last stages of the processing chain involve Fourier transforms as long as the coordinate system is Cartesian. If the coordinates are spherical or cylindrical, the functions  $\cos$  and  $\sin$  are replaced by Legendre or Bessel functions, and the chain no longer includes Fourier transforms. The example we have examined is representative of what physicists do when processing

## 4. THE FOURIER TRANSFORM

---

signals: they perform calculations based on pure mathematics and physical considerations, which then provide solid theoretical foundations justifying the subsequent signal processing operations. In such an approach, computing the Fourier transform of the measured field does not "come out of a hat," but is justified by physical theory. In my opinion, this approach is the only one that can be beneficial. When signal processing operations lack genuine theoretical justification, they "often lead to poor results"! Within this intellectual framework, the role of signal processing is to **master and implement, on incomplete and noisy data, a series of numerical calculations that best reproduce those of the underlying theory.**

### 4.4 Break: The Hartley Transform

At the beginning of this section, we saw that a real function,  $s(t)$ , can be expressed as follows,

$$s(t) = \int_0^{+\infty} S_{\cos}(u) \cos(2\pi ut) du + \int_0^{+\infty} S_{\sin}(u) \sin(2\pi ut) du \quad (4.29)$$

From this, one can arrive at the classical expression for the Fourier transform through Euler's identities and some algebraic manipulations. In the expression (4.29), the functions  $S_{\cos}$  and  $S_{\sin}$  are real, and the Fourier transform, which is a complex function, has symmetry properties,

$$S(-u) = S^*(u) \quad (4.30)$$

that render negative frequencies redundant. It is legitimate to question the utility of complicating matters by using a complex Fourier transform when "half of it" is unnecessary. If you are averse to this complexity, you might be attracted to the transform introduced by Ralph Hartley (Hartley, 1942), which Ronald Bracewell\* ardently supports. This transform is only applicable to real functions and can be easily derived from the above expression using the following elementary properties,

$$\cos(-2\pi ut) = \cos(+2\pi ut) \quad (4.31)$$

$$\sin(-2\pi ut) = -\sin(+2\pi ut) \quad (4.32)$$

Thus, we have,

$$\int_0^{+\infty} S_{\cos}(u) \cos(2\pi ut) du = \int_{-\infty}^{+\infty} H_{\cos}(u) [\cos(2\pi ut) + \sin(2\pi ut)] du, \quad (4.33)$$

---

\*Ronald Newbold Bracewell (July 22, 1921 – August 12, 2007) was an Australian astronomer and physicist involved in the SETI program.

where we defined,

$$H_{\cos}(u) = \frac{1}{2}S_{\cos}(|u|) \quad (4.34)$$

Similarly,

$$\int_0^{+\infty} S_{\sin}(u) \sin(2\pi ut) du = \int_{-\infty}^{+\infty} H_{\sin}(u) [\cos(2\pi ut) + \sin(2\pi ut)] du, \quad (4.35)$$

where

$$H_{\sin}(u \geq 0) = \frac{1}{2}S_{\sin}(u); \text{ and; } H_{\sin}(u \leq 0) = -\frac{1}{2}S_{\sin}(-u) \quad (4.36)$$

Finally, the signal  $s(t)$  can be written as,

$$s(t) = \int_{-\infty}^{+\infty} H(u) \text{cas}(2\pi ut) du \quad (4.37)$$

which we shall call the inverse Hartley transform, where the cas function is given by,

$$\text{cas}(2\pi ut) \equiv \cos(2\pi ut) + \sin(2\pi ut) \quad (4.38)$$

and where the Hartley transform,  $H(u) = H_{\cos}(u) + H_{\sin}(u)$ , can be obtained via the direct transform of  $s(t)$ ,

$$H(u) = \int_{-\infty}^{+\infty} s(t) \text{cas}(2\pi ut) dt \quad (4.39)$$

In addition to the fact that it uses only real functions, the Hartley transform possesses symmetry properties that allow the construction of very fast numerical transformation algorithms. These algorithms are at least as fast, and certainly simpler, than specialized fast Fourier transform algorithms for real signals. Furthermore, the basis functions, cas, are real functions that have almost the same interesting properties in mathematical physics as the cos and sin functions.

## 5 Fourier Series

### 5.1 Theoretical Foundations and Definitions

The solution of partial differential equations sometimes involves conditions on the boundaries of a finite domain. For example, this occurs when studying the Earth's normal modes and assuming that the normal stresses on its surface are zero. In such cases, the values that certain separation constants can take are no longer real numbers but integers. This is known as mode selection. The solution then becomes,

$$s(t) = S_0 + \sum_{n=1}^{\infty} S_{\cos,n} \cos\left(\frac{2\pi nt}{T}\right) + S_{\sin,n} \sin\left(\frac{2\pi nt}{T}\right) \quad (5.1)$$

where  $T$  is the duration (or length) of the domain between the boundaries where conditions are imposed. The above equation is called a Fourier series, and the coefficients  $S_{\cos,n}$  and  $S_{\sin,n}$  can be computed using the orthogonality properties of the eigenfunctions,

$$\int_0^T \cos\left(\frac{2\pi mt}{T}\right) \sin\left(\frac{2\pi nt}{T}\right) dt = 0 \quad \forall(m,n) \quad (5.2)$$

$$\int_0^T \cos\left(\frac{2\pi mt}{T}\right) \cos\left(\frac{2\pi nt}{T}\right) dt = \begin{cases} 0 & m \neq n \\ T/2 & m = n \end{cases} \quad (5.3)$$

and,

$$\int_0^T \sin\left(\frac{2\pi mt}{T}\right) \sin\left(\frac{2\pi nt}{T}\right) dt = \begin{cases} 0 & m \neq n \\ T/2 & m = n \end{cases} \quad (5.4)$$

The Fourier coefficients are then given by,

$$S_0 = \frac{1}{T} \int_0^T s(t) dt \quad (5.5)$$

$$S_{\cos,n} = \frac{2}{T} \int_0^T s(t) \cos\left(\frac{2\pi nt}{T}\right) dt \quad (5.6)$$



---

and,

$$S_{\sin,n} = \frac{2}{T} \int_0^T s(t) \sin\left(\frac{2\pi nt}{T}\right) dt \quad (5.7)$$

Just as with the Fourier transform discussed in the previous section, more compact forms can be obtained using Euler's identities and complex notation,

$$s(t) = \sum_{n=-\infty}^{+\infty} S_n \exp\left(\frac{2i\pi nt}{T}\right) \quad (5.8)$$

where,

$$S_n = \frac{1}{T} \int_0^T s(t) \exp\left(-\frac{2i\pi nt}{T}\right) dt \quad (5.9)$$

Note that this time the sum extends over  $n \in \mathbb{Z}$ .

## 5.2 Example: Vibrations of a Taut String

We will focus on calculating the small amplitude vibrations,  $v(x,t)$ , of a taut string fixed at its ends. The partial differential equation relevant to this problem is the wave equation for one spatial dimension:

$$\frac{\partial^2}{\partial x^2} v(x,t) - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} v(x,t) = 0 \quad (5.10)$$

where  $c$  is the wave propagation speed. Assuming,

$$v(x,t) = l(x)m(t) \quad (5.11)$$

the separation of variables provides,

$$\frac{1}{l(x)} \frac{d^2}{dx^2} l(x) = \frac{1}{c^2 m(t)} \frac{d^2}{dt^2} m(t) \quad (5.12)$$

which must be satisfied for all pairs  $(x,t)$ . Introducing the separation constant  $\alpha$ , we obtain the

system,

$$\frac{d^2}{dx^2}l(x) - \alpha l(x) = 0 \quad (5.13)$$

and,

$$\frac{d^2}{dt^2}m(t) - \alpha c^2 m(t) = 0 \quad (5.14)$$

If  $\alpha > 0$ , we find,

$$l(x, \alpha > 0) = L_+(\alpha) \exp(+\sqrt{\alpha}x) + L_-(\alpha) \exp(-\sqrt{\alpha}x), \quad (5.15)$$

and,

$$m(t, \alpha > 0) = M_+(\alpha) \exp(+\sqrt{\alpha}ct) + M_-(\alpha) \exp(-\sqrt{\alpha}ct) \quad (5.16)$$

When  $\alpha \leq 0$ ,

$$l(x, \alpha \leq 0) = L_{\cos}(\alpha) \cos(\sqrt{|\alpha|x}) + L_{\sin}(\alpha) \sin(\sqrt{|\alpha|x}), \quad (5.17)$$

and,

$$m(t, \alpha \leq 0) = M_{\cos}(\alpha) \cos(\sqrt{|\alpha|}ct) + M_{\sin}(\alpha) \sin(\sqrt{|\alpha|}ct). \quad (5.18)$$

Physical considerations specific to the problem must now be used to select acceptable solutions. We will only consider undamped vibrations, which allows us to eliminate the evanescent solutions  $m(t, \alpha > 0)$  and, consequently,  $l(x, \alpha > 0)$ . The solutions corresponding to

$$\alpha \leq 0$$

are acceptable but must be subject to the boundary conditions of the string, which we will assume are located at  $x = 0$  and  $x = L$ . At these points, the vibrations must vanish, and the acceptable solutions

must satisfy,

$$l(0, \alpha \leq 0) = l(L, \alpha \leq 0) = 0 \quad (5.19)$$

which is only satisfied by,

$$L_{\sin}(\alpha) \sin(\sqrt{|\alpha|}x) \quad (5.20)$$

when,

$$\sqrt{|\alpha|} = \frac{k\pi}{L}; \text{ with } k \in \mathbb{N}^* \quad (5.21)$$

The boundary condition of the string prevents a continuous variation of  $\alpha$ , and only discrete values are permitted. This is called mode selection. Ultimately, the most general acceptable solution is of the form,

$$v(x, t) = \sum_{k=1}^{+\infty} \sin\left(\frac{k\pi x}{L}\right) \left[ V_{\cos, k} \cos\left(\frac{k\pi ct}{L}\right) + V_{\sin, k} \sin\left(\frac{k\pi ct}{L}\right) \right] \quad (5.22)$$

where the coefficients  $V_{\cos, k}$  and  $V_{\sin, k}$  need to be determined. This can be done by assuming the shape and velocity of the string at time  $t = 0$ . For example, if,

$$v(x, 0); \text{ known and } ; \left. \frac{\partial}{\partial t} v(x, t) \right|_{t=0} = 0 \quad (5.23)$$

we have  $V_{\sin, k} = 0$ , due to the initial velocity condition being zero, and  $V_{\cos, k}$  such that,

$$v(x, 0) = \sum_{k=1}^{+\infty} V_{\cos, k} \sin\left(\frac{k\pi x}{L}\right) \quad (5.24)$$

This expression is a Fourier series, and the coefficients,

$$V_{\cos, k} = \frac{2}{L} \int_0^L v(x, 0) \sin\left(\frac{k\pi x}{L}\right) dx \quad (5.25)$$

## 6. PROPERTIES OF THE FOURIER TRANSFORM

---

The acceptable solution given the initial conditions is therefore,

$$v(x, t) = \sum_{k=1}^{+\infty} V_{\cos, k} \sin\left(\frac{k\pi x}{L}\right) \cos\left(\frac{k\pi ct}{L}\right) \quad (5.26)$$

Note that this solution can be written as,

$$v(x)|_{t=t_0} = \sum_{k=1}^{+\infty} V_k(t_0) \sin\left(\frac{k\pi x}{L}\right) \quad (5.27)$$

where we have introduced the time-varying Fourier coefficients,

$$V_k(t_0) \equiv V_{\cos, k} \cos\left(\frac{k\pi ct_0}{L}\right) \quad (5.28)$$

which indicates that at any time  $t = t_0$ , the shape of the string is a Fourier series. Similarly,

$$v(t)|_{x=x_0} = \sum_{k=1}^{+\infty} V_k(x_0) \cos\left(\frac{k\pi ct}{L}\right) \quad (5.29)$$

where we have defined,

$$V_k(x_0) \equiv V_{\cos, k} \sin\left(\frac{k\pi x_0}{L}\right) \quad (5.30)$$

indicates that the vibrations at any point  $x = x_0$  on the string are also a Fourier series with frequencies dependent on the length of the string\*.

## 6 Properties of the Fourier Transform

The Fourier transform has many properties, which are listed in the book by [Bracewell et Bracewell \(1986\)](#). Here, we will only mention those that will be frequently used in the following sections of this book.

---

\*Hence the famous question posed by Mark Kac: "Can we hear the shape of a drum?" ([Kac, 1966](#))

---

## 6.1 Linearity

This property is a direct consequence of the linearity of function integration:

$$\mathcal{F} \alpha f(t) + \beta g(t) = \alpha \mathcal{F} f(t) + \beta \mathcal{F} g(t), \quad (6.1)$$

where  $\alpha$  and  $\beta$  are constants.

## 6.2 Symmetries

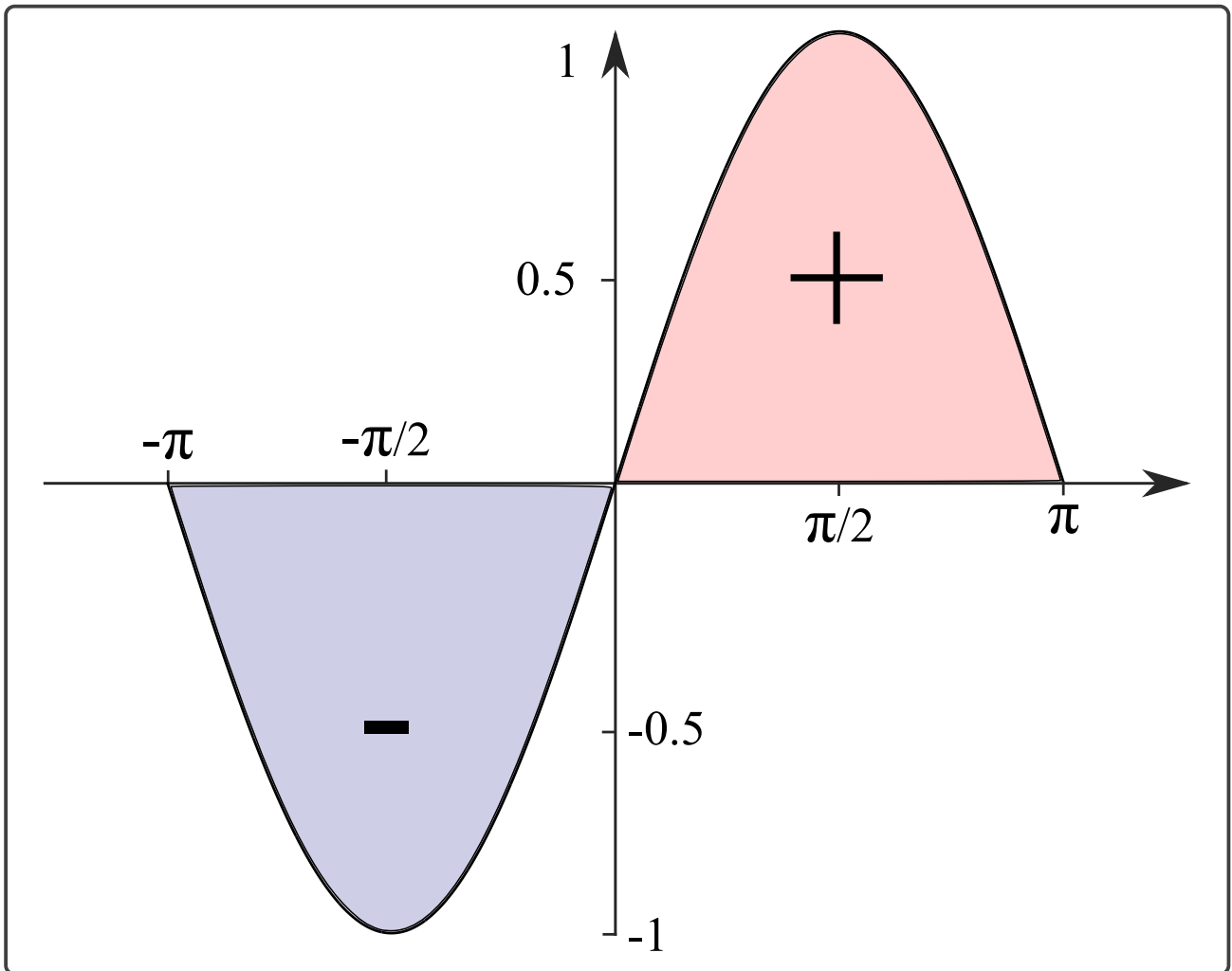
The symmetry properties of the Fourier transform are very useful for deducing and verifying certain results. Consider, for example, a real and even function,

$$f_p(-t) = f_p(t) \in \mathbb{R} \quad (6.2)$$

The Fourier transform of such a function is given by,

$$\begin{aligned} F_p(u) &= \int_{-\infty}^{+\infty} f_p(t) \exp(-2i\pi ut) dt \\ &= \int_{-\infty}^{+\infty} f_p(t) \cos(-2\pi ut) dt + i \underbrace{\int_{-\infty}^{+\infty} f_p(t) \sin(-2\pi ut) dt}_{=0} \\ &= \int_{-\infty}^{+\infty} f_p(t) \cos(-2\pi ut) dt \\ &= F_p(-u) \end{aligned} \quad (6.3)$$

where it is verified that  $F_p(u)$  is even and real. Indeed, since  $f_p(t)$  is an even function and  $\sin(-2\pi ut)$  is an odd function, their product is an odd function, whose integral over the period is zero. To illustrate this, let's take  $f_p(t)$  constant and equal to 1; it is indeed an even function. We then find ourselves in the trivial case shown in Figure (1.4), where we sum two "signed" areas that cancel out. Similar calculations show that a real odd function,  $f_i(-t) = -f_i(t)$ , has a purely imaginary and odd Fourier transform. Thus, we can say that [the Fourier transform preserves parity](#).



**Figure 1.4:** Sine function between  $-\pi$  and  $\pi$ . In blue and red, the areas (negative and positive) of the sine function illustrating the integration of the sine over a period. The sum of these two areas is zero.

Any real function  $f(t)$  can always be written as the sum of an odd function,

$$f_i(t) = [f(t) - f(-t)]/2 \quad (6.4)$$

and an even function,

$$f_p(t) = [f(t) + f(-t)]/2 \quad (6.5)$$

The linearity of the Fourier transform then establishes that the transform,

$$F(u) = F_p(u) + F_i(u) \quad (6.6)$$

is complex and satisfies,

$$F(-u) = F^*(u) \tag{6.7}$$

where  $*$  denotes the complex conjugate. The information corresponding to negative frequencies is redundant as it can be deduced from the information about positive frequencies. This property is utilized in numerical analysis, where specialized Fourier transform programs for real functions are found. All symmetries are summarized in the following formulas,

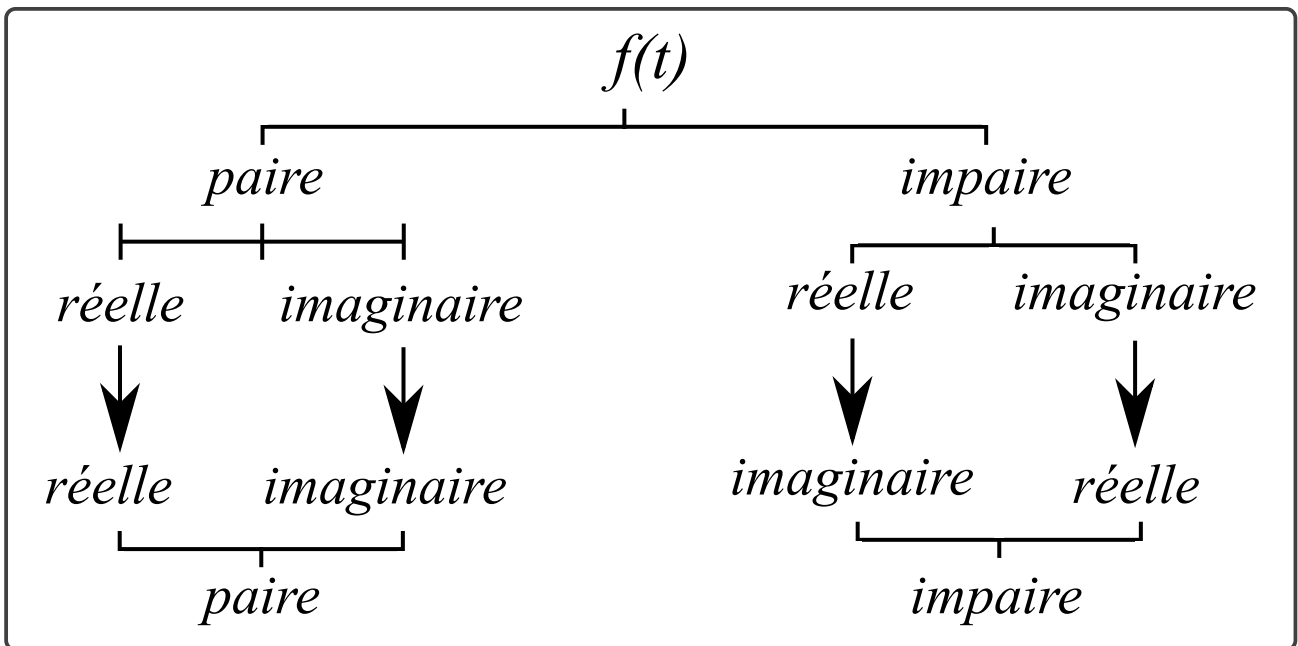


Figure 1.5: Symmetries of the Fourier transform for different types of functions

### 6.3 Similarity

This property, which is easily demonstrated by performing a change of variable in the integral defining the Fourier transform, expresses the duality that exists between a function and its Fourier transform:

$$\mathcal{F} [f(\alpha t)](u) = \frac{1}{|\alpha|} F(u/\alpha). \tag{6.8}$$

This property indicates that the narrower the temporal support of a function, the wider the frequency support of its Fourier transform. Denis Gabor first illustrated this in his famous 1946 paper, *Theory of Communication* (Gabor, 1946), by introducing the notion of Heisenberg boxes. We will not delve into the details of these boxes, also called time-frequency atoms when dealing with

## 6. PROPERTIES OF THE FOURIER TRANSFORM

---

time-frequency transforms such as the Fourier transform or wavelet transforms; we will simply describe these boxes. For more details, readers are encouraged to refer to Chapter 4 of Stéphane Mallat's book, *A Wavelet Tour of Signal Processing* (Mallat, 1999).

A brief preliminary discussion is necessary before describing these atoms. The linear operator  $L$ , whatever it may be, associates to any function  $g \in \mathbb{L}^2(\mathbb{R})$  the following value:

$$Lg(\gamma) = \int_{-\infty}^{+\infty} g(t)\phi^*\gamma(t)dt = \langle g, \phi\gamma \rangle \quad (6.9)$$

The Parseval's theorem provides the following extension to the above expression:

$$Lg(\gamma) = \int_{-\infty}^{+\infty} g(t)\phi\gamma(t)dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{g}(u)\hat{\phi}_\gamma(u)du \quad (6.10)$$

With these two relations established, we can now briefly describe these boxes. A Fourier atom  $\phi_\gamma$  is constructed using a window  $f$  that can be translated in time by  $t'$  and also modulated in frequency by  $u$ , giving:

$$\phi_\gamma(t) = \exp(iut)f(t-t') \quad (6.11)$$

Relation (6.10) shows that the information contained in  $\langle g, \phi_\gamma \rangle$  depends only on the spread of  $\phi_\gamma$  in time and frequency:

$$|\phi_\gamma|^2 = \int_{-\infty}^{+\infty} |\phi_\gamma(t)|^2 dt = 1 \quad (6.12)$$

$|\phi_\gamma(t)|^2$  can be interpreted as a probability density centered at:

$$t_\gamma = \int_{-\infty}^{+\infty} t|\phi_\gamma(t)|^2 dt \quad (6.13)$$

and whose spread  $\sigma_t^2(\gamma)$  is measured by the variance:

$$\sigma_t^2(\gamma) = \int_{-\infty}^{+\infty} (t-t_\gamma)^2 |\phi_\gamma(t)|^2 dt \quad (6.14)$$



---

Plancherel's formula ensures the following relation:

$$\int_{-\infty}^{+\infty} |\hat{\phi}_\gamma(u)|^2 du = 2\pi|\phi_\gamma|^2, \quad (6.15)$$

Thus, we can naturally write the median frequency and the spread of the box in frequency as follows:

$$u_\gamma = \frac{1}{2\pi} \int_{-\infty}^{+\infty} u |\hat{\phi}_\gamma(u)|^2 du \quad (6.16)$$

and:

$$\sigma_u^2(\gamma) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} (u - u_\gamma)^2 |\hat{\phi}_\gamma(t)|^2 du \quad (6.17)$$

We then obtain a rectangle (Figure (1.6)) whose area is given by the product of the variances in frequency and time. Heisenberg's uncertainty theorem shows that the area of this rectangle is greater than or equal to 1/2, so:

$$\sigma_t \sigma_u \geq 1/2. \quad (6.18)$$

Thus, it is clear that [the narrower the temporal support of a function, the wider the frequency support of its transform.](#)

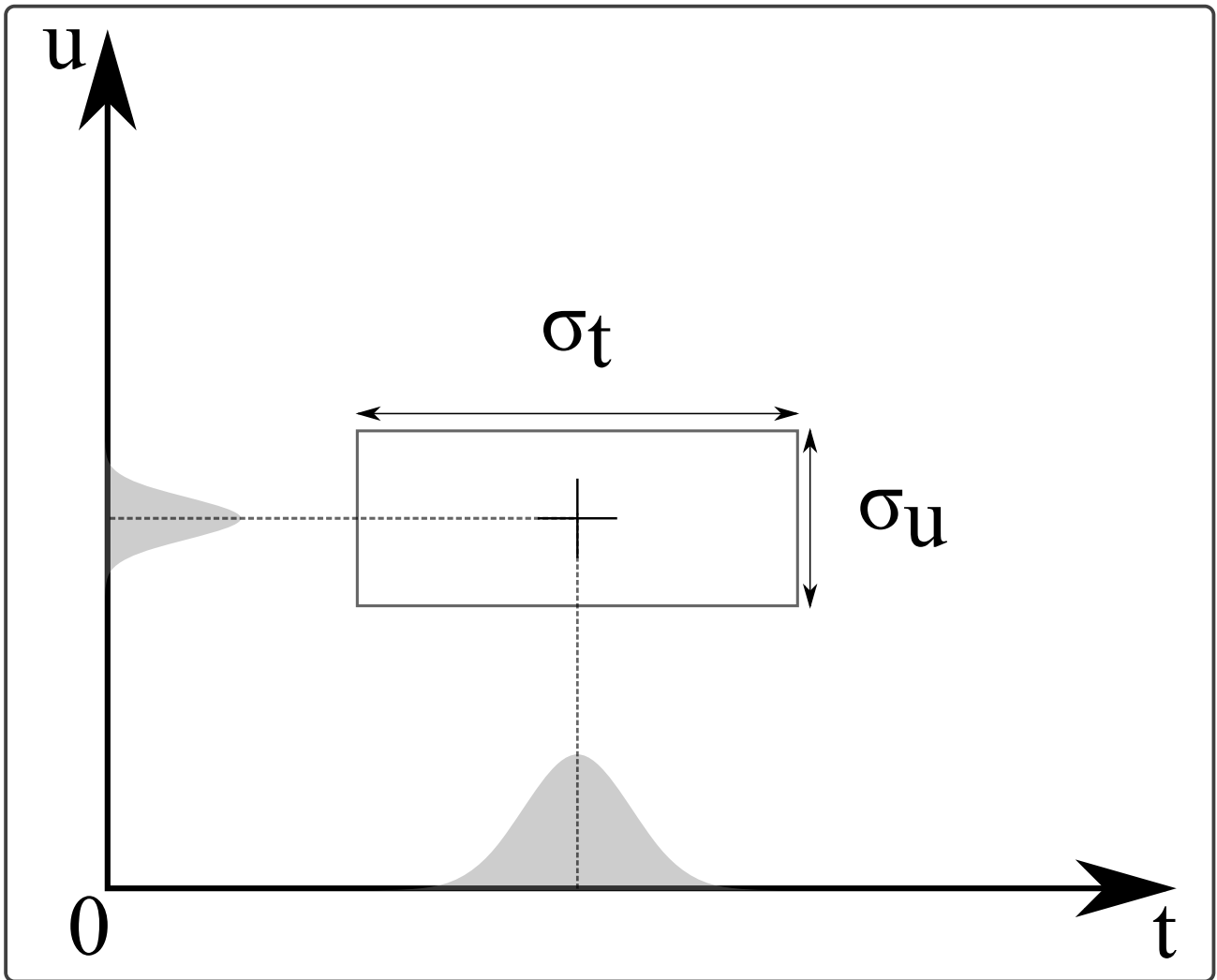


Figure 1.6: Heisenberg box schematizing the time-frequency duality of a Fourier atom

## 6.4 Translation

This property is also derived by a simple change of variable. It expresses that translating a signal results in a phase shift in the Fourier dual space:

$$\mathcal{F} f(t - t_0) = \exp(-2i\pi u t_0) \mathcal{F} f(t) \quad (6.19)$$

Reversing the application, it indicates that a frequency shift is equivalent to a time modulation:

$$\exp(2i\pi u_0 t) f(t) = \mathcal{F}^{-1} F(u - u_0) \quad (6.20)$$

## 6.5 Differentiation

This property allows for the easy determination of the Fourier transforms of derivatives of a function. For the first derivative:

$$\begin{aligned}
 \mathcal{F}\left[\frac{d}{dt}f(t)\right](u) &= \mathcal{F}\left[\lim_{\xi \downarrow 0} \frac{f(t+\xi) - f(t)}{\xi}\right](u) \\
 &= \lim_{\xi \downarrow 0} \frac{\mathcal{F}[f(t+\xi) - f(t)](u)}{\xi} \\
 &= \lim_{\xi \downarrow 0} \frac{[\exp(2i\pi u\xi) - 1]F(u)}{\xi} \\
 &= 2i\pi uF(u)
 \end{aligned} \tag{6.21}$$

The generalization to the  $n^{\text{th}}$  derivative is immediate:

$$\mathcal{F}\left[\frac{d^n}{dt^n}f(t)\right](u) = (2i\pi u)^n F(u) \tag{6.22}$$

Note also that the right-hand side of this expression remains valid when  $n$  is not an integer but is a positive real number. This allows for the definition of the notion of non-integer differentiation of a function, which is useful for studying fractals and abrupt variations that occur in certain signals. Non-integer derivatives are also useful for studying wave propagation in highly heterogeneous media where properties vary randomly.

## 7 Multidimensional Fourier Transforms

### 7.1 Example: Extension of Potential Fields

The example of extending potential fields seen previously in the two-dimensional case can be extended to three dimensions. The calculations naturally lead to a two-dimensional Fourier transform. In the three-dimensional case, the potential must satisfy:

$$\frac{\partial^2}{\partial x^2}v(x,y,z) + \frac{\partial^2}{\partial y^2}v(x,y,z) + \frac{\partial^2}{\partial z^2}v(x,y,z) = 0 \tag{7.1}$$

Assuming that the sources are located in the lower half-space, a similar reasoning to that used for the two-dimensional case leads to an acceptable solution:

$$v(x,y,z) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} V(u_x, u_y) \exp[2i\pi(u_x x + u_y y)] \times \exp(-2\pi z \sqrt{u_x^2 + u_y^2}) du_x du_y \tag{7.2}$$

Knowledge of the field in the plane  $z = 0$  provides:

$$V(u_x, u_y) = \int \int_{-\infty}^{+\infty} v(x, y, 0) \exp[-2i\pi(u_x x + u_y y)] dx dy \quad (7.3)$$

which is a two-dimensional Fourier transform. The inverse transform is given by:

$$v(x, y, 0) = \int \int_{-\infty}^{+\infty} V(u_x, u_y) \exp[+2i\pi(u_x x + u_y y)] du_x du_y. \quad (7.4)$$

The processing chain for the three-dimensional extension is the same as for the two-dimensional extension; it suffices to replace the one-dimensional Fourier transforms with their two-dimensional versions.

### 7.2 General Definitions

The generalization to  $n$  dimensions leads to:

$$V(\vec{u}) = \int \cdots \int_{\mathbb{R}^n} v(\vec{x}) \exp[-2i\pi \vec{u} \bullet \vec{x}] d\vec{x} \quad (7.5)$$

for the direct Fourier transform, and

$$v(\vec{x}) = \int \cdots \int_{\mathbb{R}^n} V(\vec{u}) \exp[+2i\pi \vec{u} \bullet \vec{x}] d\vec{u} \quad (7.6)$$

for the inverse Fourier transform, where  $\bullet$  denotes the dot product.

### 7.3 Sign Conventions in Space-Time

The multidimensional Fourier transform we have defined is applicable to both spatial coordinates and time. However, it is wise to adopt a sign convention that differentiates the time dimension from the spatial dimensions.

$$V(\vec{u}, u_t) = \int \cdots \int_{\mathbb{R}^4} v(\vec{x}, t) \exp[-2i\pi(\vec{u} \bullet \vec{x} - u_t t)] d\vec{x} dt \quad (7.7)$$

---

for the direct Fourier transform, and

$$v(\vec{x}, t) = \int \cdots \int_{\mathbb{R}^4} V(\vec{u}, u_t) \exp[+2i\pi(\vec{u} \bullet \vec{x} - u_t t)] d\vec{u} du_t \quad (7.8)$$

for the inverse Fourier transform. This definition of the Fourier transform is frequently used in seismology.

---

---

# CHAPTER 2

---

## CONVOLUTION AND CORRELATION

<b>1</b>	<b>Convolution</b> . . . . .	<b>46</b>
1.1	Where Do We Encounter Convolutions? . . . . .	46
1.2	Spatial Convolution . . . . .	50
1.3	Convolution and Probability . . . . .	51
1.4	Properties of Convolution . . . . .	51
1.5	Fourier Transform of a Convolution . . . . .	52
1.6	Differentiation of a Convolution . . . . .	54
<b>2</b>	<b>Correlation</b> . . . . .	<b>54</b>

---

# 1 Convolution

The convolution of two functions,  $f(t)$  and  $g(t)$ , is defined by the integral:

$$[f * g](t) \equiv \int_{-\infty}^{+\infty} f(\tau)g(t - \tau)d\tau = \int_{-\infty}^{+\infty} f(t - \tau)g(\tau)d\tau \quad (1.1)$$

where we use the classic notation  $*$  for the convolution operator. Convolution is frequently encountered in signal processing because it appears in:

- linear systems theory,
- Green's function theory when solving partial differential equations,
- probability theory, where it is used to compute the distribution of sums of independent random variables.

The origins of convolution are as fundamental as those of the Fourier transform, and we will see that these two mathematical operations have remarkable properties with respect to each other. Before establishing these main properties, and as we did for the Fourier transform, we will first explore the "domain" of convolution.

## 1.1 Where Do We Encounter Convolutions?

### Temporal Convolution

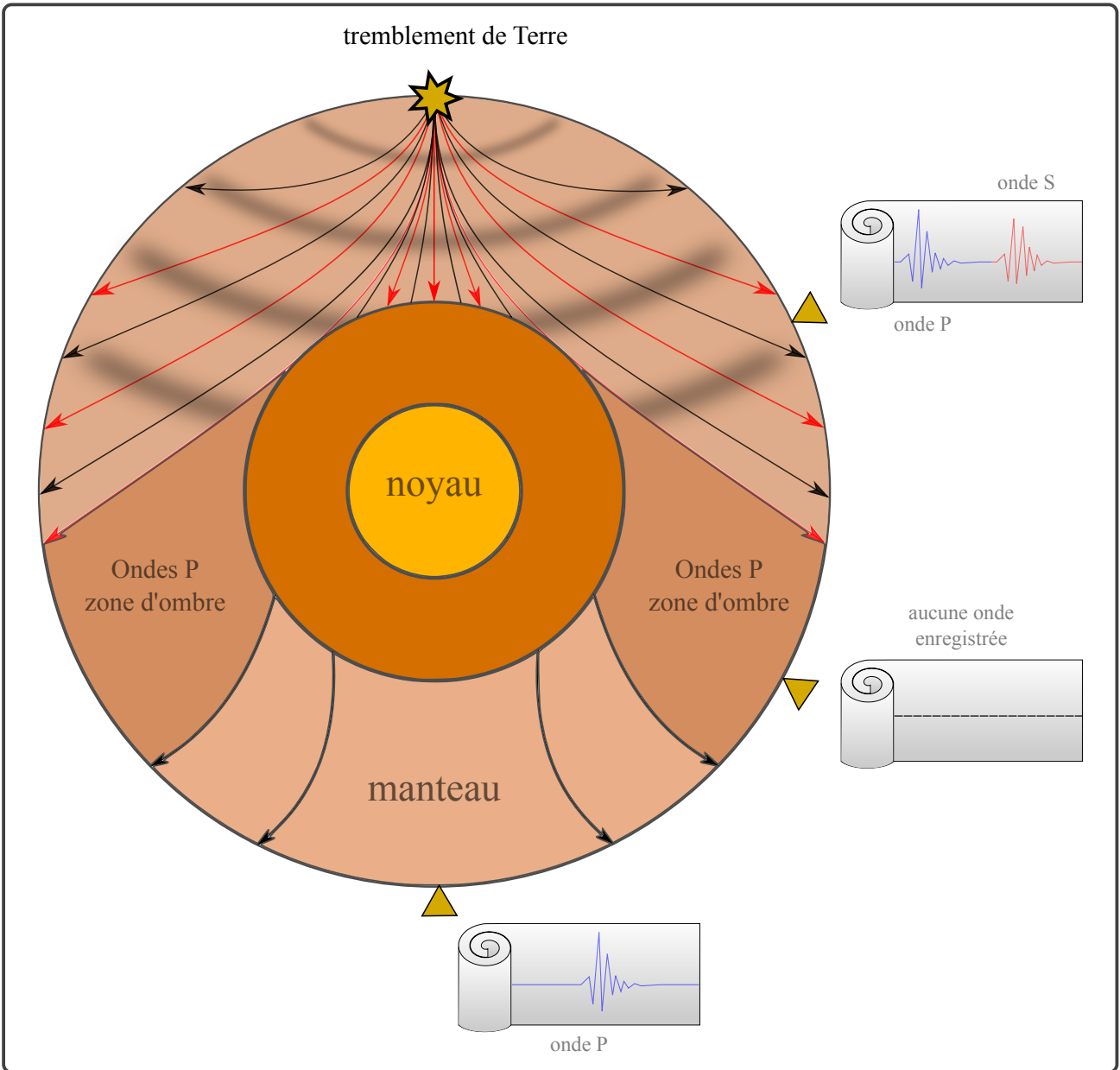
The concept of temporal convolution is closely related to the notions of [linearity and time \(or space\) invariance \(or stationarity\)](#). One of the main tasks of physicists is to study systems through which signals pass. A system is characterized by a functional  $\mathcal{G}$  that associates an input signal  $e(t)$  with an output signal  $s(t)$ ,

$$e(t) \mapsto_{\mathcal{G}} s(t) \equiv \mathcal{G}[e(t)](t) \quad (1.2)$$

The system in question can be the very object of the study, and its characteristics can be examined by injecting specific signals and observing the results. This approach is used when emitting electromagnetic or elastic waves into the Earth to study its structure (see [Figure 2.1](#)). In other cases, it is the input signal that interests the physicist, and the system serves as a [pair of glasses](#) through which the phenomenon  $e(t)$  is viewed. This occurs whenever measurements are made using an instrument, whether it is an astronomer looking at the sky through a telescope or a geophysicist

# 1. CONVOLUTION

recording ground vibrations with a seismometer. Many problems in experimental physics are of this nature, and their solutions are more or less easy to find depending on the complexity of the systems involved. The simplest systems one can imagine are linear time-invariant systems. These very simple systems arise in problems where the underlying physics is linear or as first-order approximations of nonlinear systems. A linear system satisfies the following relationships:



**Figure 2.1:** Diagram of Earth's structure obtained from seismic wave tomography. The yellow triangles represent the seismometers used by geophysicists.

$$\mathcal{G}[\alpha_1 e_1(t) + \alpha_2 e_2(t)](t) = \alpha_1 \mathcal{G}[e_1(t)](t) + \alpha_2 \mathcal{G}[e_2(t)](t), \quad (1.3)$$



and

$$s(t) = \mathcal{G}[e(t)](t) \implies s(t - \xi) = \mathcal{G}[e(t - \xi)](t) \quad (1.4)$$

These two properties allow us to establish that if the input signal consists of two signals of the same shape, with different amplitudes, occurring at different times, then,

$$f_1.e(t - t_1) + f_2.e(t - t_2) \mapsto_{\mathcal{G}} f_1.s(t - t_1) + f_2.s(t - t_2) \quad (1.5)$$

Of course, this can be generalized further,

$$\sum f_i.e(t - t_i) \mapsto_{\mathcal{G}} \sum f_i.s(t - t_i) \quad (1.6)$$

and even, in the limiting case where the input signals are infinitesimally close, forming a *continuum*,

$$\int_{-\infty}^{+\infty} f(\tau)e(t - \tau)d\tau \mapsto_{\mathcal{G}} \int_{-\infty}^{+\infty} f(\tau)s(t - \tau)d\tau \quad (1.7)$$

The integrals above are convolution integrals. Suppose now that the signals  $e(t - \tau)$  in the left integral are impulses,  $\delta(t - \tau)$ , as brief as we want\*. In this case, somewhat like representing a function by a juxtaposition of *sticks* of different heights, the integral† becomes,

$$\int_{-\infty}^{+\infty} f(\tau)\delta(t - \tau)d\tau = f(t) \quad (1.8)$$

Let,

$$g(t) \equiv \mathcal{G}[\delta(t)](t) \quad (1.9)$$

be the system's impulse response. Then,

$$\mathcal{G}[f(t)](t) = \int_{-\infty}^{+\infty} f(\tau)g(t - \tau)d\tau \quad (1.10)$$

---

\*The limit process, that is, an infinitely brief impulse, is discussed in the section on the Dirac impulse.

†Which we will revisit as the "sampling formula" in the section on the Dirac impulse.

## 1. CONVOLUTION

---

This expression shows that [the response of a linear and time-invariant system is equal to the convolution product of the input signal with the system's impulse response](#). The system is entirely characterized by its impulse response. The time-dependent system cannot respond before being excited, and its impulse response is causal, that is, such that,

$$g(t < 0) = 0 \quad (1.11)$$

Many physical systems can be reasonably well represented by linear time-invariant systems. This is the case for many electronic circuits, optical setups, and mechanical assemblies. In seismology, the Earth is often considered an elastic medium and, therefore, linear and invariant. This approximation forms the basis for interpreting seismic recordings. To illustrate our points, we invite the reader to use the program [ex\\_convolution.m](#) in which the convolution of [Ricker](#) and [chirp](#) is performed on random reflectivities. Figure (2.2) provides an example.

The [Ricker](#) wavelet, sometimes called the [Mexican hat](#), is defined by the following relation,

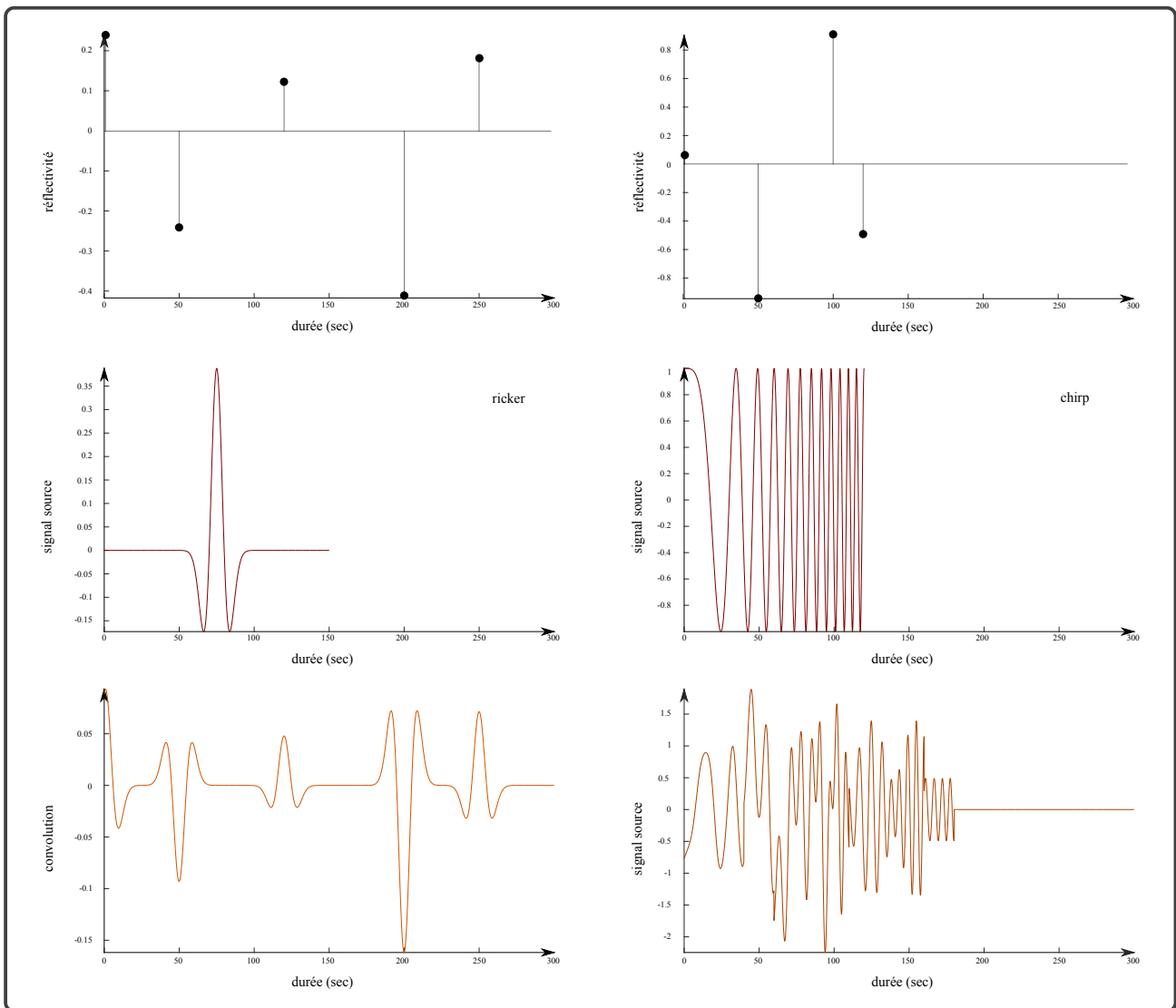
$$r(t) = (1 - 2\pi^2 f^2 t^2) e^{-\pi^2 f^2 t^2} \quad (1.12)$$

It is also found in the form,

$$r(t) = \frac{2}{\sqrt{3\sigma\pi^{1/4}}} \left(1 - \frac{t^2}{\sigma^2}\right) e^{-\frac{t^2}{2*\sigma^2}} \quad (1.13)$$

The function [ricker.m](#) provides an implementation of the Ricker wavelet. The [chirp](#), which in English means "tweet", is a pseudo-periodic signal of duration  $T$ , modulated in frequency ( $\Delta f$ ) around a carrier frequency ( $f_0$ ) and also modulated in amplitude. The function [chirp\\_lin.m](#) provides an implementation of a particular case where the frequency ramp is linear and the envelope modulation remains constant. This signal is defined as follows,

$$c(t) = \text{Re}\left\{Ae^{2\pi i\left(f_0 + \frac{\Delta f}{2T}.t - \frac{\Delta f}{2}.t\right).t}\right\} \quad \text{avec } A = 1 \quad \text{et } \forall t \in \{0, T\} \quad (1.14)$$



**Figure 2.2: Temporal convolutions.** This example is frequently encountered in seismics; the two signals at the top are the impulse responses of the subsurface. This is what would be obtained if we were capable of emitting a Dirac impulse using seismic sources. In practice, seismic sources are Ricker impulses (left center) or sweeps or chirps (right center). The resulting seismic traces (bottom) are the convolution of the impulse response (top) with the sources (middle). Note that the **sweep**, which has a long duration, completely masks the events present in the impulse response. How to retrieve them? See the section on correlation and Figure ??.

## 1.2 Spatial Convolution

Spatial convolutions, which involve functions depending on spatial coordinates, are very common as they appear in the theory of **Green's functions** (a theory extensively covered in a dedicated book by **Roach, 1982**) applied to solving partial differential equations. The Green's function represents the field created by a point source: a point mass in gravitation, a point charge in electrostatics, a dipole in magnetostatics, etc. For example, the gravitational potential created by a unit point mass located

## 1. CONVOLUTION

---

at the origin is given by:

$$g(x, y, z) = \frac{G}{\sqrt{x^2 + y^2 + z^2}} \quad (1.15)$$

As you know, the potential  $v(x, y, z)$  of multiple masses is equal to the sum of the potentials caused by each mass, *ie*,

$$v(x, y, z) = \iiint_{\mathbb{R}^3} \rho(\xi, \theta, \zeta) g(x - \xi, y - \theta, z - \zeta) d\xi d\theta d\zeta \quad (1.16)$$

This is a three-dimensional spatial convolution where  $\rho(\xi, \theta, \zeta)$  is the spatial mass density distribution. Analogous expressions are also encountered in potential theory, electromagnetism, etc. Spatial convolutions are also seen in seismic tomography for weakly diffracting media where the [Born approximation](#) can be applied. This approximation originates from quantum mechanics for very sparse scattering potentials ([Cohen-Tannoudji et al., 1998](#)). In first-order Born approximation, only the incident wave and the waves scattered by a single interaction, a single scatterer, are considered and studied ([Hudson et Heritage, 1981](#)). More generally, it pertains to perturbation theory in mathematics.

### 1.3 Convolution and Probability

Convolution appears in probability theory as follows. Let  $\alpha$  and  $\beta$  be two independent random variables with respective probability densities  $\alpha$  and  $\beta$ . The probability density  $g(\gamma)$  of the sum  $\gamma = \alpha + \beta$  is given by the convolution:

$$g(\gamma) = \int_{-\infty}^{+\infty} a(\gamma - \xi) b(\xi) d\xi. \quad (1.17)$$

We will see later that this property, combined with the Central Limit Theorem, explains why the normal distribution holds a special place in statistics.

### 1.4 Properties of Convolution

#### Commutativity, Associativity, Distributivity

Convolution is commutative:

$$f * g = g * f \quad (1.18)$$

associative:

$$(f * g) * h = f * (g * h) \quad (1.19)$$

and also distributive with respect to addition:

$$f * (g + h) = f * g + f * h \quad (1.20)$$

These properties are immediate consequences of those of integration and are easily established. Note, however, that it is due to the fact that one of the functions is "flipped" – meaning that the integration variable appears with a negative sign (see equation 1.1) – in the convolution integral that convolution is commutative. [Without this flipping, commutativity does not hold.](#)

## 1.5 Fourier Transform of a Convolution

The Fourier transform of a convolution product is obtained by explicitly writing out the following integrals:

$$\begin{aligned} \mathcal{F}[f * g](u) &= \int_{-\infty}^{+\infty} \left\{ \int_{-\infty}^{+\infty} f(\xi) g(t - \xi) d\xi \right\} \exp(-2i\pi ut) dt \\ &= \int_{-\infty}^{+\infty} f(\xi) \left\{ \int_{-\infty}^{+\infty} g(t - \xi) \exp(-2i\pi ut) dt \right\} d\xi \\ &= \int_{-\infty}^{+\infty} f(\xi) \mathcal{F}[g(t - \xi)](u) d\xi \\ &= G(u) \int_{-\infty}^{+\infty} f(\xi) \exp(-2i\pi u\xi) d\xi \\ &= G(u)F(u) \end{aligned} \quad (1.21)$$

[The Fourier transform of a convolution product is equal to the product of the Fourier transforms \(Plancherel's theorem\),](#)

$$\mathcal{F}[(f * g)(t)](u) = F(u)G(u). \quad (1.22)$$

The dual of the previous theorem indicates that:

$$\mathcal{F}[f(t)g(t)](u) = [F * G](u). \quad (1.23)$$

## 1. CONVOLUTION

---

Applying this theorem to the specific case where  $g(t) = f^*(t)$ , we obtain:

$$\begin{aligned}\mathcal{F} [f(t)f^*(t)](u) &= \mathcal{F} [|f(t)|^2](u) \\ &= F(u) * F^*(-u)\end{aligned}\tag{1.24}$$

which can be written as:

$$\int_{-\infty}^{+\infty} |f(t)|^2 \exp(-2i\pi ut) dt = \int_{-\infty}^{+\infty} F(v)F^*(u-v)dv\tag{1.25}$$

By setting  $u = 0$ :

$$\int_{-\infty}^{+\infty} |f(t)|^2 dt = \int_{-\infty}^{+\infty} |F(u)|^2 du\tag{1.26}$$

This important relation is known as the [Rayleigh-Parseval theorem](#); it indicates that [the energy of the signal is conserved by the Fourier transform](#). The simple form of the Fourier transform of a convolution product has significant consequences. From an analytical perspective, the simplification is substantial since one transitions from an integral formulation to a straightforward product of functions. This property, combined with the fact that convolution is a frequently encountered mathematical operation, greatly enhances the role of the Fourier transform in signal processing. Many calculations are simpler when performed *via* the Fourier transform. For example, as seen in probability theory, the probability density  $p_s(x)$  of a sum of  $N$  independent random variables is given by the convolution chain:

$$p_s(x) = p_1(x) * p_2(x) * \dots * p_N(x)\tag{1.27}$$

which, after Fourier transform, becomes:

$$P_s(u) = P_1(u) \times P_2(u) \times \dots \times P_N(u),\tag{1.28}$$

where the Fourier transforms  $P_i(u)$  are called the characteristic functions of the probability densities  $p_i(x)$ .

## 1.6 Differentiation of a Convolution

We have:

$$\begin{aligned}
 \mathcal{F}^{-1} \left\{ \mathcal{F} \left[ \frac{d}{dt} [f * g](t) \right] (u) \right\} (t) &= \mathcal{F}^{-1} [2i\pi u F(u) G(u)] (t) \\
 &= \mathcal{F}^{-1} [2i\pi u F(u)] (t) * \mathcal{F}^{-1} [G(u)] (t) \\
 &= \left( \frac{d}{dt} f(t) \right) * g(t)
 \end{aligned} \tag{1.29}$$

and also:

$$\begin{aligned}
 \mathcal{F}^{-1} \left\{ \mathcal{F} \left[ \frac{d}{dt} [f * g](t) \right] (u) \right\} (t) &= \mathcal{F}^{-1} [2i\pi u F(u) G(u)] (t) \\
 &= \mathcal{F}^{-1} [F(u)] (t) * \mathcal{F}^{-1} [2i\pi u G(u)] (t) \\
 &= f(t) * \left( \frac{d}{dt} g(t) \right)
 \end{aligned} \tag{1.30}$$

which simplifies to:

$$\frac{d}{dt} [f * g](t) = \left( \frac{d}{dt} f(t) \right) * g(t) = f(t) * \left( \frac{d}{dt} g(t) \right), \tag{1.31}$$

which should not be confused with the differentiation of a simple product of functions.

## 2 Correlation

The cross-correlation of two functions  $f(t)$  and  $g(t)$  is defined by,

$$\begin{aligned}
 r_{f,g}(l) &= f(t) \diamond g(t) \\
 &\equiv \int_{-\infty}^{+\infty} f^*(t) g(t+l) dt \\
 &= \int_{-\infty}^{+\infty} f^*(t-l) g(t) dt \\
 &= f^*(-t) * g(t)
 \end{aligned} \tag{2.1}$$

and can be interpreted as a convolution where one of the functions is not "reversed." The variable  $l$  represents the time shift between the function and its replica. [Cross-correlation is not commutative](#),

## 2. CORRELATION

---

$$r_{f,g}(l) = f^*(-t) * g(t) \neq f(t) * g^*(-t) = r_{g,f}(l) \quad (2.2)$$

The Fourier transform of the cross-correlation is easily calculated using the theorems discussed earlier,

$$\begin{aligned} R_{f,g}(u) &\equiv \mathcal{F} [r_{f,g}(l)] (u) \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f^*(t)g(t+l) \exp(-2i\pi ul) dt dl \\ &= \int_{-\infty}^{+\infty} f^*(t) dt \int_{-\infty}^{+\infty} g(t+l) \exp(-2i\pi ul) dl \\ &= G(u) \int_{-\infty}^{+\infty} f^*(t) \exp(2i\pi ut) dt \\ &= G(u) \left[ \int_{-\infty}^{+\infty} f(t) \exp(-2i\pi ut) dt \right]^* \\ &= F^*(u)G(u), \end{aligned} \quad (2.3)$$

where the property  $z_1^* z_2 = (z_1 z_2^*)^*$  has been used to transition from the fourth to the fifth line.

Note that,

$$F^*(u)G(u) = [F(u)G^*(u)]^* \quad (2.4)$$

which implies,

$$r_{f,g}(l) = r_{g,f}(-l) \quad (2.5)$$

The autocorrelation is such that its Fourier transform is,

$$R_{f,f}(u) = |F(u)|^2 \quad (2.6)$$

The energy spectrum of a function is equal to the Fourier transform of the autocorrelation of the function. This relationship between autocorrelation and the energy spectrum is known as the Wiener-Khinchin theorem when  $f(t)$  is a stochastic process\*. In analytical calculations, such processes are generally defined by their autocorrelation function, and the Wiener-Khinchin theorem

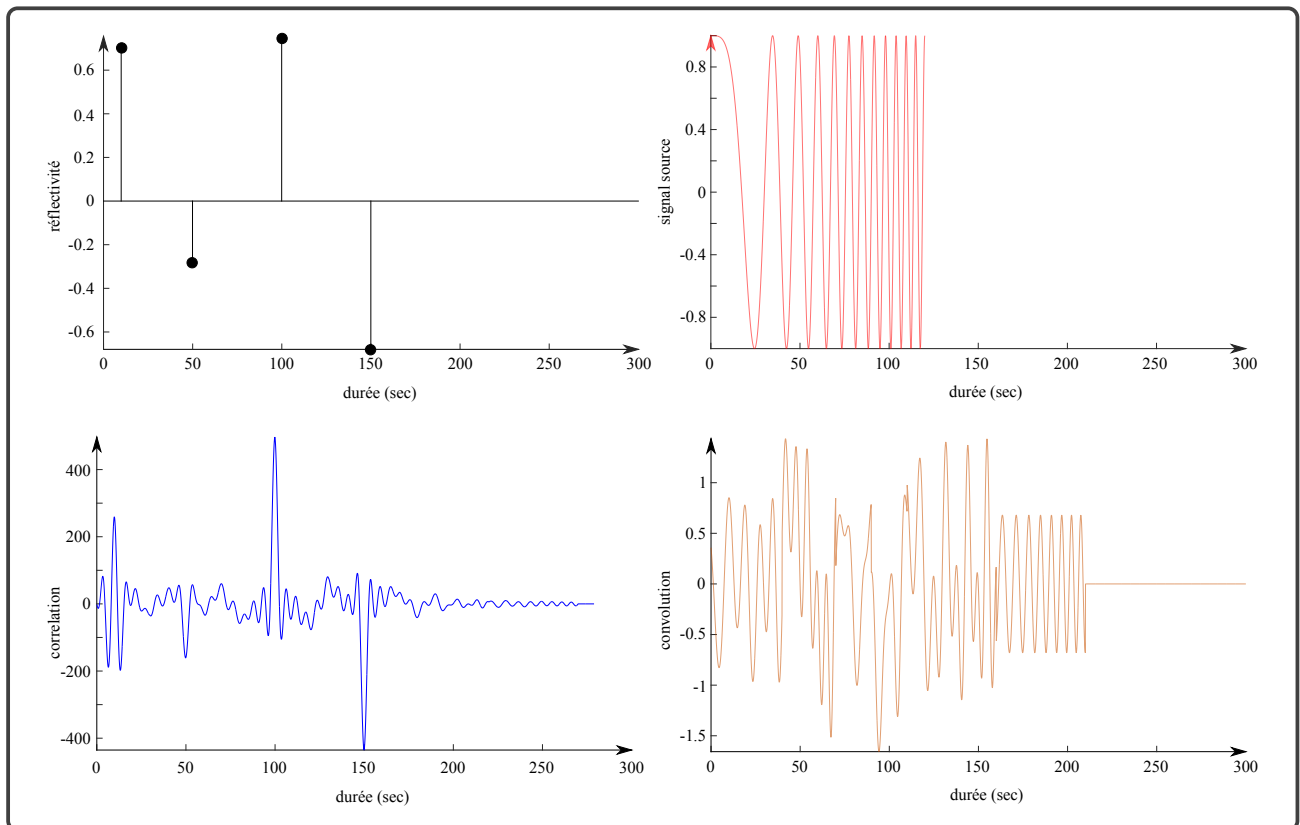
---

\*For more details, see the chapter on stochastic processes.



allows the deduction of the energy spectrum, though it does not provide information about the phase.

Cross-correlation (Figure 2.3) represents the power or energy if the two functions  $f$  and  $g$  are physically associated, such as: intensity and voltage (power), magnetic and electric fields (Poynting vector), or force and velocity.



**Figure 2.3: Example of cross-correlation. The convolution of an ideal seismic impulse response (top left) with a sweep (top right) emitted by a vibratory truck produces a trace (bottom right) in which arrivals are indistinguishable. The cross-correlation (bottom left) between this trace and the sweep helps to better discern the arrivals. This operation is routinely performed in seismic surveys when vibratory trucks are used as sources. Its effectiveness is due to the very specific shape (frequency sweep) of the sweep.**

---

---

# CHAPTER 3

---

## THE HILBERT TRANSFORM

1	Definition . . . . .	57
2	Formulae: Hilbert Transforms . . . . .	60

### 1 Definition

The linear system whose transfer function – that is, the Fourier transform of the impulse response – is given by

$$G(u) = -i \operatorname{sgn}(u) \tag{1.1}$$

has the sole effect of advancing the phases by  $\pi/2$  and is called a quadrature filter. The impulse response (1.2) allows us to obtain the system's response to an input  $e(t)$ .

$$g(t) = \frac{1}{\pi t} \tag{1.2}$$

This response is generally expressed in the following form,

$$s(t) = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{e(\tau)}{t - \tau} d\tau. \tag{1.3}$$

---

By definition,  $s(t)$  is called the [Hilbert](#) transform of  $e(t)$ , in honor of David Hilbert (1862-1943), born in Königsberg where he lived. He studied and began his career there until 1895, when he moved to Göttingen. His research covered a vast range of topics, including number theory, the theory of proof, algebraic geometry, variational calculus, and integral equations. His work on the development of arbitrary functions into series of orthogonal functions is particularly relevant for this course. We will use the following notation,

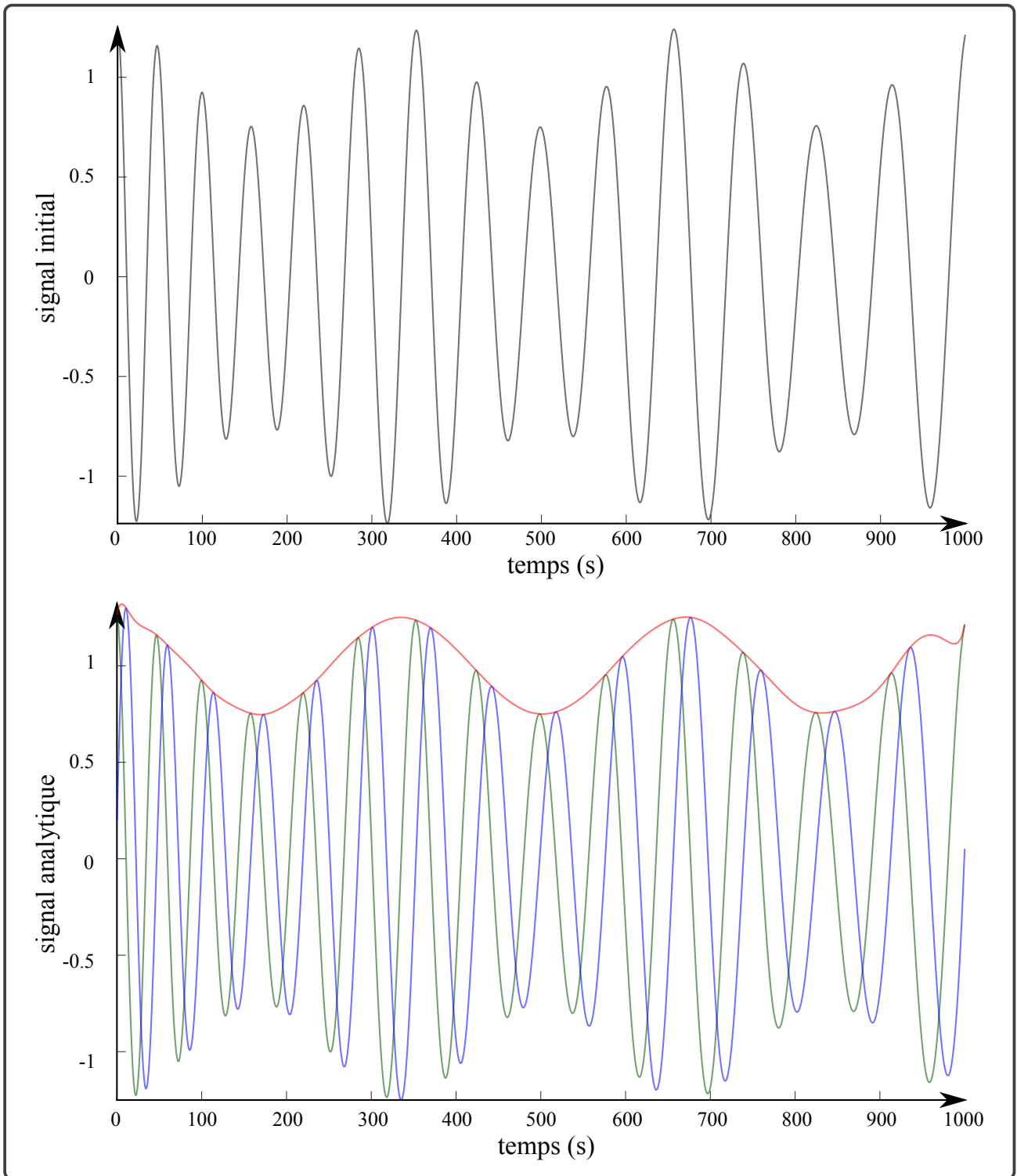
$$F_{Hi}(t) \equiv \mathcal{H}[f](t) \equiv \frac{1}{\pi t} * f(t). \quad (1.4)$$

The [Hilbert](#) transform is used when studying causal signals. Non-stationary signals are often analyzed via their analytic signal, which is computed using the [Hilbert](#) transform. This transform therefore allows us to compute the analytic signal,

$$f_a(t) \equiv f(t) + i\mathcal{H}[f](t) \quad (1.5)$$

associated with  $f(t)$ . The magnitude of the analytic signal provides the envelope of  $f(t)$  (figure 3.1). An analytic signal is the complex equivalent of a real signal where all positive and zero frequencies are doubled, and negative frequencies are canceled. The program [ex\\_hilbert\\_transform.m](#) performs this computation and produces the images in figure (3.1). It uses the subfunction [hilbert\\_transform.m](#), which allows the user to choose either the [hilbert](#) function – native to Matlab® – or to more explicitly develop the Hilbert transform algorithm.

## 1. DEFINITION



**Figure 3.1: Envelope Calculation by Hilbert Transform.** The initial signal (top, in gray) has an analytic function which is a complex function whose real part (bottom, in green) is equal to the initial signal itself, and whose imaginary part (bottom, in blue) is the Hilbert transform of the initial signal. The envelope (bottom, in red) is obtained by calculating the magnitude of the analytic signal.

---

## 2 Formulae: Hilbert Transforms

$$\cos(2\pi u_0 t) \mapsto +\sin(2\pi u_0 t) \quad (2.1)$$

$$\sin(2\pi u_0 t) \mapsto -\cos(2\pi u_0 t) \quad (2.2)$$

$$\frac{\sin(t)}{t} \mapsto \frac{1 - \cos(t)}{t} \quad (2.3)$$

$$\frac{1}{(1+t^2)} \mapsto \frac{t}{(1+t^2)} \quad (2.4)$$

$$\exp(-\alpha|t|) \cos(2\pi u_0 t) \mapsto \exp(-\alpha|t|) \sin(2\pi u_0 t) \quad (2.5)$$

---

---

# CHAPTER 4

---

## USEFUL FUNCTIONS IN FOURIER ANALYSIS

1	Catalogue of Useful Functions . . . . .	62
2	Window (the "scissors") . . . . .	62
3	Cardinal Sine . . . . .	64
4	Triangle . . . . .	65
5	Exponential Functions . . . . .	67
5.1	Exponential Decaying to Infinity . . . . .	67
5.2	Gaussian . . . . .	67
6	Dirac Delta Function (the "photo") . . . . .	68
7	Sign Function . . . . .	72
8	Heaviside Distribution (the switch) . . . . .	73
9	Dirac Comb (the camera) . . . . .	75
10	Sine and Cosine Functions . . . . .	77
11	Form: Fourier Transforms . . . . .	77

---

# 1 Catalogue of Useful Functions

The previous chapters introduced us to the [Fourier](#) transformation through mathematical physics, that is, from an idealistic perspective where we did not question the feasibility of performing the calculations we developed using real signals. To delve deeper, it is now necessary to establish a link between this idealistic viewpoint and practical application. This link consists of the more or less rigorous answers to the inevitable questions that arise when dealing with real signals. While one can indeed pose numerous questions, the following are ubiquitous:

- Infinity, present in  $F(u) = \int_{-\infty}^{+\infty} f(t) \exp(-2i\pi ut) dt$ , does not exist in the computer since the signal I have is of finite duration. How will this integral be evaluated? What errors will I incur?
- Truncating the above integral is not sufficient because, even for a limited duration, I do not know the signal at all times but only at certain instances. What do I lose by not knowing the signal densely? How are my calculations affected?
- What I measure is not the signal of interest but the signal plus noise, which consists of measurement errors and other unwanted signals. What is the impact of this noise on my calculations?

The quality of the answers to these questions directly controls the analytical power of the different methods that will be used. Before discussing these answers in detail in the following chapters, it is necessary to have a set of "tools" that will allow us to "mathematize" the questions we pose. These tools will be functions or distributions that act as "scissors," "switches," "cameras," etc. With these tools, we will be able to mathematically articulate the transition from the ideal of mathematical physics to the reality of numerical processing. Some of the functions we will consider in this chapter are not functions in the strict sense and can only be rigorously manipulated in the sense of distributions, which the Anglo-Saxons call "[generalized functions](#)". We will emphasize their physical significance and how these mathematical entities appear as physical limits\*.

## 2 Window (the "scissors")

The window  $\Pi(t)$ , also known as the "rectangular" or "boxcar" function, is defined by,

$$\Pi(t) = \begin{cases} 0 & |t| > 1/2 \\ 1 & |t| \leq 1/2 \end{cases} \quad (2.1)$$

---

\*See, for example, the discussion concerning the [Dirac](#) impulse.

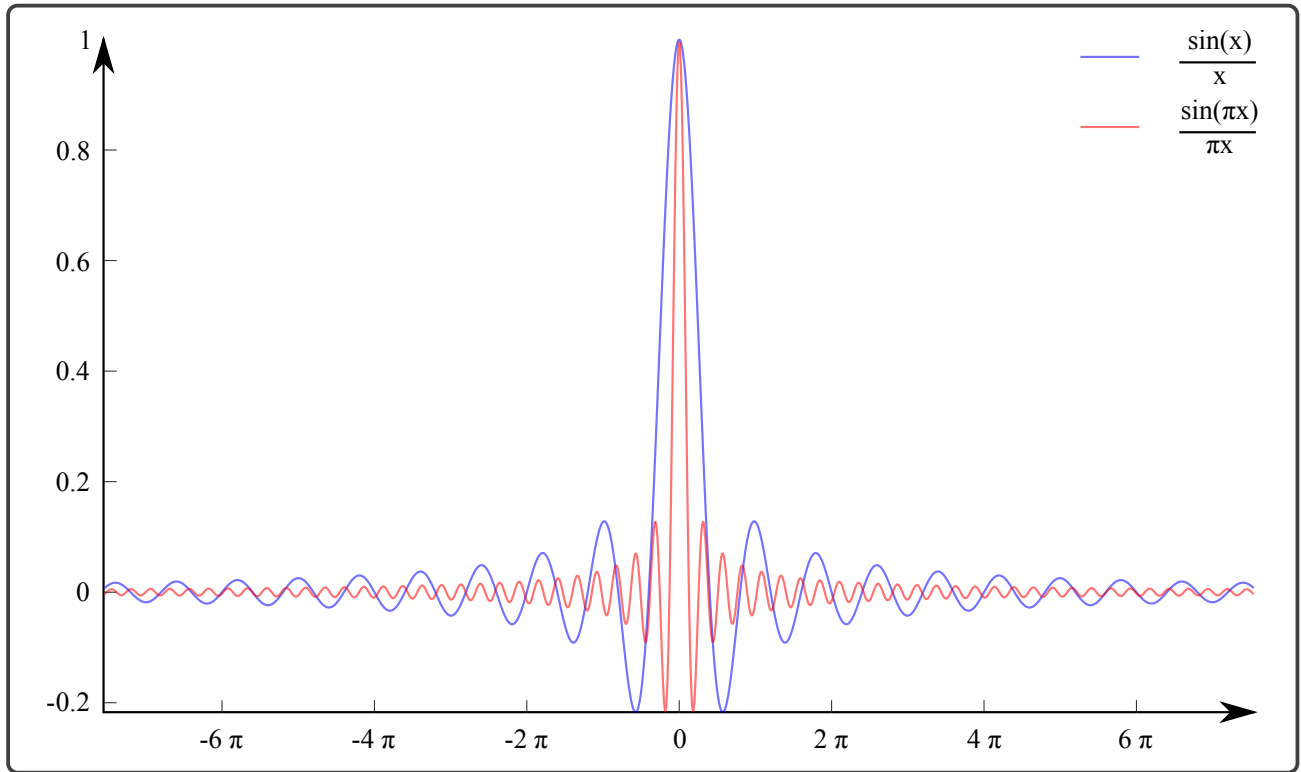
## 2. WINDOW (THE "SCISSORS")

is one of the fundamental functions that we will continually use to symbolize the truncation of signals. We can thus view it as the pair of "scissors" in the toolbox that we are filling. Its **Fourier** transform,

$$\mathcal{F}[\Pi(t)](u) = \frac{\sin(\pi u)}{\pi u} \equiv \text{sinc}(u) \quad (2.2)$$

is easy to obtain by direct integration. The sinc function, called "sine cardinal," is illustrated in figure (4.1), and will be discussed in more detail in chapter (3). Note the presence of  $\pi$  in its expression, in accordance with the definition by **Bracewell et Bracewell (1986)**, which many authors overlook. The **Fourier** transform of a window with width  $T$ , amplitude  $\alpha$ , and centered at  $t_0$  is obtained by using the theorems seen in the previous chapter,

$$\begin{aligned} \mathcal{F}\left[\alpha\Pi\left(\frac{t-t_0}{T}\right)\right](u) &= \alpha\mathcal{F}\left[\Pi\left(\frac{t-t_0}{T}\right)\right](u) \\ &= \alpha\exp(-2i\pi ut_0)\mathcal{F}\left[\Pi\left(\frac{t}{T}\right)\right](u) \\ &= \alpha T\exp(-2i\pi ut_0)\text{sinc}(uT) \end{aligned} \quad (2.3)$$



**Figure 4.1: Sine cardinal functions. In blue, the most commonly known unnormalized form, and in red, the normalized form.**

Without encroaching too much on the following chapters, it is good to justify the use of the



window function now. One might think that these scissors are unnecessary to express the fact that a signal is known only for a limited duration, and it is simpler to write that the **Fourier** transform of such a signal is,

$$\mathcal{F}[\textit{truncated signal}](u) = \int_{\textit{beginning}}^{\textit{end}} (\textit{signal}) \times \exp(-2i\pi ut) dt \quad (2.4)$$

where we simply take the endpoints of the signal's observation interval as the limits of the integral. This calculation is correct and provides the same result as one would obtain using the window, but it has a major drawback: it implies a redefinition of the **Fourier** transform operator. Such redefinition is rigorously discouraged, which is why it is better to write,

$$\mathcal{F}[\textit{truncated signal}](u) = \mathcal{F}[(\textit{signal}) \times \Pi(\textit{of the appropriate duration})](u) \quad (2.5)$$

where we can use the symbolic notation  $\mathcal{F}$  since we retain the initial definition of the **Fourier** transform.

### 3 Cardinal Sine

We have already encountered this function, which is the **Fourier** transform of the window function. It plays a role in the interpolation and filtering of signals. The cardinal sine, defined by,

$$\textit{sinc}(u) \equiv \frac{\sin(\pi u)}{\pi u} \quad (3.1)$$

is such that,

$$\left\{ \begin{array}{l} \textit{sinc}(0) = 1 \\ \textit{sinc}(n) = 0 \quad (n \in \mathbb{Z}) \\ \int_{-\infty}^{+\infty} \textit{sinc}(t) dt = 1 \end{array} \right. \quad (3.2)$$

Using the duality properties of the [Fourier](#) transform, it is directly shown that,

$$\mathcal{F}[\text{sinc}(t)](u) = \Pi(u) \tag{3.3}$$

The importance of the cardinal sine comes from the fact that its [Fourier](#) transform is zero outside the interval  $[-1/2; 1/2]$ . We will see, in the chapter on filtering, that convolution by a cardinal sine is a low-pass filtering. We will also see, in the chapter on sampling, that the cardinal sine allows, under certain conditions, the interpolation of signals for which only discrete values are known. Finally, note that for signals  $f(t)$  such that  $F(u) = 0$  outside the interval  $[-1/2; 1/2]$  we have,

$$F(u)\Pi(u) = F(u) \tag{3.4}$$

which, after inverse [Fourier](#) transform, gives,

$$[f * \text{sinc}](t) = f(t) \tag{3.5}$$

For such signals with bounded spectra, the cardinal sine is the identity element of convolution.

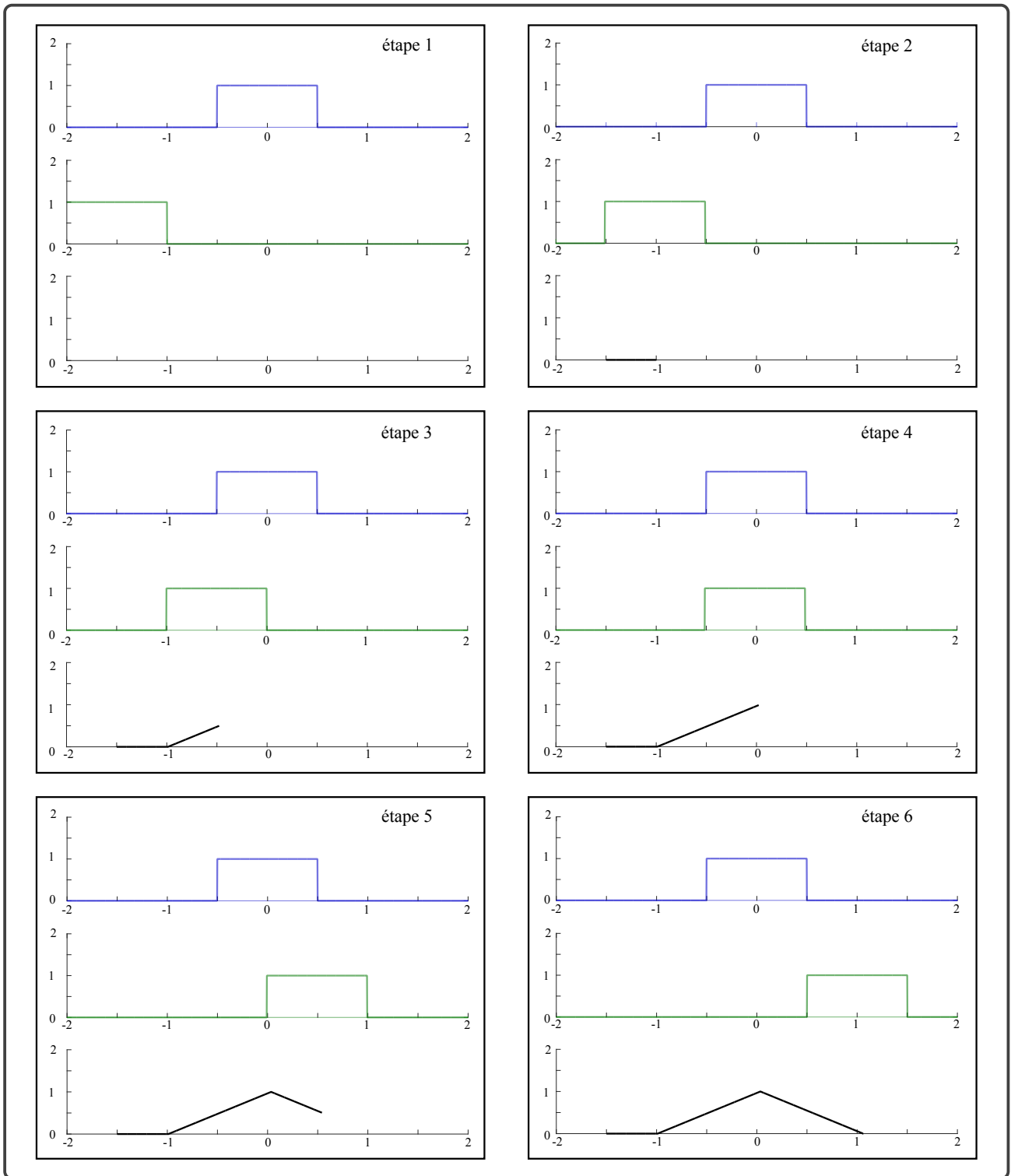
## 4 Triangle

The triangle function defined by the following relation,

$$\Lambda(t) = \begin{cases} 0 & \text{si } |t| > 1 \\ 1 - |t| & \text{si } |t| \leq 1 \end{cases} \tag{4.1}$$

frequently appears in calculations as it is the self-convolution of the window function (Figure 4.2). The program [ex\\_autoconv\\_fenetre.m](#) demonstrates the result of the self-convolution product of a rectangular function in the form of an animation. This observation immediately shows that,

$$\begin{aligned} \mathcal{F}[\Lambda(t)](u) &= \mathcal{F}\{[\Pi * \Pi](t)\}(u) \\ &= \text{sinc}^2(u) \end{aligned} \tag{4.2}$$



**Figure 4.2:** From top to bottom, and from left to right, different stages of the self-convolution product of a rectangular function (blue curve) by itself (green curve). As can be seen, the result is the triangle function (black curve).

## 5 Exponential Functions

### 5.1 Exponential Decaying to Infinity

First, let's consider the case of the function  $\exp(-|t|)$ , which often appears in the analysis of signals generated by damping or diffusion processes. Its [Fourier](#) transform is computed by direct integration.

Let us first note that,

$$\begin{aligned}\mathcal{F}[\exp(-|t|)](u) &= \int_{-\infty}^{+\infty} \exp(-|t|) \exp(-2i\pi ut) dt \\ &= \int_{-\infty}^{+\infty} \exp(-|t|) \cos(-2i\pi ut) dt \\ &= 2 \int_0^{+\infty} \exp(-|t|) \cos(-2i\pi ut) dt \\ &= 2 \operatorname{Re} \left[ \int_0^{+\infty} \exp(-|t|) \exp(-2i\pi ut) dt \right]\end{aligned}\tag{5.1}$$

Therefore, we have,

$$\begin{aligned}\mathcal{F}[\exp(-|t|)](u) &= 2 \operatorname{Re} \left[ \int_0^{+\infty} \exp[(-2i\pi u - 1)t] dt \right] \\ &= 2 \operatorname{Re} \left( \frac{1}{2i\pi u + 1} \right) \\ &= \frac{2}{(2\pi u)^2 + 1}.\end{aligned}\tag{5.2}$$

We will revisit this Fourier transform when we study [Butterworth](#) filters.

### 5.2 Gaussian

The case of the Gaussian function  $\exp(-\pi t^2)$  is interesting for different reasons. This function is important in statistics where it represents the [Gauss](#) distribution underlying the least squares methods. We have already seen the role of convolution in probability theory. Moreover, the Gaussian is often used in distribution theory and plays a part in defining [Heisenberg](#)'s uncertainty principle (see chapter [6.3](#)), whose implications we will explore in signal processing. The calculation of the

---

Fourier transform is clever,

$$\begin{aligned}\mathcal{F}[\exp(-\pi t^2)](u) &= \int_{-\infty}^{+\infty} \exp(-\pi t^2) \exp(-2i\pi ut) dt \\ &= \exp(-\pi u^2) \int_{-\infty}^{+\infty} \exp[-\pi(t+iu)^2] dt \\ &= \exp(-\pi u^2) \int_{-\infty}^{+\infty} \exp(-\pi \xi^2) d\xi \\ &= \exp(-\pi u^2),\end{aligned}\tag{5.3}$$

where we have used the property  $\int_{-\infty}^{+\infty} \exp(-\pi \xi^2) d\xi = 1$ . Note that the Gaussian function is its own Fourier transform.

## 6 Dirac Delta Function (the "photo")

The Dirac delta function,  $\delta(t)$ , is named in honor of [Paul Dirac](#) (1902-1984) who was born in Bristol and studied electrical engineering at the University. In 1923, he moved to Cambridge as a student and published, two years later, his paper on the fundamental laws of quantum mechanics ([Dirac, 1925](#)). This work was based on recent results by [Heisenberg](#) and extended them significantly. Other papers followed, and in 1933, at the age of 31, [Dirac](#) received the Nobel Prize in Physics. In 1937, he married [Margit Wigner](#), sister of the eminent physicist [Eugen Wigner](#). Between 1930 and 1940, [Dirac](#) focused on developing quantum electrodynamics; his last papers concerned general relativity. It was in 1926 that [Dirac](#) introduced his famous "function"  $\delta(t)$ , which is zero everywhere except at the origin and has an integral equal to 1, to represent a unit impulse at  $t = 0$  with no effect for  $t \neq 0$ .  $\delta(t)$  is not a function in the usual sense, as a function that is zero outside the origin has an integral of zero. The Dirac delta function was empirically manipulated for a long time until it found a rigorous mathematical justification within the framework of distribution theory developed by [Laurent Schwartz](#) in 1950 ([Schwartz, 1950](#)).

This function is not a classical function and can only be formally defined in the sense of distributions. It is difficult to enumerate all the roles played by this distribution, which is encountered in numerous calculations. The attribute "photo" attached to the Dirac delta function is there to remind us that it allows us, thanks to the sampling formula, to mathematically express the fact "that we sample the value of a signal at a given instant". But the Dirac delta function is more than that, as we will see. Historically, the notion of an impulse was introduced by physicists before mathematicians invented distributions. It should be noted that the impulse is in line with other

## 6. DIRAC DELTA FUNCTION (THE "PHOTO")

---

physical idealizations such as point mass, point charge, infinitely thin layers, etc., which are easily manageable in calculations but physically unrealizable. We will approach the Dirac delta function in this way: as an ideal that we can never exactly achieve but can approximate closely enough to be useful. In this context, "sufficiently close" is reached when we can no longer measure the duration of the impulse, or when the response time of the excited system is so much longer than the duration of the excitation that it doesn't matter. Thus, according to this definition, the same **stimulus** may or may not be considered an impulse; it will be up to you to judge based on the overall characteristics of the excited system. When dealing with an impulse, it is not important to specify its duration, and we will write that the excitation occurred at the origin of time,

$$\delta(t) = 0 \forall t \neq 0 \quad (6.1)$$

However, the integral of the impulse represents what the excited system will dissipate and must be defined,

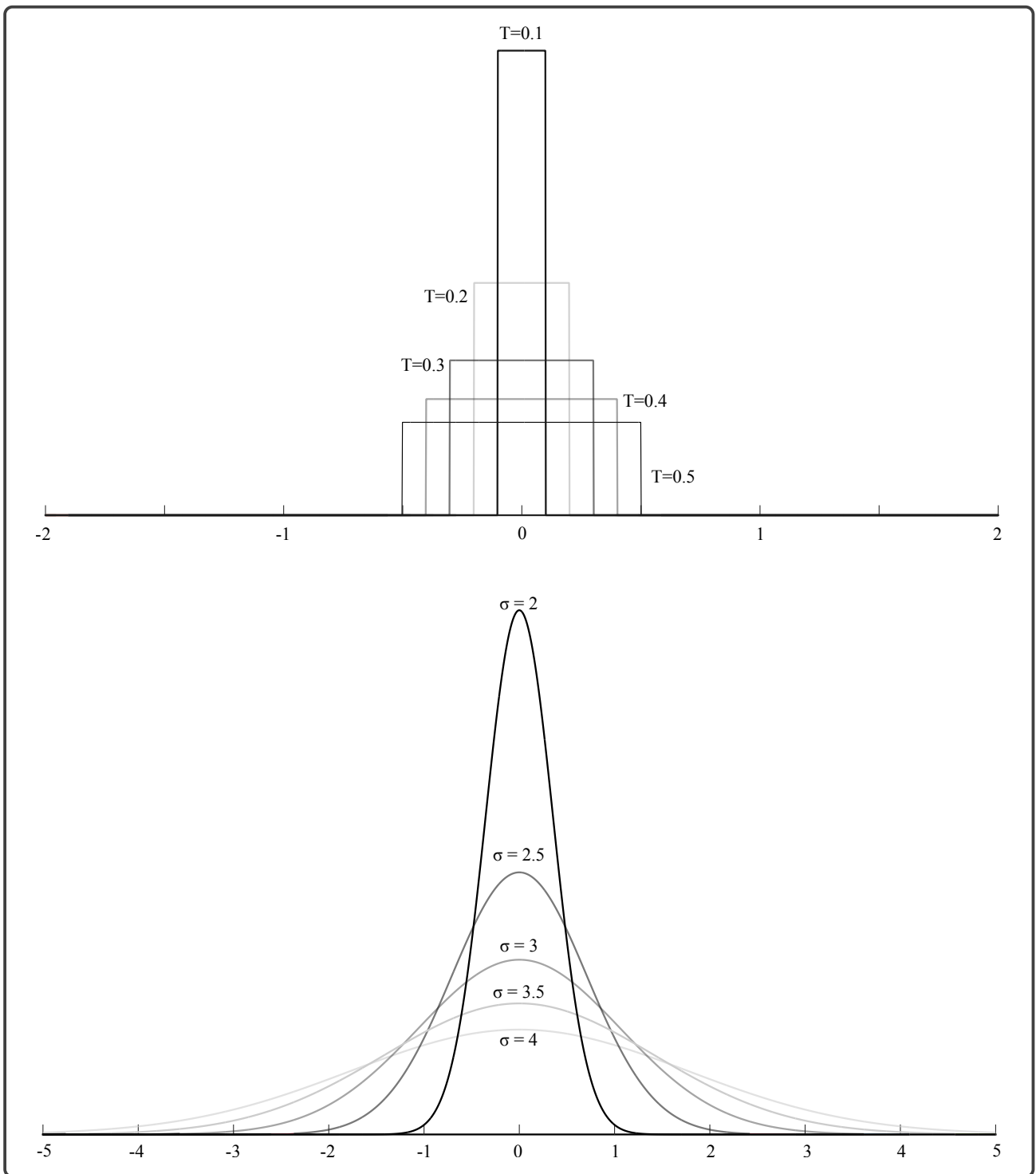
$$\int_{-\infty}^{+\infty} \delta(t) dt = 1 \quad (6.2)$$

The fundamental properties of the Dirac delta function can be established by representing  $\delta(t)$  as the limit of classical functions localized around the origin (see figure 4.3 obtained with the program [porte\\_gaussienne\\_vers\\_dirac.m](#)). One can, for example, use the window function,

$$\delta(t) = \lim_{T \downarrow 0} \frac{1}{T} \Pi\left(\frac{t}{T}\right) \quad (6.3)$$

where the Gaussian function,

$$\delta(t) = \lim_{\sigma \downarrow 0} \frac{1}{\sqrt{\pi}\sigma} \exp\left(-\frac{t^2}{\sigma^2}\right) \quad (6.4)$$



**Figure 4.3: Top: Evolution of the window function as  $T$  approaches 0. Bottom: Evolution of the Gaussian function as the standard deviation  $\sigma$  approaches 0. As can be seen in both cases, the temporal support decreases, the functions thin out and tend to infinity, thus approaching Dirac distributions.**

Let's demonstrate the sampling formula using this type of representation of the Dirac delta

## 6. DIRAC DELTA FUNCTION (THE "PHOTO")

---

function,

$$\begin{aligned}
 \int_{-\infty}^{+\infty} \delta(t) f(t) dt &= \lim_{T \downarrow 0} \int_{-\infty}^{+\infty} \frac{1}{T} \Pi\left(\frac{t}{T}\right) f(t) dt \\
 &= \lim_{T \downarrow 0} \frac{1}{T} \int_{-T/2}^{+T/2} f(t) dt \\
 &= \lim_{T \downarrow 0} \frac{1}{T} \int_{-T/2}^{+T/2} \left[ f(0) + t f^{(1)}(0) + \frac{t^2}{2} f^{(2)}(0) + \dots \right] dt \\
 &= \lim_{T \downarrow 0} \left[ f(0) + \frac{T^2}{24} f^{(2)}(0) + \dots \right] \\
 &= f(0).
 \end{aligned}$$

By generalizing this result, it is easy to establish that the [Dirac](#) delta function is the identity element of convolution,

$$[\delta * f](t) = f(t) \tag{6.5}$$

We deduce the translation formula,

$$\delta(t - t_0) * f(t) = f(t - t_0) \tag{6.6}$$

The [Plancherel](#) theorem immediately provides the Fourier transform of the Dirac delta function,

$$\mathcal{F}[\delta * f](u) = F(u) \implies \mathcal{F}[\delta(t)](u) = 1. \tag{6.7}$$

Among the many properties of the Dirac delta function, we cite,

$$\delta(-t) = \delta(t) \tag{6.8}$$

$$t\delta(t) = 0 \tag{6.9}$$

$$f(t)\delta(t - t_0) = f(t_0)\delta(t_0 - t) \tag{6.10}$$



---

and

$$\delta(\alpha t) = \frac{1}{|\alpha|} \delta(t) \quad (6.11)$$

The last property, which means that the Dirac delta function is a homogeneous distribution, is useful when dealing with Dirac combs that we will see later. It ensures the consistency of the Fourier transform of the Dirac delta function,

$$\begin{aligned} \delta(\alpha t) &= \mathcal{F}^{-1} \{ \mathcal{F} \{ \delta(\alpha t) \} \} (t) \\ &= \mathcal{F}^{-1} \left[ \frac{1}{|\alpha|} \right] (t) \\ &= \frac{1}{|\alpha|} \delta(t) \end{aligned} \quad (6.12)$$

## 7 Sign Function

This function is defined by,

$$\text{sgn}(t) = \begin{cases} +1 & \text{si } t > 0 \\ 0 & \text{si } t = 0 \\ -1 & \text{si } t < 0 \end{cases} \quad (7.1)$$

It has a [Fourier](#) transform that can only be calculated in the sense of distributions. To do this, let's introduce the function,

$$\phi_T(t) = \exp\left(-\frac{|t|}{T}\right) \text{sgn}(t) \quad (7.2)$$

of which the sign function is a limiting case,

$$\lim_{T \rightarrow \infty} \phi_T(t) = \text{sgn}(t). \quad (7.3)$$

We have,

$$\begin{aligned} \mathcal{F} [\phi_T(t)](u) &= \int_{-\infty}^{+\infty} \exp\left(-\frac{|t|}{T}\right) \operatorname{sgn}(t) \exp(-2i\pi ut) dt \\ &= -\int_{-\infty}^0 \exp\left(\frac{t}{T}\right) \exp(-2i\pi ut) dt + \int_0^{+\infty} \exp\left(-\frac{t}{T}\right) \exp(-2i\pi ut) dt \\ &= \frac{-4i\pi u}{1/T^2 - (2i\pi u)^2} \end{aligned} \quad (7.4)$$

When  $T \rightarrow +\infty$ , we obtain a limiting [Fourier](#) transform which is that of the sign function,

$$\mathcal{F} [\operatorname{sgn}(t)](u) = \frac{1}{i\pi u} \quad (7.5)$$

## 8 Heaviside Distribution (the switch)

The Heaviside step function is named after [Oliver Heaviside](#) (1850-1925) who was born in London. In his youth, he was interested in experiments on electricity, and he published his first scientific paper at the age of twenty-two. His work concerned the transmission of electrical signals in transatlantic cables. A controversy arose between him and the specialists who did not believe in his technique of reducing attenuation by using inductances judiciously placed along the cable. It was during this time that [Heaviside](#) established the telegrapher's equation,  $\frac{1}{C} \frac{\partial^2 V}{\partial x^2} = L \frac{\partial^2 V}{\partial t^2} + R \frac{\partial V}{\partial t}$  where  $C$ ,  $L$ , and  $R$  are the capacitance, inductance, and resistance of the line, respectively. Mainly concerned with the problems of transmitting electromagnetic signals over long distances, he predicted, simultaneously with [A.E. Kennelly](#) from Harvard University, the existence of the ionosphere. His studies on transient signals led him to develop a clever mathematical formalism that was a precursor to our current symbolic calculus, using [Fourier](#) and [Laplace](#) transforms. It was in this context that he invented the famous distribution now bearing his name,

$$H(t) = \begin{cases} +1 & \text{si } t > 0 \\ 1/2 & \text{si } t = 0 \\ 0 & \text{si } t < 0 \end{cases} \quad (8.1)$$

This distribution is essential for representing discontinuities such as those caused by the opening or closing of a circuit and for selecting the causal part of a signal. Additionally, convolution with

$H(t)$  allows for integrating a signal,

$$\begin{aligned} [H * f](t) &= \int_{-\infty}^{+\infty} f(\xi) H(t - \xi) d\xi \\ &= \int_{-\infty}^t f(\xi) d\xi, \end{aligned} \quad (8.2)$$

from which we deduce that,

$$\frac{d}{dt} (H * f) = \frac{d}{dt} H * f = f, \quad (8.3)$$

which shows that,

$$\frac{d}{dt} H(t) = \delta(t) \quad (8.4)$$

A rigorous demonstration of this result requires the use of distributions,

$$\begin{aligned} \frac{d}{dt} H(t) &= \lim_{h \downarrow 0} \frac{H(t+h) - H(t)}{h} \\ &= \lim_{h \downarrow 0} \frac{1}{h} \Pi\left(\frac{t+h/2}{h}\right) \\ &= \delta(t) \end{aligned} \quad (8.5)$$

Letting,

$$H(t) = \frac{1}{2} [1 + \operatorname{sgn}(t)] \quad (8.6)$$

the results from the previous sections immediately provide,

$$\mathcal{F} [H(t)](u) = \frac{1}{2} \delta(u) + \frac{1}{2i\pi u} \quad (8.7)$$

The second term on the right-hand side is equal to the inverse of the differentiation operator we encountered in the first chapter; it is the integration operator such that,

$$\begin{aligned} \mathcal{F} \left[ \int_{-\infty}^{+\infty} f(\xi) d\xi \right] (u) &= \mathcal{F} [H * f](u) \\ &= \frac{1}{2i\pi u} F(u) \end{aligned} \quad (8.8)$$

"Neglecting" the Dirac impulse in this expression is equivalent to ignoring a potential constant of integration."

## 9 Dirac Comb (the camera)

This distribution is extremely important for describing signal sampling. It is defined as a sequence of Dirac impulses occurring at a cadence of  $\tau = 1$ ,

$$\text{shah}(t) \equiv \sum_{n=-\infty}^{+\infty} \delta(t-n) \quad (9.1)$$

The main properties of this distribution are:

$$\text{shah}(t+n) = \text{shah}(t) \quad (n \in \mathbb{Z}) \quad (9.2)$$

which indicates that shah is periodic with period 1,

$$\int_{n-1/2}^{n+1/2} \text{shah}(t) dt = 1 \quad (n \in \mathbb{Z}) \quad (9.3)$$

and

$$\text{shah}(t) = 0 \quad (t \notin \mathbb{Z}) \quad (9.4)$$

which are directly established from the fundamental properties of the Dirac impulse. Furthermore,

$$\text{shah}\left(\frac{t}{\tau}\right) = |\tau| \sum_{n=-\infty}^{+\infty} \delta(t-n\tau) \quad (9.5)$$

This last relation is demonstrated using the homogeneity property of the Dirac impulse,

$$\begin{aligned} \text{shah}\left(\frac{t}{\tau}\right) &= \sum_{n=-\infty}^{+\infty} \delta\left(\frac{t}{\tau} - n\right) \\ &= \sum_{n=-\infty}^{+\infty} \delta\left(\frac{t-n\tau}{\tau}\right) \\ &= |\tau| \sum_{n=-\infty}^{+\infty} \delta(t-n\tau) \end{aligned} \quad (9.6)$$

The **Fourier** transform of the comb function can be calculated using a trick involving writing  $\text{shah}(t)$ , which is 1-periodic, as a **Fourier** series,

$$\text{shah}(t) = \sum_{n=-\infty}^{+\infty} \alpha_n \exp(2i\pi nt) \quad (9.7)$$

where the coefficients are,

$$\begin{aligned} \alpha_n &= \int_{-1/2}^{1/2} \text{shah}(t) \exp(-2i\pi nt) dt \\ &= \int_{-1/2}^{1/2} \delta(t) \exp(-2i\pi nt) dt \\ &= \exp(-2i\pi nt)|_{t=0} \\ &= 1 \end{aligned} \quad (9.8)$$

Thus,

$$\text{shah}(t) = \sum_{n=-\infty}^{+\infty} \exp(+2i\pi nt) \quad (9.9)$$

The **Fourier** transform of the comb is then,

$$\begin{aligned} \mathcal{F}[\text{shah}(t)](u) &= \sum_{n=-\infty}^{+\infty} \mathcal{F}[\delta(t-n)](u) \\ &= \sum_{n=-\infty}^{+\infty} \exp(-2i\pi nu) \\ &= \sum_{n=-\infty}^{+\infty} \exp(+2i\pi nu) \\ &= \text{shah}(u) \end{aligned} \quad (9.10)$$

The **Dirac** comb is its own **Fourier** transform.

## 10 Sine and Cosine Functions

The calculation of the [Fourier](#) transforms of these functions involves distributions, and we have:

$$\begin{aligned}\mathcal{F} [\cos(2\pi u_0 t)](u) &= \mathcal{F} \left[ \frac{\exp(-2i\pi u_0 t) + \exp(+2i\pi u_0 t)}{2} \right](u) \\ &= \frac{1}{2} \int_{-\infty}^{+\infty} \exp[-2i\pi(u+u_0)t] dt + \frac{1}{2} \int_{-\infty}^{+\infty} \exp[-2i\pi(u-u_0)t] dt \quad (10.1) \\ &= \frac{1}{2} \delta(u+u_0) + \frac{1}{2} \delta(u-u_0).\end{aligned}$$

An analogous reasoning yields,

$$\mathcal{F} [\sin(2\pi u_0 t)](u) = \frac{i}{2} \delta(u+u_0) - \frac{i}{2} \delta(u-u_0) \quad (10.2)$$

## 11 Form: Fourier Transforms

$$t \exp(-t) \mathbb{H}(t) \mapsto \frac{1}{(1+2i\pi u)^2} \quad (11.1)$$

$$\frac{1}{2} \Pi(t+1) + \frac{1}{2} \Pi(t-1) \mapsto \cos(2\pi u) \operatorname{sinc}(u) \quad (11.2)$$

$$\Pi(t) * \operatorname{sgn}(t) \mapsto -i \frac{\operatorname{sinc}(u)}{\pi u} \quad (11.3)$$

$$\Delta(t) * \operatorname{sgn}(t) \mapsto -i \frac{\operatorname{sinc}^2(u)}{\pi u} \quad (11.4)$$

$$\frac{1}{\sqrt{|t|}} \mapsto \frac{1}{\sqrt{|u|}} \quad (11.5)$$

---


$$|\cos(\pi t)| \mapsto \frac{1}{2} \text{shah}(u) \left[ \text{sinc}\left(u + \frac{1}{2}\right) + \text{sinc}\left(u - \frac{1}{2}\right) \right] \quad (11.6)$$

$$\exp(-\pi t^2) \cos(2\pi t) \mapsto \frac{1}{2} \exp\left[-\pi(u+1)^2\right] + \frac{1}{2} \exp\left[-\pi(u-1)^2\right] \quad (11.7)$$

$$J_1(2\pi t)/2t \mapsto \sqrt{1-u^2} \Pi(u/2) \quad (11.8)$$

$$J_0(2\pi t) \mapsto \pi^{-1} \Pi(u/2) (1-u^2)^{-1/2} \quad (11.9)$$

$$\tanh(\pi t) \mapsto -\text{icosech}(\pi u) \quad (11.10)$$

---

---

# CHAPTER 5

---

## SAMPLING

<b>1</b>	<b>Sampling</b> . . . . .	<b>80</b>
1.1	Signal truncation . . . . .	80
1.2	Discretization . . . . .	82
1.3	Correct Discretization: Shannon Interpolation . . . . .	83
1.4	Incorrect Discretization: Spectral Aliasing . . . . .	84
1.5	Analog-to-Digital Conversion: Quantization . . . . .	86



---

# 1 Sampling

We will now address a very important part of the course, and what we will see in this chapter constitutes one of the "launching pads" necessary for practical applications in signal processing. Analog signal processing, or continuous processing, is becoming increasingly rare, although it should be noted here that analog modification of signals still exists at the sensor level; however, digital signal processing is becoming more frequent due to the increasing power of computers and the flexibility allowed by digital processing, which permits operations that are unachievable by analog means, *eg*, non-causal filtering. Sampling is an essential step in digital signal processing; for a signal to be "digested" by the computer, it must be presented as a finite sequence (*id*, of limited duration) of values (*id*, discrete) coded on a certain number of bits. The operations of truncation, discretization, and quantization will modify the theoretical expressions we have seen so far (*eg*, the bounds of the **Fourier** integral will not be infinite) and the role of this chapter is to examine the main effects of sampling and their impact on the theoretical expressions seen so far.

## 1.1 Signal truncation

In many cases, the signal we wish to study is not known in its entirety, but only for a limited duration. The question that then arises is to what extent the sample we possess is representative of the total, unknown signal. We can represent the truncation of a signal using the window,

$$s_T(t) = s(t) \Pi\left(\frac{t-t_0}{T}\right) \quad (1.1)$$

where  $s_T(t)$  is the truncated part of the total signal  $s(t)$ . By transitioning into the dual space of **Fourier**,

$$S_T(u) = T \times S(u) * [\exp(-2i\pi ut_0) \text{sinc}(uT)] \quad (1.2)$$

which shows that the **Fourier** transform of the truncated signal is a degraded version of that of the total signal. The degradation results from the convolution by sinc, which has the effect of "mixing" the values of  $S(u)$ . When the observation period is long, the central lobe of the sinc function is very narrow, and the degradation is minimal; however, according to the similarity principle, if the recording window is short, the central lobe is wide and the frequency resolution is poor. To better understand

## 1. SAMPLING

---

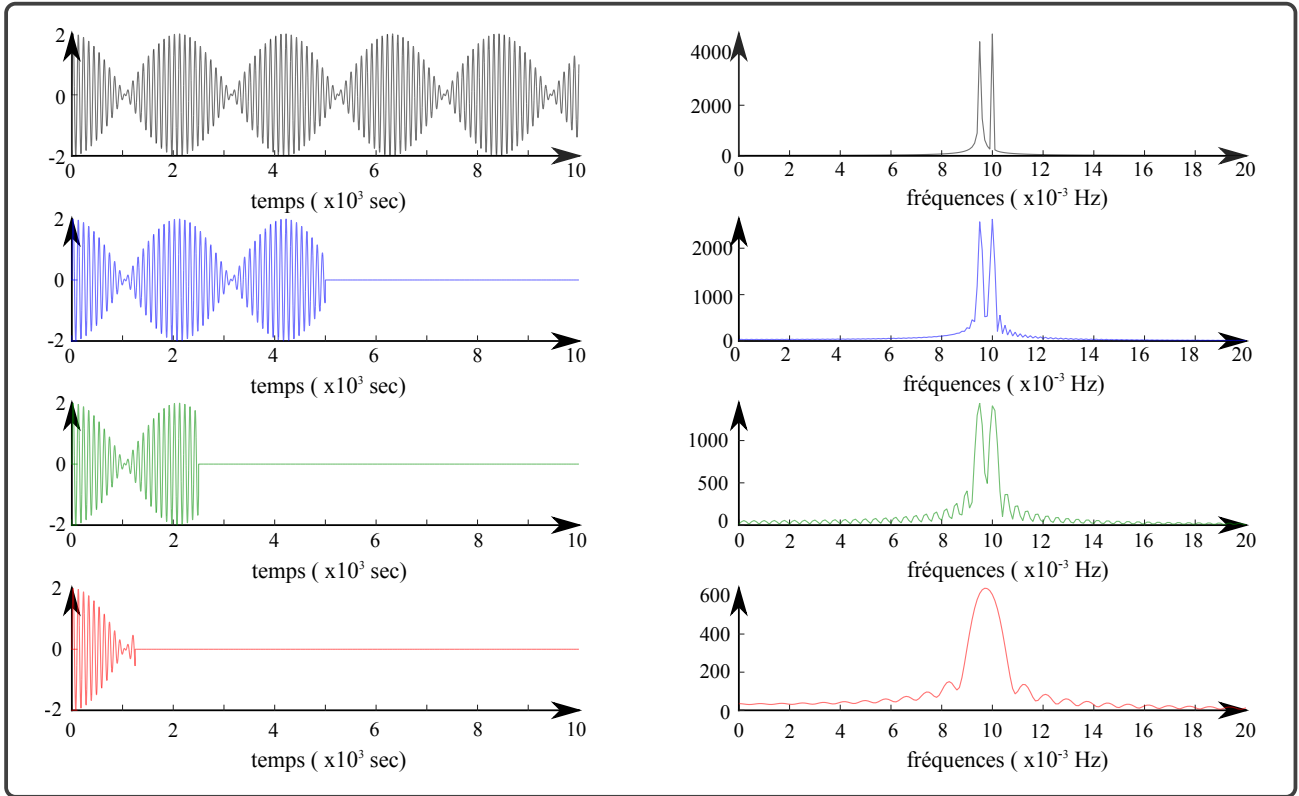
this, let us consider the simple case  $s(t) = \cos(2\pi u_0 t)$  for which  $s_T(t) = \cos(2\pi u_0 t) \Pi(t/T)$ . We then have,

$$S_T(u) = (T/2) \{ \text{sinc}[(u - u_0)T] + \text{sinc}[(u + u_0)T] \} \quad (1.3)$$

that is to say, the two Dirac impulses of  $S(u)$  are replaced by two sinc functions, which no longer allow for an infinitely precise determination of the frequency  $u_0$ . By analogy with the resolution of an optical instrument, we can define the frequency resolution as being equal to the half-width of the central lobe of the sinc functions,

$$\delta u \approx \frac{1}{T} \quad (1.4)$$

The longer the observation period, the better the resolution. The function [ex\\_troncature.m](#) illustrates the influence of truncation on frequency resolution. The results are shown in the figure [5.1](#).



**Figure 5.1:** Illustration of the effects of truncation on frequency resolution  $\delta u$ . The two frequencies become indistinguishable when the truncation no longer reveals the presence of beats in the signal. At that point, the truncated signal can be interpreted as a single damped sinusoid instead of two sinusoids producing beats. We indeed find that the amplitude of the Fourier transform is also modified and corresponds to half the duration of the truncation window (*cf* equation 1.3).

## 1.2 Discretization

### Spectral duplication

Discretization involves replacing the continuous signal  $s(t)$  with the sequence of values taken by the signal at multiples of the discretization interval  $\tau$ . The thus discretized signal constitutes a set of discrete values that can be represented by,

$$s_{\tau}(t) = s(t) \text{shah}\left(\frac{t}{\tau}\right) \quad (1.5)$$

This representation, which uses the product of a function with a distribution, is not very rigorous and only makes sense when it appears under an integral. This representation allows us to compute

the Fourier transform of the discretized signal,

$$\begin{aligned}
 S_\tau(u) &= \tau S(u) * \text{shah}(\tau u) \\
 &= \sum_{n=-\infty}^{+\infty} S(u) * \delta\left(u - \frac{n}{\tau}\right) \\
 &= \sum_{n=-\infty}^{+\infty} S\left(u - \frac{n}{\tau}\right)
 \end{aligned} \tag{1.6}$$

which shows us that  $S_\tau(u)$  consists of an infinite number of duplicates of  $S(u)$ , spaced at intervals of  $u_e = \tau^{-1}$ . The Fourier transform of a discretized signal is therefore a periodic function with period  $\tau^{-1}$ .

### 1.3 Correct Discretization: Shannon Interpolation

This theorem, established by Claude Shannon (1916-2001) while he was an engineer at Bell Laboratories (Shannon *et al.*, 1951), forms the foundation of discrete signal processing and information theory. If the signal has a bounded spectrum, meaning that  $S(u) = 0$  when  $|u| > u_c$ , the duplicates will not overlap if the sampling frequency is such that,

$$u_e > 2u_c \tag{1.7}$$

that is,

$$u_c < u_N \tag{1.8}$$

where the Nyquist frequency  $u_N = u_e/2$ . This condition, known as the Shannon sampling theorem, intuitively expresses the fact that the period of a periodic phenomenon can only be determined if the phenomenon is observed more than twice per period. Note that strictly sampling twice per period is insufficient; sample  $\sin(2\pi t)$  from  $t = 0$  and you will see! When this condition is satisfied, it is possible to recover the Fourier transform of the total signal,

$$S(u) = S_\tau(u) \Pi\left(\frac{u}{u_e}\right) \tag{1.9}$$

whence,

$$s(t) = \tau^{-1} s_\tau(t) * \text{sinc}(t/\tau) \quad (1.10)$$

that is,

$$\begin{aligned} s(t) &= \frac{1}{\tau} \left[ s(t) \text{shah} \left( \frac{t}{\tau} \right) \right] * \text{sinc} \left( \frac{t}{\tau} \right) \\ &= \frac{1}{\tau} \left[ \tau \sum_{n=-\infty}^{+\infty} s(t) \delta(t - n\tau) \right] * \text{sinc} \left( \frac{t}{\tau} \right) \\ &= \sum_{n=-\infty}^{+\infty} [s(t) \delta(t - n\tau)] * \text{sinc} \left( \frac{t}{\tau} \right) \\ &= \sum_{n=-\infty}^{+\infty} \int_{-\infty}^{+\infty} \text{sinc} \left( \frac{t - \xi}{\tau} \right) s(\xi) \delta(\xi - n\tau) d\xi \\ &= \sum_{n=-\infty}^{+\infty} s(n\tau) \text{sinc} \left( \frac{t - n\tau}{\tau} \right), \end{aligned} \quad (1.11)$$

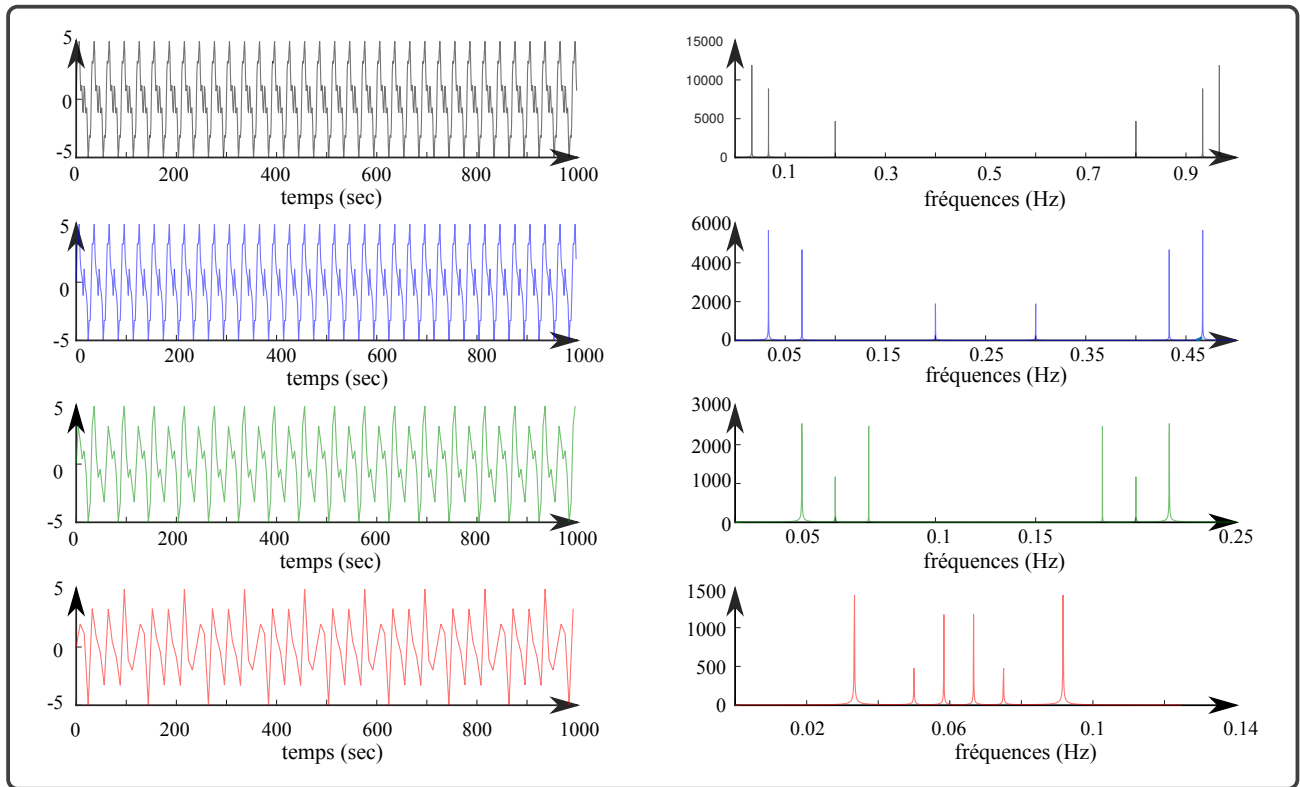
where the transition to the last line uses the sampling formula. The final equality is known as the "Shannon interpolation formula" and allows for the recovery of the continuous signal from the discrete series  $s_\tau(t)$ . It is verified that for  $t = k\tau$ ,  $\text{sinc}(k-) = \delta_k^n$  and that the interpolation formula correctly yields  $s(k\tau)$ .

## 1.4 Incorrect Discretization: Spectral Aliasing

The phenomenon of spectral aliasing is an artifact that occurs when the discretization of a signal does not satisfy the [Shannon](#) sampling theorem. In this case,

$$\tau^{-1} = u_e < 2u_c \quad (1.12)$$

and the *duplicates* overlap. The function [ex\\_replielement.m](#) illustrates the influence of discretization. The results are shown in [Figure 5.2](#).



**Figure 5.2: Illustration of the effects of discretization and spectral aliasing on the apparent frequencies of three sinusoids with frequencies of 0.2 Hz, 0.067 Hz, and 0.033 Hz. The sampling periods, from bottom to top (from black to red), are 1 second (1 Hz), 2 seconds (0.5 Hz), 4 seconds (0.25 Hz), and 8 seconds (0.125 Hz). The initial signal (top left) is discretized with a one-second interval and results from the superposition of sinusoids, including two high-frequency ones, as shown by the magnitude of the Fourier transform of the signal (top right). Under-sampling with a two-second interval (blue curves) violates the Shannon condition and causes spectral aliasing. Under-sampling with a four-second interval (green curves) further distorts the spectral lines.**

It is thus impossible to recover  $S(u)$  as we did previously (Figure 5.2). A sinusoidal signal with a frequency  $u_0 > u_e/2$  will be converted into a signal with an apparent frequency  $u_a = u_0 - mu_e$  where  $m$  is the integer such that  $|u_a| < u_e/2$ . This phenomenon is analogous to a stroboscopic effect, where the apparent rotational speed of a mechanical part depends on the ratio between the actual rotational speed and the strobe light frequency. Spectral aliasing is a very serious problem because it transfers energy from high frequencies to low frequencies, resulting in an unacceptable spectrum (Figure 5.2). Before sampling a signal, one must either ensure that it does not contain significant energy outside the interval  $[-u_e/2; +u_e/2]$ , or filter the signal with a low-pass filter to remove high frequencies before the discretization process. You might think that aliasing can only occur during analog-to-digital conversions at the sensor level. This is incorrect, and experience shows that aliasing often occurs within the computer when, for practical reasons, "one only takes one point out of five because it will be sufficient and takes up less space"! A final example of spectral

---

aliasing, which leads to incorrect interpretations, is shown in Figure 5.3. This was obtained from the function `ex_shannon.m`. As can be seen, we sampled at 100 Hz four sinusoids with frequencies of 0.5 Hz, 99.5 Hz, 100.5 Hz, and 200.5 Hz. Despite these different frequencies, the waveforms (top) are rigorously identical, and their respective Fourier spectra (bottom) suggest that these four signals are the same and beat at 0.5 Hz.

**Vocabulary** – In this book, we use the term [spectral duplication](#) to refer to the phenomenon that occurs when discretizing a signal using the comb function. The term [aliasing](#) is used to describe what happens when discretization does not satisfy the Shannon condition. In many texts, especially those written in English, [you will encounter the term "aliasing," which has a dual meaning as it can refer either to spectral duplication or to aliasing.](#)

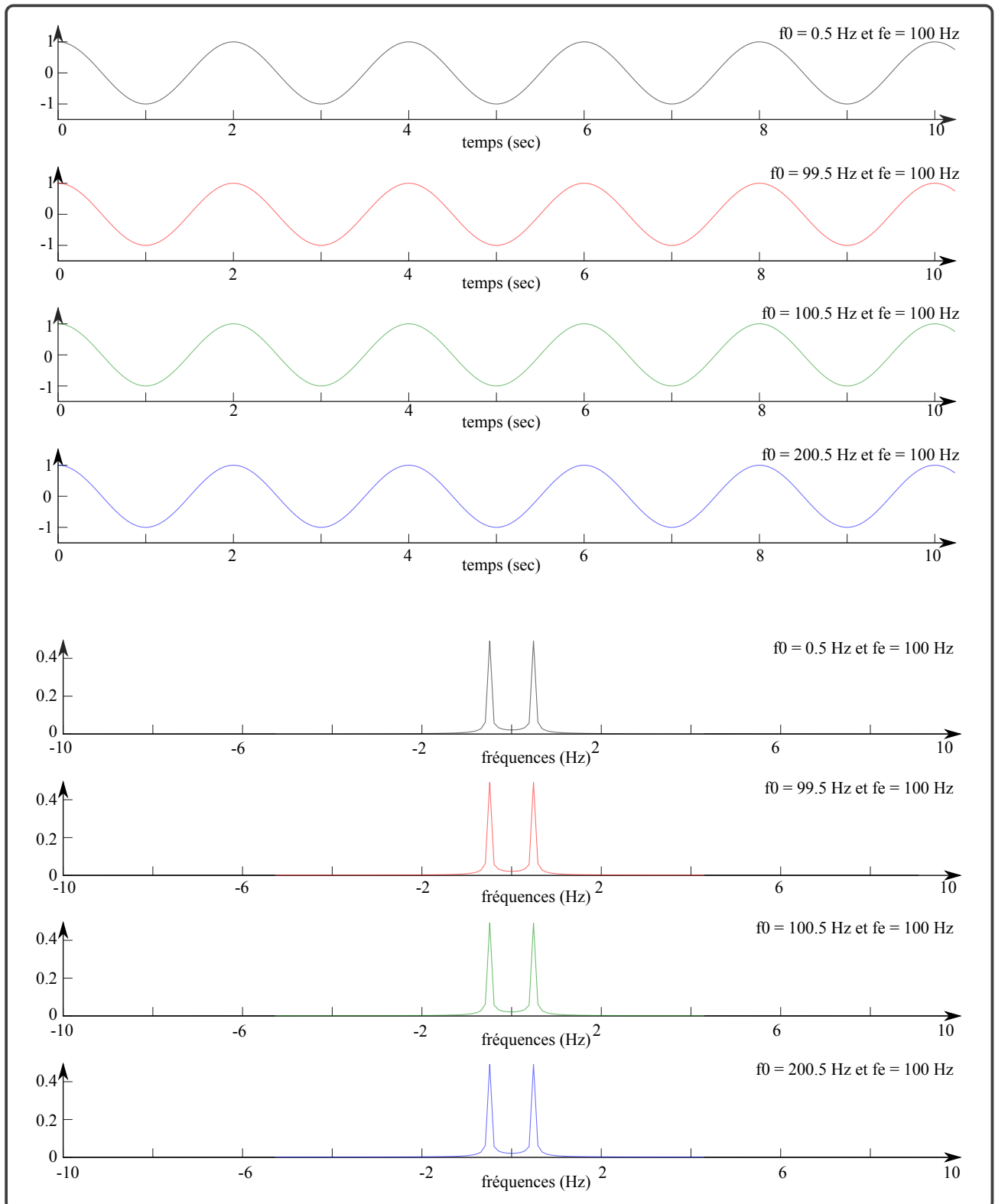
## 1.5 Analog-to-Digital Conversion: Quantization

Quantization occurs during the analog-to-digital conversion, which provides a signal generally encoded in base 2. The smallest value that can be encoded is 1, and if the encoding is done with  $n$  bits, the largest value is  $2^n - 1$ . The encoding process will reduce the infinite number of possible values that the analog signal can take to a finite and relatively small number of digital values; we will see later that this process is accompanied by the generation of quantization noise. An encoding can be characterized by its dynamic range,

$$\text{dynamique (dB)} \equiv 20 \log 2^n \tag{1.13}$$

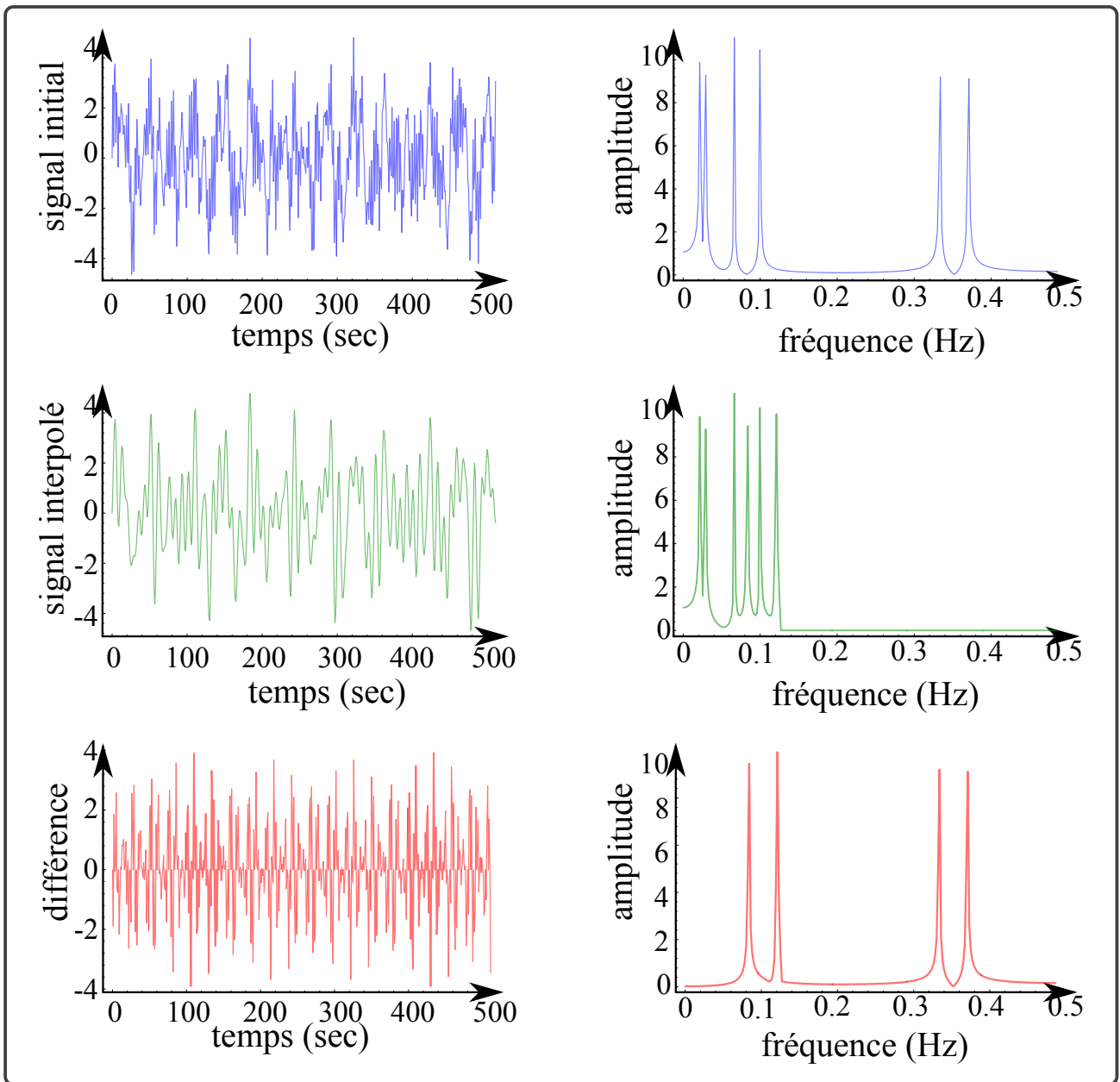
For example, a 12-bit converter has a dynamic range of approximately 72 dB. This is the ratio between the smallest value and the largest value that can be converted.

# 1. SAMPLING



**Figure 5.3: Example of spectral aliasing. At the top, the identical waveforms of four sinusoids with different frequencies, whose spectra are also identical because, in 3 out of 4 cases (red, green, and blue curves), we do not meet the Shannon sampling theorem.**





**Figure 5.4:** Loss of information due to spectral aliasing. The signal in the middle left is constructed by under-sampling the initial signal (top left) with a two-second interval followed by Shannon interpolation at a one-second interval. Although the constructed signal is sampled as finely as the initial signal, the aliasing that occurred during under-sampling has not been eliminated, as shown by the amplitude spectrum in the middle right. The difference (bottom right) between this spectrum and that of the initial signal (top right) shows that the interpolated signal no longer contains the original high frequencies between 0.3 and 0.4 Hz, which have been aliased around 0.1 Hz. This is also evident in the difference (bottom left) between the two signals, which contains the high-frequency oscillations missing in the interpolated signal.

---

---

# CHAPTER 6

---

## THE Z-TRANSFORM

1	The Utility of the Z-Transform . . . . .	90
2	Formulary: Z-Transforms . . . . .	91

# 1 The Utility of the Z-Transform

Instead of a formal mathematical approach, we will present the Z-transform as a convenient notation for manipulating the [Fourier](#) transforms of signals discretized at a constant interval,

$$s_\tau(t) = s(t) \text{shah}\left(\frac{t}{\tau}\right) \quad (1.1)$$

for which,

$$\begin{aligned} S_\tau(u) &= \tau \int_{-\infty}^{+\infty} \left\{ \sum_{n=-\infty}^{+\infty} s(t) \delta(t - n\tau) \right\} \exp(-2i\pi ut) dt \\ &= \tau \sum_{n=-\infty}^{+\infty} \int_{-\infty}^{+\infty} [s(t) \exp(-2i\pi ut)] \delta(t - n\tau) dt \\ &= \tau \sum_{n=-\infty}^{+\infty} s(n\tau) \exp(-2i\pi un\tau) \end{aligned} \quad (1.2)$$

where the factor  $\tau$  ensures correct scaling and equivalence between this expression and the continuous [Fourier](#) transform. The Z-transform is simply obtained by making the following variable change,

$$Z \equiv \exp(-2i\pi u\tau) \quad (1.3)$$

Some authors use the conjugate definition,  $Z \equiv \exp(+2i\pi u\tau)$ . In any case, this allows expression (1.2) to be rewritten in the form of a Z-transform,

$$\begin{aligned} S(Z) &= \tau \sum_{n=-\infty}^{+\infty} s(n\tau) Z^n \\ &= \tau [\dots + s(-\tau) Z^{-1} + s(0) + s(\tau) Z + s(2\tau) Z^2 + \dots] \end{aligned} \quad (1.4)$$

which is a polynomial in  $Z$ . In fact, this definition can be extended to any complex  $Z$ , but the choice we have made is appropriate because it allows for the equivalence between the Z-transform and the [Fourier](#) transform of discrete signals. The above expressions show that the Z-transform is merely a way of writing the Z, but the choice we have made is appropriate because it allows for the equivalence between the Z-transform and the [Fourier](#) transforms of discrete signals. When performing calculations involving the Fourier transforms of such signals, it is up to you to decide if using the Z-transform notation is useful or not. We encourage the reader to consult Jon [Claerbout's](#)

## 2. FORMULARY: Z-TRANSFORMS

---

book, *Fundamentals of Geophysical Data Processing* (Claerbout, 1985), to see numerous applications of the Z-transform. It is clear that all properties of the Fourier transform are preserved for the Z-transform, whose primary interest lies in the manipulation of discrete signals. For example, just as

$$\exp(-2i\pi u\tau)S(u) \tag{1.5}$$

est la transformée de **Fourier** du signal  $s(t)$  retardé d'un pas de temps,

$$ZS(Z) \tag{1.6}$$

is the Z-transform of the discrete signal  $s_\tau(t)$  delayed by the same time step: Z can be considered as the unit delay operator. Similarly, the Z-transform of the convolution of two discrete signals is equal to the product of their Z-transforms.

## 2 Formulary: Z-Transforms

$$[a_0, a_1, a_2, \dots] \mapsto a_0 + a_1Z + a_2Z^2 + \dots \tag{2.1}$$

$$[1, 1, 1, \dots] \mapsto \frac{1}{1-Z} \tag{2.2}$$

$$[0, 1, 1, \dots] \mapsto \frac{Z}{1-Z} \tag{2.3}$$

$$[0, 1, 2, 3, \dots, n, \dots] \mapsto \frac{Z}{(1-Z)^2} \tag{2.4}$$

$$[0, 1, 4, 9, \dots, n^2, \dots] \mapsto \frac{Z(1+Z)}{(1-Z)^3} \tag{2.5}$$

---


$$[0, 1, 8, 27, \dots, n^3, \dots] \mapsto \frac{3Z^2(1+Z)}{(1-Z)^4} + \frac{Z(1+2Z)}{(1-Z)^3} \quad (2.6)$$

$$[1, \exp(-\alpha), \exp(-2\alpha), \dots] \mapsto \frac{1}{1 - \exp(-\alpha)Z} \quad (2.7)$$

$$[0, 1 - \exp(-\alpha), 1 - \exp(-2\alpha), \dots] \mapsto \frac{Z(1 - \exp(-\alpha)Z)}{(1-Z)(1 - \exp(-\alpha)Z)} \quad (2.8)$$

$$[0, \exp(-\alpha), 4\exp(-2\alpha), \dots] \mapsto \frac{(1 + \exp(-\alpha)Z)\exp(-\alpha)Z}{(1 - \exp(-\alpha)Z)^3} \quad (2.9)$$

$$[0, \exp(-\alpha), 2\exp(-2\alpha), \dots] \mapsto \frac{\exp(-\alpha)Z}{(1 - \exp(-\alpha)Z)^2} \quad (2.10)$$

---

---

# CHAPTER 7

---

## THE DISCRETE FOURIER TRANSFORM

<b>1</b>	<b>The Discrete Fourier Transform</b> . . . . .	<b>94</b>
1.1	Discretization of the Fourier Transform . . . . .	94
1.2	The Fast Fourier Transform Algorithm . . . . .	95

---

# 1 The Discrete Fourier Transform

## 1.1 Discretization of the Fourier Transform

Just as we discussed the issue of discretizing time-domain signals, we will now address the discretization of their Fourier transforms. To be correct, the discretization of time-domain signals must satisfy the Shannon criterion. Some remarks based on the duality of the Fourier transform will allow us to establish an equivalent rule without performing any calculations. We have seen that the discretization of a signal can only be rigorously achieved if its Fourier transform has limited support. By duality, we infer that the discretization of the Fourier transform can only be done for signals with limited temporal support. Poor temporal discretization leads to spectral aliasing: poor frequency sampling will cause aliasing of the signal. There will be no temporal aliasing if the frequency discretization satisfies the dual Shannon criterion.

$$\nu < \frac{1}{T}, \quad (1.1)$$

where  $T$  is the duration of the temporal support of the signal and  $\nu$  is the frequency sampling interval. We have also seen that temporal discretization induces a periodicization of the Fourier transform: frequency discretization induces a temporal periodicization. Strictly speaking, it is only possible to discretize the signal and its Fourier transform without violating the Shannon criterion and its dual if the signal is periodic. In practice, a discrete Fourier transform will therefore always be the Fourier transform of a periodic signal. This limitation is severe and should never be forgotten. The previous reasoning allows us to derive the formula for the discrete Fourier transform. It can also be obtained using a more intuitive approach by revisiting the Fourier transform of a discretized signal,

$$S_\tau(u) = \tau \sum_{n=-\infty}^{+\infty} s(n\tau) \exp(-2i\pi un\tau), \quad (1.2)$$

that is, after truncating (centered at the origin) the signal to a duration  $T$ ,

$$S_{\tau,T}(u) = T \operatorname{sinc}(Tu) * \left[ \tau \sum_{n=-\infty}^{+\infty} s(n\tau) \exp(-2i\pi un\tau) \right] \quad (1.3)$$

The frequency resolution can be taken as half the width of the central lobe of the sinc function,

$$v = \frac{1}{T} \tag{1.4}$$

The range of useful frequencies being  $[0; \tau^{-1}]$ , the discrete frequencies are found to be

$$u_k = kv \quad (k = 0, 1, \dots, N - 1) \tag{1.5}$$

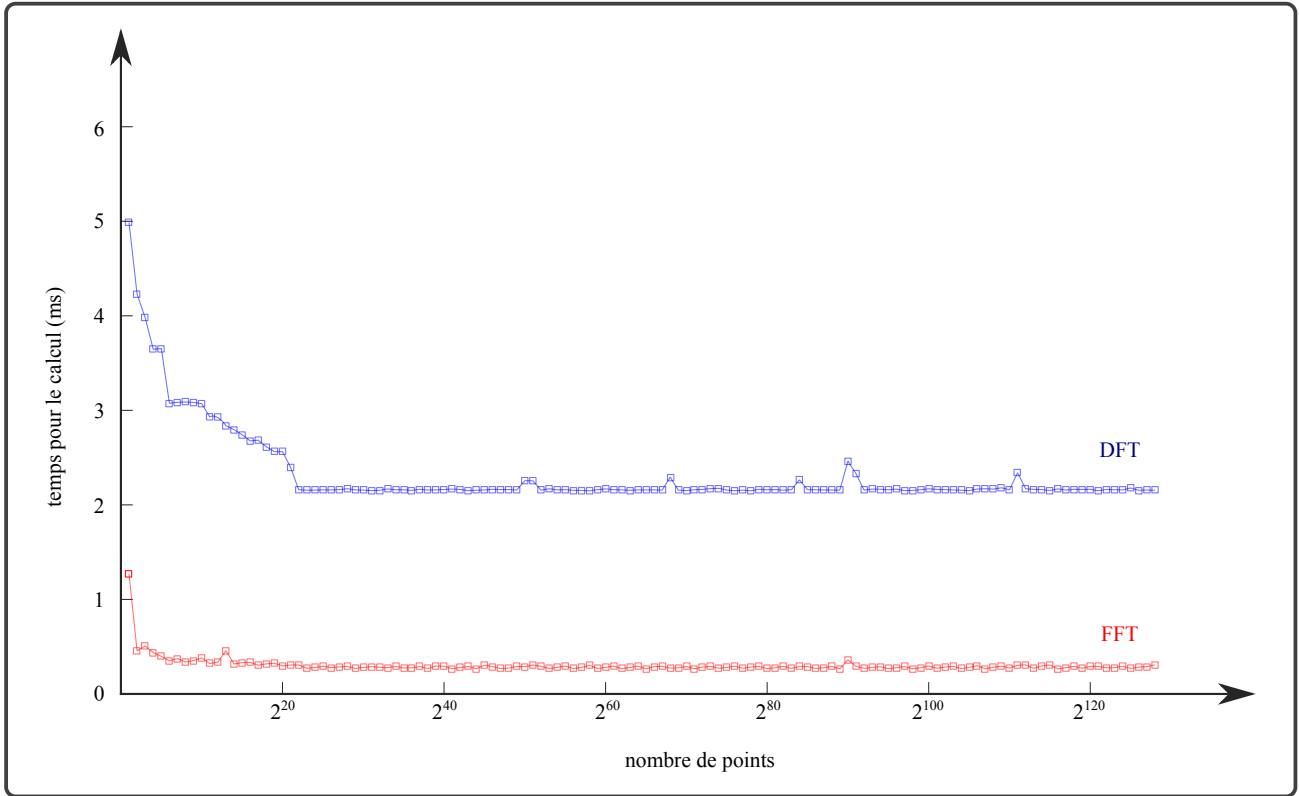
where  $N$  is the number of points in the truncated discrete signal. Under these conditions, the discrete **Fourier** transform is given by

$$S_{\tau,T}(kv) = \tau \sum_{n=0}^{N-1} s(n\tau) \exp\left(\frac{-2i\pi kn}{N}\right) \quad (k = 0, 1, \dots, N - 1) \tag{1.6}$$

## 1.2 The Fast **Fourier** Transform Algorithm

The algorithm we will examine in this section was a revolution in numerical analysis and signal processing; it is considered one of the ten greatest algorithms of the 20<sup>th</sup> century (Cipra, 2000). It allows for the rapid computation of discrete **Fourier** transforms of digitized signals, which, at the time of its discovery, made many previously impractical analysis methods feasible. To give you an idea of the algorithm's power, we will cite only the test results by Jon Claerbout (Claerbout, 1992), where his "slow" program—i.e., the one implementing the double loop of the discrete **Fourier** transform with  $N^2$  iterations—takes 153 seconds to compute the discrete **Fourier** transform of a 1024-point signal, while the program using the fast algorithm takes only 0.7 seconds! The program `ex_dft_vs_fft.m` compares computation times between a fast **Fourier** transform and a discrete **Fourier** transform on Matlab®. In the program, we use the `fft` function indiscriminately, but it is important to know that if the signal whose spectrum we want to compute does not have a dimension that is a power of 2, the classical algorithm implementing the double loop is used. Even though software and computers have made enormous advancements since Claerbout's test, Figure (7.1) still shows up to a factor of 5 difference in computation time between the two algorithms. The fast Fourier transform has been widely used since the famous 1965 article by James Cooley and John Tukey (Cooley et Tukey, 1965), although the algorithm was originally conceived by Carl Friedrich Gauss in 1805, and has been adapted several times since, including notable work by Cornelius Lanczos in 1942 (Danielson et Lanczos, 1942).





**Figure 7.1:** Comparison of computation times between the DFT and the FFT. The Fourier transform is performed on a signal consisting of  $2^N$  points in red and  $2^N - 1$  points in blue. It is observed that the so-called fast algorithm is nearly 5 times faster.

Let us denote,

$$W \equiv \exp(-2i\pi/N) \tag{1.7}$$

the discrete Fourier transform then takes the form,

$$S_{\tau,T}(kv) = \tau \sum_{n=0}^{N-1} s(n\tau) W^{kn} \quad (k = 0, 1, \dots, N-1) \tag{1.8}$$

which, when adopting a matrix notation, becomes

$$\begin{bmatrix} S(0) \\ S(v) \\ S(2v) \\ S(3v) \\ \vdots \end{bmatrix} = \tau \begin{bmatrix} 1 & 1 & 1 & 1 & \dots \\ 1 & W & W^2 & W^3 & \dots \\ 1 & W^2 & W^4 & W^6 & \dots \\ 1 & W^3 & W^6 & W^9 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \cdot \begin{bmatrix} s(0) \\ s(\tau) \\ s(2\tau) \\ s(3\tau) \\ \vdots \end{bmatrix} \tag{1.9}$$

## 1. THE DISCRETE **FOURIER** TRANSFORM

---

The inverse matrix has the same structure and is obtained by replacing  $W$  with  $W^{-1}$ , so

$$\begin{bmatrix} s(0) \\ s(\tau) \\ s(2\tau) \\ s(3\tau) \\ \vdots \end{bmatrix} = \tau \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots \\ 1 & W^{-1} & W^{-2} & W^{-3} & \cdots \\ 1 & W^{-2} & W^{-4} & W^{-6} & \cdots \\ 1 & W^{-3} & W^{-6} & W^{-9} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \cdot \begin{bmatrix} S(0) \\ S(\nu) \\ S(2\nu) \\ S(3\nu) \\ \vdots \end{bmatrix}, \quad (1.10)$$

which amounts to computing the inverse discrete **Fourier** transform,

$$s(n\tau) = \nu \sum_{k=0}^{N-1} S_{\tau,T}(k\nu) W^{-kn} \quad (n = 0, 1, \dots, N-1). \quad (1.11)$$

Here we recognize that the matrix  $W$  is a **Vandermonde** matrix. The above matrix equations require  $N^2$  multiplications and as many additions, which quickly becomes enormous, explaining the significance of the work by **Cooley** and **Tukey** in 1965. Their algorithm made it possible to compute discrete **Fourier** transforms very quickly by reducing the calculation of a discrete **Fourier** transform of length  $N$  to that of two transforms of length  $N/2$ . In fact, **J. Claerbout** notes that **Vern Herbert** of *Chevron Standard Ltd.* had already programmed this as early as 1962. In practice, the discrete signal  $s(n\tau)$  is decomposed into two interleaved signals  $^1s$  and  $^2s$  such that

$$\begin{cases} s'(n) \equiv s(2n\tau) & (n = 0, 1, \dots, N/2-1) \\ s''(n) \equiv s[(2n+1)\tau] & (n = 0, 1, \dots, N/2-1) \end{cases} \quad (1.12)$$

We then obtain,

$$\begin{aligned} S_{\tau,T}(k\nu) &= \tau \sum_{n=0}^{N/2-1} s(2n\tau) W^{2kn} + \tau \sum_{n=0}^{N/2-1} s[(2n+1)\tau] W^{k(2n+1)} \\ &= \tau \sum_{n=0}^{N/2-1} s'(n) (W^2)^{kn} + \tau W^k \sum_{n=0}^{N/2-1} s''(n) (W^2)^{nk} \end{aligned} \quad (1.13)$$

for  $k = 0, 1, \dots, N/2-1$ , the two sums are the discrete **Fourier** transforms of the interleaved signals,

$$S_{\tau,T}(k\nu) = \frac{S'_{2\tau,T}(k\nu)}{2} + \frac{S''_{2\tau,T}(k\nu)}{2} W^k \quad (1.14)$$

For  $k = 0, 1, \dots, N/2 - 1$ , and setting  $k = l + N/2$ , the two sums can be written as

$$\begin{aligned}
& S_{\tau,T} \left[ \left( l + \frac{N}{2} \right) \mathbf{v} \right] \\
&= \tau \sum_{n=0}^{N/2-1} s'(n) (W^2)^{n(l+\frac{N}{2})} + \tau W^{l+N/2} \sum_{n=0}^{N/2-1} s''(n) (W^2)^{n(l+\frac{N}{2})} \\
&= \tau \sum_{n=0}^{N/2-1} s'(n) W^{nN} (W^2)^{nl} + \tau W^{N/2} W^l \sum_{n=0}^{N/2-1} s''(n) W^{nN} (W^2)^{nl}
\end{aligned} \tag{1.15}$$

but  $W^{lN} = 1$  and  $W^{N/2} = -1$ , which allows us to obtain the simplified form

$$S_{\tau,T} \left[ \left( k + \frac{N}{2} \right) \mathbf{v} \right] = \frac{S'_{2\tau,T}(k\mathbf{v})}{2} - \frac{S''_{2\tau,T}(k\mathbf{v})}{2} W^k \tag{1.16}$$

with  $k = 0, 1, \dots, N/2 - 1$ . Thus, the computation of the discrete **Fourier** transform of a series with  $N$  values has been reduced to that of two transforms of interlaced series with  $N/2$  values each. If  $N = 2^p$ , this reduction can be performed  $p$  times, starting the process by computing the discrete **Fourier** transforms of  $N$  series containing only one value, then of series with 2 values, then 4, and so forth, up to the complete series. Overall, the number of operations is significantly reduced: the algorithm described enables the calculation of the transform of a series of  $N$  values with only  $2Np$  operations, compared to  $2N^2$  for the direct algorithm using the matrix form.

To illustrate what we have just discussed, the following subroutine, TFR, written in **Fortran** – which stands for **F**ormula **T**ranslator – implements the Fast **Fourier** Transform (FFT) algorithm. The program computes the direct transform when the variable `dirinv=1` and the inverse transform when `dirinv=-1`. The complex values of the signal are provided in the array `signal`, and the number of values, `n`, must be an integer power of 2. Other programs can be found in some of the books cited in the bibliographic references ([Claerbout \(1992\)](#), [Claerbout \(1985\)](#), [Kanasewich \(1981a\)](#), and [Press et al. \(1986\)](#)).

```
subroutine TFR(dirinv,signal,n)

  integer n,i,j,k,m,istep
  real dirinv,scale,arg
  complex signal(n),cplx,cw,cdel,ct

  scale=1./sqrt(float(n))
  do i =1,n
    signal(i)=signal(i)*scale
  end do
  j=1
  k=1

  do i=1,n
    if(i.le.j) then
      ct=signal(j)
      signal(j)=signal(i)
      signal(i)=ct
    end if
    m=n/2
    do while (j.gt.m.and.m.gt.1)
      j=j-m
      m=m/2
    end do
    j=j+m
  end do

  do while(k.ge.n)
    istep=2*k
    cw=1.
    arg=dirinv*3.14159265/float(k)
    cdel=cplx(cos(arg),sin(arg))
    do m=1,k
```

```

do i=m,n,istep
    ct=cw*signal(i+k)
    signal(i+k)=signal(i)-ct
    signal(i)=signal(i)+ct
    cw=cw*cdel
end do
k=istep
end do

return

end

```

Go deeper ..., but not too deep The relations (1.9) and (1.10) illustrate that the Discrete Fourier Transform (DFT) is essentially the product of a well-known Vandermonde matrix, which contains all the frequencies necessary for the decomposition (resp. reconstruction) of our signal  $s$  (resp.  $S$ ) via the coefficients of the sinusoids that facilitate these transformations. These coefficients are the unknowns. Therefore, we can view the direct (and inverse) Fourier transform as an inverse problem, which could be expressed using the notation from William Menke's book, *Geophysical Data Analysis: Discrete Inverse Theory* (Menke, 1984), as follows for the direct transform,

$$S = Ws \tag{1.17}$$

and,

$$s = W^{-1}S \tag{1.18}$$

for the inverse transform. We will not delve into the details of inverse problem theory (linear, non-linear, gradient, conditioning, *etc*), but will instead illustrate the relations (1.9) and (1.10) using the program `ex_fourier_coeff.m`, which calls the functions `fourier_coeff.m` and `fourier_reconstruct.m`.

Figure (7.2) illustrates the results. At the top, we have the Fourier spectra of a window function (shown in the two figures below in black line), ranging from 0 to the Nyquist frequency (0.05 Hz).

## 1. THE DISCRETE FOURIER TRANSFORM

---

The blue curve was obtained using Matlab's native `fft` function, while the red curve was obtained by inverting relation (1.9). As can be seen, the two spectra are identical. The two figures below correspond to the reconstruction of the window function, which means inverting relation (1.10), using a limited number of frequencies—i.e., not using all the frequencies previously calculated. For the calculation of  $S$ , it is evident that the number of frequencies to be computed, as previously discussed, must be at least equal to the number of points in the window  $s$ . For the inverse operation, and as we have also seen, to accurately reconstruct the original signal, we need to use at least as many frequencies as there are points in the signal. Here we illustrate two cases where, out of the 4096 points of the window  $s$ , thus requiring at least 4096 frequency values, we keep only 10 (center figure) and 100 (bottom figure). The sum of all reconstructed components should theoretically be exactly equal to the original signal (according to the Plancherel theorem, cf relation 6.15, and Shannon's reconstruction formula, cf relation 1.11). However, the window function is one of the rare cases where this is not possible, which explains the apparent oscillations, also known as the Gibbs effect, because its spectrum is not of bounded support! Remember that this function has a cardinal sine as its Fourier transform, cf figure (4.1).

We could have intuitively predicted this, as ultimately, this function is discontinuous and transitions from 0 to 1 in an infinitesimally small amount of time. Invoking Heisenberg's uncertainty principle, cf relation (6.18), implies that the frequency required to describe this jump must be infinite, which is physically and numerically impossible.

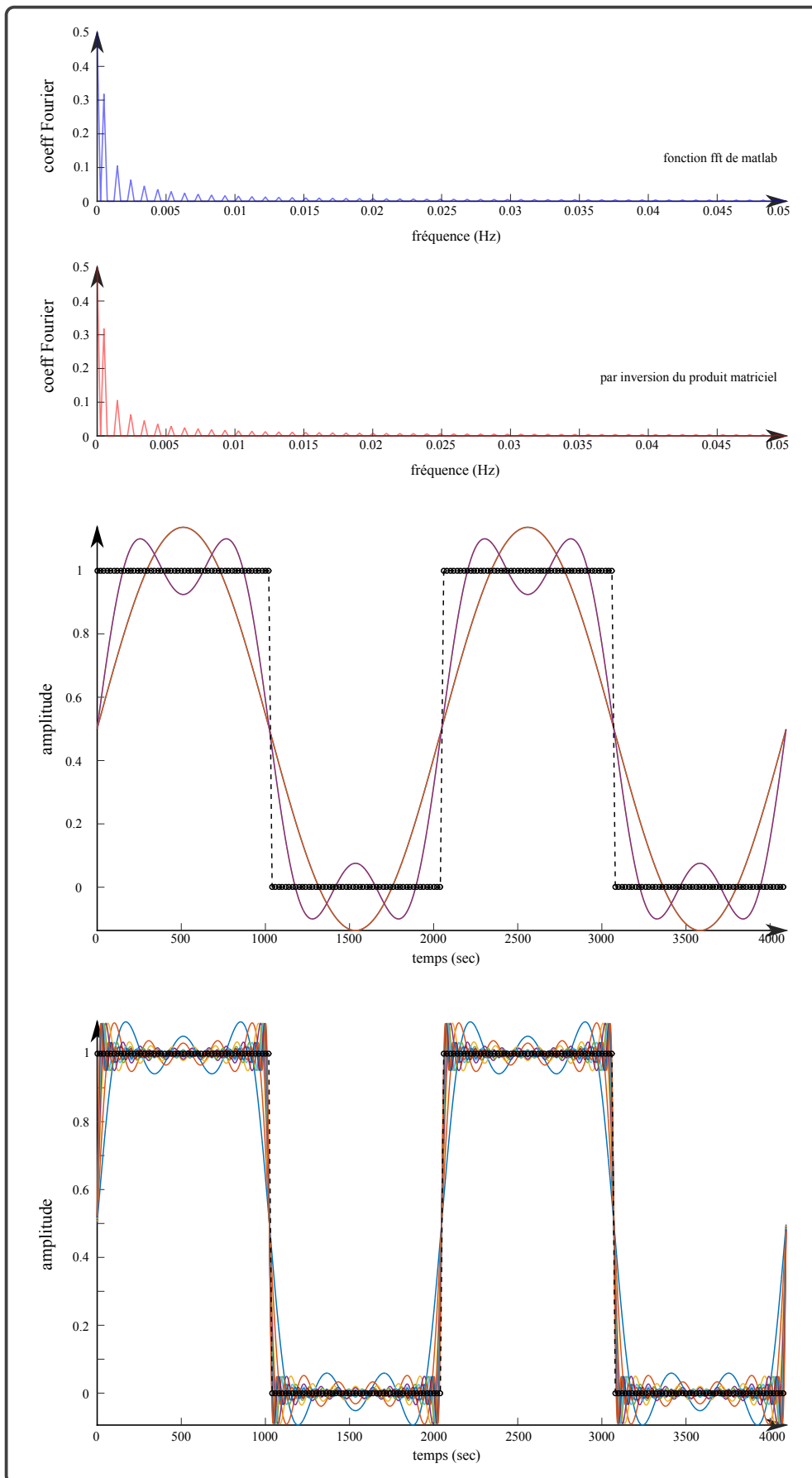


Figure 7.2: Direct and inverse Fourier transforms obtained by inverting the Vandermonde matrix.

---

---

# CHAPTER 8

---

## STOCHASTIC PROCESSES

1	Definition of Stochastic Processes . . . . .	104
2	1/f Noise . . . . .	105
3	White Noise . . . . .	107
4	Brownian Noise . . . . .	109
5	Pink Noise . . . . .	110
6	Black Noise . . . . .	111
7	Stable Laws (Gauss, Cauchy, <i>etc</i> ) . . . . .	111



---

Until now, all the calculations we have performed have been within a deterministic framework, and we have always implicitly assumed that the signals we dealt with were perfectly known. In practice, this view is insufficient, and it is necessary to account for the fact that the signals being processed contain a certain amount of noise. Generally, the reasons given for the presence of noise include measurement uncertainties, electronic noise, *etc*. When signals are noisy, they can no longer be treated deterministically; a probabilistic approach must be adopted, in which the signal under study is considered as a sample drawn from the set of all possible signals. This set is called a stochastic process, with the particular signal being a realization of that process. This chapter does not provide a detailed exposition on stochastic processes; it is merely a general overview meant to introduce a few key terms. For a thorough presentation on the subject, we refer the reader to [Athanasios Papoulis's](#) book, *Probability, Random Variables, and Stochastic Processes* (Papoulis, 1984), listed in the bibliography.

## 1 Definition of Stochastic Processes

Such a process is characterized by its moments, among which the most useful are the mean,

$$\mu_x(t) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N x_n(t) \quad (1.1)$$

and the autocorrelation,

$$r_{x,x}(t, t + \tau) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N x_n(t) x_n(t + \tau) \quad (1.2)$$

where the signals  $x_n(t)$  represent realizations of the process  $\{x\}(t)$ . From a practical standpoint, completely arbitrary stochastic processes are not very useful because one rarely has a large number of realizations to calculate the aforementioned statistical attributes. It is generally to circumvent this difficulty that particularly simple processes are introduced, which we will briefly discuss below. [Before discussing these, it is essential to keep in mind that these simplified processes have the immense advantage of being easily manipulable but also the significant drawback of often being too idealized to be realistic!](#) This certainly explains why many theoretically sophisticated signal processing methods are rarely applicable in practice, as the signals they are supposed to handle do not exist,

- when the statistical moments of a process do not depend on time, the process is said to be

stationary in the strict sense;

- when only the mean and the autocorrelation are time-independent, the process is said to be second-order stationary or weakly stationary.

It is worth noting right away that such processes are rare in practice, probably because they are information-poor: many geophysical signals owe their richness to their non-stationarity—the most illustrative example is certainly seismic signals. Consequently, it is primarily the noise itself that may be well described by stationary stochastic processes, rather than the signals as a whole. Among stationary processes are ergodic processes, where moments can be computed from a single sample by replacing ensemble sums with integrals over the values taken over time by a single realization,

$$\mu_x = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(t) dt \quad (1.3)$$

for the mean and,

$$r_{x,x}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(t)x(t+\tau) dt \quad (1.4)$$

for the autocorrelation. The energy spectrum of a stochastic process is computed using the [Wiener-Khinchine](#) theorem, which we have already discussed,

$$|X(u)|^2 = \mathcal{F}[r_{x,x}(\tau)](u). \quad (1.5)$$

## 2 1/f Noise

It is common for the energy spectra of geophysical signals to follow power law distributions,

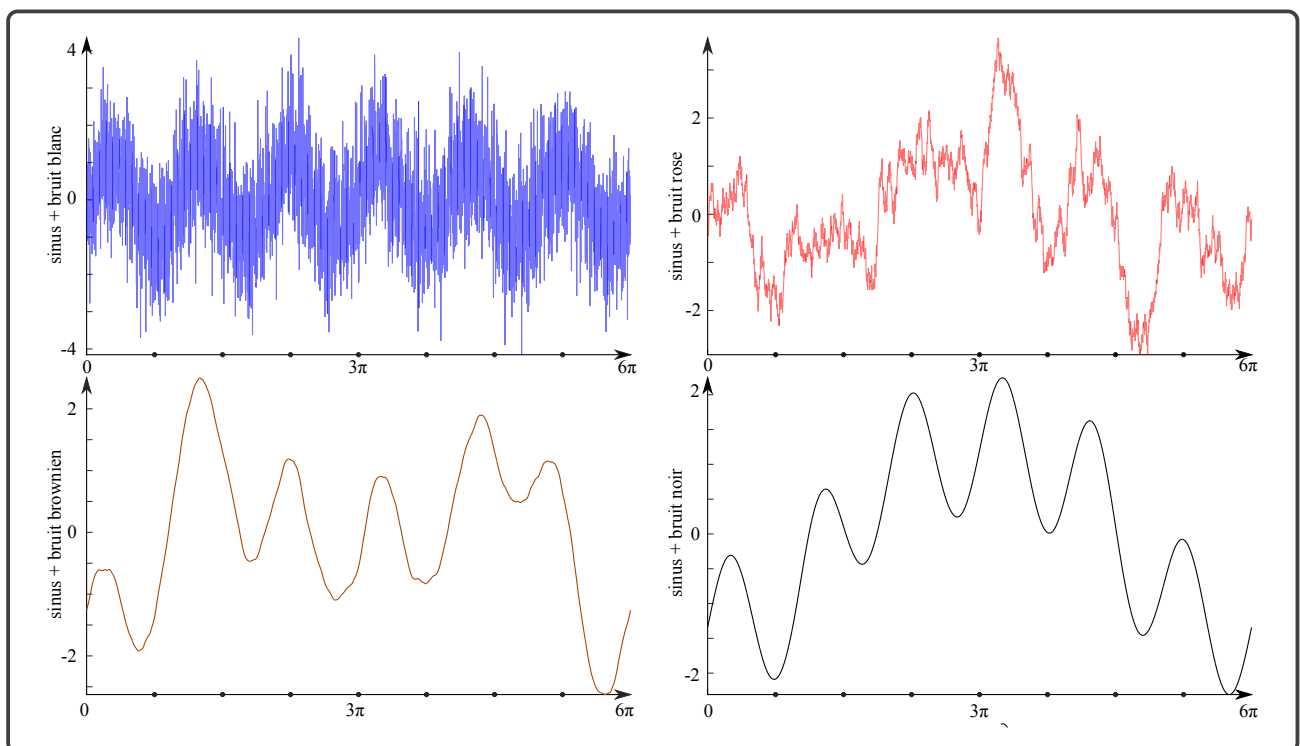
$$E(u) = E_0 u^{-\beta} \quad (2.1)$$

where, in general,  $\beta \in [0; 4]$ . Such signals are referred to as "1/f noises", and examples include the topography of young oceanic floors, geoid undulations, temporal variations of the Earth's magnetic field, ... Such noises are invariant under scale changes, meaning that whether one contracts or dilates the time scale, the energy spectrum retains its power-law form with the same exponent. Thus, 1/f noises appear similar at all scales; they are statistically self-similar. There are numerous articles on

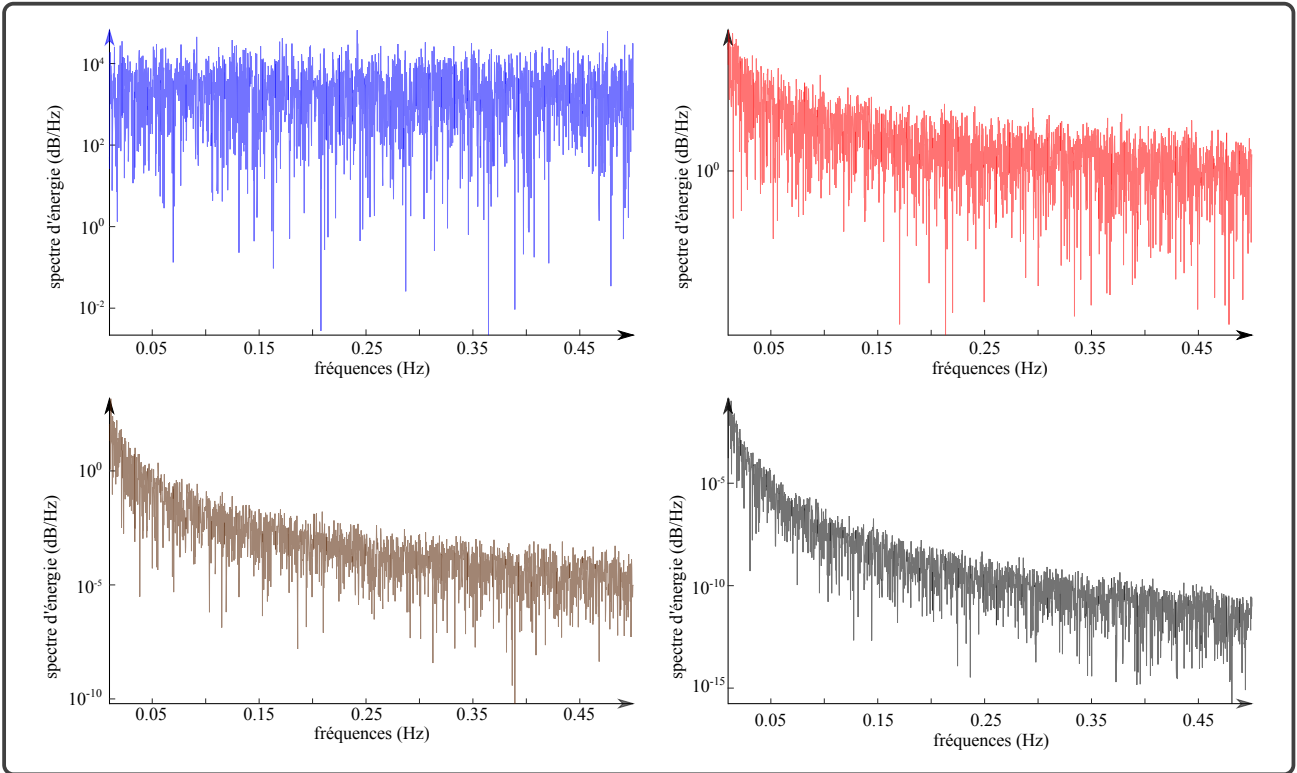
this subject, such as the one by [Jérémy Kasdin](#) from 1995, *Discrete Simulation of Colored Noise and Stochastic Processes and  $1/f^\alpha$  Power Law Noise Generation* (Kasdin, 1995). Here, we will only mention three specific types of such noises.

- **pink "noise"** has a constant energy per octave band, in contrast to white noise, whose spectrum is constant across all frequencies. For this type of noise, the coefficient  $\beta$  has a value of 1;
- **brownian "noise"**, also known as **brownian** motion in honor of the Scottish botanist [Robert Brown](#) (1773–1858), who first described in 1828 ([Brown, 1828](#)) the very irregular movements of large particles within pollen grains. For this type of noise, also referred to as red noise, the coefficient  $\beta$  is equal to 2;
- **black "noise"**, named by analogy to the thermal radiation of a black body, has a spectrum that decreases even more rapidly, and its coefficient  $\beta$  is equal to 3.

To illustrate all that we have discussed, the program [ex\\_bruit.m](#), which utilizes the sub-function [fct\\_bruit\\_couleur.m](#), adds the various types of noise we have mentioned to a sinusoidal function. Figures (8.1) and (8.2) demonstrate the nature and effect of these noises on a given signal.



**Figure 8.1:  $1/f$  Noise + Sine Wave.** These noises have identical phase spectra, which gives them correlated morphologies and allows us to observe that the black noise is a "smoother" version of the brownian noise, which in turn is a "smoother" version of the pink noise, etc



**Figure 8.2: 1/f Noise.** Representation of the energy spectra, in logarithmic scale, for white noise (blue curve), pink noise (red curve), brownian noise (brown curve), and black noise (black curve).

### 3 White Noise

White noise (blue curves in Figures 8.1 and 8.2),  $\{b\}(t)$ , is a stationary ergodic stochastic process in which successive values exhibit no correlation,

$$r_{b,b}(\tau) = a^2 \delta(\tau) \tag{3.1}$$

where  $a$  is a constant that sets the noise energy,

$$|B(u)|^2 = a^2 \tag{3.2}$$

which is uniformly distributed across the frequency axis (cf blue curve in figure 8.2), hence the term "white noise" by analogy with physical optics. With a correlation distance of zero, past values provide no information for predicting future values. A common example of nearly white noise is quantization noise (figures 8.3 and 8.4), which is generated during coding operations where

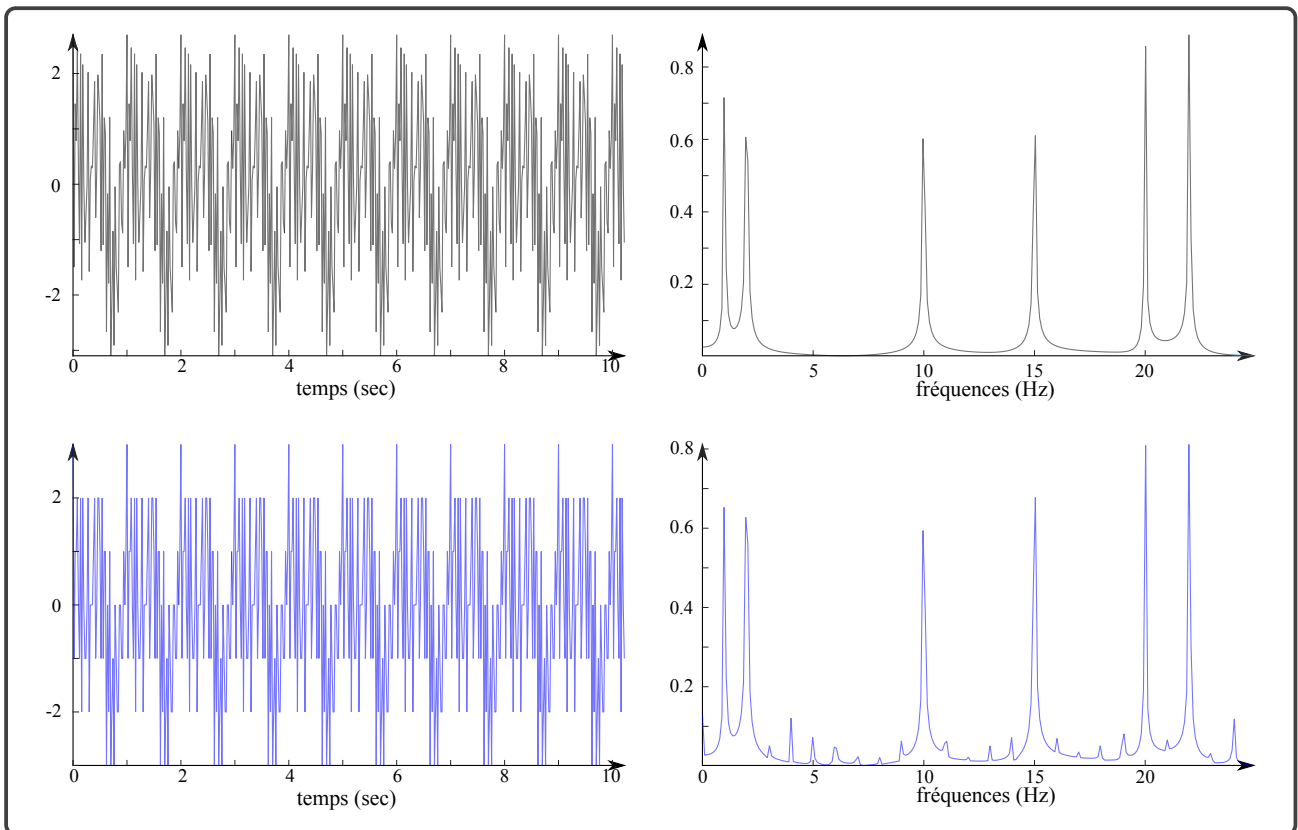
analog signals are transformed into digital signals through discretization. During such coding, analog values are assigned their digital counterparts, leading to a "rounding error," which constitutes the quantization noise  $\{q\}(t)$  whose probability density is approximately,

$$\mathcal{P}\{q\} = \begin{cases} 1/\delta q & \text{si } q \in [-\delta q/2; \delta q/2] \\ 0 & \text{si } q \notin [-\delta q/2; \delta q/2] \end{cases} \quad (3.3)$$

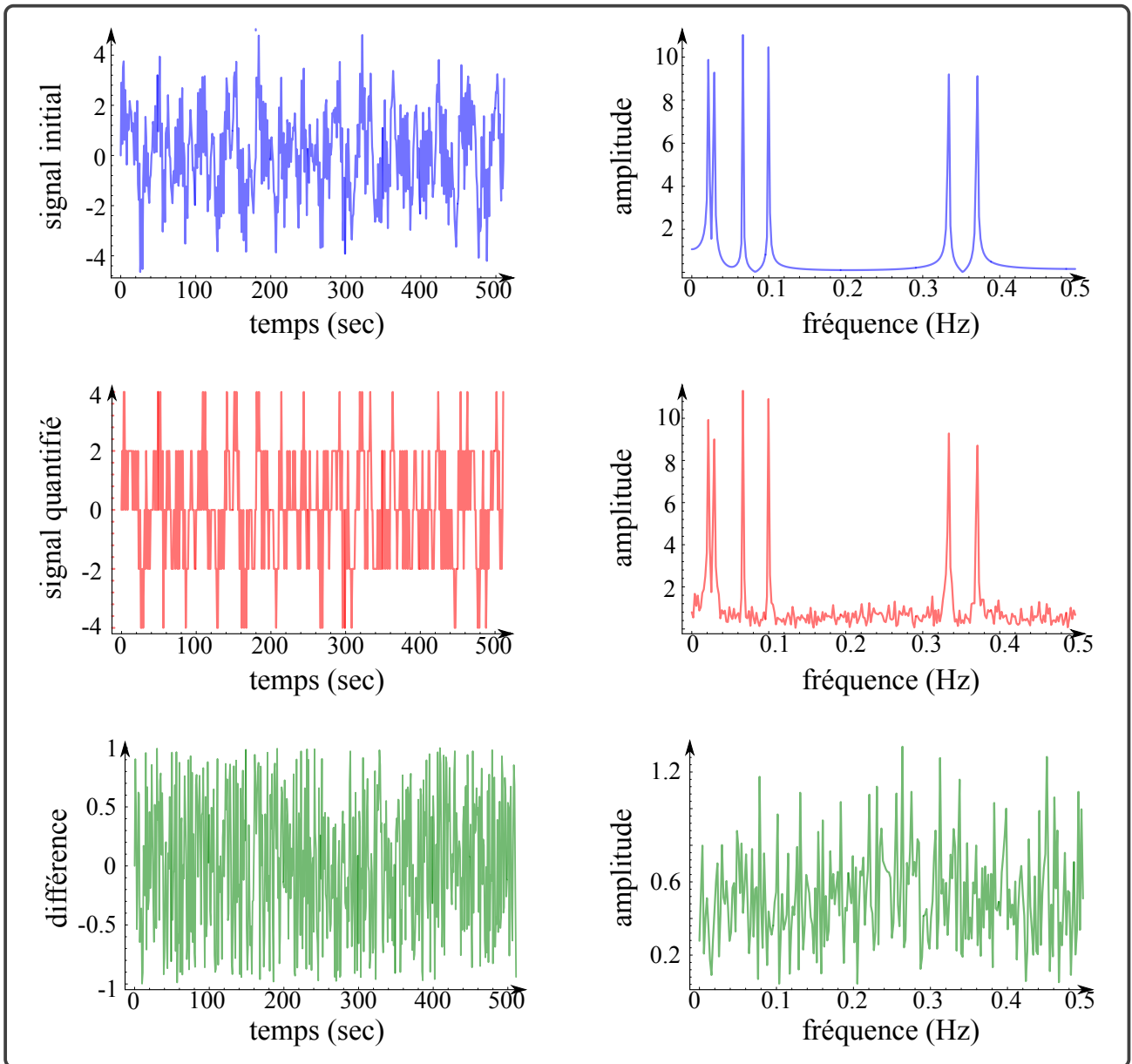
where  $\delta q$  is the quantization increment. The total energy (variance) of the noise,

$$\begin{aligned} r_{q,q}(0) &= \int_{-\infty}^{+\infty} \mathcal{P}\{q\} q^2 dq \\ &= \delta q^2 / 12 \end{aligned} \quad (3.4)$$

is uniformly distributed among the coefficients of the discrete **Fourier** transform of the sampled signal. Thus, if the series contains 256 values and  $\delta q = 0.10$ , the average level of the energy spectrum is approximately  $3 \times 10^{-6}$ .



**Figure 8.3: Quantization noise. The initial signal (top left). Its severe quantization (middle left) produces an amplitude spectrum (middle right) that is noisy compared to the initial spectrum (top right). Curves obtained with the program `ex_bruit_quantification.m`.**



**Figure 8.4: Quantization noise.** The initial signal (top left) is still our "favorite" previously shown in several figures. Its severe quantization (middle left) yields an amplitude spectrum (middle right) that is noisy compared to the initial spectrum (top right). The difference between the two signals (bottom left) is the quantization noise, whose spectrum (bottom right) is close to that of white noise.

## 4 Brownian Noise

These noises are related to the "white" noise paradigm mentioned earlier. Even when limited to a few octaves, practically realizable "white" noises are very useful for describing stochastic processes such as those involved in [brownian](#) motion. If, in such motion, the increments  $dx(t)$  are derived from

---

"white" noise, then the position,

$$x(t) = \int_0^t dx(\xi) d\xi \quad (4.1)$$

will be a **brownian** noise (see Figures 8.1 and 8.2, brown curves), such that,

$$|X(u)|^2 = X_0 u^{-2} \quad (4.2)$$

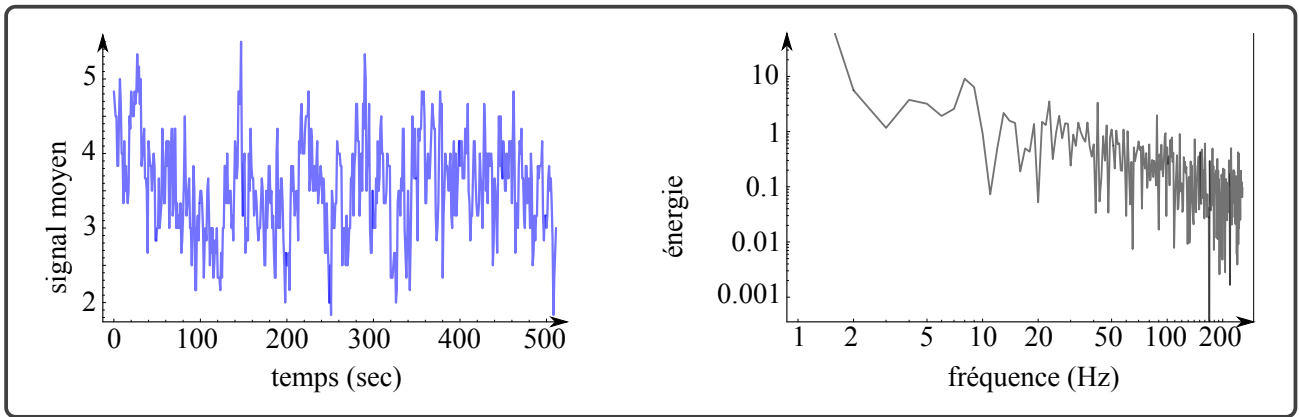
We will encounter **brownian** noises when the signal under study is the sum of random increments.

## 5 Pink Noise

These noises (Figures 8.1 and 8.2, red curves) are such that,

$$E(u) = E_0 u^{-1} \quad (5.1)$$

They are encountered in a wide range of situations, leading to the assertion that they play, with respect to  $1/f$  noises, a role similar to that of the normal distribution with respect to statistical distributions. These noises have been noted for their aesthetic properties, and some authors have pointed out that many musical sounds exhibit "pink" spectra. Electronic noises generated by semiconductors are also "pink". Although not the only method, "pink" noises are easily created by superimposing relaxation processes with sufficiently different time constants (*cf* Figure 8.5).



**Figure 8.5: Pink noise generated by the superposition of simple processes with different time constants. In this example, the processes are six dice, with the first being rolled at each time increment, the second only every other time, the third every fourth time, etc The average of the values displayed by the dice is calculated at each time increment, producing the noise on the left, whose amplitude spectrum (on the right) is reasonably pink. Thus, it is seen that the superposition of a small number of processes with different time constants easily produces pink noise. Such situations likely occur frequently in Nature.**

## 6 Black Noise

These noises (figures 8.1 and 8.2, brown curves) correspond to signals for which  $\beta > 2$ , and they are often associated with "catastrophic" geophysical processes such as the floods of the Nile, whose level variations have a spectrum where  $\beta = 2.8$ . "Black" noises have the particularity of possessing statistical persistence in accordance with the famous law of series; thus, the floods of the Nile occur in successive years as shown by [Harold Edwin Hurst](#) (Hurst, 1951). The Hurst exponent,

$$H \equiv \frac{\log(R/\sqrt{S})}{\log(T)} \quad (6.1)$$

where  $R$ ,  $S$ , and  $T$  are respectively the maximum range, the variance, and the observation duration of the signal, allows for the measurement of the persistence of a statistical phenomenon. Moreover,  $\beta = 2H + 1$ .

## 7 Stable Laws (Gauss, Cauchy, etc)

We have already reported that the probability density,  $h$ , of the sum of independent random variables is given by the convolution of the individual distributions  $f$  and  $g$ . Therefore, in general, we have

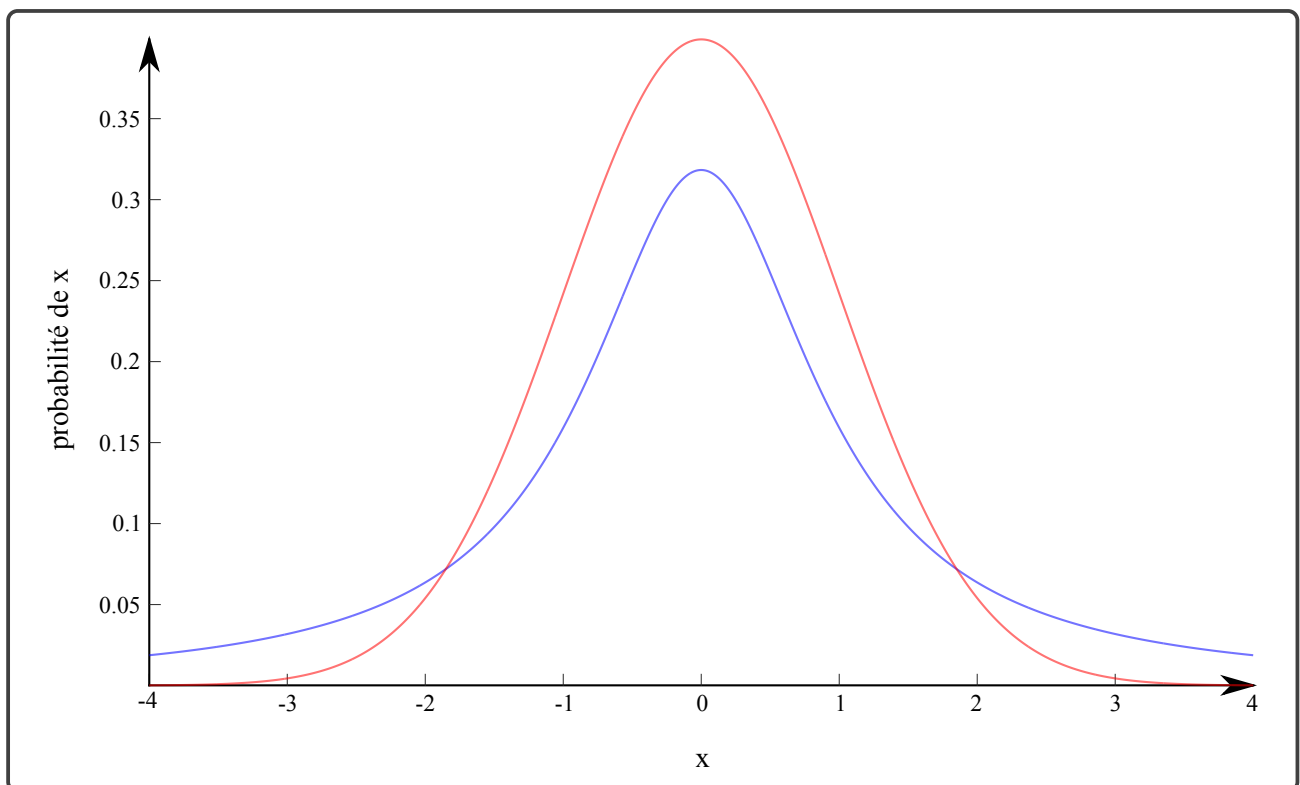


$$h(x+y) = f(x) * g(y) \tag{7.1}$$

where the forms of the functions are *a priori* arbitrary and different from one another. It is interesting to search for distributions that are invariant with respect to the above convolution; that is, distributions that yield the same distribution after convolution, up to a dilation and a translation. We seek functions,

$$f_{x+y}(x+y) = f_x(x) * f_y(y) \tag{7.2}$$

Such distributions have been termed 'stable laws' by the French mathematician [Paul Lévy](#) (1886–1971), and they serve as probabilistic attractors. The most well-known stable law is certainly [the normal, or Gaussian, distribution](#) (Figure 8.6),



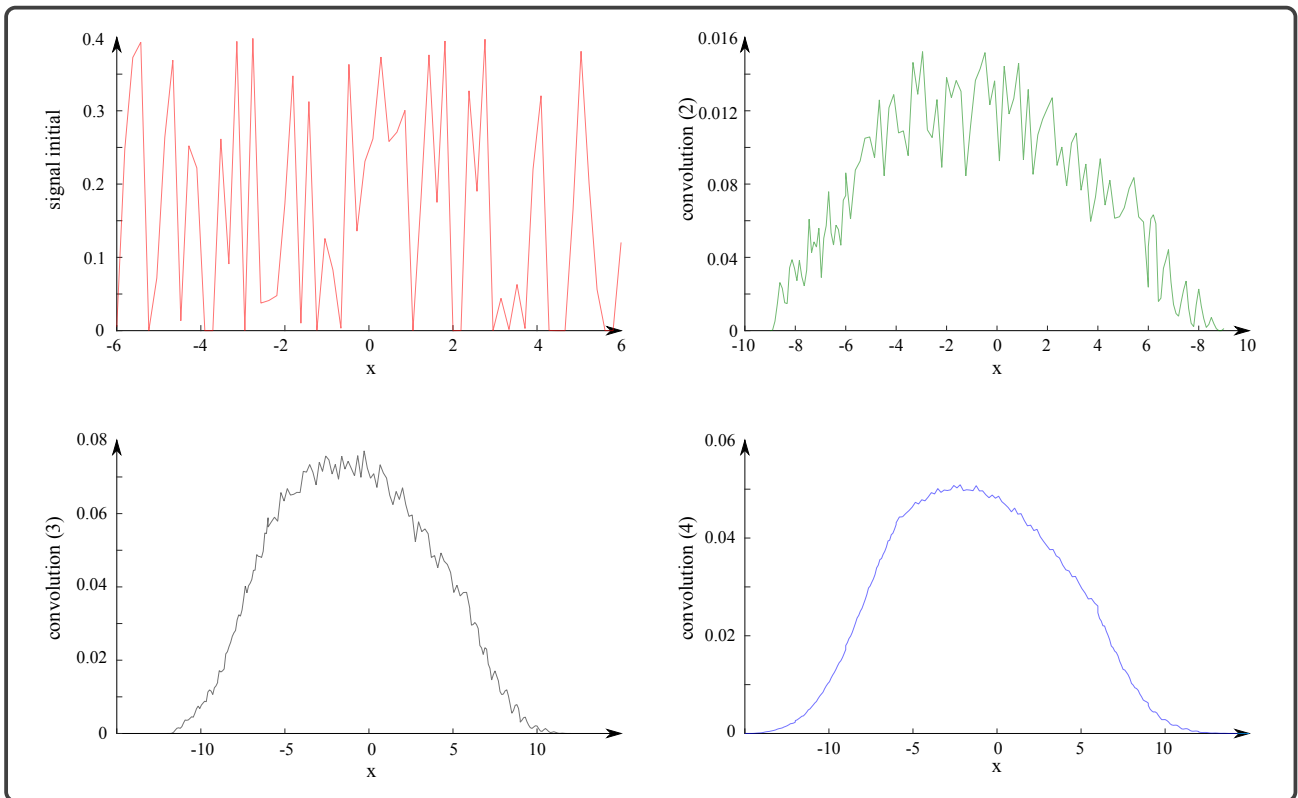
**Figure 8.6: Two stable laws. The distributions of Gauss (red) and Cauchy (blue). Although these two distributions do not appear very different at first glance, the resulting statistical consequences are dramatically so, as you will see later in Figures ?? and ?. These two curves were produced using the program `ex_lois_stables.m`, which calls the sub-functions `fct_cauchy.m` and `fct_normal.m`.**

$$N(x | \mu_x, \sigma_x^2) = \frac{1}{\sigma_x \sqrt{2\pi}} \exp \left[ -\frac{(x - \mu_x)^2}{2\sigma_x^2} \right] \quad (7.3)$$

where  $\mu_x$  and  $\sigma_x^2$  are the mean and the variance, respectively. We directly establish that,

$$\mathcal{P} \{z = x + y\} = N(z | \mu_z = \mu_x + \mu_y, \sigma_z^2 = \sigma_x^2 + \sigma_y^2) \quad (7.4)$$

where it is always assumed that  $x$  and  $y$  are independent random variables. The convolution of normal laws is thus a normal law whose variance is equal to the sum of the variances and whose mean is equal to the sum of the means. The normal law is well-known because it can be obtained as the limiting distribution of an infinite sum of independent variables whose distributions have finite variances; this is the consequence of the central limit theorem (Figure 8.7).



**Figure 8.7:** Illustration of the central limit theorem. The initial distribution (top left) has a finite variance and differs markedly from the Gaussian distribution with the same variance. Convolution of this initial distribution with itself (top right) yields a distribution that is already closer to the Gaussian. The triple autocorrelation (bottom left) and quadruple autocorrelation (bottom right) show that convergence is very rapid. Redo this figure by changing the initial distribution using the program `ex_theoreme_central_limite.m`; you will see that the convergence is striking in almost all cases.

It follows that the normal distribution is often used to describe the probabilistic behavior of physical measurements, with the reasoning being that these measurements incorporate a multitude of disturbances, whose sum is likely to conform to a normal statistic. While it is true that many autocorrelated distributions converge rapidly to the normal law, this should not be regarded as an absolute generalization, and there are cases where this is not the case.

The normal distribution is not the only stable law; in fact, there are infinitely many, including the [Cauchy](#) distribution (Figure 8.6, blue curve)

$$C(x | m_x, s_x) = \frac{s_x/\pi}{(x - m_x)^2 + s_x^2} \quad (7.5)$$

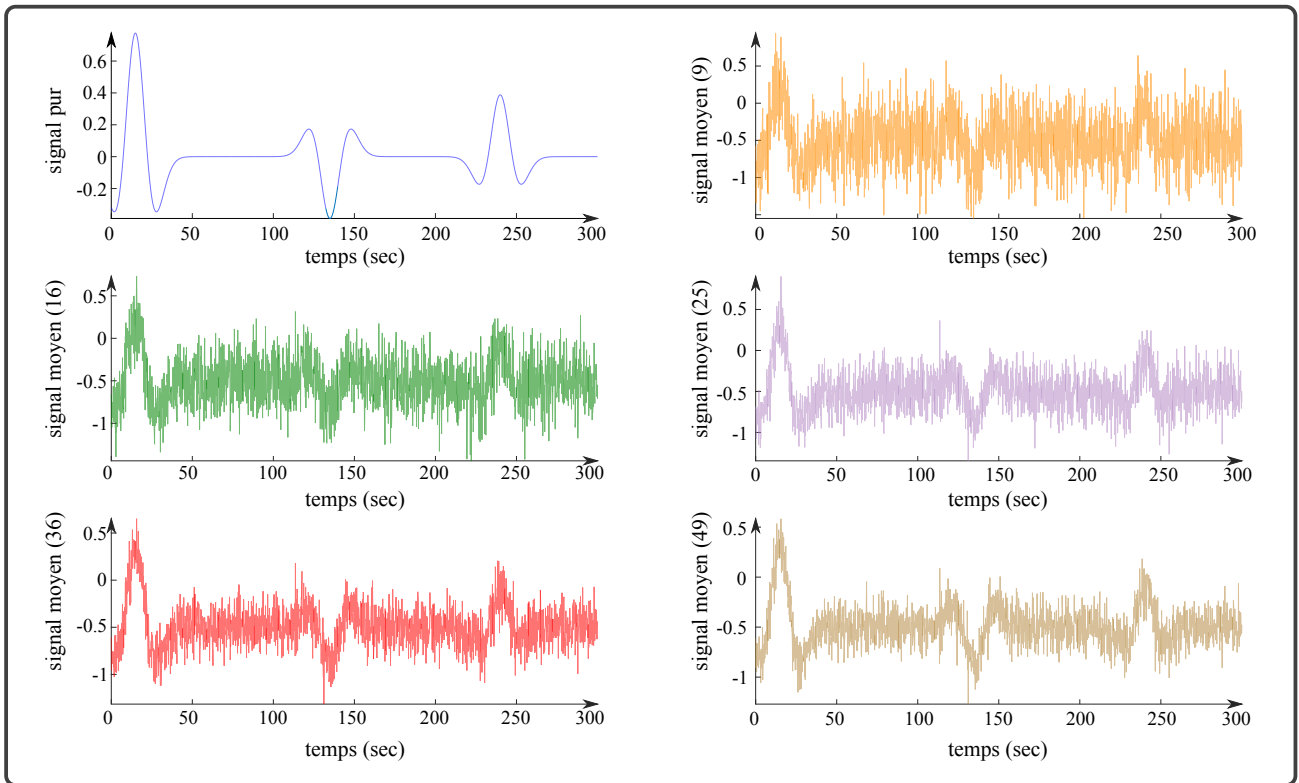
It is interesting to note that the mean and variance are not defined for this distribution. But that is not the worst part; in fact, it is easily shown that,

$$\mathcal{P} \left\{ z = \frac{x+x}{2} \right\} = C(z | m_z = m_x, s_z = s_x) \quad (7.6)$$

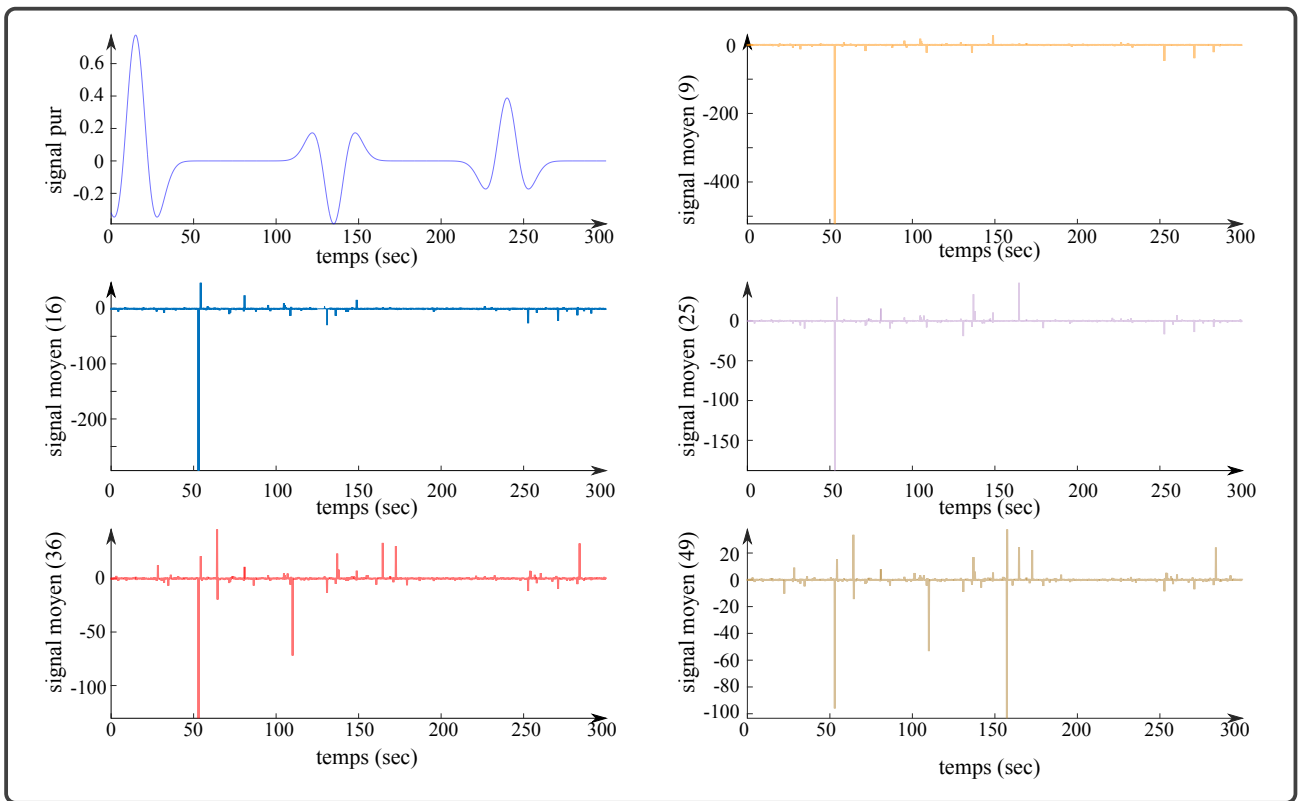
that is, the mean of two variables following a [Cauchy](#) distribution also follows the same [Cauchy](#) distribution, and therefore exhibits the same dispersion around the median. This result is a significant issue because it indicates that, with respect to the [Cauchy](#) statistic, "unity does not strengthen" (Figures 8.8 and 8.9 obtained using the program [ex\\_stack\\_cauchy\\_gauss.m](#)). This would not be problematic if the [Cauchy](#) distribution were not common; unfortunately, this is not the case. For example, the ratio of two independent variables with identical distributions follows the [Cauchy](#) statistic\*

---

\*Admittance calculators, coherence functions, and other transfer functions: beware!



**Figure 8.8: Gaussian noise. Addition of traces to increase the signal-to-noise ratio. In this example, the noise is white and Gaussian. The goal is to recover the initial signal (top left) by calculating the average of a certain number of noisy realizations. The average of 9 realizations (top right) allows for the identification of the first two events of the pure signal. Averages taken over more realizations (middle and bottom) improve the signal-to-noise ratio and allow for the recovery of the pure signal arrivals. An average taken over an infinite number of realizations converges stochastically to the pure signal.**



**Figure 8.9: Cauchy noise. Addition of traces to increase the signal-to-noise ratio. The disaster becomes apparent when summing the traces; the signal-to-noise ratio increases dramatically!**

---

---

# CHAPTER 9

---

## TIME-FREQUENCY DUALITY

1	Measuring signal duration . . . . .	119
2	The Uncertainty Principle in Signal Processing . . . . .	121
2.1	Deterministic Approach . . . . .	121
3	Causal signal duality . . . . .	123
4	Minimum Delay Signals . . . . .	124
4.1	Utility of Minimum Delay Signals . . . . .	124
5	Case of Continuous Signals . . . . .	125
6	Case of Discrete Signals . . . . .	125
7	The Cepstral Domain . . . . .	129

We will examine some correspondences that exist between a function and its **Fourier** transform. These correspondences provide a better understanding of the **Fourier** transformation and facilitate obtaining certain quick results about the function from its transform and *vice-versa*. For example, it is easily shown that,

$$\int_{-\infty}^{+\infty} f(t) dt = F(0) \quad (0.1)$$

higher-order moments can be obtained by applying the "reverse differentiation theorem",

$$\mathcal{F} [(-2i\pi t)^n f(t)](u) = F^{(n)}(u) \quad (0.2)$$

from which,

$$\int_{-\infty}^{+\infty} t^n f(t) dt = \frac{F^{(n)}(0)}{(-2i\pi)^n} \quad (0.3)$$

Two particularly interesting cases are the first and second-order moments, which, when normalized by the zero-order moment, allow the calculation of the barycentric abscissa of a function,

$$\langle t \rangle_f \equiv \frac{\int_{-\infty}^{+\infty} t f(t) dt}{\int_{-\infty}^{+\infty} f(t) dt} = -\frac{F^{(1)}(0)}{2i\pi F(0)} \quad (0.4)$$

and the quadratic mean abscissa, which can also be viewed as a reduced moment of inertia,

$$\langle t^2 \rangle_f \equiv \frac{\int_{-\infty}^{+\infty} t^2 f(t) dt}{\int_{-\infty}^{+\infty} f(t) dt} = -\frac{F^{(2)}(0)}{4\pi^2 F(0)} \quad (0.5)$$

The quadratic mean abscissa of a convolution product is easily calculated by noting that,

$$\begin{aligned} \int_{-\infty}^{+\infty} t^2 [f * g](t) dt &= -\left. \frac{\mathcal{F} [f * g]^{(2)}(u)}{4\pi^2} \right|_{u=0} \\ &= -\left. \frac{F^{(2)}G + 2F^{(1)}G^{(1)} + FG^{(2)}}{4\pi^2} \right|_{u=0} \end{aligned} \quad (0.6)$$

and,

$$\int_{-\infty}^{+\infty} [f * g](t) dt = F(0) G(0), \quad (0.7)$$

from which the desired result follows,

$$\begin{aligned} \langle t^2 \rangle_{f*g} &= -\frac{1}{4\pi^2} \left( \frac{F^{(2)}(0)}{F(0)} + \frac{G^{(2)}(0)}{G(0)} + 2 \frac{F^{(1)}(0)}{F(0)} \frac{G^{(1)}(0)}{G(0)} \right) \\ &= \langle t^2 \rangle_f + \langle t^2 \rangle_g + 2 \langle t \rangle_f \times \langle t \rangle_g. \end{aligned} \quad (0.8)$$

If one of the two functions has its barycentric abscissa at the origin, we recover the rule of additivity of variances. We invite the reader to focus on the section related to this analogy in the work by [Bracewell et Bracewell \(1986\)](#),

$$\langle t^2 \rangle_{f*g} = \langle t^2 \rangle_f + \langle t^2 \rangle_g. \quad (0.9)$$

## 1 Measuring signal duration

We have already considered the notion of the duration of a signal when introducing the [Dirac](#) impulse. This notion was unambiguous because we used the window function, whose bounded support defines the duration unequivocally. Signals with unbounded support have infinite duration; however, many of these signals have most of their energy concentrated in time, and it is then possible to associate a finite duration with them, which can be termed the effective duration. The challenge is to choose a method for calculating this duration; the simplest approach is to adopt the duration of a window that has the same ordinate at the origin and the same zero-order moment as the signal in question,

$$\begin{aligned} D_f &\equiv \frac{\int_{-\infty}^{+\infty} f(t) dt}{f(0)} \\ &= \frac{F(0)}{\int_{-\infty}^{+\infty} F(u) du} \\ &= \frac{1}{D_F}, \end{aligned} \quad (1.1)$$

from which a first duality relationship follows,

$$D_f \times D_F = 1 \quad (1.2)$$



to be compared with the similarity theorem (see paragraph 6.3), which states that a dilation of the time axis corresponds to a contraction of the frequency axis. However, this definition of duration is not satisfactory as it is not invariant under translation; such invariance can be achieved by using the autocorrelation function, which is known to have a maximum at the origin. One then defines,

$$\begin{aligned}
 D_{f \diamond f} &\equiv \frac{\int_{-\infty}^{+\infty} [f \diamond f](\tau) d\tau}{[f \diamond f](0)} \\
 &= \frac{\int_{-\infty}^{+\infty} f^*(\tau) d\tau \int_{-\infty}^{+\infty} f(\tau) d\tau}{\int_{-\infty}^{+\infty} f^*(\tau) f(\tau) d\tau} \\
 &= \frac{|F(0)|^2}{\int_{-\infty}^{+\infty} |F(u)|^2 du} \\
 &= \frac{1}{D_{|F|^2}}
 \end{aligned} \tag{1.3}$$

where the infinite limits of the integral in the numerator of the first line allow, through a change of variable, a rewriting in the form of a product of two integrals. Thus, a new **correlation-energy duality** relationship has been obtained,

$$D_{f \diamond f} \times D_{|F|^2} = 1 \tag{1.4}$$

The duration of the autocorrelation has the drawback of involving the signal indirectly; therefore, several authors have introduced an alternative definition based on the calculation of the 'moment of inertia' and the 'center of gravity' of the signal,

$$\left( D_{|f|^2}^2 \right) \equiv \frac{\int_{-\infty}^{+\infty} t^2 |f(t)|^2 dt}{\int_{-\infty}^{+\infty} |f(t)|^2 dt} - \left( \frac{\int_{-\infty}^{+\infty} t |f(t)|^2 dt}{\int_{-\infty}^{+\infty} |f(t)|^2 dt} \right)^2 \tag{1.5}$$

This definition is satisfactory in many cases and often yields results that align with intuition. We will see that adopting this definition leads to an uncertainty relation identical to that of **Heisenberg** in quantum physics.

## 2 The Uncertainty Principle in Signal Processing

### 2.1 Deterministic Approach

This principle, universal in Physics, also applies in signal processing whenever the [Fourier](#) transform is involved, thus allowing an analogy with wave phenomena. Before addressing the uncertainty relation as it is known in Quantum Physics, let us consider the case of a truncated sinusoid using a window. In this case, the [Fourier](#) transform of the sinusoid of infinite duration, composed of two [Dirac](#) impulses, is replaced by two sinc functions. The intuition\* suggests that the uncertainty in the frequency of the sinusoid is something like half the width of the central lobe of the sinc function,

$$\delta u \approx \frac{1}{T} \tag{2.1}$$

where  $T$  is the duration of the truncated signal. This definition is identical to that used by optical scientists, who define the resolution of an optical instrument as the radius of the first dark ring in the diffraction pattern; it corresponds to the duration  $D_f$  discussed in the previous section. Adopting the quadratic duration  $D_{|f|^2}$  allows us to obtain the [Heisenberg](#) uncertainty relation—which has already been addressed in this book, *cf* relation (6.18)—in honor of its originator.

[Werner Heisenberg](#) (1901-1976) was born in Würzburg (Germany) and studied theoretical physics in Munich ([Sommerfeld](#)) and Göttingen ([Born](#)). He published a dozen articles on atomic physics between 1922 and 1924, and during the same period (1923), he completed his doctoral thesis on turbulent flows. It was in 1924 that he met [Wolfgang Pauli](#) and [Niels Bohr](#), and the following year he laid the foundations for a new mechanics of atomic systems. The continuation of this work led to the famous uncertainty principle (1927) ([Heisenberg, 1927](#)), which earned him the Nobel Prize in 1932. After the war, [Heisenberg](#) reconstructed the [Max Planck](#) Institute of Physics; he then worked on a relativistic quantum field theory (with [Pauli](#)), a theory of nuclear structure motivated by the discovery of the neutron (1932), and a meson field theory (with [Yukawa](#) in 1935) which was confirmed by the discovery of the meson (1947) in cosmic rays. He briefly returned to the study of turbulent flows in 1948

Let us restrict ourselves to the case of a signal whose barycentric abscissa, as well as that of its

---

\*In this regard, reading the passage concerning this issue in the "[Feynman Lectures on Physics](#)" ([Feynman et al., 2011](#)) is instructive.

Fourier transform, are at the origin. We then have,

$$\begin{aligned}
\left(D_{|f|^2}^2 \times D_{|F|^2}^2\right)^2 &= \frac{\int_{-\infty}^{+\infty} t^2 |f(t)|^2 dt}{\int_{-\infty}^{+\infty} |f(t)|^2 dt} \times \frac{\int_{-\infty}^{+\infty} u^2 |F(u)|^2 du}{\int_{-\infty}^{+\infty} |F(u)|^2 du} \\
&= \frac{\int_{-\infty}^{+\infty} |tf(t)|^2 dt}{\int_{-\infty}^{+\infty} |f(t)|^2 dt} \times \frac{\int_{-\infty}^{+\infty} |uF(u)|^2 du}{\int_{-\infty}^{+\infty} |F(u)|^2 du} \\
&= \frac{\int_{-\infty}^{+\infty} |tf(t)|^2 dt \int_{-\infty}^{+\infty} |f^{(1)}(t)|^2 dt}{4\pi^2 \left(\int_{-\infty}^{+\infty} |f(t)|^2 dt\right)^2},
\end{aligned} \tag{2.2}$$

where we have used the Parseval theorem and then the differentiation relation. By employing the Schwarz inequality (cf relation 2.7),

$$\begin{aligned}
\left(D_{|f|^2}^2 \times D_{|F|^2}^2\right)^2 &\geq \frac{\left(\int_{-\infty}^{+\infty} [tf^*(t) f^{(1)}(t) + tf(t) f^{(1)*}(t)] dt\right)^2}{(4\pi)^2 \left(\int_{-\infty}^{+\infty} |f(t)|^2 dt\right)^2} \\
&= \frac{\left(\int_{-\infty}^{+\infty} t [f(t) f^*(t)]^{(1)} dt\right)^2}{(4\pi)^2 \left(\int_{-\infty}^{+\infty} |f(t)|^2 dt\right)^2} \\
&= \frac{\left(\int_{-\infty}^{+\infty} f(t) f^*(t) dt\right)^2}{(4\pi)^2 \left(\int_{-\infty}^{+\infty} |f(t)|^2 dt\right)^2} \\
&= \frac{1}{(4\pi)^2}
\end{aligned} \tag{2.3}$$

where the transition from the second to the third line involves integration by parts. A final evident simplification leads to the Heisenberg relation,

$$D_{|f|^2}^2 \times D_{|F|^2}^2 \geq \frac{1}{4\pi} \tag{2.4}$$

The equality is achieved (optimal time-frequency resolution) by the Gaussians that we will encounter in the chapter on wavelets. More generally, the uncertainty principle indicates that it is illusory to claim an infinitely good resolution simultaneously in time and frequency; the observation is unavoidable, the "Dirac monochromatic" does not exist!

The Schwarz inequality is demonstrated as follows. Let  $F(u)$  and  $G(u)$ , be two functions, and a

real constant  $\varepsilon$ . We have,

$$\int_{-\infty}^{+\infty} |F(u) + \varepsilon G(u)|^2 du > 0. \quad (2.5)$$

After expansion, this expression becomes,

$$\int_{-\infty}^{+\infty} |F(u)|^2 du + \varepsilon \int_{-\infty}^{+\infty} [F^*(u)G(u) + F(u)G^*(u)] du + \varepsilon^2 \int_{-\infty}^{+\infty} |G(u)|^2 du > 0, \quad (2.6)$$

which is a quadratic polynomial in  $\varepsilon$  that, to remain always positive, must have a non-positive discriminant, that is to say, such that,

$$\left[ \int_{-\infty}^{+\infty} [F^*(u)G(u) + F(u)G^*(u)] du \right]^2 \leq 4 \int_{-\infty}^{+\infty} |F(u)|^2 du \int_{-\infty}^{+\infty} |G(u)|^2 du \quad (2.7)$$

which is the sought inequality.

### 3 Causal signal duality

Causal signals, which are identically zero at negative times, are common in signal processing. They can be expressed in the form,

$$f_c(t) = H(t) f(t) \quad (3.1)$$

and thus we have,

$$F_c(u) = \frac{1}{2} \left[ \delta(u) - \frac{i}{\pi u} \right] * F(u) \quad (3.2)$$

$$= \frac{1}{2} \left[ F(u) - i \left( \frac{1}{\pi u} \right) * F(u) \right] \quad (3.3)$$

Using the definition of the [Hilbert](#) transform, relation (1.3), one obtains the [Bayard-Bode](#) relation,

$$F_c(u) = \frac{1}{2} [F(u) - i \mathcal{H}[F](u)], \quad (3.4)$$

which indicates that the [Fourier](#) transform of a causal signal has an imaginary part equal to the negative of the [Hilbert](#) transform of the real part. This property is used to rapidly compute the

---

numerical [Hilbert](#) transform of signals using the fast [Fourier](#) transform algorithm.

## 4 Minimum Delay Signals

### 4.1 Utility of Minimum Delay Signals

The objective of this section is to provide some clarifications regarding a class of signals frequently encountered in geophysics, particularly in seismic deconvolution, known as "minimum delay signals" or "minimum phase signals." This class of signals is somewhat shrouded in mystery, and the numerous conversations we have had with "specialists" on the subject lead us to believe that a straightforward presentation of these signals is not without value. These signals can be introduced in an extremely formal manner\*, but we prefer to adopt an approach more connected to physical principles. The basic principle justifying the use of minimum delay signals is to observe that, when excited by a source of energy, physical systems arrange to reemit this energy as quickly as the laws describing their behavior allow. This mode of operation relies on the principles of least action, which form the foundation of physics. It turns out that this principle of optimal energy restitution can serve as an additional constraint, proving very useful for regularizing certain signal processing problems. In practice, it is necessary to have a "measure" of a signal's duration that allows quantifying the "rapidity" of energy restitution. We have seen that several choices are possible, and we will adopt the quadratic measure†.

$$R(f) \equiv 4\pi^2 \int_0^{+\infty} t^2 |f(t)|^2 dt. \quad (4.1)$$

Now, the problem we wish to solve is as follows: let  $\{f_n(t)\}$  be a collection of causal signals, all of which have the same amplitude spectrum  $|F(u)|$ . The objective is to find, within this collection, the unique signal  $f_{\min}(t)$  such that,

$$R(f_{\min}) \text{ MINIMUM} \quad (4.2)$$

This signal is referred to as the minimum-delay signal associated with the collection. Given that the amplitude spectrum is fixed, the signals  $f_n(t)$  differ by their phase spectra  $\phi_n(u)$ . It follows that the minimum-energy-delay constraint, which operates in the time domain, should be accompanied by a condition on the phase in the frequency domain. Hence the term "minimum-phase signal".

---

\*For example, in the work by [E.R. Robinson](#), "[Seismic Deconvolution](#)".

†Le  $4\pi^2$  is included merely to simplify some of the expressions that will follow.

## 5 Case of Continuous Signals

Let us begin with the case of continuous signals whose **Fourier** transform

$$F_n(u) = |F(u)| \exp[i\phi_n(u)] \quad (5.1)$$

The energy-delay measurement then takes the form,

$$\begin{aligned} R(f_n) &= \int_{-\infty}^{+\infty} |F_n^{(1)}(u)|^2 du \\ &= \int_{-\infty}^{+\infty} \left| \left[ |F(u)|^{(1)} + i|F(u)|\phi_n^{(1)}(u) \right] \exp[i\phi_n(u)] \right|^2 du \\ &= \int_{-\infty}^{+\infty} \left[ |F(u)|^{(1)} \right]^2 du + \int_{-\infty}^{+\infty} |F(u)|^2 \left[ \phi_n^{(1)}(u) \right]^2 du \end{aligned} \quad (5.2)$$

Energy recovery is as fast as possible when the last term on the right-hand side is minimized,

$$\int_{-\infty}^{+\infty} \left[ \phi_{\min}^{(1)}(u) \right]^2 du \text{ MINIMUM} \quad (5.3)$$

The signal with minimal energy-delay must also be of minimal phase variation, that is, as least dispersed as possible.

## 6 Case of Discrete Signals

Consider now a discrete signal comprising  $L + 1$  values,

$$s = \{s_0, s_1, \dots, s_L\} \quad (6.1)$$

The factorization of the  $Z$ -transform of this signal shows that it can be generated by convolving  $L$  dipoles,

$$\{s_0, s_1, \dots, s_L\} = \{\alpha_1, \beta_1\} * \{\alpha_2, \beta_2\} * \dots * \{\alpha_L, \beta_L\} \quad (6.2)$$

The amplitude spectrum of this signal is equal to the product of the amplitude spectra of the

dipoles,

$$|\mathcal{F}\{\mathbf{s}_0, s_1, \dots, s_L\}| = \prod_{l=1}^L |\alpha_l + \beta_l Z| \quad (6.3)$$

Noting that,

$$|\alpha_l + \beta_l Z| = |\beta_l + \alpha_l Z| \quad (6.4)$$

we observe that the  $2^L$  signals generated by convolving the  $L$  dipoles, whether inverted or not, have  $|\mathcal{F}\{\mathbf{s}_0, s_1, \dots, s_L\}|$  as their amplitude spectrum. Each dipole offers the alternative,

$$\{\alpha_l, \beta_l\} \text{ ou } \{\beta_l, \alpha_l\} \quad (6.5)$$

depending on whether it is inverted or not. Among these two dipoles, the one with the largest absolute value for the first term is the minimum-delay dipole. For example, among,

$$\{1, -2\} \text{ et } \{-2, 1\} \quad (6.6)$$

it is the dipole  $\{-2, 1\}$  that is of minimum delay. We then observe that among the  $2^L$  signals that can be created from the initial dipoles, there is one that corresponds to the particular case where all the dipoles are of minimum delay. This signal, which we will denote by  $\{\mathbf{s}_0, s_1, \dots, s_L\}_{\min}$ , is called the minimum-delay signal associated with  $\{\mathbf{s}_0, s_1, \dots, s_L\}$ . It has the distinguishing feature of possessing an amplitude spectrum identical to that of the initial signal.

Now let us examine the phase spectrum of the minimum-delay signal. Knowing that,

$$\arg(\mathcal{F}\{\mathbf{s}_0, s_1, \dots, s_L\}) = \sum_{l=1}^L \arg(\alpha_l + \beta_l Z) \quad (6.7)$$

we are led to examine the phase spectra of the dipoles generating the signal. The phase spectrum of the minimum-delay dipole,  $\{\alpha_l, \beta_l\}$ , is given by,

$$\begin{aligned} \phi_{\min, l}(u) &\equiv \arg(\alpha_l + \beta_l Z) \\ &= -\arctan \left[ \frac{\beta_l \sin(2\pi u \tau)}{\alpha_l + \beta_l \cos(2\pi u \tau)} \right] \end{aligned} \quad (6.8)$$

and,

$$\begin{aligned}\phi_{\min,l}^{(1)}(u) &\equiv \frac{d}{du} \arg(\alpha_l + \beta_l Z) \\ &= -\frac{2\pi\tau [\beta_l^2 + \alpha_l \beta_l \cos(2\pi u\tau)]}{\alpha_l^2 + \beta_l^2 + 2\alpha_l \beta_l \cos(2\pi u\tau)}\end{aligned}\quad (6.9)$$

An identical calculation applied to the inverted dipole\*,  $\{\beta_l, \alpha_l\}$ , yields,

$$\begin{aligned}\phi_{\max,l}^{(1)}(u) &\equiv \frac{d}{du} \arg(\beta_l + \alpha_l Z) \\ &= -\frac{2\pi\tau [\alpha_l^2 + \alpha_l \beta_l \cos(2\pi u\tau)]}{\alpha_l^2 + \beta_l^2 + 2\alpha_l \beta_l \cos(2\pi u\tau)} \\ &= \phi_{\min,l}^{(1)}(u) - \frac{2\pi\tau (\alpha_l^2 - \beta_l^2)}{\alpha_l^2 + \beta_l^2 + 2\alpha_l \beta_l \cos(2\pi u\tau)}\end{aligned}\quad (6.10)$$

Note that,

$$|\alpha_l| > |\beta_l| \implies \alpha_l^2 + \alpha_l \beta_l \cos(2\pi u\tau) > 0 \quad (6.11)$$

and,

$$\alpha_l^2 + \beta_l^2 + 2\alpha_l \beta_l \cos(2\pi u\tau) > 0 \quad (6.12)$$

These inequalities reveal that,

$$\phi_{\max,l}^{(1)}(u) < 0 \quad (6.13)$$

that is, the phase  $\phi_{\max,l}^{(1)}(u)$  of the maximum-delay dipole is a monotonically decreasing function. This is not the case for the phase of the minimum-delay dipole, which can be either increasing or decreasing. The triangular inequality

$$|a + b| \leq |a| + |b| \quad (6.14)$$

---

\*Also known as the maximum-delay dipole.



applied to the relation,

$$\phi_{\min,l}^{(1)}(u) = \phi_{\max,l}^{(1)}(u) + \frac{2\pi\tau(\alpha_l^2 - \beta_l^2)}{\alpha_l^2 + \beta_l^2 + 2\alpha_l\beta_l \cos(2\pi u\tau)} \quad (6.15)$$

yields,

$$\left| \phi_{\min,l}^{(1)}(u) \right| \leq \left| \phi_{\max,l}^{(1)}(u) \right| + \frac{2\pi\tau(\alpha_l^2 - \beta_l^2)}{\alpha_l^2 + \beta_l^2 + 2\alpha_l\beta_l \cos(2\pi u\tau)} \quad (6.16)$$

Since,

$$\frac{2\pi\tau(\alpha_l^2 - \beta_l^2)}{\alpha_l^2 + \beta_l^2 + 2\alpha_l\beta_l \cos(2\pi u\tau)} > 0 \quad (6.17)$$

$$\left| \phi_{\min,l}^{(1)}(u) \right| < \left| \phi_{\max,l}^{(1)}(u) \right| \quad (6.18)$$

Returning to the case of the complete signal, we have observed that the phase,

$$\arg(\mathcal{F}\{\mathbf{s}_0, s_1, \dots, s_L\}) = \sum_{l=1}^L \phi_l(u) \quad (6.19)$$

which immediately gives,

$$\frac{d}{du} \arg(\mathcal{F}\{\mathbf{s}_0, s_1, \dots, s_L\}) = \sum_{l=1}^L \phi_l^{(1)}(u) \quad (6.20)$$

L'inégalité triangulaire permet d'obtenir que,

$$\left| \frac{d}{du} \arg(\mathcal{F}\{\mathbf{s}_0, s_1, \dots, s_L\}) \right| \leq \sum_{l=1}^L \left| \phi_l^{(1)}(u) \right|, \quad (6.21)$$

and, using the results obtained for the minimum-delay dipole, it follows directly that,

$$\left| \frac{d}{du} \arg(\mathcal{F}\{\mathbf{s}_0, s_1, \dots, s_L\}_{\min}) \right| < \left| \frac{d}{du} \arg(\mathcal{F}\{\mathbf{s}_0, s_1, \dots, s_L\}) \right| \quad (6.22)$$

This inequality indicates that the minimum-delay signal associated with a collection of signals

generated by  $L$  dipoles is the signal with the slowest phase variation, that is, the signal with the least possible dispersion.

## 7 The Cepstral Domain

The title of this subsection is not the result of typographical dyslexia, but indeed introduces one of the most intriguing aspects of the time-frequency duality. The "cepstral" domain (Oppenheim et Schafer, 2004) is the realm of homomorphic deconvolution, where "quefreny", "liftering", "sispha", and "alanysis" reign. It involves transforming a time-domain signal into another domain, analogous to time, using the properties of real and complex logarithms. The real cepstrum utilizes only the amplitude of the signal's spectrum, and by neglecting its phase, it becomes impossible to reconstruct all the initial information. With the complex logarithm (Oppenheim, 1965), it becomes possible to accurately reconstruct both the phase and amplitude of the original signal. The "cepstral transform"  $\mathcal{C}(\tau)$  of a time-domain signal  $y(t)$  is given by the following relation,

$$\mathcal{C}(\tau) = \mathcal{F}(\ln(|\mathcal{F}(y(t))|)) \quad (7.1)$$

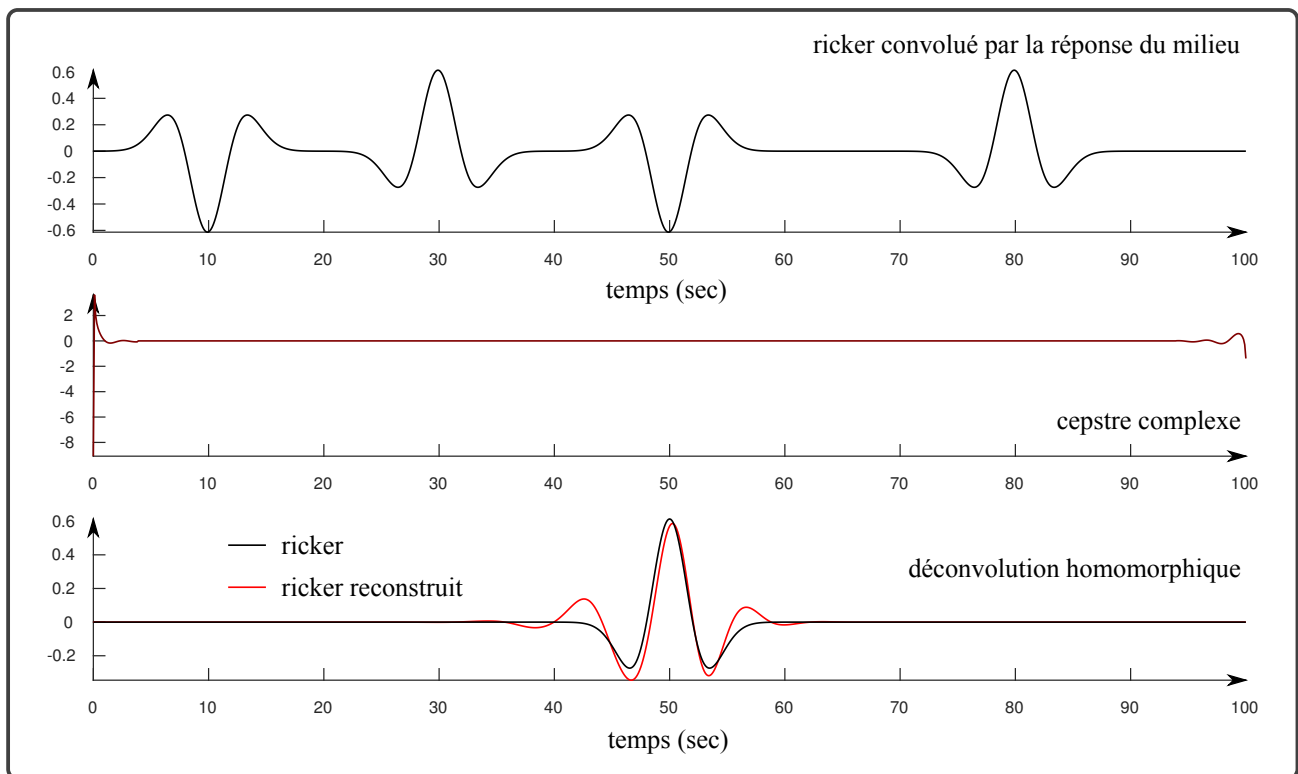
This definition (7.1), from an algorithmic perspective, can be expressed in the form,

$$\begin{aligned} \hat{y}(u) &= \mathcal{F}(y(t)) \\ \mathcal{L}y(u) &= \ln(r_{\hat{y}}) + \sqrt{-1} * \phi_{\hat{y}} \\ \mathcal{C}(\tau) &= \mathcal{R}(\mathcal{F}^{-1}(\mathcal{L}y)) \end{aligned} \quad (7.2)$$

$\mathcal{F}$  and  $\mathcal{F}^{-1}$  are the direct and inverse Fourier transforms, respectively, whose magnitude and phase are represented by  $r$  and  $\phi$ .

The idea here is to use the fact that a recorded signal, such as a seismic wave, results from the convolution of a source (*eg* a Ricker wavelet) with the impulse response of the medium (*eg* a distribution of reflectors). Since cepstral analysis allows us to transition from the data space to a space where the cepstra of the two convolved elements are simply superimposed and added, if their supports are "sufficiently distant", it will be possible to separate and reconstruct either the propagated source or the medium's response by simply canceling out a part of the total cepstrum. One way to approach this is to consider the following situation: the wave propagates through a medium rich in reflectors. The more reflectors there are, the less the cepstral supports will be overlapping. The

program `ex_deconv_homo.m` illustrates this concept."



**Figure 9.1: Homomorphic Deconvolution.** The first figure illustrates the signal to be analyzed. It simply results from the convolution of a Ricker wavelet, shown in black in the bottom figure, with a sequence of positive (+1) or negative (-1) reflectors. The cepstrum, obtained using the relations (7.2), is shown in the center in brown. By canceling out the part of this curve corresponding to the Green's function of the medium, it is possible to reconstruct the source through an inverse cepstral transform. The result of the deconvolution (red Ricker in the bottom figure) is superimposed on the original source (black Ricker).

---

---

# CHAPTER 10

---

## LINEAR FILTERING

1	<b>Filters and Z-Transforms</b> . . . . .	132
2	<b>Operator and Filters in Numerical Analysis</b> . . . . .	135
3	<b>Narrowband Filters</b> . . . . .	136
3.1	Recursiveness and Infinite Impulse Response . . . . .	136
4	<b>Filter Stability</b> . . . . .	141
4.1	Back to narrow band filter . . . . .	141
4.2	The general case . . . . .	142
5	<b>Butterworth Filters</b> . . . . .	144
6	<b>General Overview</b> . . . . .	144
7	<b>The Bilinear Transformation</b> . . . . .	145
8	<b>An example</b> . . . . .	147
9	<b>Wiener Filters</b> . . . . .	150
10	<b>Wiener Filtering in the Frequency Domain</b> . . . . .	150

This chapter deals exclusively with linear filtering applied to signals through a convolution operation (Kanasewich, 1981b). We have already encountered this type of filtering when studying linear systems, where the output signal is a filtered version of the input signal. A linear filter is fully characterized by its transfer function, which is the Fourier transform of its impulse response. The magnitude of the transfer function, known as the gain, indicates which frequencies will be attenuated, preserved, or amplified. Traditionally, examining the gain allows filters to be classified as low-pass, high-pass, band-pass, or all-pass; however, geophysics also employs numerous filters that do not fit these categories, such as potential field extension operators, pole reduction filters, etc The ideal low-pass filter is a rectangular function,

$$\Pi\left(\frac{u}{2u_b}\right) \quad (0.1)$$

with an impulse response that is a sinc function (cf figure (4.1))

$$2u_b \text{sinc}(2u_b t) \quad (0.2)$$

High-pass or band-pass filters can be constructed in a similar manner, and they all share the drawback of having an oscillatory impulse response with a decay that is slower the more abrupt the cutoff of their gain. These ideal filters are impractical and their discretization makes them perform poorly in practice. In particular, the "rectangular" filters exhibit oscillations near the edges known as the Gibbs phenomenon. Most of the time, the filters used have a real impulse response, and often it is required that they be additionally non-phase-shifting, which is not possible if the filter is causal

## 1 Filters and Z-Transforms

We will see how the Z-transform allows us to study and practically design digital filters applicable to sampled signals. Consider the discrete convolution,

$$s_n \equiv s(n\tau) = \sum_{k=-\infty}^{+\infty} f_k e_{n-k} \quad (1.1)$$

where  $f_k$  is a discrete filter whose characteristics we wish to determine. After applying the Z-transform, the convolution becomes

$$S(Z) = F(Z)E(Z) \quad (1.2)$$

For example, filtering,

$$s_n = \frac{1}{5} \sum_{k=-\infty}^{+\infty} e_{n-k} \quad (1.3)$$

corresponding to a moving average over 5 values has the filter\*

$$f = \left\{ \frac{1}{5}; \frac{1}{5}; \frac{1}{5}; \frac{1}{5}; \frac{1}{5} \right\} \quad (1.4)$$

whose Z-transform is,

$$F(Z) = \frac{1}{5} (Z^{-2} + Z^{-1} + 1 + Z + Z^2) \quad (1.5)$$

It is then possible to calculate the gain of the filter, given by,

$$|F(Z)| = \sqrt{F(Z)F^*(1/Z)} \quad (1.6)$$

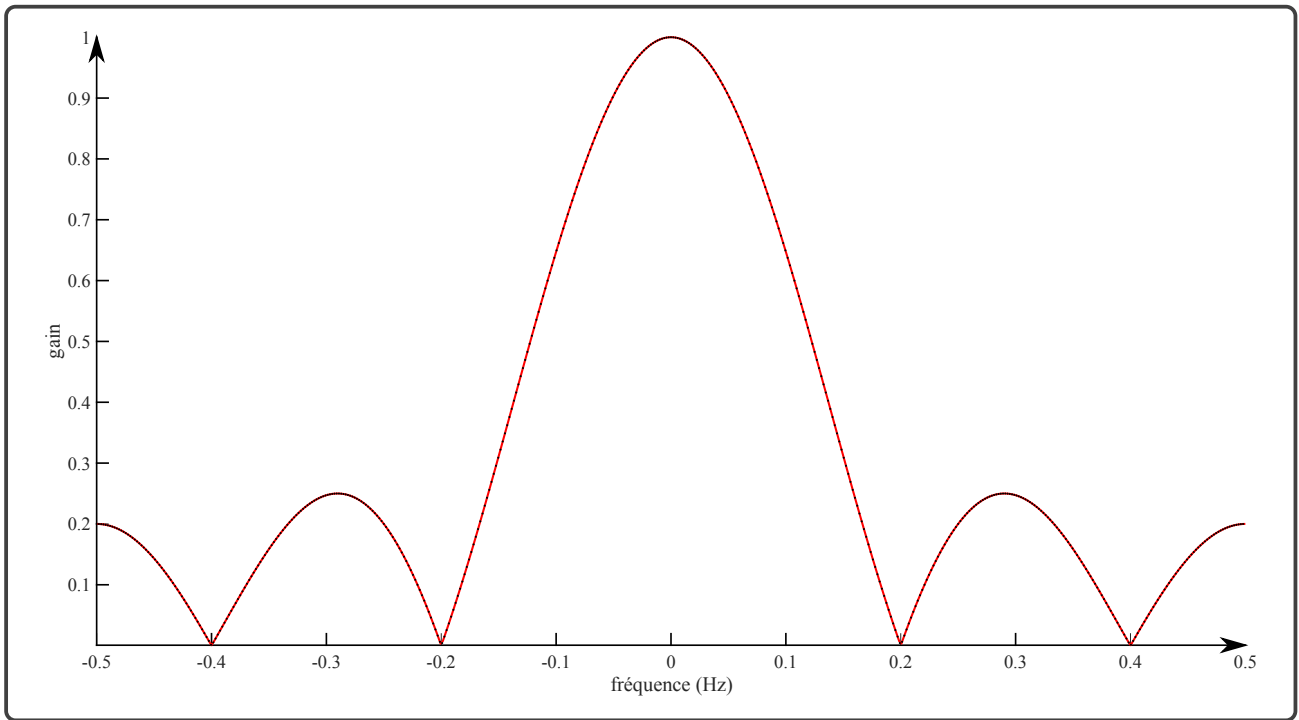
which, after expansion on the unit circle of [Fourier](#), *ie*, in the particular case where  $Z=e^{2i\pi u\tau}$ ,

$$|F(u)| = \sqrt{1 + \frac{8}{5} \cos(2\pi u\tau) + \frac{6}{5} \cos(4\pi u\tau) + \frac{4}{5} \cos(6\pi u\tau) + \frac{2}{5} \cos(8\pi u\tau)} \quad (1.7)$$

The expression (1.7) allows the calculation of the gain for any frequency within the [Shannon](#) interval  $[-1/2\tau; +1/2\tau]$  and a comparison with the gain of the ideal continuous filter consisting of a rectangular pulse of 5 seconds duration when  $\tau = 1$ .

---

\*In the following, we will denote the value of discrete filters at time zero in boldface.



**Figure 10.1:** Gain of the discrete filter  $(1,1,1,1,1)/5$ . Comparison between the gain given by expression (1.7), in red, and that of a perfect filter (black dashed lines) obtained by a 5-second duration rectangular pulse. ([ex\\_gain\\_transformeeZ.m](#))

One can take any filter,

$$f = \{+864; -144; +186; -55; -79; +4; +4\} \quad (1.8)$$

calculate its Z-transform,

$$F(Z) = 864 - 144Z + 186Z^2 - 55Z^3 - 79Z^4 + 4Z^5 + 4Z^6 \quad (1.9)$$

and factorize it,

$$F(Z) = (4 + Z)(4 - Z)(-3 + 2iZ)(-3 - 2iZ)(2 - Z)(3 + Z) \quad (1.10)$$

in order to decompose the initial filter as a cascade of dipoles. This operation allows for easier study of the filter characteristics based on those of the dipoles; thus, the stability analysis of the overall filter can be performed. If one of the dipoles is numerically unstable, the entire filter will also be unstable. The filters we have just discussed consist of a sequence, more or less long, of numerical values that are convolved with the signal to be processed. This is why they are called finite impulse

response filters\* as opposed to infinite impulse response filters† which we will now encounter.

## 2 Operator and Filters in Numerical Analysis

Finite difference operators are widely used filters in numerical analysis for solving partial differential equations. There is a whole range of filters that approximate the ideal operator to varying degrees or possess particular qualities (causal, anti-causal, *etc*). Two widely used second derivative operators are,

$$\{1; -2; 1\} \tag{2.1}$$

and,

$$\left\{ -\frac{1}{12}; \frac{15}{12}; -\frac{28}{12}; \frac{15}{12}; -\frac{1}{12} \right\}. \tag{2.2}$$

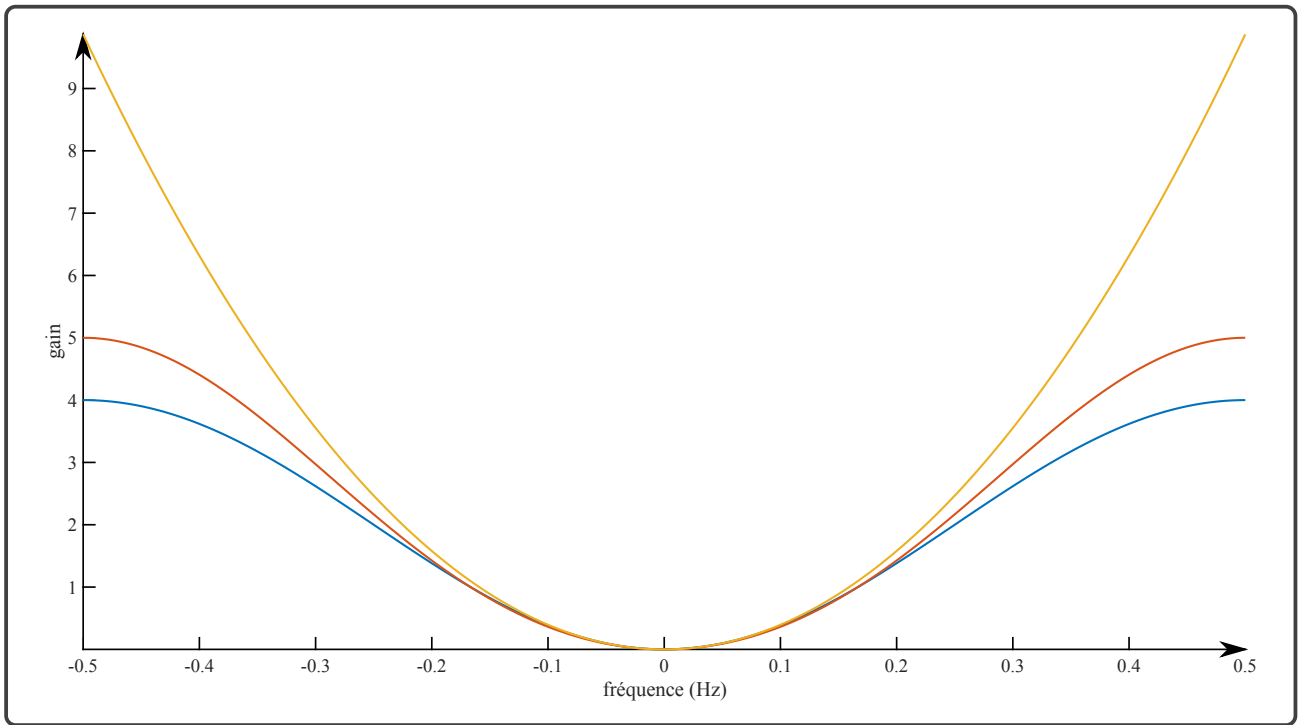
The Z-transform allows for the calculation of the gain and phase of these filters (*cf* figure 10.2) and to deduce their characteristics. In particular, it is easy to define the frequency domain of validity for the discrete operator. The program [ex\\_gain\\_tz.m](#) produces the following results,

---

\*"FIR" in Anglo-Saxon terminology.

†"IIR" in Anglo-Saxon terminology.





**Figure 10.2: Gains of finite difference operators.** This example shows the gains (solid lines) of the two second derivative operators,  $\{1; -2; 1\}$  and  $\{-\frac{1}{12}; \frac{15}{12}; -\frac{28}{12}; \frac{15}{12}; -\frac{1}{12}\}$ . The yellow curve represents the gain of the ideal operator,  $(2\pi u)^2$ , which is best approximated by the 5-term filter (red curve). Both filters are non-phase-shifting as they are centered on the origin. The calculations assume a unit sampling step, and you will notice that proper use of the filters is only possible if limited to low frequencies, approximately Nyquist/5 for the less efficient operator, which requires sampling the signals with a finer step than that dictated by the **Shannon rule**.

### 3 Narrowband Filters

#### 3.1 Recursiveness and Infinite Impulse Response

The band-pass filter with the narrowest bandwidth is the one that only retains a particular frequency,

$$F(u) = \delta(u + u_0) + \delta(u - u_0) \tag{3.1}$$

$$f(t) = \exp(-2i\pi u_0 t) + \exp(+2i\pi u_0 t) \tag{3.2}$$

### 3. NARROWBAND FILTERS

---

A causal discretization of this filter provides,

$$\begin{aligned}
 F(Z) &= 1 + ZZ_0 + (ZZ_0)^2 + (ZZ_0)^3 + \dots \\
 &+ 1 + Z/Z_0 + (Z/Z_0)^2 + (Z/Z_0)^3 + \dots \\
 &= 1/(1 - ZZ_0) + 1/(1 - Z/Z_0)
 \end{aligned} \tag{3.3}$$

with  $Z_0 = \exp(-2i\pi u_0 \tau)$ . Note that writing the  $Z$ -transform of the filter as a ratio of polynomials allows for the manipulation of an infinite impulse response. The above expression shows that calculating the filter's gain will pose a numerical problem at  $u = \pm u_0$ . This is because the values for which the denominator of  $F(Z)$  is zero, known as the poles, lie on the unit circle. When  $Z$  traverses this circle, the poles are encountered, leading to numerical issues. The desired filter is not realizable in its current form and must be modified to eliminate these numerical difficulties. The solution is to place the poles just off the unit circle so that the gain is no longer infinite. The trade-off is that the filter will no longer be as perfect as initially desired (*cf* figure 10.3). Thus, let us set

$$Z'_0 = (1 - \varepsilon)Z_0 \tag{3.4}$$

and,

$$Z''_0 = (1 - \varepsilon)/Z_0 \tag{3.5}$$

with  $\varepsilon > 0$ , we then obtain the modified filter,

$$\begin{aligned}
 F'(Z) &= \frac{1}{1 - ZZ'_0} + \frac{1}{1 - ZZ''_0} \\
 &= \frac{\alpha_0 + \alpha_1 Z}{1 + \beta_1 Z + \beta_2 Z^2}
 \end{aligned} \tag{3.6}$$

with,

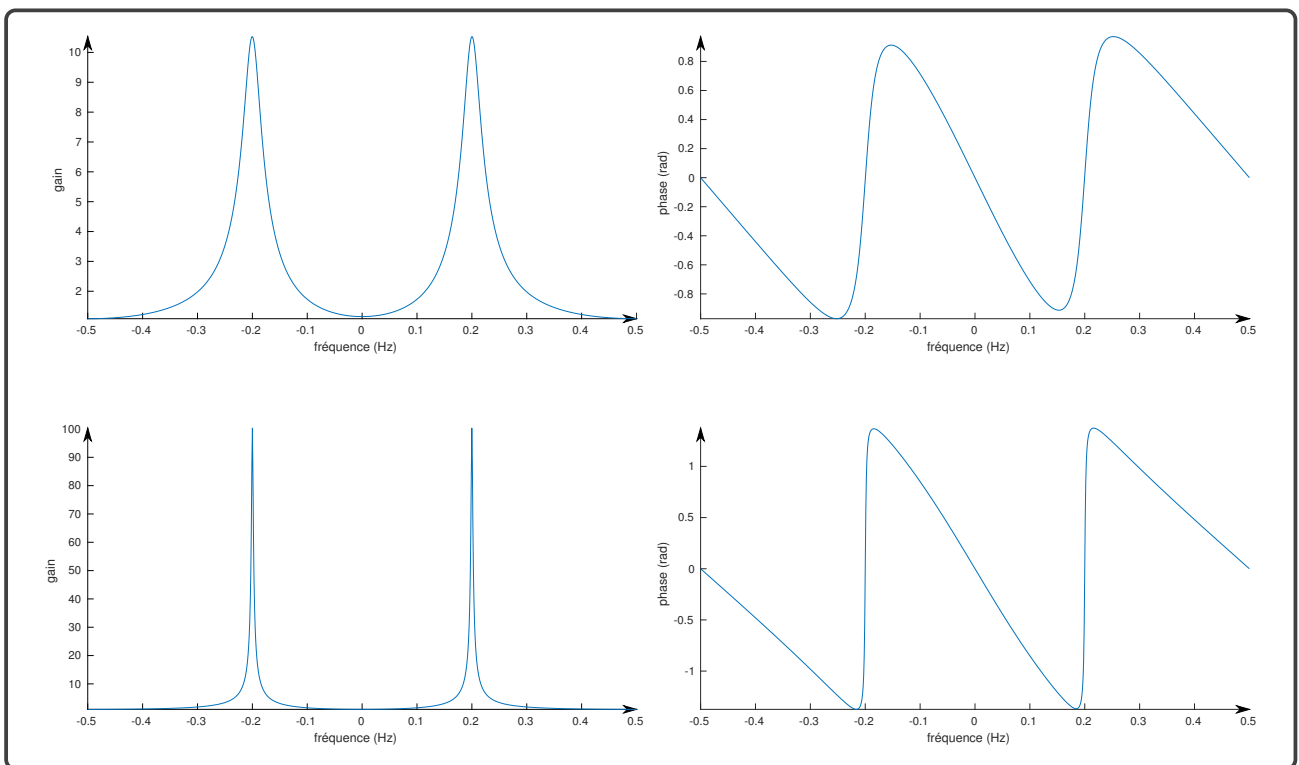
$$\begin{cases}
 \alpha_0 &= 2 \\
 \alpha_1 &= -2(1 - \varepsilon) \cos(2\pi u_0 \tau) \\
 \beta_1 &= -2(1 - \varepsilon) \cos(2\pi u_0 \tau) \\
 \beta_2 &= (1 - \varepsilon)^2
 \end{cases} \tag{3.7}$$

The filtering operation can be expressed as a product of  $Z$ -transforms,

$$S(Z) = E(Z)F'(Z) \quad (3.8)$$

that is, using the filter's expression,

$$S(Z) = E(Z)(\alpha_0 + \alpha_1 Z) - ZS(Z)(\beta_1 + \beta_2 Z) \quad (3.9)$$



**Figure 10.3: Gain and phase of the narrowband filter. The upper filter was constructed with  $\epsilon = 0.1$ , and the lower one with  $\epsilon = 0.01$ . ([ex\\_bande\\_etroite.m](#))**

Récrivons cette expression en développant chaque terme,

$$\begin{array}{cccccc}
 S(Z) & \alpha_0 E(Z) & \alpha_1 Z E(Z) & \beta_1 Z S(Z) & \beta_2 Z^2 S(Z) & \\
 \Downarrow & \Downarrow & \Downarrow & \Downarrow & \Downarrow & \\
 \left[ \begin{array}{c} s_0 \\ + \\ s_1 Z \\ + \\ s_2 Z^2 \\ + \\ \vdots \end{array} \right] & = & \left[ \begin{array}{c} \alpha_0 e_0 \\ + \\ \alpha_0 e_1 Z \\ + \\ \alpha_0 e_2 Z^2 \\ + \\ \vdots \end{array} \right] & + & \left[ \begin{array}{c} \alpha_1 e_0 Z \\ + \\ \alpha_1 e_1 Z^2 \\ + \\ \vdots \end{array} \right] & - & \left[ \begin{array}{c} \beta_1 s_0 Z \\ + \\ \beta_1 s_1 Z^2 \\ + \\ \vdots \end{array} \right] & - & \left[ \begin{array}{c} \beta_2 s_0 Z^2 \\ + \\ \vdots \end{array} \right] & . & (3.10)
 \end{array}$$

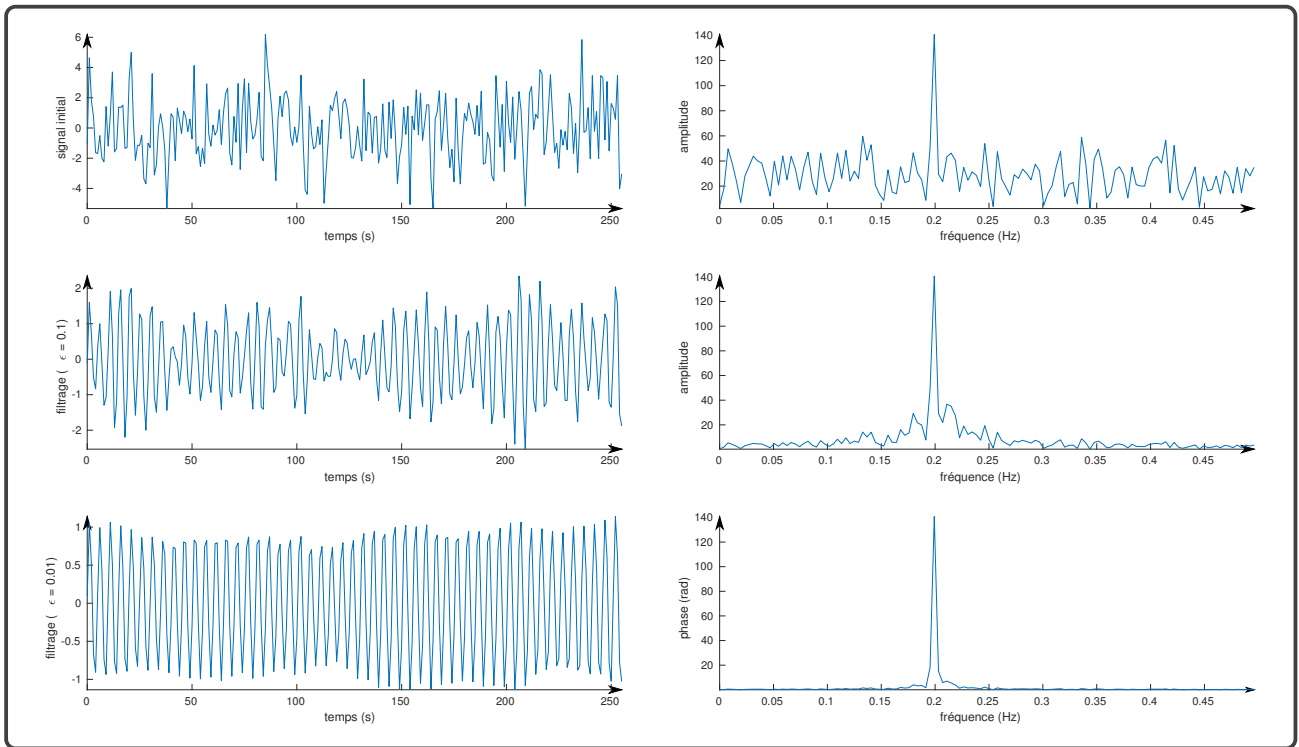
Since the equality must hold for all  $Z$ , it is necessary that it holds individually for each power of  $Z$ , that is, for each term in the expression above. This leads to the recursive expressions,

$$\left\{ \begin{array}{l} s_0 = \alpha_0 e_0 \\ s_1 = \alpha_0 e_1 + \alpha_1 e_0 - \beta_1 s_0 \\ s_n = \alpha_0 e_n + \alpha_1 e_{n-1} - \beta_1 s_{n-1} - \beta_2 s_{n-2} \end{array} \right. \quad (3.11)$$

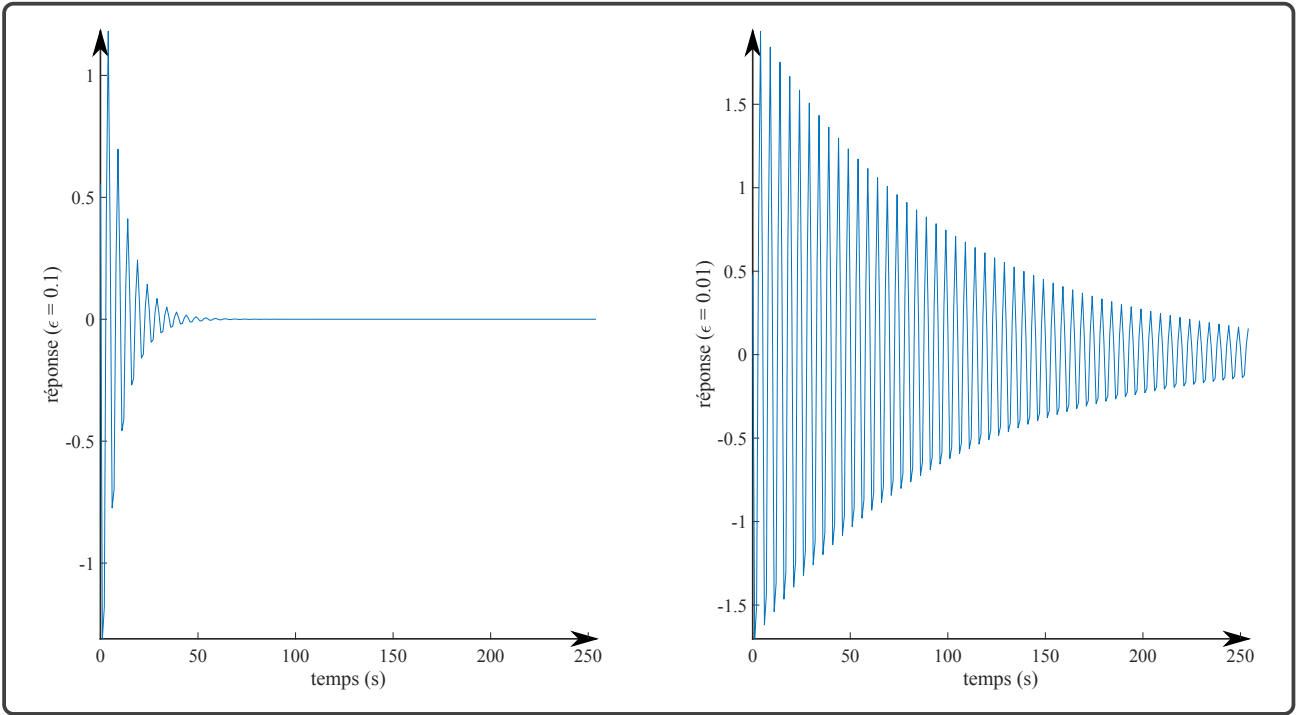
**An example** The filter can be applied using these recursive formulas, one advantage of which is speed: the above operation requires only 7 operations (additions and multiplications), whereas filtering by convolution using the filter in its non-recursive form,

$$\begin{aligned}
 F'(Z) &= 1 + ZZ'_0 + (ZZ'_0)^2 + (ZZ'_0)^3 + \dots \\
 &+ 1 + ZZ''_0 + (ZZ''_0)^2 + (ZZ''_0)^3 + \dots
 \end{aligned} \quad (3.12)$$

requires significantly more operations. For example, if  $\varepsilon = 0.05$ , truncating the filter to a coefficient equal to 10% of  $f_0$  requires extending to  $n = 45$ , which costs 90 operations! The narrowband filter we constructed allows for isolating a spectral component within a strong noise. This is illustrated in Figure 10.4, where, in particular, the boundary effects associated with the recursive formula are noticeable. These boundary effects become more pronounced as the filter's impulse response is longer (*cf* Figure 10.5).



**Figure 10.4:** Applications of the narrowband filter. The initial signal (top left) consists of a pure frequency (0.20 Hz) and Gaussian white noise with unit variance. The amplitude spectrum (top right) clearly shows the spectral component and the noise level. The first filtering attempt (middle left) removes a significant portion of the noise (middle right). The filter used is the one at the top of Figure 10.3. The second attempt (bottom left) was performed with the filter closer to the ideal shown at the bottom of Figure 10.3. The spectral analysis of the filtered signal (bottom right) shows that the noise has indeed been further eliminated. However, the filtered signal (bottom left) shows a very disturbing boundary effect. The more or less significant nature of these boundary effects can be understood by examining the impulse response of the filters (*cf* Figure 10.5).



**Figure 10.5:** Impulse responses of the narrowband filters shown in Figure 10.3. The filter constructed with  $\varepsilon = 0.1$  has a much shorter impulse response (on the left) compared to the one (on the right) constructed with  $\varepsilon = 0.01$ . These differences in duration explain the more or less significant boundary effects that occur when implementing recursive filters.

## 4 Filter Stability

### 4.1 Back to narrow band filter

In the previous section, we modified the ideal filter by setting,

$$F'(Z) = \frac{1}{1 - ZZ_0'} + \frac{1}{1 - ZZ_0''}. \quad (4.1)$$

By expanding the terms,

$$\frac{1}{1 - ZZ_0'} = 1 + (1 - \varepsilon)ZZ_0 + (1 - \varepsilon)^2 (ZZ_0)^2 + \dots \quad (4.2)$$

and,

$$\frac{1}{1 - ZZ_0''} = 1 + (1 - \varepsilon)Z/Z_0 + (1 - \varepsilon)^2 (Z/Z_0)^2 + \dots, \quad (4.3)$$

We observe that convergence is achieved only if  $\varepsilon > 0$ , meaning that the poles of the filter must

lie outside the unit circle. Otherwise, the series do not converge and the filter is unstable.

## 4.2 The general case

The general problem of filter stability can lead to rather lengthy algebraic developments. However, a sufficient condition to ensure the stability of recursive filters,

$$F(Z) = \frac{N(Z)}{D(Z)} \quad (4.4)$$

can be easily obtained by expressing the denominator in the form,

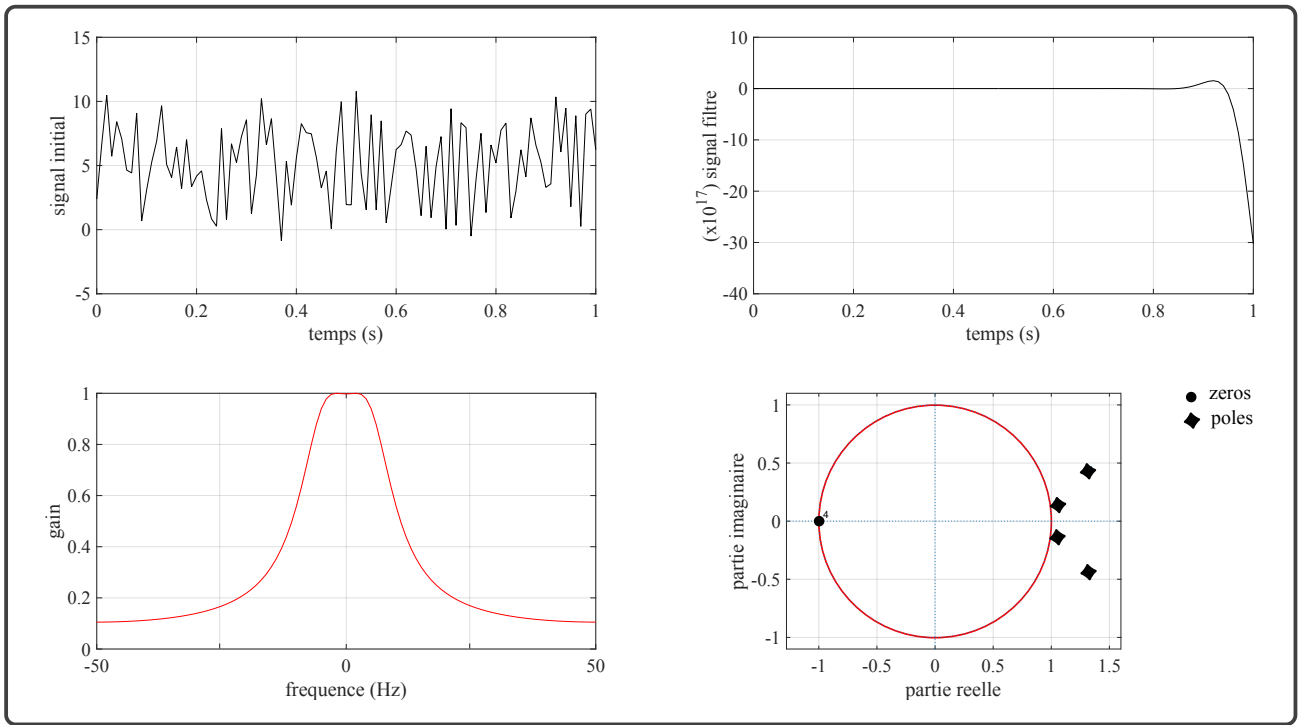
$$D(Z) = D_0 \prod_{l=1}^L (1 - ZZ_l) \quad (4.5)$$

which yields,

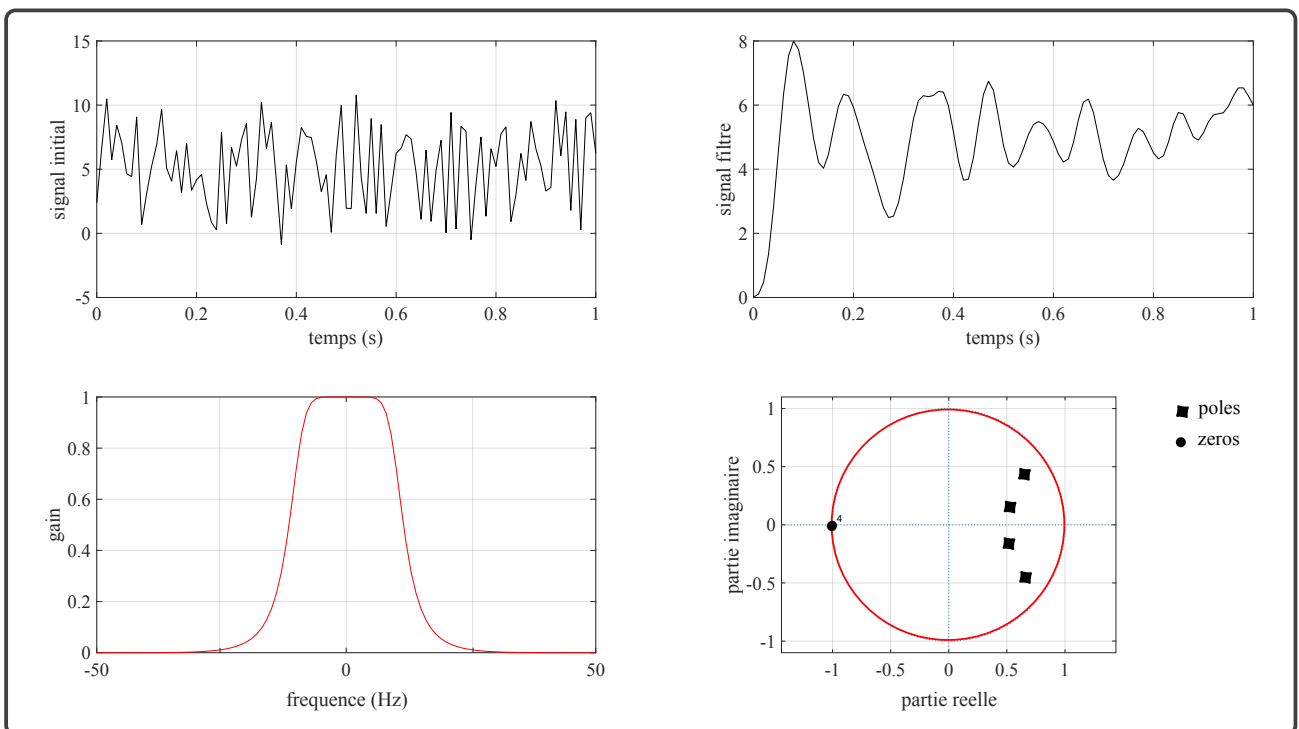
$$\begin{aligned} F(Z) &= \frac{N(Z)}{D_0} \prod_{l=1}^L (1 - ZZ_l)^{-1} \\ &= \frac{N(Z)}{D_0} \prod_{l=1}^L [1 + ZZ_l + (ZZ_l)^2 + \dots]. \end{aligned} \quad (4.6)$$

It is clear that the filter will be stable only if all the series within the product converge, that is, if all the poles  $Z_l^*$  of the filter are outside the unit circle. This means that all dipoles  $\{1; -Z_l\}$  must satisfy  $|Z_l| < 1$ . Such dipoles are said to be of minimum phase, and their convolution is as well. An unstable filter is unusable even though its gain might perfectly meet expectations (Figure 10.6); however, it can be stabilized by making  $D(Z)$  a minimum-phase filter (Figure 10.7). Figures obtained using the program [ex\\_filtre\\_stable\\_instable.m](#)

## 4. FILTER STABILITY



**Figure 10.6:** An unstable low-pass filter. This filter, designed to retain only the low frequencies of the initial signal (top left), is unstable and produces an unusable filtered signal (top right) showing exponential numerical divergence. Although the filter's gain (bottom left) is as expected, instability occurs because some poles are inside the unit circle (bottom right).



**Figure 10.7:** A stable low-pass filter. This filter is stable, as shown by the filtered signal (top right). It was obtained from the unstable filter in Figure 10.6 by making  $D(Z)$  a minimum-phase filter, which does not alter the gain (bottom left) but places all the poles outside the unit circle (bottom right)



---

## 5 Butterworth Filters

## 6 General Overview

The discretization and truncation of signals make it impossible to realize ideal band-pass filters\*. A good approximation of these filters can be obtained using Butterworth filters, whose low-pass gain function (Figure 10.8) is given by the following relation,

$$|F(u)|^2 = \frac{1}{1 + (u/u_c)^{2n}} \quad (6.1)$$

approaches a window function as the order  $n \rightarrow +\infty$ . Furthermore

$$\begin{cases} |F(\pm u_c)|^2 & = & 1/2 \\ |F(0)|^2 & = & 1 \end{cases} \quad (6.2)$$

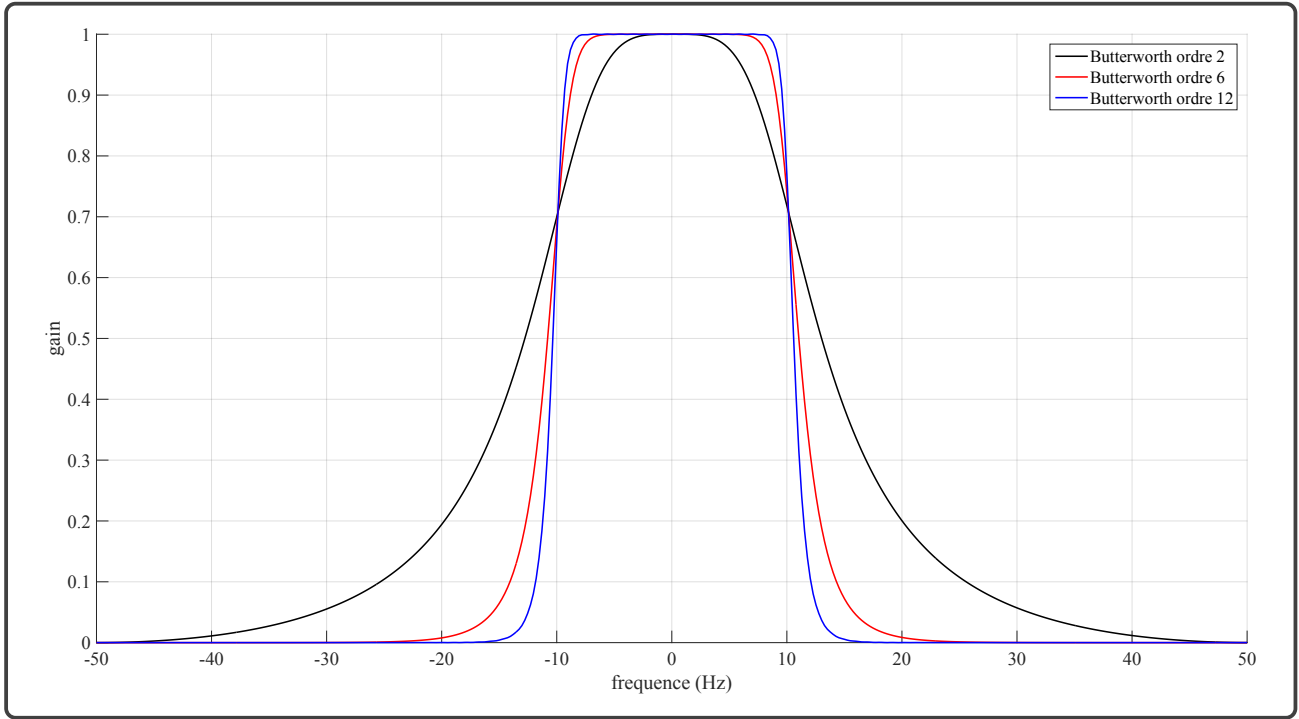
the attenuation at  $u = \pm u_c$  is  $10 \log(1/2) = -3 \text{ dB}$ , which defines the filter's bandwidth as  $[-u_c; +u_c]$ . Outside this band, the higher the filter order, the more rapid the roll-off. For example, a roll-off of at least 48 dB per octave in the range  $[u_c; 2u_c]$  is achieved for orders such that,

$$20 \log \left( \frac{|F(u_c)|}{|F(2u_c)|} \right) \geq 48, \quad (6.3)$$

which justifies the choice of  $n \geq 9$ , which provides a minimum attenuation of 51 dB

---

\*That is, filters constructed using window functions.



**Figure 10.8:** Examples of Butterworth low-pass filters (`ex_butter_2_6_12.m`) for orders of 2, 6, and 12. The higher the order, the closer the filter is to a window function.

The construction of a high-pass filter is easily achieved using a low-pass filter and a passthrough filter\*

$$\begin{aligned}
 |F(u)|^2 &= 1 - \frac{1}{1 + (u/u_c)^{2n}} \\
 &= \frac{(u/u_c)^{2n}}{1 + (u/u_c)^{2n}}
 \end{aligned}
 \tag{6.4}$$

Similarly, a band-pass filter is the intersection of a low-pass filter and a high-pass filter,

$$|F(u)|^2 = \left[ \frac{1}{1 + (u/u_h)^{2n}} \right] \times \left[ \frac{(u/u_b)^{2n}}{1 + (u/u_b)^{2n}} \right]
 \tag{6.5}$$

with the passband being  $[u_b; u_h]$ .

## 7 The Bilinear Transformation

Applying filters via a recursive formula is recommended when the volume of data to be processed is large or when real-time filtering is required. This raises the issue of obtaining the recursive formula

---

\*That is, a filter with an impulse response equal to the Dirac delta function.

corresponding to a filter for which we only know *a priori* the gain. The problem is as follows: given the magnitude of the filter's **Fourier** transform, how can we compute the coefficients of the same filter in the physical space to be able to use a recursive formula? If the general form of the recursive formula is,

$$s_n = \sum_{k=0}^M \alpha_k e_{n-k} + \sum_{l=1}^L \beta_l s_{n-l} \quad (7.1)$$

then the Fourier transform of the filter can be written as,

$$F(u) = \frac{\sum_{k=0}^M \alpha_k Z^k}{1 - \sum_{l=1}^L \beta_l Z^l} \quad (7.2)$$

Obtaining the coefficients  $\alpha_k$  and  $\beta_l$  thus requires that  $F(u)$  be expressed as a ratio of two polynomials in  $Z$ , which is not always straightforward since the variable change  $u \rightarrow Z$  is rarely exact. It is then necessary to use an approximate correspondence between  $u$  and  $Z$  via the development

$$\begin{aligned} -2i\pi u\tau &= \ln(Z) \\ &= -2 \left[ \frac{1-Z}{1+Z} + \frac{1}{3} \left( \frac{1-Z}{1+Z} \right)^3 + \frac{1}{5} \left( \frac{1-Z}{1+Z} \right)^5 + \dots \right] \end{aligned} \quad (7.3)$$

of which the first term provides the bilinear approximation,

$$u \approx \frac{1}{i\pi\tau} \frac{1-Z}{1+Z} \quad (7.4)$$

which is valid (with an error of less than 5%) only for,

$$|u| \leq \frac{1}{10\tau} \quad (7.5)$$

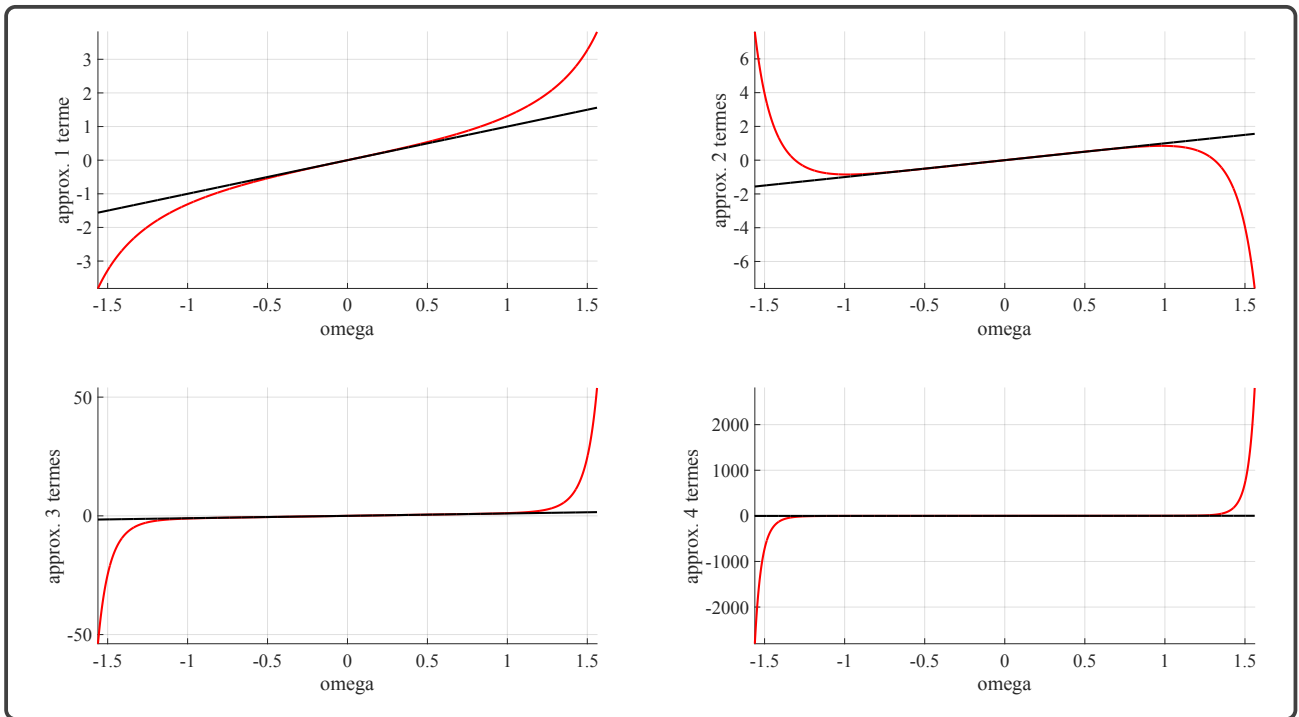
which is a much more restrictive condition than that of **Shannon** discussed in the chapter on sampling (Figure 10.9). From a practical standpoint, a filter constructed using the bilinear approximation will only function correctly for frequencies adhering to this constraint; otherwise, the filter will exhibit performance different from what was specified during its design. It is possible to mitigate this limitation by adopting higher-order approximations\*, but this will result in a longer recursive

---

\*Note that an odd order should always be chosen.

## 8. AN EXAMPLE

formula; thus, a trade-off must be found ([ex\\_bilinear\\_4\\_termes.m](#)).



**Figure 10.9:** The first four terms of the "bilinear approximation". The bilinear transformation, in the strict sense, has a validity range restricted to approximately Nyquist/4 (top left), as shown by the comparison with the line of slope 1 (black). The approximation using only the first two terms of the expansion (top right) has a broader validity range but is practically unusable due to spectral aliasing caused by the function not being bijective. The approximation with three terms (bottom left) is practically usable and has a validity range significantly larger than the classical approximation. *etc*

## 8 An example

The bilinear approximation makes the transformation  $u \rightarrow Z$  straightforward; for example, in the case of a first-order band-pass filter,

$$|F(u)|^2 = \left[ \frac{u_h^2}{u^2 + u_h^2} \right] \times \left[ \frac{u^2}{u^2 + u_b^2} \right] \quad (8.1)$$

pour lequel on peut choisir,

$$F(u) = \left[ \frac{-iu_h}{u - iu_h} \right] \times \left[ \frac{u}{u - iu_b} \right], \quad (8.2)$$

the variable transformation yields,

$$F(Z) \approx \frac{\alpha_0 + \alpha_2 Z^2}{1 + \beta_1 Z + \beta_2 Z^2} \quad (8.3)$$

and the recursive formula,

$$s_n = \alpha_0 e_n + \alpha_2 e_{n-2} - \beta_1 s_{n-1} - \beta_2 s_{n-2} \quad (8.4)$$

which requires only 7 operations. The gain and phase of this filter are shown in Figure 10.10. As you can see in the same figure, the fourth-order filter has a gain that is evidently closer to the ideal window. This higher-order filter has a more complex  $Z$ -transform,

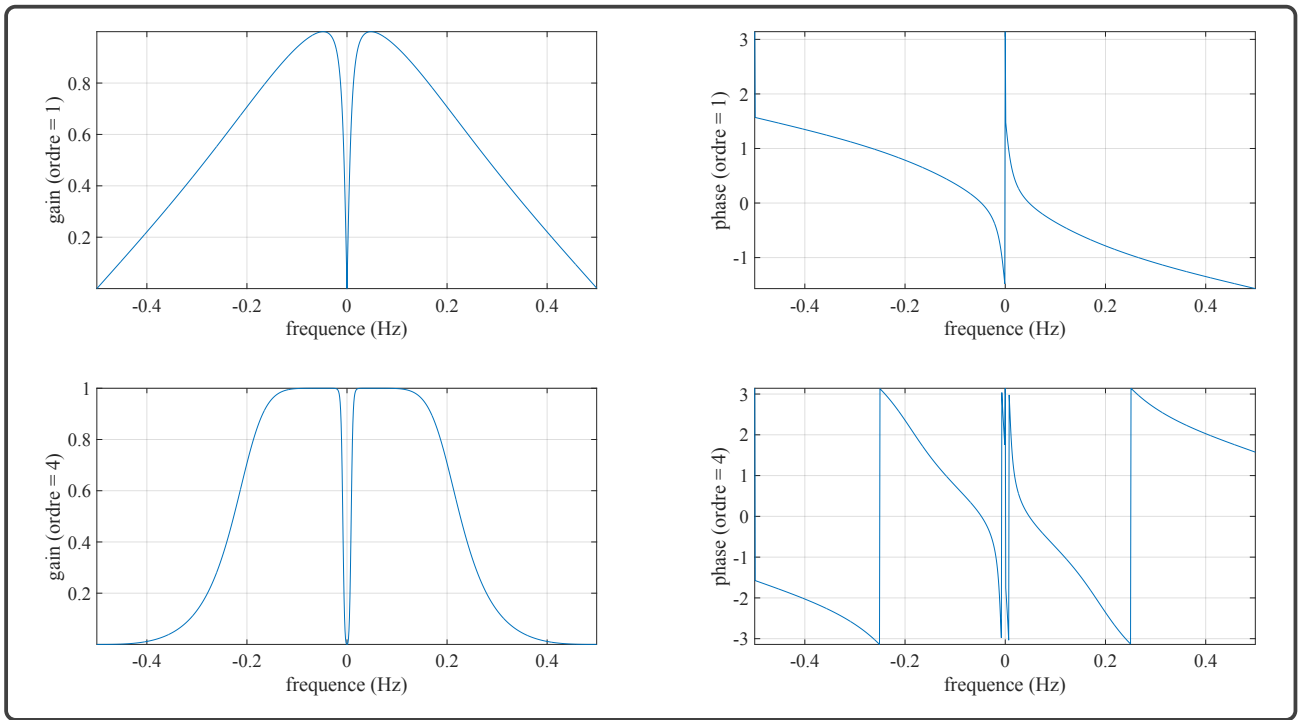
$$F(Z) = \frac{\alpha_0 + \alpha_2 Z^2 + \alpha_4 Z^4 + \alpha_6 Z^6 + \alpha_8 Z^8}{1 + \beta_1 Z + \beta_2 Z^2 + \beta_3 Z^3 + \beta_4 Z^4 + \beta_5 Z^5 + \beta_6 Z^6 + \beta_7 Z^7 + \beta_8 Z^8}, \quad (8.5)$$

and the corresponding recursive formula,

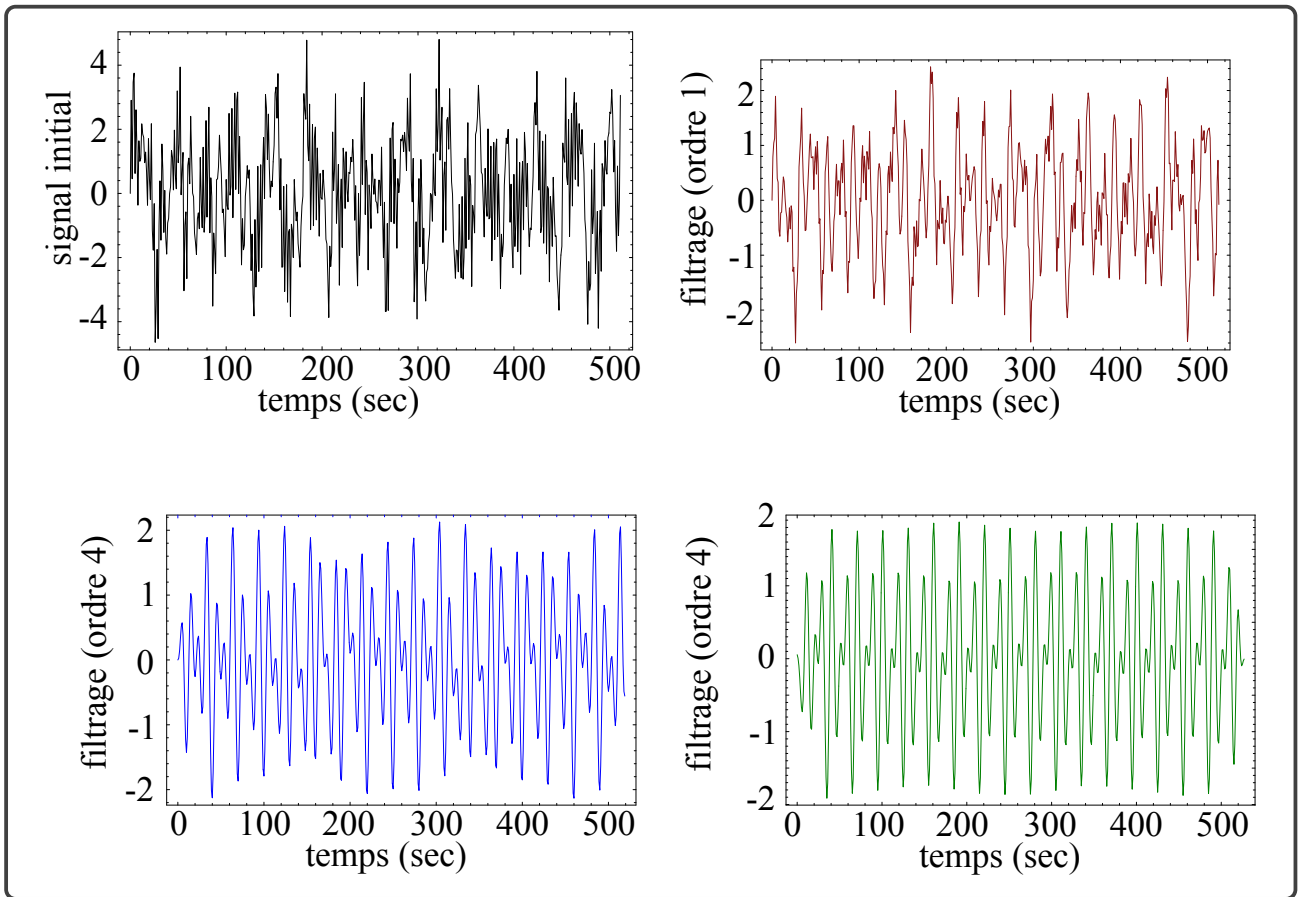
$$s_n = \sum_{k=0}^4 \alpha_{2k} e_{n-2k} - \sum_{l=1}^8 \beta_l s_{n-l} \quad (8.6)$$

and the corresponding recursive formula is longer than that of the first-order filter. This is the price to pay for achieving a filter that is closer to the ideal. Note that the filters we have just created were based on the magnitude of their Fourier transform without concern for the phase, which, in this specific example, is not zero (Figure 10.10). These filters are thus phase-shifting (Figure 10.11). A non-phase-shifting filter (Figure 10.11) can be easily realized by applying the phase-shifting filter in a forward and backward manner; the gain of the resulting filter is then equal to the square of that of the initial phase-shifting filter. Note that the ability to perform non-phase-shifting filtering requires a backward filtering operation, which is anti-causal. This aligns with what we observed at the beginning of this chapter, namely that a non-phase-shifting filter is necessarily anti-causal.

## 8. AN EXAMPLE



**Figure 10.10: Butterworth band-pass filters. The filter at the top is a first-order filter with a passband of 0.04-0.14 Hz, while the filter at the bottom is of the same passband but of fourth order.**



**Figure 10.11: Band-pass filtering.** The initial signal (top left) contains spectral lines between 0.04 Hz and 0.14 Hz, which we will isolate using the filters shown in Figure 10.10. The filtering performed with the first-order filter is shown in the top right, and that with the fourth-order filter is shown in the bottom left. Both of these filtrations were performed in a forward-only manner and are therefore phase-shifting. A non-phase-shifting filtration using the fourth-order filter is shown in the bottom right.

## 9 Wiener Filters

### 10 Wiener Filtering in the Frequency Domain

Unlike the recursive filtering we have just discussed, where the processed signal  $d(t)$  is deterministic, Wiener filtering accounts for the presence of noise  $b(t)$  in the signal to be filtered,

$$s(t) = d(t) + b(t) \tag{10.1}$$

The problem is to construct a linear filter that, when applied to  $s(t)$ , provides an output as close as possible to  $d(t)$ . In the case of Wiener filtering, 'as close as possible' means 'in the least squares sense,' and the desired filter  $f_W$  is such that,

$$f_W * s \stackrel{\mathbf{L}_2}{=} d \quad (10.2)$$

The obtained filter will be optimal in the probabilistic sense if the noise statistics are Gaussian. Otherwise, a different norm would need to be adopted for optimization. The equality above can be expressed as,

$$\int_{-\infty}^{+\infty} |[f_W * s](t) - d(t)|^2 dt \text{ MINIMUM} \quad (10.3)$$

in the time domain, or

$$\int_{-\infty}^{+\infty} |F_W(u)S(u) - D(u)|^2 du \text{ MINIMUM} \quad (10.4)$$

in the frequency domain. By expanding this latter expression,

$$\begin{aligned} & \int_{-\infty}^{+\infty} \left[ |D(u)|^2 |F_W(u) - 1|^2 + |F_W(u)|^2 |B(u)|^2 \right] du \\ & + \int_{-\infty}^{+\infty} D(u)B^*(u) \left[ |F_W(u)|^2 - F_W^*(u) \right] du \\ & + \int_{-\infty}^{+\infty} D^*(u)B(u) \left[ |F_W(u)|^2 - F_W(u) \right] du \\ & \text{MINIMUM} \end{aligned} \quad (10.5)$$

An important simplification of this expression occurs if we assume that the noise and the deterministic component are uncorrelated. The last two integrals are identically zero, and the filter must be such that,

$$\int_{-\infty}^{+\infty} \left[ |D(u)|^2 |F_W(u) - 1|^2 + |F_W(u)|^2 |B(u)|^2 \right] du \text{ MINIMUM} \quad (10.6)$$

By requiring that  $F_W \in \mathbf{R}$ , the minimization condition becomes,

$$\frac{d}{dF_W(u)} \int_{-\infty}^{+\infty} \left[ |D(u)|^2 (F_W(u) - 1)^2 + F_W(u)^2 |B(u)|^2 \right] du \quad (10.7)$$



---

and the filter,

$$F_W(u) = \frac{|D(u)|^2}{|D(u)|^2 + |B(u)|^2} \quad (10.8)$$

The filter can only be constructed if the energy spectra of the components  $b(t)$  and  $d(t)$  are known (Figure 10.12), which is generally possible only through prior information. This reflects the ongoing ambiguity in signal processing that we mentioned in the introduction. While this information may seem difficult to obtain, it is important to remember that the filter was constructed through a minimization process (least squares) that nullifies the first derivative of the cost function. Thus, errors in the filter definition will only start to manifest in the second order, which mitigates their impact (Figure 10.12). The deterministic signal (top left) is a pure sinusoid (0.1 Hz) and is contaminated by Gaussian noise (second row left, black curve) before filtering (last row left, black curve). The energy spectra of the noise (second row right) and the deterministic signal (top right) are used to construct the gain of the Wiener filter (fourth row right). The filtered signal is shown at the bottom left (fourth row). The robustness of Wiener filtering can be appreciated in this figure, where the filter gain (left) was constructed by replacing the noise energy spectrum with its average value. It is observed that the filtered signal (right) is not significantly affected by this simplification. Figure (10.12) was obtained using the function [ex\\_wiener.m](#).

## 10. WIENER FILTERING IN THE FREQUENCY DOMAIN

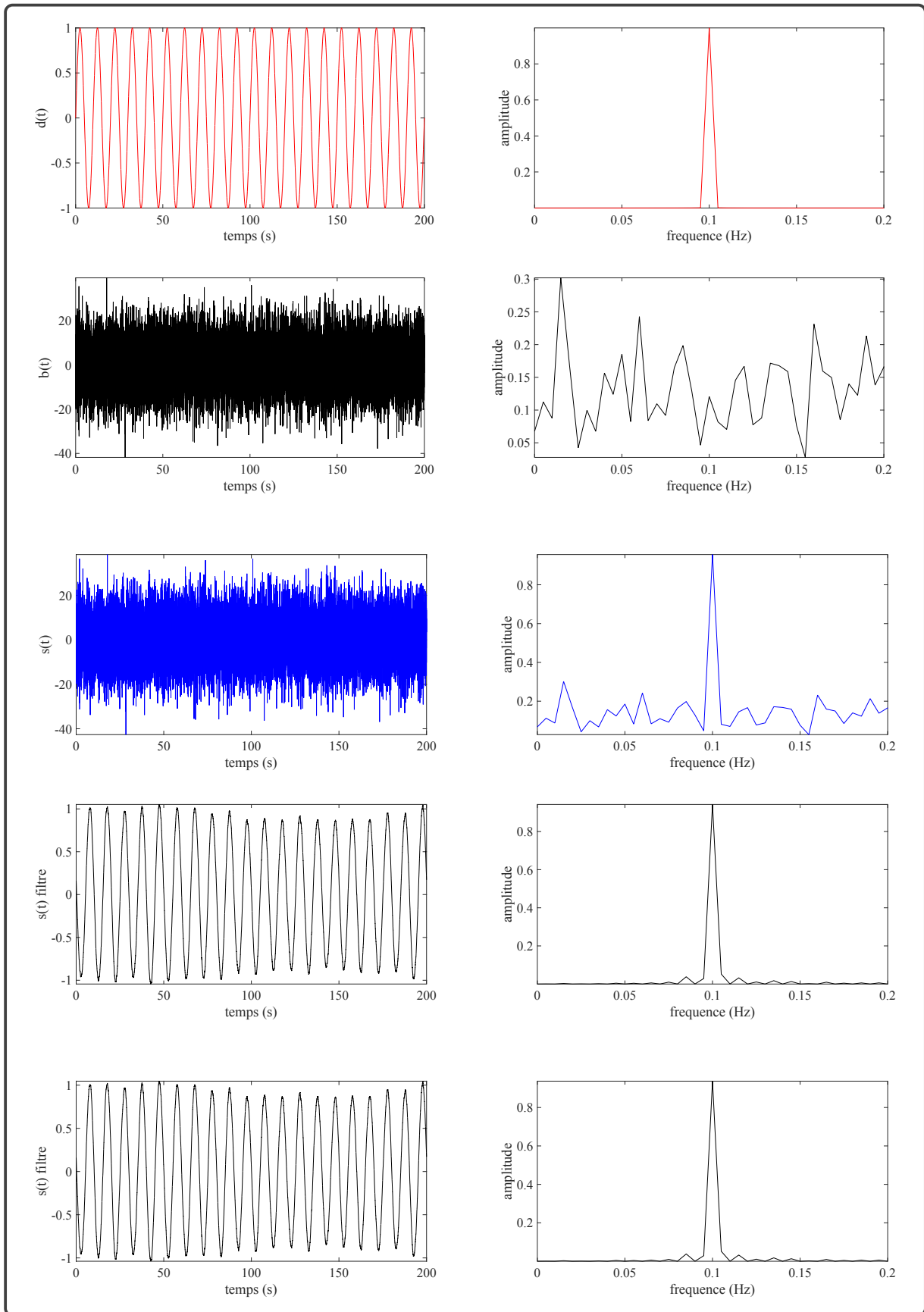


Figure 10.12: Wiener Filtering



---

---

# CHAPTER 11

---

## SPECTRAL ANALYSES

<b>1</b>	<b>The spectral analysis models</b> . . . . .	<b>156</b>
1.1	<i>Prony, Hildebrand, Pisarenko and Schuster</i> : Trigonometric series . . . . .	156
1.2	<i>Burg, Pisarenko, . . .</i> : autoregressive models . . . . .	158
<b>2</b>	<b>Discrete Fourier Transform Analysis</b> . . . . .	<b>160</b>
2.1	Schuster's periodogram . . . . .	160
2.2	Signal truncation effects . . . . .	162
2.3	Apodization windows . . . . .	164
2.4	Impact of a trend . . . . .	167
2.5	Statistical issues . . . . .	168
<b>3</b>	<b>Autoregressive model analysis</b> . . . . .	<b>170</b>
3.1	The prediction error filter . . . . .	171
3.2	Prediction error filter utility . . . . .	172

---

# 1 The spectral analysis models

This chapter deals with the problem of spectral analysis, which is the study of the distribution of the energy of a signal as a function of frequency. This distribution law is known as the energy spectrum, which is defined as the square of the modulus of the [Fourier](#) transform of the signal,

$$E(u) \equiv |S(u)|^2 \tag{1.1}$$

The simplest method of calculating the energy spectrum of a sampled signal is to use the Discrete [Fourier](#) Transform. Although effective, this method has certain drawbacks which have led to the development of alternative techniques whose main advantage is the ability to achieve very fine frequency resolutions. This is not a miraculous violation of the Uncertainty Principle discussed earlier, but rather a consequence of the fact that these techniques are autoregressive and implicitly extrapolate the analysed signal beyond the observation interval, thereby increasing the frequency resolution. However, this extrapolation comes with restrictive assumptions\* that limit the applicability of these methods to certain categories of signals. We refer the interested reader to the excellent article by [Kay et Marple \(1981\)](#) for a critical review of these methods. In general, any spectral analysis method is based on fitting a model to the data and calculating a spectrum from the parameters of that model. Seen in this light, it is clear that spectral analysis falls within the scope of inverse problem theory.

## 1.1 *Prony, Hildebrand, Pisarenko and Schuster* : Trigonometric series

The oldest model was proposed in 1795 by Baron de [Prony \(1795\)](#) , not for spectral analysis, but to describe the behaviour of certain gases. This model,

$$s_{\text{PRONY}}(t) = \sum_{m=1}^M S_m \exp(\alpha_m t) \exp(2i\pi u_m t) \tag{1.2}$$

is composed of damped sinusoids and has strong links with [Fourier](#) analysis. This model is very general, with adjustable parameters

$$\{M, S_m, \alpha_m, u_m\} \quad (m = 1, \dots, M) \tag{1.3}$$

---

\*which many authors conveniently overlook!

## 1. THE SPECTRAL ANALYSIS MODELS

---

make the inverse problem highly non-linear. Note that even the number\*  $M$  of elements in the sum is *a priori* unknown. This is an inverse problem where the exact number of parameters is not known. The [Prony](#) model is often used in signal processing, and the inverse problem is generally simplified and not treated in a non-linear way. The solutions obtained are approximate and have biases that become more significant as the signal-to-noise ratio deteriorates. Other models used in spectral analysis can be considered as simplified versions of the [Prony](#) model. For example, the model of [Hildebrand \(1956\)](#),

$$s_{\text{HILDEBRANT}}(t) = \sum_{m=1}^M S_m \exp(2i\pi u_m t) \quad (1.4)$$

is obtained by setting  $\alpha_m = 0$  in the [Prony](#) model. The adjustable parameters are,

$$\{M, S_m, u_m\} \quad (m = 1, \dots, M) \quad (1.5)$$

The model used in the method of [Pisarenko \(1973\)](#),

$$s_{\text{PISARENKO}}(t) = \sum_{m=1}^M S_m \exp(2i\pi u_m t) + \sigma_b b(t) \quad (1.6)$$

is very similar to [Hildebrand](#), but explicitly takes into account that the data are contaminated by white noise,  $b(t)$ , whose variance,  $\sigma_b^2$ , is part of the set of adjustable parameters,

$$\{M, \sigma_b, S_m, u_m\} \quad (m = 1, \dots, M) \quad (1.7)$$

The solution provided by [Pisarenko](#) involves working from the autocorrelation function of the data and does not allow for the recovery of phases. One only has access to the energy spectrum. The models of [Hildebrand](#) and [Pisarenko](#) are suited for representing data with a line spectrum. In contrast, due to the presence of the damping coefficients,  $\alpha_m \neq 0$ , the [Prony](#) model allows for the analysis of continuous spectra, which may also contain lines.

All of these models are highly non-linear, and estimating their parameters poses significant challenges. The suboptimal solutions typically computed are often unsatisfactory when the data are noisy. Estimating the order,  $M$ , of these models can be done more or less accurately and is undoubtedly a critical stage of these techniques. This probably explains the popularity of the [Schuster](#)

---

\*Called the order of the model.

model,

$$s_{\text{SCHUSTER}}(n\tau) = \sum_{k=0}^{N-1} S_k \exp(2i\pi kn/N) \quad (n = 0, 1, \dots, N-1) \quad (1.8)$$

whose frequencies, fixed *a priori*, correspond to the number  $N$  of available data\*. The set of parameters

$$\{S_k\} \quad (k = 0, \dots, N-1) \quad (1.9)$$

is reduced to those that appear linearly in the [Prony](#) model. We will see that this model, fitted to the data by least squares, gives a spectral analysis by discrete [Fourier](#) transform.

## 1.2 [Burg, Pisarenko,...](#): autoregressive models

We will now delve into the realm of spectral analysis using autoregressive models. Many methods employ such models, with one of the most popular being the maximum entropy analysis method. The simplest way to understand the role of autoregressive models is to start with the [Fourier](#) transform,

$$S_\tau(u) = \tau \sum_{n=-\infty}^{+\infty} s(n\tau) Z^n \quad (1.10)$$

of the discrete signal  $s_n \equiv s(n\tau)$ . When the signal is truncated, we have seen that the sum in the above equation is bounded,

$$S_{\tau,T}(u) = \tau \sum_{n=0}^{N-1} s_n Z^n \quad (1.11)$$

where  $T = N\tau$ . The associated energy spectrum,

$$E_{\tau,T}(u) = \tau \left| \sum_{n=0}^{N-1} s_n Z^n \right|^2, \quad (1.12)$$

is represented by a finite number of terms, which poses problems for analytical representation if the true spectrum contains lines. A better representation of such a spectrum can be achieved by

---

\*Assumed to be sampled at a constant interval  $\tau$ .

## 1. THE SPECTRAL ANALYSIS MODELS

---

using an autoregressive model of the type,

$$s_n = - \sum_{m=1}^M b_m s_{n-m} \quad (1.13)$$

where  $M$  is the model order\*.

This justification for autoregressive models can be further supported by noting that a discrete monochromatic signal leads directly to an autoregressive expression AR(2),

$$\begin{aligned} s_n &= \sin(2\pi u_0 \tau n) \\ &= 2 \cos(2\pi u_0 \tau) \sin(2\pi u_0 \tau (n-1)) - \sin(2\pi u_0 \tau (n-2)) \\ &= 2 \cos(2\pi u_0 \tau) s_{n-1} - s_{n-2} \end{aligned} \quad (1.14)$$

The initialisation of this recursive formula is necessary when the signal is truncated, and two initial values,  $e_0$  and  $e_1$ , must be provided. These values determine the amplitude and phase of the sine wave to be generated. The recursive formula is,

$$\begin{cases} s_0 &= e_0 \\ s_1 &= 2 \cos(2\pi u_0 \tau) s_0 + e_1 \\ s_n &= 2 \cos(2\pi u_0 \tau) s_{n-1} - s_{n-2} \quad (n > 1) \end{cases} \quad (1.15)$$

and its Z-transform provides,

$$S(Z) = \frac{e_0 + e_1 Z}{1 - 2 \cos(2\pi u_0 \tau) Z + Z^2} \quad (1.16)$$

which is none other than the Z-transform of the narrowband filter studied in the chapter on filtering. Extending this to a signal composed of  $M$  frequencies,

$$s_n = \sum_{l=1}^M S_l \exp(2i\pi u_l \tau n) \quad (n \geq 0) \quad (1.17)$$

it's not much more difficult. Indeed,

$$s_{n-m} = \sum_{l=1}^M S_l \exp[2i\pi u_l \tau (n-m)] \quad (n \geq 0), \quad (1.18)$$

---

\*Such models are often referred to in the technical literature by the notation AR(M)



and by multiplying both sides of this equation by a coefficient  $b_m$  and summing  $M + 1$  such equations,

$$\begin{aligned} \sum_{m=0}^M b_m s_{n-m} &= \sum_{m=0}^M b_m \sum_{l=1}^M S_l [\exp(2i\pi u_l \tau)]^{n-m} \\ &= \sum_{l=1}^M S_l [\exp(2i\pi u_l \tau)]^{n-M} \sum_{m=0}^M b_m [\exp(2i\pi u_l \tau)]^{M-m} \end{aligned} \quad (1.19)$$

valid for  $n \geq M$ . Let's choose the coefficients  $b_m$  such that,

$$b_0 = 1 \quad (1.20)$$

and,

$$\sum_{m=0}^M b_m [\exp(2i\pi u_l \tau)]^{M-m} = 0, \quad (1.21)$$

we obtain the recursive formula directly,

$$s_n = - \sum_{m=1}^M b_m s_{n-m} \quad (n \geq M) \quad (1.22)$$

Line spectra can thus be modelled by autoregressive models, for which the task now is to determine the parameters  $b_m$ .

## 2 Discrete Fourier Transform Analysis

### 2.1 Schuster's periodogram

This technique involves the representation of the observed signal,

$$s_n^{obs} \equiv s(n\tau) \quad (n = 0, 1, \dots, N-1) \quad (2.1)$$

using the model,

$$s_n^{mod} = \sum_{k=0}^{N-1} S_k \exp(2i\pi kn/N) \quad (n = 0, 1, \dots, N-1), \quad (2.2)$$

## 2. DISCRETE FOURIER TRANSFORM ANALYSIS

consisting of  $N$  sinusoids with frequencies that are multiples of  $\nu = 1/N\tau$ . Note that this model is highly constrained: the frequencies are fixed *a priori* in both value and number, and the nature of the functions is also predetermined; they are sinusoids and nothing else. The only adjustable parameters are the  $S_k$ , which allow the amplitudes and phases of each sinusoid in the model to be adjusted. Several generalisations have been proposed to also adjust the number of frequencies and their values. Although these generalisations are quite legitimate, their main drawback is that they render the problem highly non-linear and practically very difficult to solve. Let us rewrite our initial model in its expanded form,

$$\begin{pmatrix} s_0^{mod} \\ \vdots \\ s_n^{mod} \\ \vdots \\ s_{N-1}^{mod} \end{pmatrix} = \begin{bmatrix} 1 & \cdots & 1 & \cdots & 1 \\ \vdots & & \vdots & & \vdots \\ 1 & \cdots & \exp\left[\frac{2i\pi nk}{N}\right] & \cdots & \exp\left[\frac{2i\pi n(N-1)}{N}\right] \\ \vdots & & \vdots & & \vdots \\ 1 & \cdots & \exp\left[\frac{2i\pi k(N-1)}{N}\right] & \cdots & \exp\left[\frac{2i\pi(N-1)^2}{N}\right] \end{bmatrix} \times \begin{pmatrix} S_0 \\ \vdots \\ S_k \\ \vdots \\ S_{N-1} \end{pmatrix}; \quad (2.3)$$

or, in a more compact form,

$$\vec{s}_{mod} = \mathbf{W} \vec{S} \quad (2.4)$$

The problem now is the computation of the components of the vector  $\vec{S}$ , so that

$$\vec{s}_{mod} \approx \vec{s}_{obs} \quad (2.5)$$

This fitting is not unique and of course depends on the criterion chosen to determine whether the model predictions are close to the observed data: a norm must be chosen. The classical choice of the  $\mathbb{L}_2$  norm leads to the optimal least squares fitting criterion, for which the best model is such that

$$\|\vec{s}_{mod} - \vec{s}_{obs}\|^2 \text{ MIMIMUM.} \quad (2.6)$$

The solution obtained by applying this criterion is

$$\begin{aligned} \vec{S} &= [\mathbf{W}^H \mathbf{W}]^{-1} \mathbf{W}^H \vec{s}_{obs} \\ &= \frac{1}{N} \mathbf{W}^* \vec{s}_{obs} \end{aligned} \quad (2.7)$$

where we have used the fact that,

$$[\mathbf{W}^H \mathbf{W}]^{-1} = \frac{1}{N} \mathbf{I} \quad (2.8)$$

and,

$$\mathbf{W}^H = \mathbf{W}^* \quad (2.9)$$

Let us rewrite this solution in its extended form,

$$S_k = N^{-1} \sum_{n=0}^{N-1} s_n^{obs} \exp(-2i\pi kn/N) \quad (k = 0, 1, \dots, N-1). \quad (2.10)$$

(2.10) is a slightly modified form of the discrete **Fourier** transform. The least-squares fitting of the **Schuster** model presented at the beginning of this section is thus equivalent to the spectral analysis method based on the discrete **Fourier** transform of the observed signal. This method, which is by far the most commonly used, is therefore very precise; in particular, it only provides optimal solutions when the noise contaminating the data is **Gaussian** and white. Otherwise\*, the solution obtained can be significantly biased, as indicated by the notable lack of robustness of the least squares criterion.

## 2.2 Signal truncation effects

The equivalence between **Schuster**'s method and the discrete **Fourier** transform allows us to make direct use of some previously established results. For example, the fact that the observed signal is a truncated version of the real signal.

$$s_T(t) = s(t) \Pi(t/T) \quad (2.11)$$

means that the computed **Fourier** transform is a degraded version of the real signal,

$$S_T(u) = TS(u) * \text{sinc}(uT) \quad (2.12)$$

From a practical point of view, this degradation manifests itself in two effects: the limitation of

---

\*for example, in the presence of outliers in the signal.

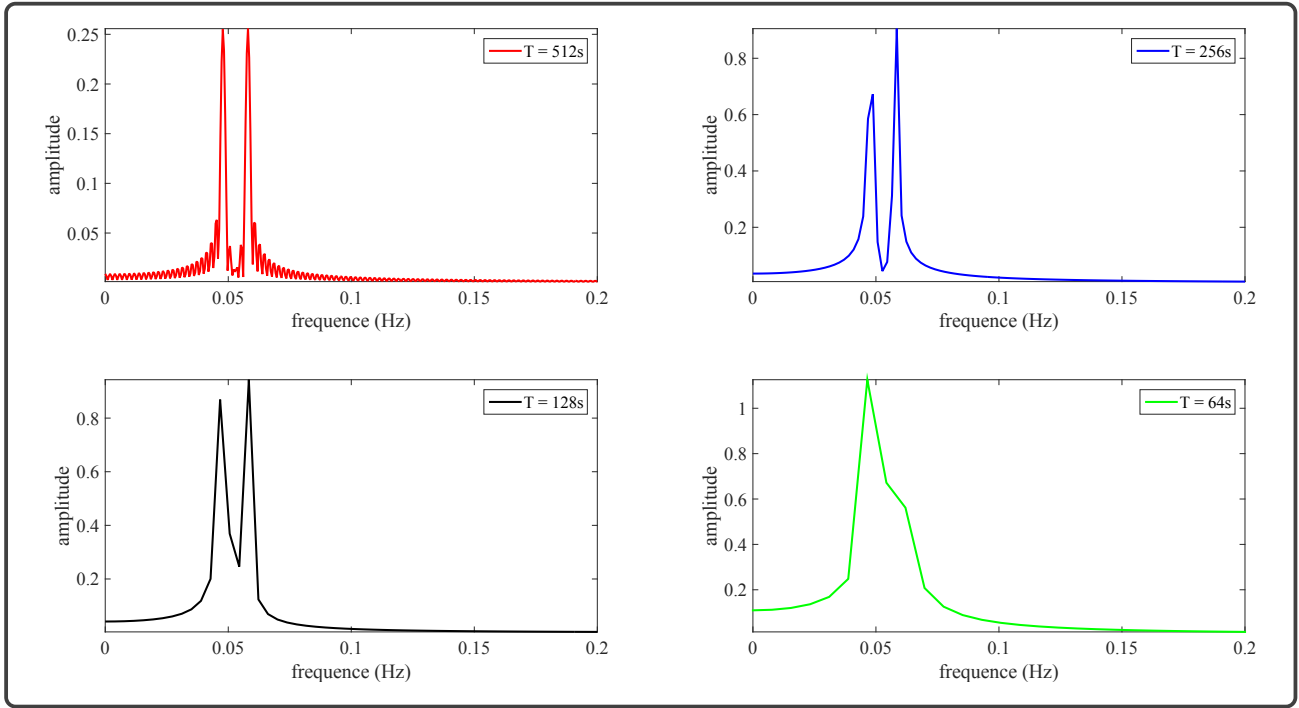
the frequency resolution and the phenomenon of *leakage*..

Frequency resolution, as we have seen, can be defined as the width of the main lobe of the sinc function,

$$\delta u \approx \frac{1}{T} \tag{2.13}$$

and it is clear that severe truncation can prevent the resolution of closely spaced spectral lines (Figure 11.1). The remedy is to increase the length of the analysed signal or to use a spectral analysis method other than *Schuster*'s. If you choose the latter solution, make sure that the 'miracle' method you intend to use is suitable for your signal.

*"leakage"* is a phenomenon caused by the secondary lobes of the sinc function that appear in the convolution described above. If the spectrum of the signal being analysed contains a mixture of large and small energy peaks, these secondary lobes can completely obscure the smaller energy peaks near the larger ones. This effect results in a transfer of energy from the original frequency to neighbouring frequencies, hence the term *"leakage"*. If the original peak is very intense, this transfer can affect a significant portion, or even all, of the calculated spectrum, so it is sometimes necessary to reduce this effect by using apodization windows.



**Figure 11.1: Signal truncation limits the frequency resolution.** The signal analysed in this example consists of two sinusoids (0.048 Hz and 0.058 Hz) sampled with  $\tau = 1$ ;s. The frequency resolution, approximately equal to  $1/T$ , is sufficient to resolve the spectral lines at  $T = 512$ ;s (top left),  $T = 256$ ;s (top right), and  $T = 128$ ;s (bottom left). However, a stronger truncation,  $T = 64$ ;s (bottom right), no longer allows the resolution of the two spectral lines. Note that the amplitude of the peaks in these spectra decreases as they widen. Note also the increasing prominence of the secondary lobes associated with the main peaks as the duration of the signal analysed decreases. This phenomenon, known as *leakage*, can be reduced by using apodization windows.

### 2.3 Apodization windows

Apodisation windows are used to reduce the leakage phenomenon. These are functions  $f(t)$  whose Fourier transform has smaller secondary lobes than those of  $\text{sinc}(u)$  (Figures 11.2 and 11.3, cf [ex\\_apodisation.m](#)). In this case, the apodised sample,

$$s_T(t) = \frac{1}{A_f} f(t/T) s(t) \quad (2.14)$$

where the normalization factor,

$$A_f = \int_{-T/2}^{+T/2} f(t/T) dt \quad (2.15)$$

corrects for the artificial attenuation introduced by the window. The resulting Fourier transform

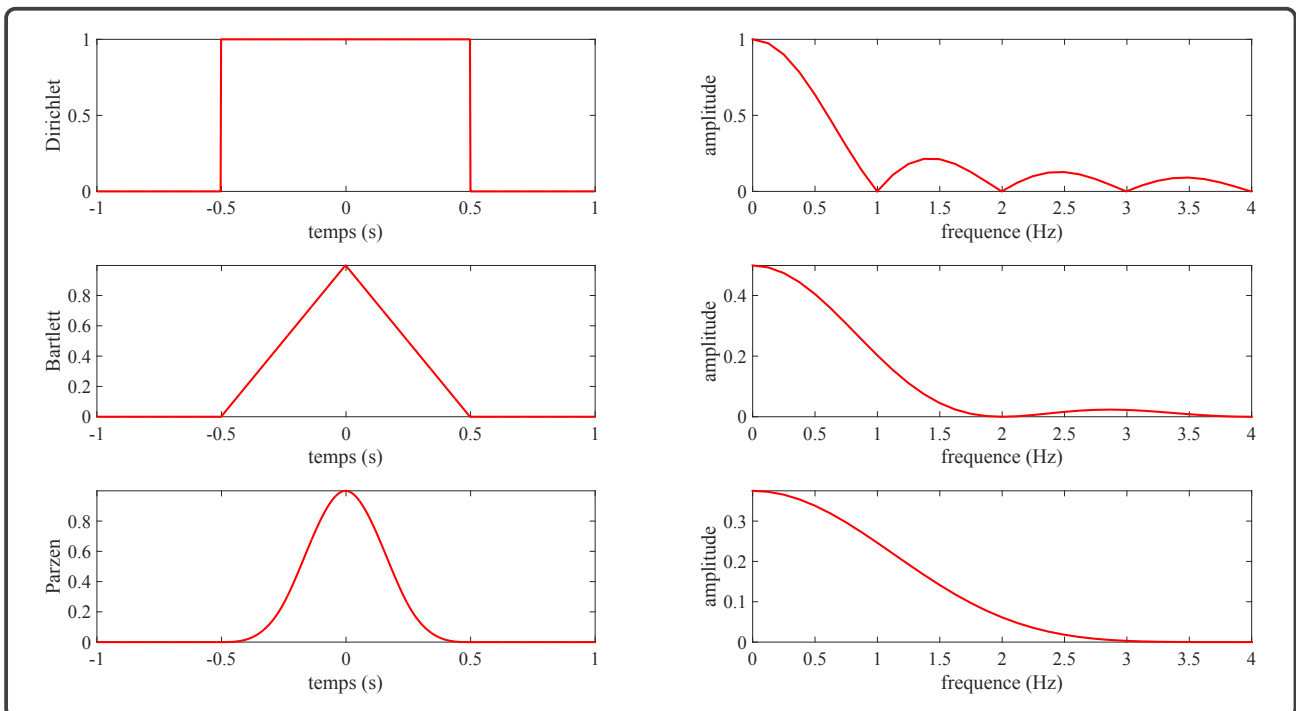
(Figure 11.4),

$$S_T(u) = \frac{T}{A_f} S(u) * F(uT). \tag{2.16}$$

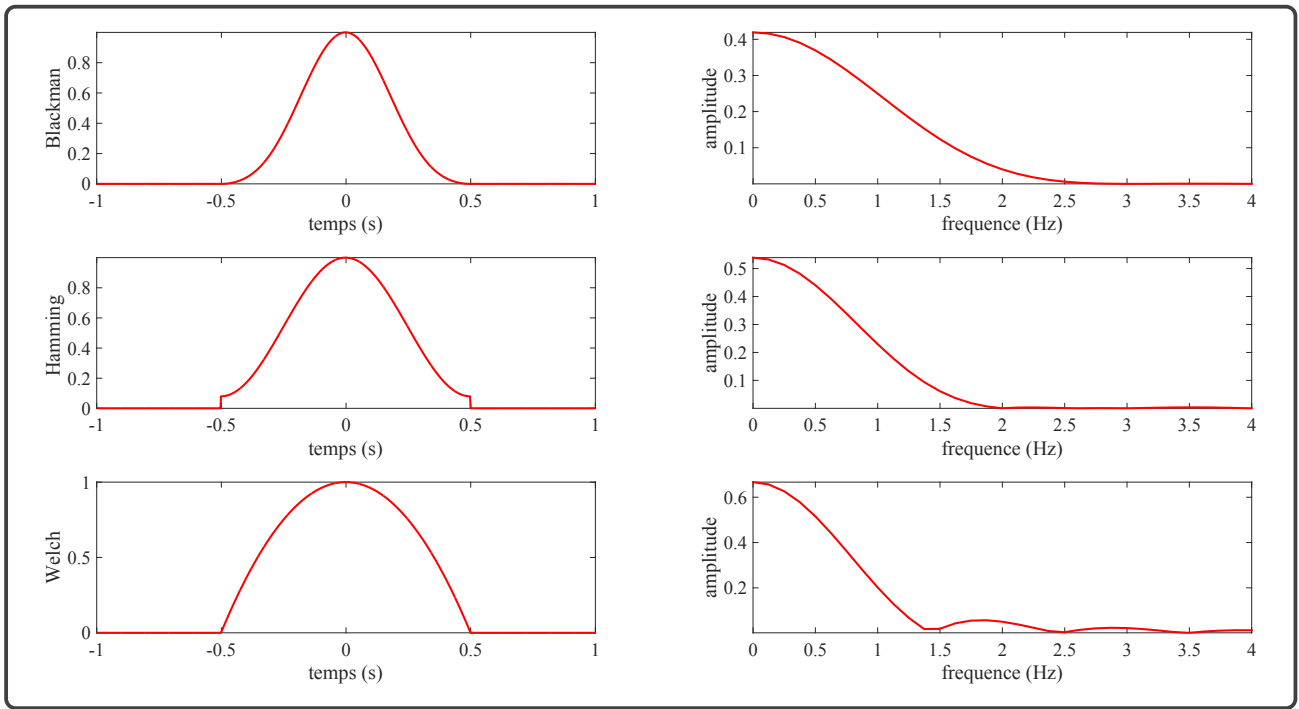
Among the many possible apodisation windows (Figures 11.2 and 11.3), all of which are zero outside the interval  $[-T/2; +T/2]$ , we can mention the **Dirichlet** window,

$$\Pi(t/T) \tag{2.17}$$

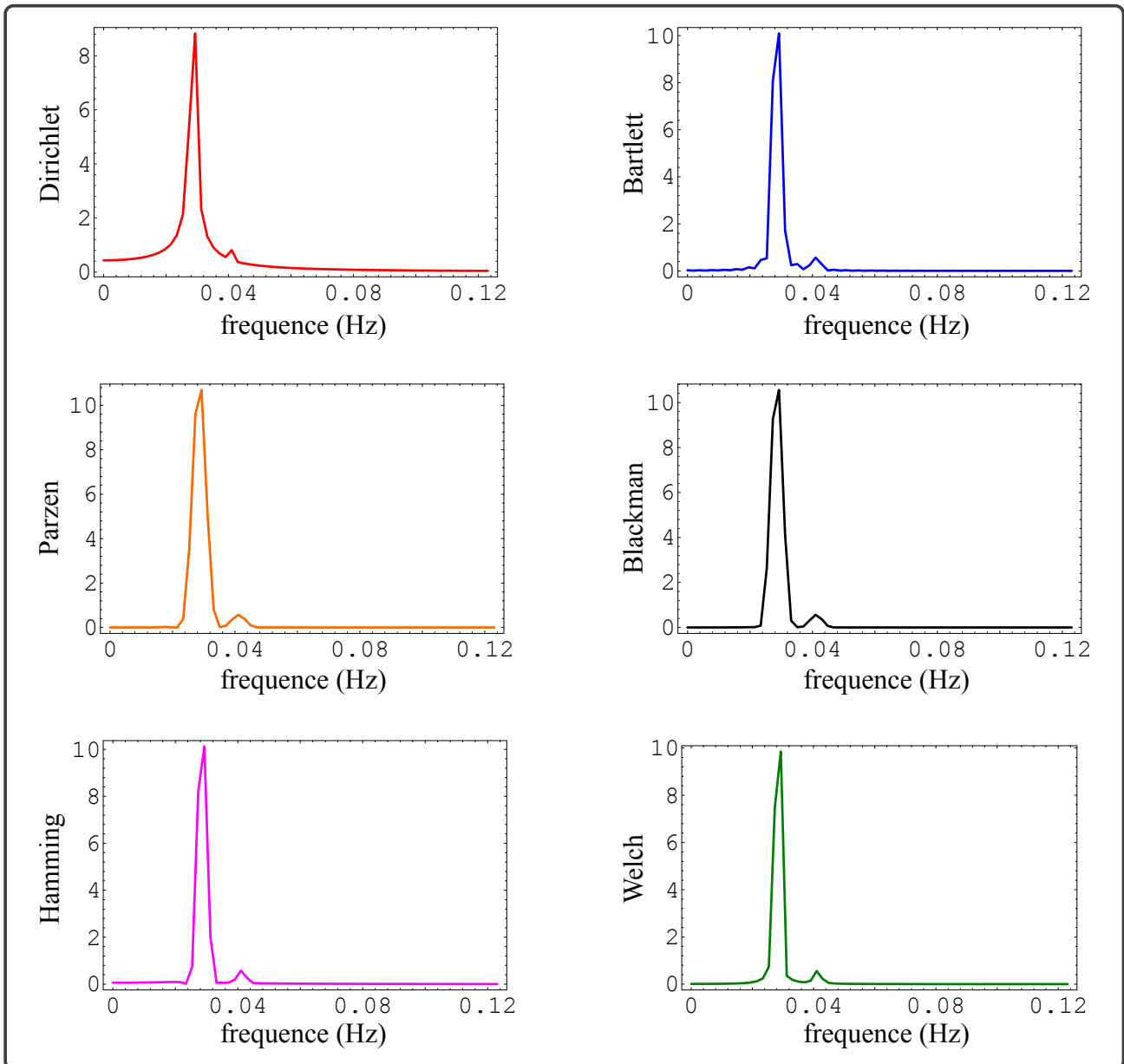
which is none other than the window discussed in the chapter on signal truncation. It is important to note that, as in this case, there is no magic solution: the reduction in "leakage" comes at the cost of a reduction in frequency resolution.



**Figure 11.2: Apodization windows.** These three windows are the **Dirichlet** (top left), **Bartlett** (middle left) and **Parzen** (bottom left) windows. They are obtained by successive auto-convolutions of the function  $\Pi(t)$ . As the number of auto-convolutions increases (from top to bottom), the window becomes smoother, and its Fourier transform (right) has attenuated secondary lobes and a wider central lobe, corresponding to a degradation in frequency resolution. From the central limit theorem illustrated in a previous chapter, you know that the limiting window obtained by this auto-convolution process is the Gaussian window, which is not very different from the **Parzen** window



**Figure 11.3: Apodization windows.** These three windows are the **Blackman** (top left), **Hamming** (middle left) and **Welch** (bottom left) windows. Note that the **Blackman** window is very similar to the **Parzen** window.



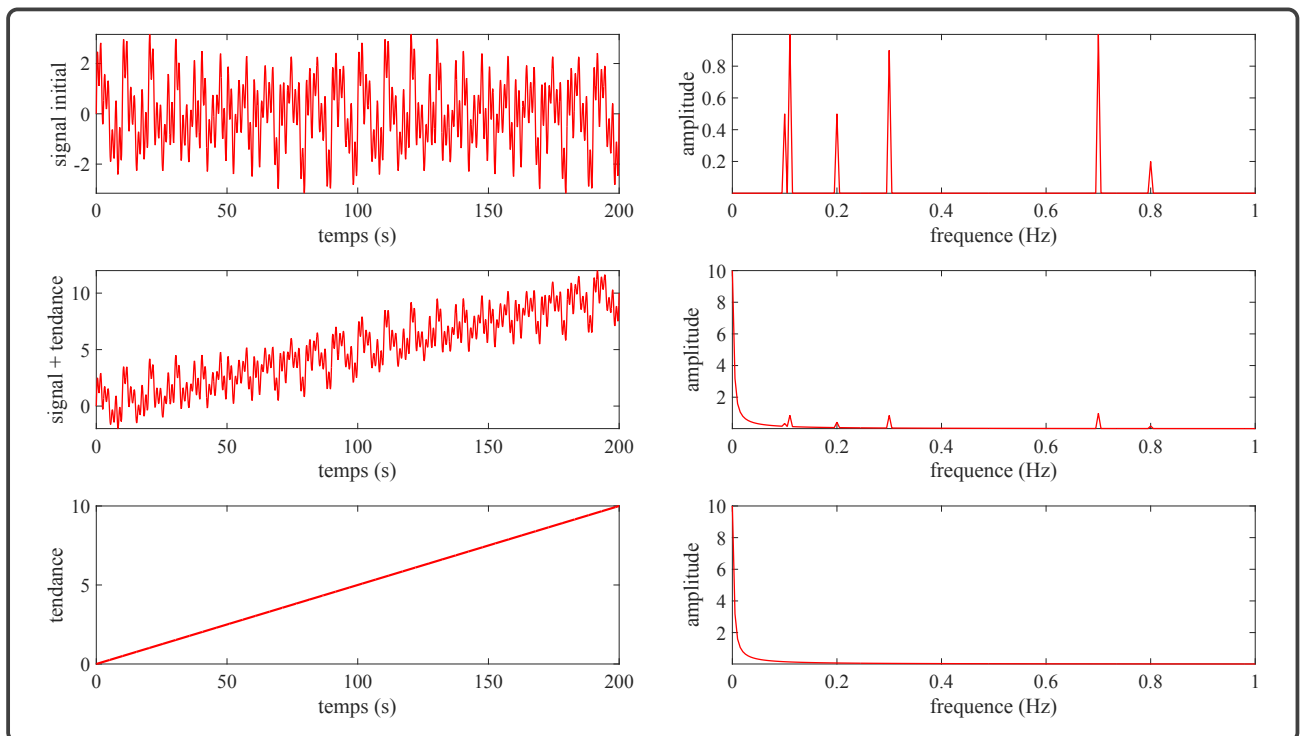
**Figure 11.4:** Effects of the apodisation window. The analysed signal consists of two sinusoids with different amplitudes (1 and 0.05), sampled with  $\tau = 1;s$  and  $T = 512;s$ . The amplitude spectra obtained after apodising the signal with windows smoother than the Dirichlet window (top left) allow a better resolution of the low amplitude spectral line.

## 2.4 Impact of a trend

We will refer to a trend as the component of the sampled signal characterised by oscillations with periods longer than the duration of the sample itself. Ideally, the energy of this trend should be entirely contained within the spectral coefficient corresponding to the zero frequency; in practice, as we have seen, "*leakage*" causes some of this energy to spill over to neighbouring frequencies. If the trend is significant, and therefore energetic, this leakage will cause significant distortion in the spectral coefficients corresponding to the lower frequencies of the spectrum (Figure 11.5). There will



also be additional effects due to the nature of the [Schuster](#) model, which can only produce signals of period  $T$ . Adopting this model implicitly assumes that the signal being analysed is itself periodic, and the presence of a trend means that this periodic signal will essentially exhibit a sawtooth pattern, with its spectrum dominating the rest. As a result, a large portion of the spectrum obtained can become contaminated and difficult to interpret. The presence of a trend in a signal is therefore an unfortunate event; its removal is necessary to obtain a usable spectrum. However, this removal is generally not straightforward and requires a good understanding of the physics of the signal to develop an appropriate model for the trend to be removed.



**Figure 11.5: Effects of the presence of a trend in the analysed signal. When the signal contains a significant trend (middle left), its spectrum (middle right) is primarily representative of that of the trend alone (bottom left). Some details that are visible in the spectrum (top right) of the signal without the trend (top left) may then be obscured.**

## 2.5 Statistical issues

We will consider the case where the signal contains white Gaussian noise. Due to the linearity of the discrete Fourier transform, the real and imaginary parts of the spectral estimates\*  $\tilde{S}_k$  will be Gaussian variables, and the coefficients of the power spectrum,

$$\tilde{E}_k = \left| \tilde{S}_k \right|^2 \tag{2.18}$$

\*The presence of a tilde indicates that we have an estimate of the parameter in question.

follow a  $\chi_2$  distribution

$$\frac{\tilde{E}_k}{E_k} = \chi_2^2 \quad (2.19)$$

where the values,

$$E_k = |S_k|^2 \quad (2.20)$$

are the true (but unknown) values. There are two degrees of freedom because the coefficients of the power spectrum are the sum of two squared Gaussian variables (the imaginary and real parts). The variance of the reduced variable  $\tilde{E}_k/E_k$  is 4 and does not decrease as the signal length increases because the number of spectral estimates increases in the same proportion. The only way to reduce the variance is to average  $M$  independent spectra.

$$\tilde{E}_k = \frac{1}{M} \sum_{m=1}^M \tilde{E}_{m,k} \quad (2.21)$$

so that,

$$\frac{\tilde{E}_k}{E_k} = \chi_{2M}^2. \quad (2.22)$$

The variance of the estimator is now reduced to  $4/M$ . At this point, it is important to note that if the signal being analysed is real, the Fourier coefficients corresponding to negative frequencies do not provide any information beyond that already contained in the positive frequencies. It is therefore illusory to hope for a further reduction in variance by extending the above sum to include negative frequencies.

If only a single signal is available, it is possible to divide it into segments to perform the averaging recommended earlier. However, in accordance with the uncertainty principle, improving the statistical resolution of the estimates  $\tilde{E}_k$  will result in a degradation of the frequency resolution. More specifically, the frequency resolution is such that,

$$\delta u \approx \frac{M}{T}, \quad (2.23)$$

and the standard deviation of the estimator is,

$$\sigma_E = \frac{2}{\sqrt{M}}. \quad (2.24)$$

The uncertainty relation is derived from these results,

$$\sigma_E \times \delta u \approx \frac{2\sqrt{M}}{T}. \quad (2.25)$$

Assuming that the noise contaminating the data is white and Gaussian, it is possible to use the previous results to calculate the bounds of the confidence intervals associated with the  $\tilde{E}_k$  estimates,

$$\frac{2M\tilde{E}_k}{\chi_{2M}^2(\alpha/2)} \leq E_k \leq \frac{2M\tilde{E}_k}{\chi_{2M}^2(1-\alpha/2)}. \quad (2.26)$$

where  $\alpha$  is the probability that the true value is not within the interval. The use of a window function  $f(t)$  results in a reduction in the number of degrees of freedom, which must be taken into account in the previous calculations. In this case,  $M$  should be replaced by,

$$M_a \approx \frac{M}{T} \int_{-T/2}^{+T/2} f(t/T) dt \quad (2.27)$$

In the case of the [Hamming](#) window, this reduction is approximately 50%\*.

### 3 Autoregressive model analysis

We have seen that a signal consisting of a sum of harmonic functions satisfies a recursive formula where the coefficients  $b_m$  determine the spectrum. We will now examine some of the ways to estimate the autoregressive coefficients for spectral analysis. There are several possible approaches, generally named after their developers. For example, the [Pisarenko](#) model, which is a sum of sinusoids, can be considered an autoregressive model. This is what Pisarenko chose to do, using the method of least squares to determine the model parameters. [Burg](#), on the other hand, takes a different approach and chooses to fit the autoregressive parameters by maximising the entropy of the discrepancy between the data and the signal reconstructed by the autoregressive model.

---

\*Some authors suggest overlapping the signal segments by the same proportion to preserve all the initial information.

#### 3.1 The prediction error filter

In practice, the estimation of the parameters  $b_m$  of the autoregressive model involves the use of a quality criterion for the fit, which may involve a number of *a priori* constraints on the nature of the signal being analysed. The criterion used by Burg involves minimising the total energy of the prediction error, defined by

$$e_n \equiv s_n + \sum_{m=1}^M b_m s_{n-m} \quad (3.1)$$

that is, to make,

$$\sum_n e_n^2 \text{ MINIMUM.} \quad (3.2)$$

The expression for the prediction error can be rewritten in the form of a convolution,

$$\{e_n\} = \{s_n\} * \{1, b_1, b_2, \dots, b_M\} \quad (3.3)$$

where the causal filter appears,

$$fep \equiv \{1, b_1, b_2, \dots, b_M\} \quad (3.4)$$

is called the prediction error filter. The coefficients  $b_m$  that minimise the energy of the prediction error are such that,

$$\begin{aligned} 0 &= \frac{1}{2} \frac{\partial}{\partial b_m} \sum_n e_n^2 \\ &= \sum_n e_n \frac{\partial e_n}{\partial b_m} \\ &= \sum_n e_n s_{n-m} \end{aligned} \quad (3.5)$$

If the number  $M$  of autoregressive coefficients is unlimited, a simple change of variable allows us to rewrite the last line in the form,

$$\sum_n e_{n+k} s_{n-l} = 0 \quad (k > 0, l \geq 0), \quad (3.6)$$

which is still true after multiplication by a constant,

$$\sum_n e_{n+k} b_l s_{n-l} = 0 \quad (k > 0, l \geq 0). \quad (3.7)$$

Of course, the sum of such expressions remains equal to zero, and in particular, we have the following,

$$\sum_n e_{n+k} \sum_{l \in \mathbb{N}} b_l s_{n-l} = 0 \quad (k > 0) \quad (3.8)$$

which can be simplified using the definition of the prediction error itself, to find that the autocorrelation

$$\sum_n e_{n+k} e_n = r_{e,e}(k > 0) = 0. \quad (3.9)$$

Since the autocorrelation is a symmetric function, we can modify the condition on  $k$  to obtain,

$$r_{e,e}(k \neq 0) = 0 \quad (3.10)$$

Cette expression montre que,

The autocorrelation function of the prediction error produced by an infinite duration prediction error filter is that of white noise.

### 3.2 Prediction error filter utility

The prediction error filter has the ability to transform a signal,  $s_n$  into white noise,  $e_n$ . In [Fourier](#) space, this is expressed by the relation,

$$\begin{aligned} S(Z) \times FEP(Z) &= E(Z) \\ &= \sigma_e \end{aligned} \quad (3.11)$$

where  $\sigma_e^2$  is the energy of the white noise  $e_n$ . This relationship allows us to obtain the [Fourier](#)

### 3. AUTOREGRESSIVE MODEL ANALYSIS

---

transform of the signal  $s_n$ .

$$\begin{aligned} S(Z) &= \frac{\sigma_e}{FEP(Z)} \\ &= \frac{\sigma_e}{1 + b_1Z + b_2Z^2 + \dots} \end{aligned} \tag{3.12}$$

In practice, the spectral division above is very unstable and generally yields poor results. Stabilization can be achieved by replacing the filter  $\{1, b_1, b_2, \dots\}$  with its associated minimum-phase filter. By doing so, the phases are destroyed, and it is only possible to recover the amplitude spectrum of the signal,

$$|S(Z)| = \frac{\sigma_e}{|\mathcal{D.M}\{1 + b_1Z + b_2Z^2 + \dots\}|} \tag{3.13}$$



---

---

# CHAPTER 12

---

## WAVELET TRANSFORM ANALYSIS

<b>1</b>	<b>Wavelets: A brief history</b>	<b>177</b>
1.1	Recent history	177
1.2	From Joseph Fourier to Dennis Gabor	178
1.3	From Dennis Gabor to Jean Morlet	181
1.4	Questions addressed in this chapter	183
<b>2</b>	<b>Continuous Wavelets – Discrete Wavelets – Orthogonal Wavelets</b>	<b>184</b>
2.1	Continuous Wavelet Transform	184
2.2	Orthogonal Wavelets	185
<b>3</b>	<b>How is the Orthogonal Wavelet Transform computed?</b>	<b>187</b>
3.1	The pyramid algorithm	187
3.2	Quadrature Mirror Filters	190
3.3	The inverse transform	190
<b>4</b>	<b>Filter, denoise and compress signals using orthogonal wavelets</b>	<b>191</b>
<b>5</b>	<b>How do you filter with the continuous wavelet transform ?</b>	<b>193</b>
5.1	The Reconstruction Formula	193
5.2	The reproducing kernel	194



---

<b>6</b>	<b>Asymptotic signal analysis</b>	<b>195</b>
6.1	Signaux asymptotiques	195
6.2	Asymptotic wavelet analysis	196
6.3	The stationary phase method	198
6.4	The Wavelet Transform Ridge	199
6.5	Use of non-asymptotic wavelets	200

# 1 Wavelets: A brief history

## 1.1 Recent history

Wavelet analysis emerged in the early 1980s and was the subject of significant mathematical research for about a decade. Following this period of emergence, wavelet analysis methods have been fundamental to numerous applications in fields as diverse as geophysics, medical imaging, astrophysics, data compression, etc. Today, theoretical work continues and is published in particular in the journal *Applied and Computational Harmonic Analysis*. A common feature of various wavelet techniques is the analysis of signals with fluctuations over a wide range of spatial or temporal scales. This analysis is performed via decompositions based on families of functions, which have the remarkable property of being derived by dilating a base function - the analysing wavelet - in such a way that all functions in a given family have the same shape. Depending on the analysing wavelet chosen, the resulting wavelet family may be orthogonal or non-orthogonal, with mathematical properties more or less appropriate to the signals being analysed.

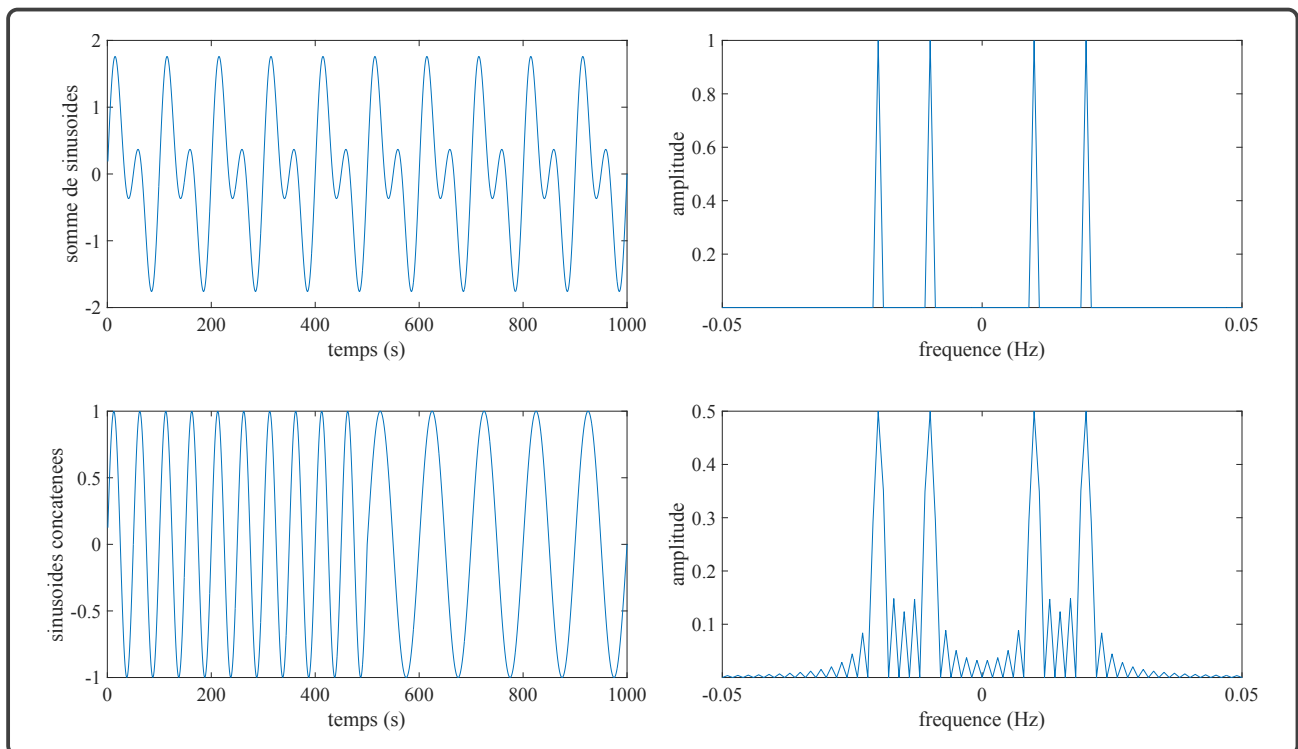
In retrospect, it has been recognised that the concept of wavelets with a constant shape was introduced by the Hungarian mathematician [Alfréd Haar](#) in the early 20th century ([Haar, 1909](#)). However, [Haar](#)'s orthogonal wavelets were not the starting point for wavelet theory in its current form. It was the work of [Jean Morlet](#) in the early 1980s that really launched the field. The wavelets proposed by [Jean Morlet](#) are non-orthogonal and are a fairly direct adaptation of Fourier analysis by segments ([Morlet et al., 1982](#)). However, it was in fact the concept of constant-shape wavelets, introduced somewhat empirically by Morlet, that served as the basis, in particular thanks to a highly theoretical paper by [Alex Grossmann](#) and [Jean Morlet](#) entitled "Decomposition of Hardy function into square-integrable wavelets of constant shape", published in an applied mathematics journal ([Grossmann et Morlet, 1984](#)). The generalisation of Morlet's wavelet transform laid the foundation for continuous wavelet transform.

A little later, the orthogonal wavelet transform was developed under the direction of [Yves Meyer](#), who was then a professor at the Centre de Recherche en Mathématiques de la Décision ([CEREMADE](#)) at the University of Paris Dauphine. The collective volume "[Fundamental Papers in Wavelet Theory](#)", published in 2006, provides an insight into the emergence of wavelet theory and shows that several fundamental foundations had already been established for some time, although they had not yet triggered the synthesis work of the 1980s ([Heil et al., 2006](#)). As is often the case in research, serendipity played a role in the history of wavelets when [Yves Meyer](#) discovered the paper by [Grossmann](#) and [Morlet](#) while waiting his turn at the photocopier in his laboratory, leafing

through journals brought in by a colleague.

## 1.2 From Joseph Fourier to Dennis Gabor

We have already noted that non-stationary signals are very common in geophysics, and that a significant part of the information they contain is embedded precisely in this non-stationarity. The **Fourier** transform, by completely neglecting the time domain, is poorly suited to the analysis of non-stationary signals. This is illustrated in the figure (12.1) obtained with the code `ondelette01.m`.



**Figure 12.1:** Illustration of the inadequacy of the **Fourier** transform for non-stationary signals: a signal consisting of two successive sinusoids (bottom left) has an amplitude spectrum (bottom right) very similar to that of a signal consisting of the superposition of the two sinusoids (top left), despite their different temporal structures (top right).

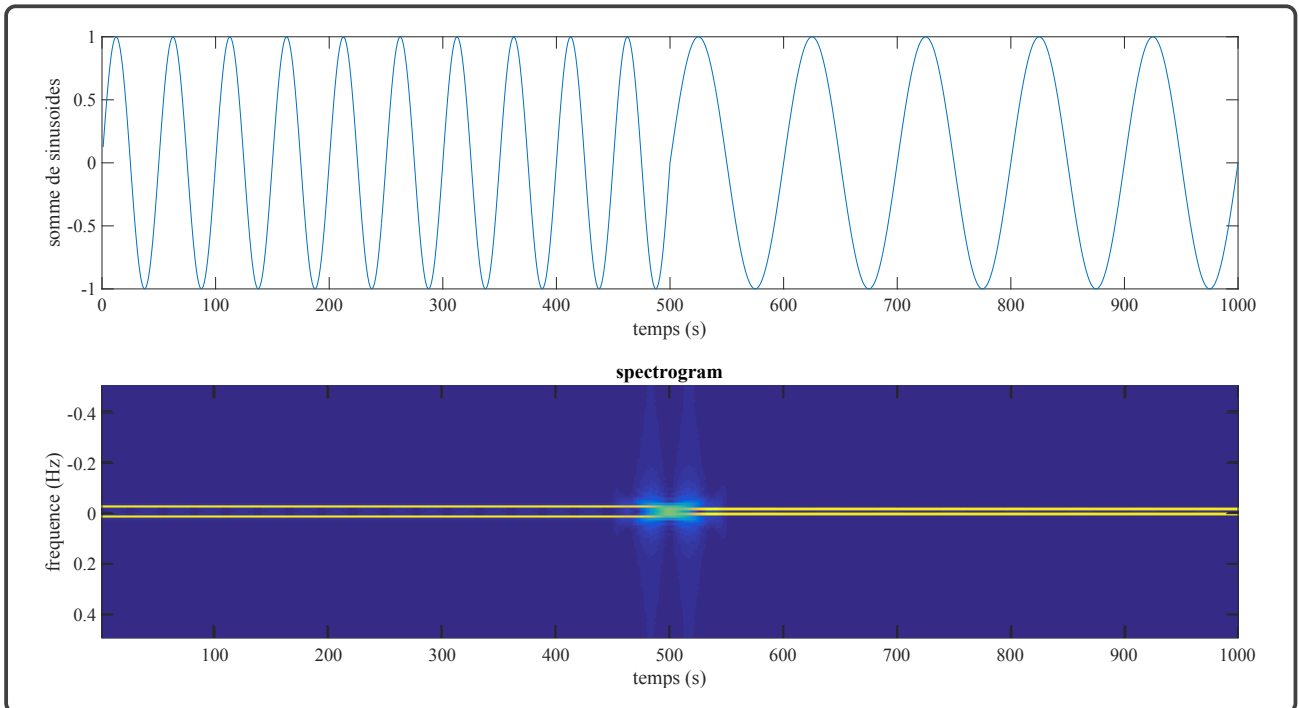
This figure (12.1) shows that the amplitude spectrum of a non-stationary signal composed of two successive sinusoids is little different from that of two superimposed sinusoids. In both cases the amplitude spectrum shows peaks at the frequencies of the sinusoids. The information about the transition from one sinusoid to another in the non-stationary signal is contained in the low amplitude peaks of the spectrum and in the phase of the **Fourier** transform. Therefore, information initially localised at a specific point on the time axis is dispersed in the frequency domain, making it difficult to retrieve. A simple solution to preserve, at least partially, the information about the transition from one sinusoid to another is to perform a **Fourier** analysis on successive segments of the signal. This

## 1. WAVELETS: A BRIEF HISTORY

was the idea of [Gabor](#) when calculating the spectrogram,

$$\mathcal{G}[w, f](u, t) = \int_{-\infty}^{+\infty} f(\tau) w(t - \tau) \exp[-2i\pi u(t - \tau)] d\tau. \quad (1.1)$$

The function  $w(t)$  is a window used to extract a segment of the signal. The code [ondelettes02.m](#) computes a simple spectrogram using a [Dirichlet](#) window to extract successive segments of the signal. The result is shown in figure (12.2), where it can be observed that the frequency and time information carried by the signal is partially recovered in the time-frequency plane representing the spectrogram



**Figure 12.2:** Example of the calculation of a simple spectrogram using a [Dirichlet](#) window to extract segments of the signal. The amplitude of the spectrogram is plotted in the time-frequency plane.

The spectrogram allows the time-frequency analysis of a signal, for example by displaying its energy  $|F(u, t)|^2$ . The choice of the window function  $w(t)$  is, *a priori*, quite flexible, but it is advantageous for this window to be optimal with respect to [Heisenberg](#)'s uncertainty principle\*. For this reason, [Gabor](#) chose the Gaussian window, which leads to the following expression for the spectrogram,

$$\mathcal{G}\left[\exp\left(-\frac{\pi t^2}{T^2}\right), f\right](u, t) = \int_{-\infty}^{+\infty} f(\tau) \exp\left(-\frac{\pi(t - \tau)^2}{T^2}\right) \exp[-2i\pi u(t - \tau)] d\tau. \quad (1.2)$$

\*See the chapter on [Time-Frequency Duality](#) for more details on the uncertainty principle.

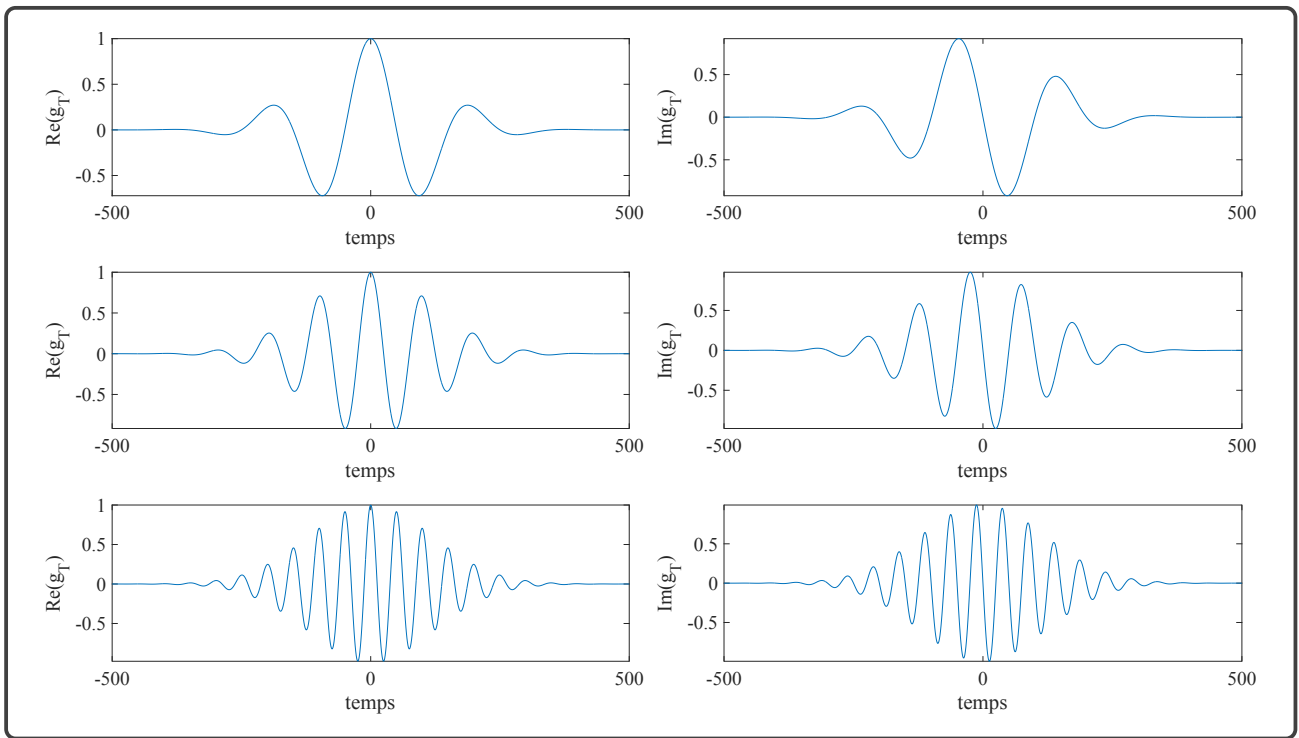
By defining the analysis function  $g_T$  as a\*,

$$g_T(u, t) \equiv \exp\left(-\frac{\pi t^2}{T^2}\right) \exp(-2i\pi ut), \quad (1.3)$$

it can be seen that the spectrogram can be rewritten in the form of a convolution product,

$$\mathcal{G}[g_T, f](u, t) = [g_T(u, \cdot) * f(\cdot)](t). \quad (1.4)$$

The code [ondelettes03.m](#) allows you to calculate the function  $g_T$ , with some examples shown in the figure (12.3)



**Figure 12.3: Analysis function  $g_T$  for three different frequencies.**

The analysis function is parameterised by the duration  $T$ , which defines the width of the window. The Gaussian in the time domain corresponds to another Gaussian in the frequency domain, and these two functions determine the time and frequency resolutions,  $\delta t$  and  $\delta u$ , that satisfy,

$$\delta t \times \delta u = \frac{1}{4\pi}, \quad (1.5)$$

\*sometimes referred to as the "gaborrette" in french

and remain constant over the whole of the  $(u, t)$  plane:

$$\delta t = \frac{T}{2\sqrt{\pi}} \text{ et } \delta u = \frac{1}{2T\sqrt{\pi}}. \quad (1.6)$$

### 1.3 From Dennis Gabor to Jean Morlet

It was in the early 1980s that a significant modification of Gabor's spectrogram was proposed by Jean Morlet, leading to the development of the wavelet transform. The modification consisted in adjusting the duration  $T$  of the window according to the frequency  $u$ . Jean Morlet chose the following setting,

$$T = \frac{\sqrt{\alpha}}{u}, \quad (1.7)$$

where  $\alpha$  is a parameter whose meaning will be discussed later. Using this new definition of the window duration, the analysis function of Gabor becomes,

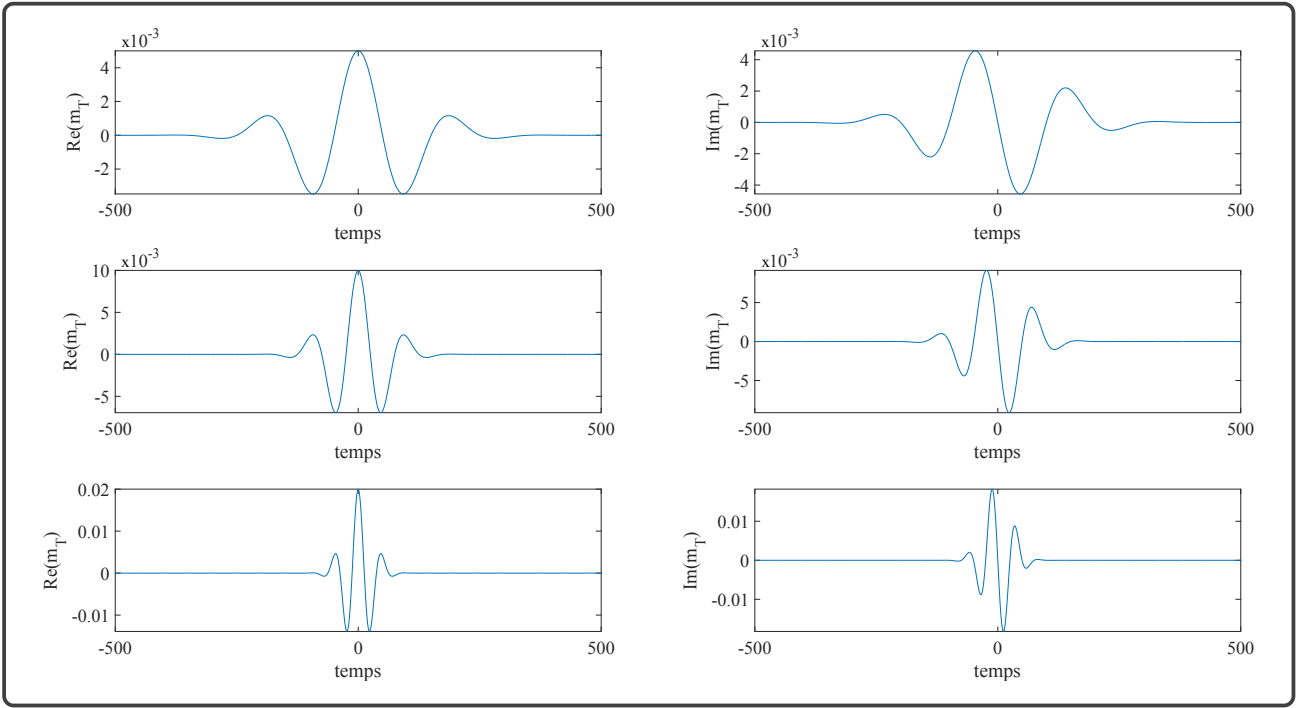
$$m_{\alpha}(u, t) = \exp\left[-\frac{\pi}{\alpha}(ut)^2\right] \exp(-2i\pi ut). \quad (1.8)$$

By performing the variable change  $u \mapsto a^{-1}$ , which introduces the dilation  $a$ , we obtain the classical expression of the normalised Morlet wavelet,

$$m_{\alpha}\left(\frac{t}{a}\right) = \frac{1}{a} \exp\left[-\frac{\pi}{\alpha}\left(\frac{t}{a}\right)^2\right] \exp\left(-\frac{2i\pi t}{a}\right). \quad (1.9)$$

The  $\alpha$  parameter allows you to adjust the ratio between the width of the Gaussian envelope and the dominant period of the wavelet, which in the case of the Morlet wavelet is  $1/a$ . For reasons we will discuss later, it is necessary that  $\alpha > 2$  for the wavelet to be considered as having zero mean. The larger  $\alpha$ , the better the frequency resolution  $\delta u$ , but at the expense of the time resolution  $\delta t$ .

The code [ondelettes04.m](#) calculates the function  $m_{\alpha}(t/a)$ , with some examples shown in figure (12.4). A comparison with figure (12.3) clearly illustrates the fundamental property of wavelets: their constant shape. All wavelets in the same family are obtained by dilating a single analysing wavelet. This property is the basis of all wavelet transforms: continuous, discrete, orthogonal, *etc*



**Figure 12.4: Morlet wavelet for three different dilations  $a$**

The **Gabor** spectrogram thus becomes the **Morlet** wavelet transform,

$$\mathcal{W} [m_\alpha, f] (a, t) = \frac{1}{a} \left[ m_\alpha \left( \frac{t}{a} \right) * f(t) \right] (t). \quad (1.10)$$

Since we are generally interested in real signals, the symmetry properties of the **Fourier** transform imply that it is sufficient to compute the **Gabor** spectrogram for  $u \geq 0$ , *ie*  $a > 0$ .

The introduction of the dilation parameter  $a$  significantly changes the properties of the wavelet transform compared to those of the **Gabor** spectrogram. In particular, the wavelet transform adapts well to non-stationarities because, whatever the time constant of a sudden change in the signal, there will always be wavelets of appropriate size to localise this change. This is due to the fact that the time resolution of the **Morlet** wavelet transform is given by,

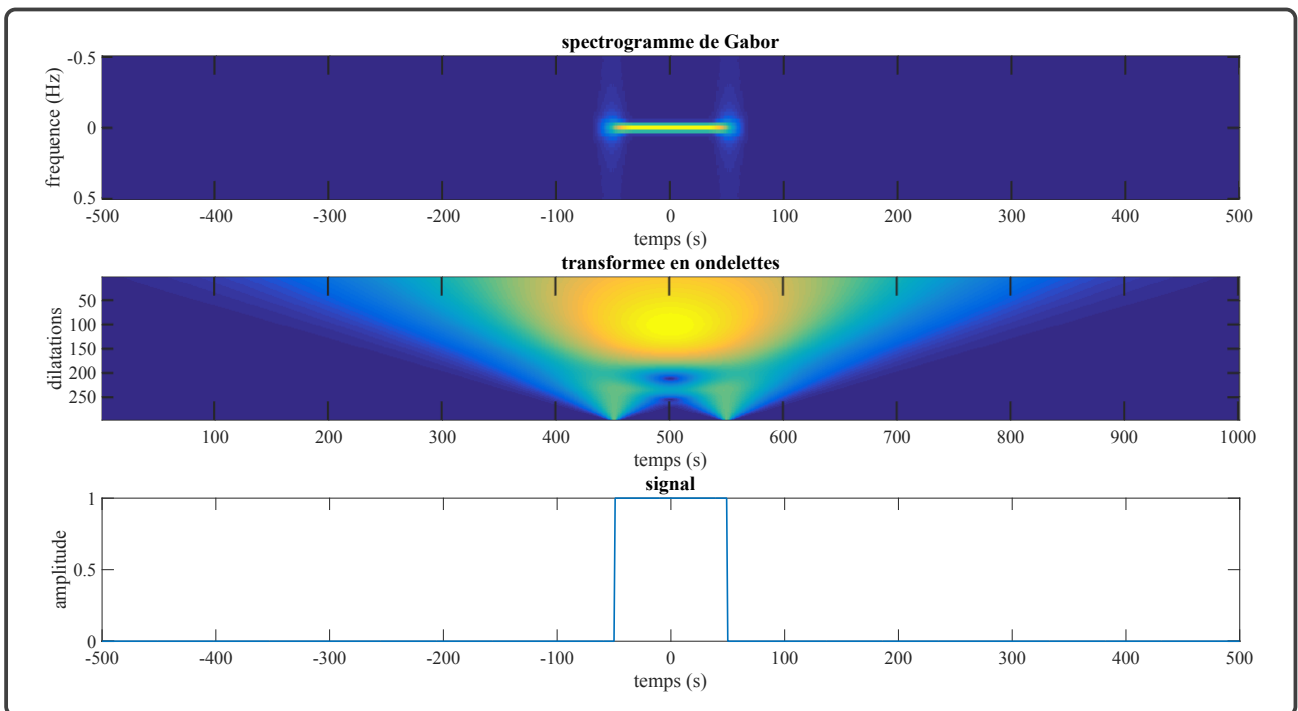
$$\delta t = \frac{a}{2\sqrt{\alpha\pi}}, \quad (1.11)$$

and is therefore not constant in the half-plane\* ( $a > 0, t$ ). Of course, in accordance with the uncertainty principle mentioned earlier, the frequency resolution varies inversely with dilation.

\*This is called the **Poincaré** half-plane.

$$\delta u = \frac{1}{2a} \sqrt{\frac{\alpha}{\pi}}. \tag{1.12}$$

This ability of the wavelet transform to adapt to the finest details of a signal has earned it the nickname 'the mathematical microscope'. This property is illustrated by the following code, which computes both the spectrogram and the wavelet transform of a Dirichlet window. The result is shown in figure (12.5), obtained using [ondelettes05.m](#).



**Figure 12.5:** The magnitude of the spectrogram (top) and the Morlet wavelet transform (middle) of a rectangular pulse.

Figure (12.5) effectively illustrates the multi-scale analysis capabilities of the wavelet transform. Wavelets with small dilation focus on the discontinuities in the signal, while wavelets with dilation matched to the width of the window correspond to a maximum amplitude in the wavelet transform.

### 1.4 Questions addressed in this chapter

Wavelet analysis has become an important field in mathematical analysis as well as in signal and image processing. Based on a strong theoretical framework, wavelet methods are used in numerous applications thanks to readily available algorithms, the most famous of which are those developed by



---

[Patrick Flandrin](#) and his colleagues\* and those from the Statistics Department of Stanford University†. These software tools will be very useful complements to the functions developed in this course.

In the remainder of this extensive chapter, we will focus specifically on the use of wavelets for signal analysis. We will explore how it is possible to [teach physics to wavelets](#), so that they allow us to extract certain information about physical systems or phenomena. For reasons that will become clear later, it is primarily the continuous wavelet transform, obtained by generalising the equation 1.10, that will enable us to achieve our goals. Therefore, in contrast to most texts, we will only moderately cover the topic of orthogonal wavelets. Due to space limitations in this short introduction, our discussion will primarily be of one-dimensional (1D) wavelets,

- \* Non-orthogonal wavelets are functions that can be chosen with considerable flexibility, allowing them to be tailored to the physical characteristics of the signals being analysed.
- \* The continuous wavelet transform allows the wavelets to be precisely localised on the events that make up the signals being analysed.
- \* The theory of the continuous wavelet transform is straightforward, and its integration into physical theories such as potential theory, wave phenomena, *etc* is more feasible than with orthogonal wavelets.

## 2 Continuous Wavelets — Discrete Wavelets — Orthogonal Wavelets

Before looking at specific aspects of wavelet analysis, we will first establish some basic principles that characterise the two main families of wavelets: continuous wavelets and orthogonal wavelets.

### 2.1 Continuous Wavelet Transform

The continuous wavelet transform is easily obtained by generalizing the [Morlet](#) wavelet transform. For reasons that will become clearer later, we choose to define the continuous wavelet transform as a convolution product,

$$\mathcal{W}[\psi, f](a, t) \equiv [f(\cdot) * \psi_a(\cdot)](t), \quad (2.1)$$

---

\*<http://perso.ens-lyon.fr/patrick.flandrin/software2.html>

†<http://www-stat.stanford.edu/wavelab/>

where the wavelet is such that,

$$\psi_a(t) \equiv \frac{1}{a} \psi\left(\frac{t}{a}\right). \quad (2.2)$$

The scale parameter  $a > 0$ , also known as the dilation, affects the analysing wavelet  $\psi(t)$  by stretching if  $a > 1$  or compressing if  $a < 1$ .

As defined above, the continuous wavelet transform is a bank of filters applied to the signal  $f$ . Since wavelets are obtained by dilation, their Fourier transforms, which are the corresponding filters, are also a family of functions generated by dilation. We will see later that the choice of wavelet is quite flexible, which allows us to give the wavelet transform special properties, including giving it physical meaning. In fact, the primary condition that a wavelet must satisfy is the admissibility condition,

$$\int_0^{+\infty} |\Psi(u)|^2 \frac{du}{u} < \infty, \quad (2.3)$$

which requires the wavelet to have a zero mean. We will discuss later that this condition is necessary to establish the reconstruction formula corresponding to the inverse wavelet transform

## 2.2 Orthogonal Wavelets

In the modern history of wavelets, orthogonal wavelets were not discovered immediately after the introduction of the continuous wavelet transform *etc* although Haar wavelets, discovered in the early 20th century, are indeed orthogonal! It is also interesting to note that Haar wavelets were used to filter signals in the 1970s (Gubbins, 1971)\*, well before the advent of wavelet theory. It is these wavelets that we will use as an example to introduce orthogonal wavelets and their main properties.

Orthogonality requires a scalar product, which we will define here as,

$$\mathcal{W}[\psi, f](a, t) \equiv [f(\cdot) * \psi_a(\cdot)](t). \quad (2.4)$$

The Haar wavelets are constructed from the function consisting of a positive rectangular window

---

\*Gubbins, D., 'Two dimensional digital filtering with Haar and Walsh transforms', *Annales de Géophysique*, 27, 85-104, 1971.

followed by a negative one,

$$\psi_H(t) = \Pi\left(t + \frac{1}{2}\right) - \Pi\left(t - \frac{1}{2}\right). \quad (2.5)$$

These functions have compact support, and an initial subset of orthogonal functions is trivially obtained by keeping only those functions whose supports are disjoint while densely covering  $\mathbb{R}$ .

$$\mathcal{H}_1 = \{\psi_H(t - 2m) \mid m \in \mathbb{Z}\}. \quad (2.6)$$

A second subset of functions which are orthogonal to each other and also orthogonal to the family  $\mathcal{H}_1$  is formed by dilating the functions in  $\mathcal{H}_1$  by a factor of  $a = 2$ ,

$$\mathcal{H}_2 = \{\psi_H\left(\frac{t}{2} - 2m\right) \mid m \in \mathbb{Z}\}. \quad (2.7)$$

The **Haar** basis is obtained by iterating this process,

$$\mathcal{H} = \bigcup_{n \in \mathcal{Z}} \mathcal{H}_{2^n}. \quad (2.8)$$

The example of **Haar** wavelets shows that orthogonality is achieved if the dilations are powers of 2. This is why the term 'octave' is often used in wavelet theory terminology. The fact that the allowed dilations are powers of 2 is a rather fundamental property that holds for most orthogonal wavelets in use. However, it is not an absolutely necessary property, since in general orthogonality can be satisfied if the dilation is given by,

$$a = q^n \text{ avec } q \in \mathbb{P}. \quad (2.9)$$

Another important property highlighted by the example of Haar wavelets is that orthogonality requires the wavelets to be translated according to a dyadic tiling when  $a = 2^n$ , triadic for  $a = 3^n$ , and so on. This constraint is of particular practical importance because it implies that the orthogonal wavelet transform is not invariant under translation. This can cause serious problems in signal analysis, since adding or removing a few values at the beginning of a signal can significantly alter its orthogonal wavelet transform.

### 3. HOW IS THE ORTHOGONAL WAVELET TRANSFORM COMPUTED?

The code `ondelettes06.m` calculates the functions of the Haar basis, as shown in the figure (12.6).

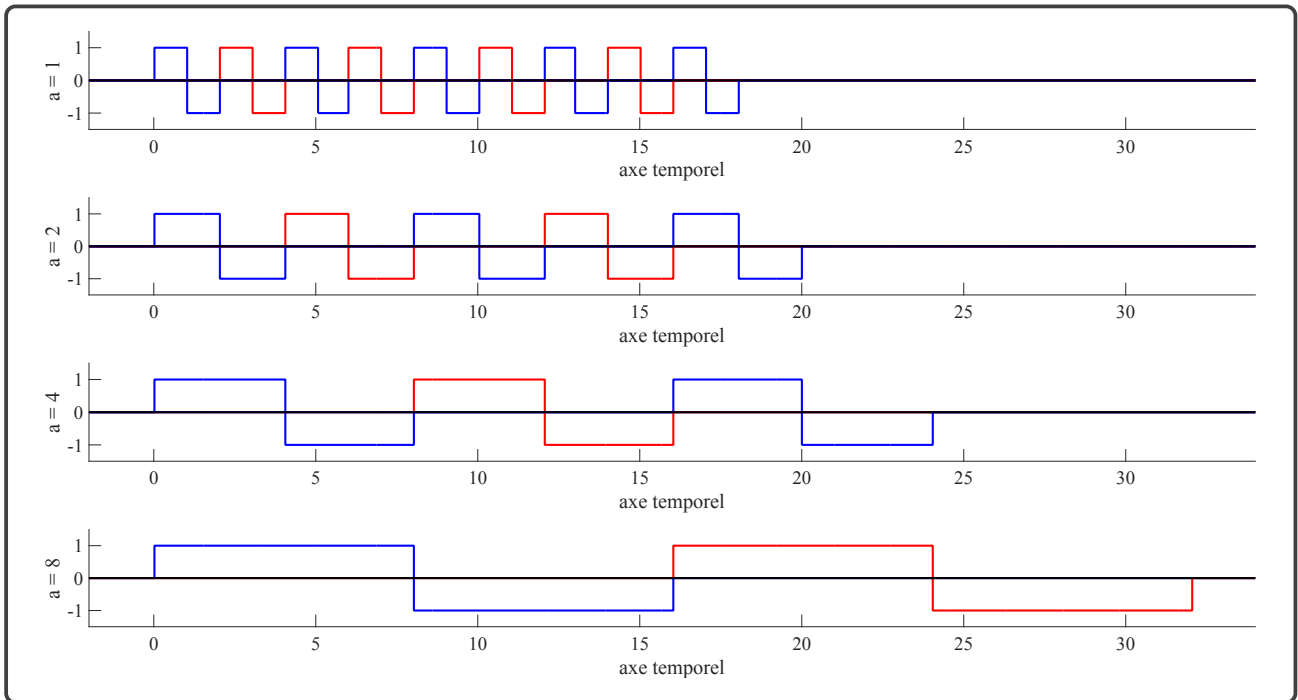


Figure 12.6: Some functions of the Haar basis for  $a = 1, 2, 4,$  and  $8$

## 3 How is the Orthogonal Wavelet Transform computed?

### 3.1 The pyramid algorithm

The very particular construction of the orthogonal wavelet transform, namely the dyadic sampling and octave discretisation of the dilations, allows a fast computation of the wavelet coefficients thanks to an algorithm proposed by Mallat (1989) and inspired by the pyramid algorithm developed in the 1970s. To understand the principle of this algorithm, let us start with the discrete version of the wavelet coefficients for the minimum dilation  $a_0 = 2^0$ ,

$$\begin{aligned}
 \mathbf{W}_0 &= \{f_1 - f_0, f_3 - f_2, f_5 - f_4 \dots\} \\
 &= \underset{\sim}{\cap}_1^2 \{f_1 - f_0, f_2 - f_1, f_3 - f_2, f_4 - f_3, f_5 - f_4 \dots\} \\
 &= \underset{\sim}{\cap}_1^2 \{+1, -1\} * \{f_0, f_1, f_2, f_3, f_4, f_5 \dots\} \\
 &= \underset{\sim}{\cap}_1^2 \{+1, -1\} * \mathbf{f}_0,
 \end{aligned} \tag{3.1}$$

where  $\mathbf{f}_0$  represents the initial signal and where the operator  $\underset{\sim}{\cap}_1^2$  denotes subsampling by 2

such that,

$$\curvearrowright_1^2 \{0, 1, 2, 3, 4, 5, \dots\} = \{0, 2, 4, \dots\}. \quad (3.2)$$

Equation (3.1) shows that wavelet coefficients can be obtained by applying a high-pass filter to the signal and then removing every other value from the filtered signal. The high-pass filter is nothing other than the dilation wavelet  $a_0$ . Let us now consider the wavelet coefficients for the dilation  $a_1 = 2^1$ ,

$$\begin{aligned} \mathbf{W}_1 &= \{(f_3 + f_2) - (f_1 + f_0), (f_7 + f_6) - (f_5 + f_4), (f_{11} + f_{10}) - (f_9 + f_8) \dots\} \\ &= \curvearrowright_1^2 \{1, -1\} * \{f_1 + f_0, f_3 + f_2, f_5 + f_4, f_7 + f_6, f_9 + f_8, f_{11} + f_{10} \dots\} \\ &= \curvearrowright_1^2 \{1, -1\} * [ \curvearrowright_1^2 \{1, 1\} * \{f_0, f_1, f_2, f_3, f_4, f_5, f_6, \dots\} ] \\ &= \curvearrowright_1^2 \{1, -1\} * \mathbf{f}_1. \end{aligned} \quad (3.3)$$

Equation (3.3) shows that the wavelet coefficients for the dilation  $a_1$  are obtained by applying the low pass filter  $1, ; 1$  and a decimation by two to obtain the signal  $\mathbf{f}_1$ , followed by a high pass filter and another decimation by two to obtain the coefficients  $\mathbf{W}_1$ . It is easy to show that the subsequent wavelet coefficients are obtained in the same way. This results in the following cascade,

$$\begin{aligned} \curvearrowright_1^2 \{+1, -1\} * \mathbf{f}_0 &\mapsto \mathbf{W}_0 \\ \curvearrowright_1^2 \{+1, +1\} * \mathbf{f}_0 &\mapsto \mathbf{f}_1 \\ \curvearrowright_1^2 \{+1, -1\} * \mathbf{f}_1 &\mapsto \mathbf{W}_1 \\ \curvearrowright_1^2 \{+1, +1\} * \mathbf{f}_1 &\mapsto \mathbf{f}_2 \\ \curvearrowright_1^2 \{+1, -1\} * \mathbf{f}_2 &\mapsto \mathbf{W}_2 \\ \curvearrowright_1^2 \{+1, +1\} * \mathbf{f}_2 &\mapsto \mathbf{f}_3 \\ &\dots \end{aligned} \quad (3.4)$$

This cascade represents the Pyramid algorithm. Note that it assumes that the initial signal contains  $2^N$  values. For the decomposition to be complete, i.e. for all the information contained in the signal to be represented in the coefficients  $\mathbf{W}$ , it is necessary to include  $\mathbf{f}_N$ , which is simply the sum of the signal values. The following code calculates the wavelet coefficients of a signal in the Haar basis,

### 3. HOW IS THE ORTHOGONAL WAVELET TRANSFORM COMPUTED?

```
function w = DirectHaar(f)
    hf = [1 -1]/sqrt(2);           % Haar wavelet = high-pass filter
    lf = [1 1]/sqrt(2);           % low-pass mirror filter
    f = f(:); nf = length(f); meanf = mean(f); w = [];
    while nf > 1
        wa = conv(f,hf);           % convolve signal with high-pass filter
        wa = wa(2:2:end);          % decimate to get wavelet coefficients
        f = conv(f,lf);           % convolve signal with low-pass filter
        f = f(2:2:end);           % decimate
        w = [w wa'];              % merge wavelet coefficients
        nf = nf/2;                 %(slow manner !)
    end
    w = [w meanf];
end
```

An example application is shown below,

```
function ondelettes07()
    close all; clc; home;
    nf = 16;
    f = randi(10,1,nf);
    disp([input signal: num2str(f)]);
    w = DirectHaar(f);
    nw = length(w);
    a = 1;                          % dilatation
    disp(-----);

    while nw > 1
        nw = nw/2;
        disp([dilatation a = num2str(a)]);
        disp([Haar coef: num2str(w(1:nw))]);
        disp(-----);
        w = w(nw+1:end);
        a = 2*a;
    end
end
```

```

end
disp(['last coefficient (mean of signal) = num2str(w(end))]);
end

```

### 3.2 Quadrature Mirror Filters

The example of the Haar wavelet decomposition illustrates a property that holds for all orthogonal wavelet bases, namely that the wavelet coefficients are obtained by the iterative application of two filters, a high-pass and a low-pass. These two filters completely define the wavelet basis and are clearly not arbitrary with respect to each other. In fact, it is necessary for the information filtered by the high-pass filter to be exactly complementary to the information filtered by the low-pass filter. Two filters with this property are called [quadrature mirror filters](#).

### 3.3 The inverse transform

Let us now see how to reconstruct the signal  $f_0$  from its Haar coefficients  $W$ . The following function reconstructs a signal from its Haar coefficients

```

function f = InverseHaar(w)

hf = [-1 1]/sqrt(2);
w = w(:);
f = w(end)*ones(size(w));
nw = length(w)/2;
H = repmat(hf,nw,length(f)/2/nw);
while nw >= 1
    W = repmat(w(1:nw),length(w)/2/nw,length(f)/nw);
    s = H' .* W';
    f = f + s(:);
    w = w(nw+1:end);
    nw = nw/2;
    if nw >= 1
        H = reshape(H,nw,length(f)/nw)/sqrt(2);
    end
end
end

```

```
end
```

## 4 Filter, denoise and compress signals using orthogonal wavelets

Filtering signals using orthogonal wavelet bases is done in the same way as other decompositions: you change the values of the wavelet coefficients and then calculate the inverse transform to reconstruct the filtered signal. It is interesting to note that the 'brutal' zeroing of certain wavelet coefficients does not produce [Gibbs](#) oscillations, unlike filtering in the [Fourier](#) basis.

The filtering performed in the following code shows an example of denoising a sinusoidal signal with a variable period. [Figure 12.7](#) shows the result. In this example, the filtering is performed by calculating the cumulative energy of the wavelet coefficients and removing those whose cumulative energy contributes less than 1% of the total energy. It is interesting to note that this filtering removes about 85% of the coefficients, which allows a significant compression of the information.

```
function ondelettes08()

    t = 1:512;
    T = linspace(20,40,length(t));
    f = sin(2*pi*t./T);
    fn = f + 10*(rand(size(f))-0.5);
    w = DirectHaar(f);
    [ws,iw] = sort(w.^2);

    nwi = length(find(w));
    cutlimit = 0.01;
    iwzero = iw(ws <= ws(end)*cutlimit);
    w(iwzero) = 0;
    nwf = length(find(w));
    disp(['ratio of compression = ' num2str(nwi/nwf)])
    ff = InverseHaar(w);

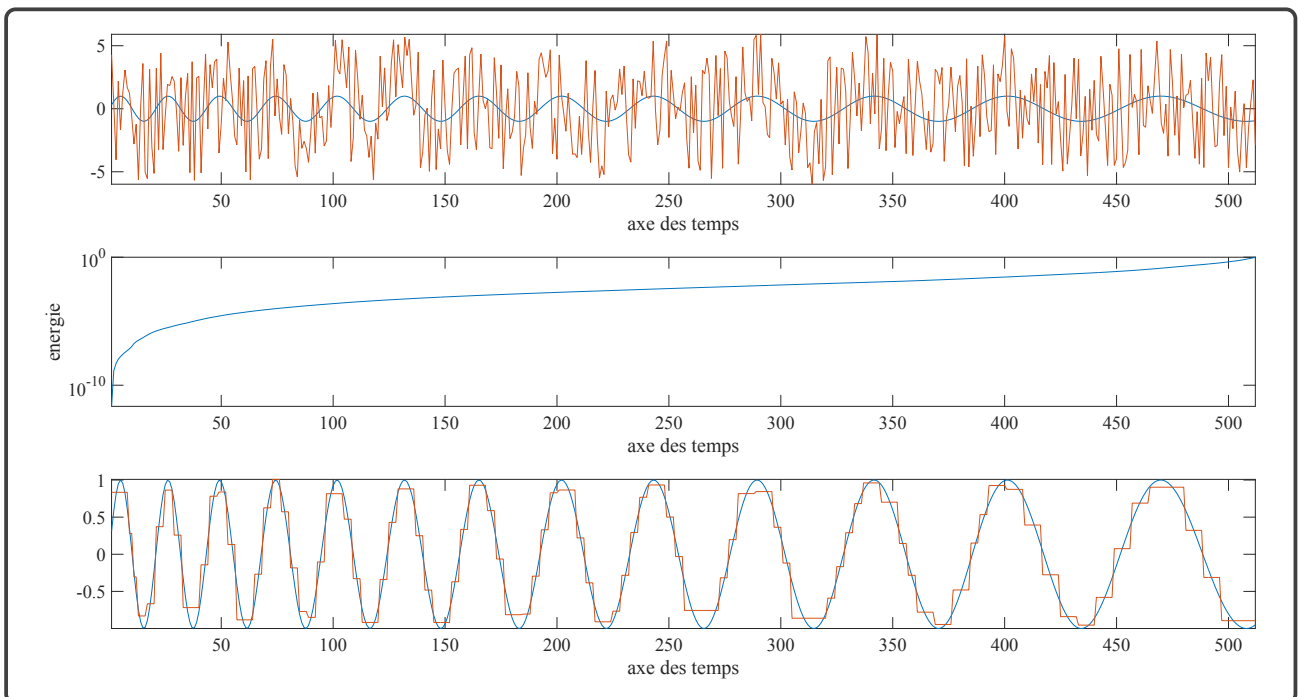
    figure
    subplot(311)
        plot(t,f,t,fn);axis tight;xlabel('time axis')
```



```

set(gca,'Fontname','Times New Roman','FontSize',18) ;
subplot(312)
semilogy(cumsum(ws)/max(cumsum(ws))) ;
axis tight;xlabel('time axis')
ylabel('energy');
set(gca,'Fontname','Times New Roman','FontSize',18) ;
subplot(313)
plot(t,f,t,ff);axis tight;xlabel('time axis')
set(gca,'Fontname','Times New Roman','FontSize',18) ;
end

```



**Figure 12.7:** Example of filtering in the Haar basis. At the top, the desired signal (in blue) and its noisy version (in green). In the middle, the cumulative energy of the wavelet coefficients. At the bottom, the reconstructed signal retaining the 15% most energetic coefficients

## 5 How do you filter with the continuous wavelet transform ?

### 5.1 The Reconstruction Formula

We will look for a reconstruction wavelet,  $\chi(t)$ , that allows us to reconstruct the signal  $f(t)$  from its transform  $\mathcal{W}[\psi, f](a, t)$ . Using a reconstruction formula of the form

$$\begin{aligned} f(t) &= \int_0^{+\infty} \mathcal{W}[\psi, f](a, t) * \chi_a(t) da \\ &= \int_0^{+\infty} f(t) * \psi_a(t) * \chi_a(t) da, \end{aligned} \quad (5.1)$$

which, after Fourier transformation, becomes,

$$F(u) = \int_0^{+\infty} F(u) \Psi(au) Q(au) da, \quad (5.2)$$

we obtain the following condition,

$$\int_0^{+\infty} \Psi(au) Q(au) da = 1. \quad (5.3)$$

This equation has a solution,

$$Q(bu) = \frac{\Psi^*(bu)}{\int_0^{+\infty} |\Psi(au)|^2 da} \quad \forall b > 0 \quad (5.4)$$

which, to be acceptable, requires,

$$0 < \int_0^{+\infty} |\Psi(au)|^2 da < \infty. \quad (5.5)$$

By setting  $v = au$ , this expression becomes

$$0 < a \int_0^{+\infty} |\Psi(v)|^2 \frac{dv}{v} < \infty. \quad (5.6)$$

Since  $a > 0$ , it can be eliminated from the above inequalities without changing the direction of

the inequalities, giving the admissibility condition in its standard form,

$$0 < C_\Psi \equiv \int_0^{+\infty} |\Psi(v)|^2 \frac{dv}{v} < \infty. \quad (5.7)$$

Taking advantage of the fact that,

$$\mathcal{F}[f^*(-t)](u) = F^*(u) \quad (5.8)$$

we obtain the expression for the reconstruction wavelet

$$\chi_a(t) = \frac{\Psi_a^*(-t)}{aC_\Psi}, \quad (5.9)$$

and the continuous reconstruction formula,

$$f(t) = \frac{1}{C_\Psi} \int_0^{+\infty} \frac{da}{a} \int_{-\infty}^{+\infty} \mathcal{W}[\Psi, f](a, \tau) \Psi_a^*(t - \tau) d\tau \quad (5.10)$$

## 5.2 The reproducing kernel

The continuous wavelet transform is complete when the entire frequency axis,  $u \in \mathbb{R}$ , is covered, *ie*, when

$$0 < \int_0^{+\infty} |\Psi(au)|^2 da < +\infty \quad \forall u \in \mathbb{R}, \quad (5.11)$$

which is automatically satisfied if the wavelet satisfies the admissibility condition discussed in the previous section. The family of wavelets,

$$\{\Psi_a(t - \tau), a \in \mathbb{R}^{+*}, \tau \in \mathbb{R}\} \quad (5.12)$$

is actually redundant, meaning that decomposing a signal over this family is redundant. As a result, the wavelet coefficients,  $\mathcal{W}[\Psi, f](a, t)$ , are correlated, which can be observed by reflexively

using the reconstruction formula,

$$\begin{aligned}
 \mathcal{W}[\psi, f](b, t) &= \mathcal{W} \left[ \psi, \int_0^{+\infty} \mathcal{W}[\psi, f](a, \cdot) * \chi_a(\cdot) da \right] (b, t) \\
 &= \int_0^{+\infty} \psi_b(\cdot) * \psi_a(\cdot) * f(\cdot) * \chi_a(\cdot) da \\
 &= \int_0^{+\infty} \mathcal{W}[\psi, f](a, \cdot) * [\psi_b(\cdot) * \chi_a(\cdot)] da \\
 &= \int_0^{+\infty} \mathcal{W}[\psi, f](a, t) * K_{b,a}(t) \frac{da}{a},
 \end{aligned} \tag{5.13}$$

where the reproducing kernel,  $K_{b,a}(t) \equiv a\psi_b(t) * \chi_a(t)$ , quantifies the redundancy of the wavelet transform. This kernel is fully defined by the analysing wavelet,

$$\begin{aligned}
 K_{b,a}(t) &= a\psi_b(t) * \chi_a(t) \\
 &= \frac{1}{C_\psi} \int_{-\infty}^{+\infty} \psi_b(\tau) \psi_a^*(\tau - t) d\tau
 \end{aligned} \tag{5.14}$$

## 6 Asymptotic signal analysis

We have just seen that the continuous wavelet transform is redundant because the information originally contained in the analysed signal is projected onto the [Poincaré](#) half-plane. We are moving from a one-dimensional space to a two-dimensional space, and it is interesting to investigate whether the projected information is 'uniformly' distributed or, conversely, 'concentrated' in preferred regions of the half-plane. This investigation can be done by noting that the wavelet transform is similar to a Fresnel-type oscillatory integral to which asymptotic approximations can be applied.

### 6.1 Signaux asymptotiques

A real signal,  $f(t)$ , can always be represented in terms of instantaneous amplitude and phase,

$$f(t) = A(t) \cos[\phi(t)]. \tag{6.1}$$

This representation admits an infinite number of solutions  $(A; \phi)$ , including the so-called canonical solution,

$$\begin{aligned}
 A_f(t) &= |Z_f(t)| \\
 \phi_f(t) &= \arg [Z_f(t)]
 \end{aligned} \tag{6.2}$$

---

where the analytic signal associated with  $f(t)$  is defined by,

$$Z_f(t) \equiv f(t) + iF_{Hi}(t) \quad (6.3)$$

The canonical solution allows for the definition of the instantaneous frequency,

$$u_f(t) = \frac{1}{2\pi} \frac{d\phi_f(t)}{dt}, \quad (6.4)$$

whose physical meaning can sometimes be confusing. This is particularly true when the signal is too slow or when the signal being analysed is the sum of two sinusoids.

We will say that a signal,

$$f(t) = A_f(t) \cos[\phi_f(t)] \quad (6.5)$$

is asymptotic if,

$$\left| \frac{d\phi_f}{dt} \right| \gg \left| \frac{1}{A_f} \frac{dA_f}{dt} \right|. \quad (6.6)$$

that is, the signal's oscillations are much faster than its envelope.

## 6.2 Asymptotic wavelet analysis

We will now focus on the wavelet transform of asymptotic signals when the analysing wavelet,  $\psi(t)$ , is itself asymptotic. We will see that the stationary phase method allows us to identify specific sets of points in the half-plane ( $a > 0, t$ ) from which we can obtain estimates of the wavelet coefficients and recover the modulation laws of the signals. Let  $f(t)$  be a locally monochromatic asymptotic real signal, with the associated analytic signal given by,

$$z_f(t) = A_f(t) \exp[i\phi_f(t)]. \quad (6.7)$$

Naturally,

$$f(t) = \text{Re}[z_f(t)]. \quad (6.8)$$

The wavelet transform of this signal is,

$$\begin{aligned} \mathcal{W}[\psi, f](a, t) &= \psi_a(t) * f(t) \\ &= \mathcal{F}^{-1}[F(u)\Psi(au)](t) \end{aligned} \quad (6.9)$$

If  $\psi(t)$  is an analytic wavelet, meaning that  $\Psi(u < 0) = 0$ , then, noting that,

$$\frac{1}{2}Z_f(u) = F(u) \text{ lorsque } u \geq 0, \quad (6.10)$$

one finds that,

$$F(u)\Psi(au) = \frac{1}{2}Z_f(u)\Psi(au) \quad (6.11)$$

The wavelet transform of the signal can therefore be expressed in terms of the wavelet transform of the corresponding analytical signal,

$$\begin{aligned} \mathcal{W}[\psi, f](a, t) &= \frac{1}{2}\mathcal{F}^{-1}[Z_f(u)\Psi(au)](t) \\ &= \frac{1}{2}\mathcal{W}[\psi, z_f](a, t) \end{aligned} \quad (6.12)$$

If the wavelet  $\psi(t)$  is an asymptotic wavelet, it can be written in the following form,

$$\psi(t) = A_\psi(t) \exp[i\phi_\psi(t)] \quad (6.13)$$

and  $\psi_a(t)$ ,

$$\psi_a(t) = \frac{1}{a}A_\psi\left(\frac{t}{a}\right) \exp\left[i\phi_\psi\left(\frac{t}{a}\right)\right] \quad (6.14)$$

Using this expression in the expression for the wavelet transform of the signal, we obtain,

$$\mathcal{W}[\psi, f](a, t) = \frac{1}{2a} \int_{-\infty}^{+\infty} A_f(\tau) A_\psi\left(\frac{t-\tau}{a}\right) \exp\left[i\phi_f(\tau) + i\phi_\psi\left(\frac{t-\tau}{a}\right)\right] d\tau \quad (6.15)$$

which is an oscillatory integral that can be approximated using the stationary phase method. This method exploits the fact that the integral takes most of its value around the points  $t_f(a, t)$  where,

$$\left. \frac{d}{d\tau} \left[ \phi_f(\tau) + \phi_\psi \left( \frac{t-\tau}{a} \right) \right] \right|_{\tau=t_f} = 0, \quad (6.16)$$

that is, where,

$$\phi_f'(t_f) = \frac{1}{a} \phi_\psi' \left( \frac{t-t_f}{a} \right). \quad (6.17)$$

The stationary phase approximation yields,

$$\begin{aligned} \mathcal{W}[\psi, f](a, t) &\simeq \frac{\sqrt{\pi} \exp \left\{ i \frac{\pi}{4} \operatorname{sgn} \left[ \phi_f''(t_f) + a^{-2} \phi_\psi'' \left( \frac{t-t_f}{a} \right) \right] \right\}}{a\sqrt{2} \sqrt{\left| \phi_f''(t_f) + a^{-2} \phi_\psi'' \left( \frac{t-t_f}{a} \right) \right|}} \\ &\quad \times A_f(t_f) A_\psi \left( \frac{t-t_f}{a} \right) \exp \left\{ i \left[ \phi_f(t_f) + \phi_\psi \left( \frac{t-t_f}{a} \right) \right] \right\} \\ &= \sqrt{\frac{\pi}{2}} \frac{\exp \left\{ i \frac{\pi}{4} \operatorname{sgn} \left[ \phi_f''(t_f) + a^{-2} \phi_\psi'' \left( \frac{t-t_f}{a} \right) \right] \right\}}{\sqrt{\left| \phi_f''(t_f) + a^{-2} \phi_\psi'' \left( \frac{t-t_f}{a} \right) \right|}} Z_f(t_f) \psi_a(t-t_f) \end{aligned} \quad (6.18)$$

### 6.3 The stationary phase method

This method was used by [Lord Kelvin](#) in 1887 to study integrals of the form,

$$I = \int_{-\infty}^{+\infty} A(t) \exp[i\phi(t)] dt \quad (6.19)$$

where  $A(t)$  and  $\phi(t)$  are regular functions. The idea behind this method is to exploit the fact that the integral takes most of its value near points where the phase  $\phi(t)$  is stationary. Suppose, without limiting the generality of our discussion, that there is only one point,  $t_0$ , for which this is true,

$$\phi'(t_0) = 0. \quad (6.20)$$

We have,

$$\begin{aligned}
 I &= \int_{-\infty}^{+\infty} A(t) \exp[i\phi(t)] dt \\
 &\simeq \int_{-\infty}^{+\infty} A(t) \exp\left[i\left(\phi(t_0) + \frac{1}{2}\phi''(t_0)(t-t_0)^2\right)\right] dt \\
 &\simeq A(t_0) \exp[i\phi(t_0)] \int_{-\infty}^{+\infty} \exp\left[\frac{i}{2}\phi''(t_0)(t-t_0)^2\right] dt \\
 &= A(t_0) \exp[i\phi(t_0)] \int_{-\infty}^{+\infty} \exp\left[\frac{i}{2}\phi''(t_0)\xi^2\right] d\xi
 \end{aligned} \tag{6.21}$$

Performing the change of variable,

$$\chi = \left| \frac{\phi''(t_0)}{2} \right|^{1/2} \xi \tag{6.22}$$

one deduces,

$$\begin{aligned}
 I &\simeq \frac{A(t_0) \exp[i\phi(t_0)]}{\sqrt{|\phi''(t_0)/2|}} \int_{-\infty}^{+\infty} \exp\left[i \operatorname{sgn}(\phi''(t_0)) \chi^2\right] d\chi \\
 &= \frac{2A(t_0) \exp[i\phi(t_0)]}{\sqrt{|\phi''(t_0)/2|}} \left\{ \int_0^{+\infty} \cos \chi^2 d\chi + i \operatorname{sgn}[\phi''(t_0)] \int_0^{+\infty} \sin \chi^2 d\chi \right\} \\
 &= \frac{2A(t_0) \exp[i\phi(t_0)]}{\sqrt{|\phi''(t_0)/2|}} \left\{ \sqrt{\frac{\pi}{8}} + i \operatorname{sgn}[\phi''(t_0)] \sqrt{\frac{\pi}{8}} \right\} \\
 &= \frac{\sqrt{2\pi} A(t_0) \exp[i\phi(t_0)]}{\sqrt{|\phi''(t_0)|}} \exp\left\{i \frac{\pi}{4} \operatorname{sgn}[\phi''(t_0)]\right\}
 \end{aligned} \tag{6.23}$$

## 6.4 The Wavelet Transform Ridge

We will define the edge,  $a_r(t)$ , of the continuous wavelet transform as the set of points  $(a, t)$  such that,

$$t_f(a_r, t) = t \tag{6.24}$$

Since,

$$\phi'_f(t_f) = \frac{1}{a} \phi'_\psi\left(\frac{t-t_f}{a}\right), \tag{6.25}$$



on the ridge,

$$a_r(t) = \frac{\phi'_\psi(0)}{\phi'_f(t)}. \quad (6.26)$$

It is therefore possible to recover the modulation law of the signal,  $\phi'_f(t)$ , from the edge of its wavelet transform. The problem now is to calculate the edge; we will use the phase to do this,

$$\Phi_{\psi}f(a,t) \equiv \arg[\mathcal{W}[\psi, f](a,t)] \quad (6.27)$$

whose estimator is unbiased, unlike the magnitude\*. The expression resulting from the stationary phase approximation is,

$$\Phi_{\psi}f(a,t) = \phi_f(t_f) + \phi_\psi\left(\frac{t-t_f}{a}\right) + Cte \quad (6.28)$$

and on the ridge,

$$\begin{aligned} \left. \frac{\partial}{\partial t} \Phi_{\psi}f(a,t) \right|_{t=t_f} &= \left[ \phi'_f(t) + \frac{\partial a}{\partial t} \times \frac{\partial \left(\frac{t-t_f}{a}\right)}{\partial a} \times \frac{\partial}{\partial \left(\frac{t-t_f}{a}\right)} \phi_\psi\left(\frac{t-t_f}{a}\right) \right]_{t=t_f} \\ &= \left[ \frac{1}{a} \phi'_\psi\left(\frac{t-t_f}{a}\right) - \left(\frac{\partial a}{\partial t}\right) \frac{t-t_f}{a^2} \phi'_\psi\left(\frac{t-t_f}{a}\right) \right]_{t=t_f} \\ &= \frac{1}{a_r} \phi'_\psi(0), \end{aligned} \quad (6.29)$$

which is the property we will use to extract the edge from the wavelet transforms. We also have,

$$\begin{aligned} \left. \frac{\partial}{\partial a} \Phi_{\psi}f(a,t) \right|_{t=t_f} &= -\frac{t-t_f}{a^2} \phi'_\psi\left(\frac{t-t_f}{a}\right) \Big|_{t=t_f} \\ &= 0. \end{aligned} \quad (6.30)$$

## 6.5 Use of non-asymptotic wavelets

The previous calculations were made assuming an asymptotic wavelet; let us see how they change when this is not the case. The main difference arises from the fact that the instantaneous amplitude

---

\*It is this stochastic behaviour that led us to choose the phase rather than the magnitude. It should be noted, however, that calculations equivalent to those we will develop for the phase can be made for the magnitude.

of the wavelet varies too rapidly to be taken out of the integral in the stationary phase approximation. Therefore we have,

$$\begin{aligned} \mathcal{W}[\psi, f](a, t) &\simeq \frac{1}{2a} A_f(t_f) \exp \left\{ i \left[ \phi_f(t_f) + \phi''_{\psi} \left( \frac{t-t_f}{a} \right) \right] \right\} \\ &\times \int_{-\infty}^{+\infty} A_{\psi} \left( \frac{t-\tau}{a} \right) \exp \left\{ \frac{i}{2} (\tau-t_f)^2 \left[ \phi''_f(t_f) + \frac{1}{a^2} \phi''_{\psi} \left( \frac{t-t_f}{a} \right) \right] \right\} d\tau. \end{aligned} \quad (6.31)$$

The calculations can be carried out in the case of the [Morlet](#) wavelet,

$$\psi_a(t) = \frac{1}{a} \exp \left( i \frac{\pi t}{a} \right) \exp \left[ -\frac{1}{2} \left( \frac{t}{2\sigma a} \right)^2 \right] \quad (6.32)$$

for which,

$$\begin{aligned} A_{\psi}(t) &= \exp \left[ -\frac{1}{2} \left( \frac{t}{2\sigma} \right)^2 \right] \\ \phi_{\psi}(t) &= \pi t \end{aligned} \quad (6.33)$$

Direct but rather lengthy calculations yield,

$$\mathcal{W}[\psi, f](a, t) \simeq \frac{\sigma \sqrt{2\pi}}{\left[ 1 + (4\pi\sigma^2 a'_r)^2 \right]^{1/4}} \exp \left[ \frac{i}{2} \arctan \left( -4\pi\sigma^2 a'_r \right) \right] Z_f(t_f) \quad (6.34)$$

where we have used the result\*,

$$\begin{aligned} &\int_{-\infty}^{+\infty} \exp \left[ -(\alpha x^2 + 2\beta x + \gamma) \right] \exp \left[ i(px^2 + 2qx + r) \right] dx \\ &= \frac{\sqrt{\pi}}{(\alpha^2 + p^2)^{1/4}} \exp \left[ \frac{\alpha(\beta^2 - \alpha\gamma) - (\alpha q^2 - 2\beta pq + \gamma p^2)}{\alpha^2 + p^2} \right] \\ &\times \exp \left\{ i \left[ \frac{1}{2} \arctan \left( \frac{p}{\alpha} \right) - \frac{p(q^2 - pr) - (p\beta^2 - 2q\alpha\beta + r\alpha^2)}{\alpha^2 + p^2} \right] \right\}. \end{aligned} \quad (6.35)$$

\*Found in the tables of [Gradshteyn and Ryzhik](#), page 485.



---

---

# CHAPTER 13

---

## SINGULAR SPECTRUM ANALYSIS

<b>1</b>	<b>Singular Spectrum Analysis (SSA)</b> . . . . .	<b>204</b>
1.1	Simple algorithm presentation . . . . .	204
1.2	To see how this works in practice . . . . .	207
<b>2</b>	<b>Analysis of the 4 stages of SSA</b> . . . . .	<b>211</b>
2.1	Embedding . . . . .	211
2.2	Singular Value Decomposition . . . . .	212
2.3	Grouping of SVD components . . . . .	214
<b>3</b>	<b>What SSA can do</b> . . . . .	<b>215</b>
3.1	Trend extraction . . . . .	215
3.2	Pseudo cycle separation . . . . .	217
3.3	Nonlinear Filtering . . . . .	221

As we have seen so far, in the series decompositions of [Fourier](#) or in wavelets, the orthogonal basis on which the signal is projected is imposed by the method. In the context of [Fourier](#) analysis, these are complex exponentials, not to mention infinite sines; for wavelets, they are specific functions that can be expanded or contracted at will, or almost. In the pragmatic approach followed in our field, we tend to lean towards [Fourier](#) analysis primarily to filter our geophysical signals; as for wavelets, they appear surprisingly in their continuous form, with the underlying notion of our contemporaries being to represent the evolution of the frequency support contained within a time series of ... over time. Although this perspective is somewhat reductive for each of the two approaches, the question arises: can we decompose our signal, for filtering, analysis, compression, *etc*, on a basis that is the most optimal and intrinsic to the original signal? As always, geophysicists have pondered this question, their intention at the time being to "fill in" gaps in a palaeoclimatic series ([Vautard et Ghil \(1989\)](#), [Vautard et al. \(1992\)](#)).

## 1 Singular Spectrum Analysis (SSA)

### 1.1 Simple algorithm presentation

Consider a discrete time series ( $\mathcal{X}$ ) of length  $N$  (with  $N > 2$ ) and, of course, non-zero,

$$\mathcal{X}_N = (x_1, \dots, x_N) \tag{1.1}$$

**Step 1: the trajectory matrix**  $\mathcal{X}$  is segmented into  $K$  sections of length  $L$  to form a matrix  $\mathbf{X}$  of dimension  $K \times N$ , where  $K = N - L + 1$ . This length  $L$  will henceforth be referred to as the analysis window  $\mathcal{L}$ , and as we will see later, after describing and discussing  $\mathbf{X}$ , it will become clear that the choice of the dimension of  $\mathcal{L}$  will dictate our decomposition. This is the first tuning parameter. This phase of embedding  $\mathcal{X}$  in  $\mathbf{X}$  is the first step of the SSA algorithm, which the Anglo-Saxons call the *embedding step*. The expression for  $\mathbf{X}$  is

$$\mathbf{X} = \begin{pmatrix} x_1 & x_2 & x_3 \cdots & x_K \\ x_2 & x_3 & x_4 \cdots & x_{K+1} \\ x_3 & x_4 & x_5 \cdots & x_{K+2} \\ \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & x_{L+2} \cdots & x_N \end{pmatrix} \tag{1.2}$$

As we can see, each column of  $\mathbf{X}$  is a segment of the realisation of  $\mathcal{X}$ , shifted or delayed by

## 1. SINGULAR SPECTUM ANALYSIS (SSA)

---

one sample. In fact, the regularity of the shift is not important; we do not have the constraint of "dt" as in the expression of the **Fourier** transform. Therefore, the column vectors of  $\mathbf{X}$  are called  *$\mathcal{L}$ -lagged vectors* and  $\mathbf{X}$  is called the  *$\mathcal{L}$ -trajectory matrix* or trajectory matrix. By construction, for any element  $(i, j)$  of  $\mathbf{X}$  we have  $x_{i,j} = x_{i+1,j+1}$ , which makes it an antidiagonal matrix defined by  $i + j = \text{constant}$ . It is a **Hankel** matrix provided it is square; otherwise it is quite easy to make it square. The values of  $\mathbf{X}$  are constant along the ascending diagonals.  $\mathbf{X}$  would be a **Toeplitz** matrix if they were constant along the descending diagonals. This **Hankel** matrix is very useful in the context of non-stationary signal decomposition, which will make it quite attractive to us later; it is also similar to an autocorrelation matrix, hence our earlier remark about the size of  $\mathcal{L}$ .

**Step 2: Singular Value Decomposition (SVD)** At this stage we are going to perform the Singular Value Decomposition, or SVD (**Golub et Reinsch, 1971**), of the matrix  $\mathbf{X}$ ; this step is a bit like going from the data space to the dual space. Let us construct  $\mathcal{S} = \mathbf{X}^t \mathbf{X}$ , the product of the transpose of  $\mathbf{X}$  with itself, to obtain a square matrix just for the purpose of using the terminology below (in fact, we could decompose  $\mathbf{X}$  directly): let  $\lambda_1, \lambda_2, \dots, \lambda_L$  be the eigenvalues of  $\mathcal{S}$ , in decreasing order of magnitude (*eg*  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_L \geq 0$ ), and  $U_1, U_2, \dots, U_L$  the orthonormal basis of the associated eigenvectors.

The rank  $d$  of  $\mathbf{X}$ , defined by  $d = \text{rank } \mathbf{X} = \max\{i | \lambda_i > 0\}$ , allows us to express  $\mathbf{X}$  as a sum of  $d$  unitary matrices using SVD,

$$\mathbf{X} = \mathbf{X}_1 + \mathbf{X}_2 + \dots + \mathbf{X}_d \quad (1.3)$$

In real life, *ie* for real signals, the rank  $d$  of  $\mathbf{X}$  is often simply the minimum of  $L$  and  $K$  ( $d = \min(L, K)$ ). The relationship (1.3) is analogous to that of the discrete **Fourier** transform; it is always possible to consider a signal as the sum of orthogonal sub-signals. Orthogonality ensures the linearity and uniqueness of the decomposition basis; in other words, energy is normally conserved from one space to another. However, it is important to note an important difference here: we are summing real numbers. Each of these unitary matrices  $\mathbf{X}_i$ , which are rank-1 matrices, is computed from the transpose of the original matrix  $\mathbf{X}$  and its eigenvalues and eigenvectors. The matrix  $i^{\text{th}}$  ( $i=1, \dots, d$ ) is defined,

$$\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^t \quad \text{avec,} \quad V_i^t = \mathbf{X}^t U_i / \sqrt{\lambda_i}. \quad (1.4)$$

**Step 3: Reconstruction** As we have just seen, the matrices  $\mathbf{X}_i$  are unitary matrices, and indeed, with the same philosophy as in the classical approach, it is possible to "group" these matrices into a physically homogeneous set, energetically homogeneous, *etc*. This is the second tuning parameter of the SSA algorithm: how to group the unitary matrices. For this purpose, the index set  $i \in \{1, \dots, d\}$  is divided into  $m$  disjoint index subsets  $\{I_1, \dots, I_m\}$ .

Let  $I$  be the set of  $p$  indices of  $i$ ,  $I = i_1, i_2, \dots, i_p$ . Since the relation (1.3) is linear, the resulting matrix  $\mathbf{X}_I$ , which groups the indices  $I$ , is expressed as follows,

$$\mathbf{X}_I = \mathbf{X}_{I_1} + \mathbf{X}_{I_2} + \dots + \mathbf{X}_{I_m} \quad (1.5)$$

We call this step the grouping of the eigentriplets  $(\lambda_s, U$  and  $V)$ . Obviously, in the limiting case where  $m = d$ , the relation (1.5) reduces rigorously to the relation (1.3), and we obtain our unitary matrices.

**Step 4: Diagonale average or Hankelization** This is the final step. Once the submatrices  $\mathbf{X}_I$  have been constructed, the task is to return to the data space, that is, to calculate the time series of length  $N$  associated with these matrices. Let  $\mathbf{Y}$  be a matrix of dimension  $L \times K$ , where for each element  $y_{i,j}$ , we have  $1 \leq i \leq L$  and  $1 \leq j \leq K$ . Let  $L^*$  be the minimum between  $L$  and  $K$  ( $\min\{L, K\}$ ), and let  $K^\times$  be the maximum between  $L$  and  $K$  ( $\max\{L, K\}$ ). We always have  $N = L + K - 1$ . Finally, let  $y_{ij}^* = y_{ij}$  if  $L < K$ , and  $y_{ij}^* = y_{ji}$  otherwise. The diagonal average, applied to the  $k^{th}$  index of the time series  $y$  associated with the matrix  $\mathbf{Y}$ , yields,

$$y_k = \begin{cases} \frac{1}{k} & \sum_{m=1}^k y_{m,k-m+1}^* & 1 \leq k \leq L^* \\ \frac{1}{L^*} & \sum_{m=1}^{L^*} y_{m,k-m+1}^* & L^* \leq k \leq K^* \\ \frac{1}{N-K+1} & \sum_{m=k-K^*+1}^{N-K^*+1} y_{m,k-m+1}^* & K^* \leq k \leq N^* \end{cases} \quad (1.6)$$

The relation (1.6) corresponds to the average of the  $k^{th}$  element along the anti-diagonal where  $i + j = k + 1$ . For  $k = 1$ ,  $y_1 = y_{1,1}$ , for  $k=2$ ,  $y_2 = (y_{1,2} + y_{2,1})/2$ , *etc*. Thus, from the matrices of step  $n^o3$ , we reconstruct the corresponding time series of length  $N$ . A note on terminology: when the diagonal mean is applied to the unitary matrices, the resulting series are called elementary series.

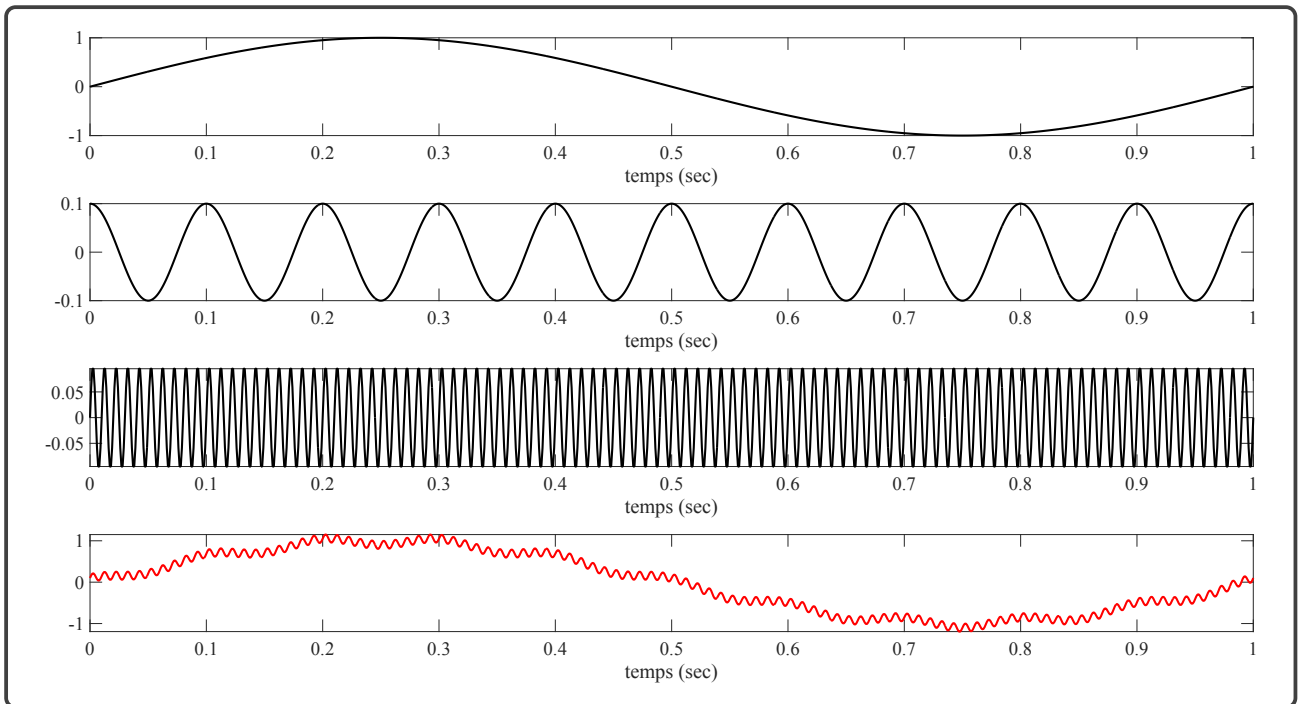
We note that nothing prevents us from extending SSA naturally from real signals to complex signals. It is sufficient to replace all transposes (symbols  $'$  in our demonstration) by complex

conjugates.

## 1.2 To see how this works in practice

For this example, obtained using the MATLAB program `ex_ssa01.m`, we consider three sinusoids with increasing frequencies of 1 Hz, 10 Hz, and 100 Hz, all sampled at 1 kHz and with different amplitudes.

We sum them, and Figure (13.1) shows the initial situation.



**Figure 13.1:** At the top is the 1 Hz sine wave, followed in descending order by the 10 Hz and 100 Hz sine waves. These three sinusoids are shown in black. The last one at the bottom, in red, is the sum of these sinusoids.

We will present (*cf* Figure 13.2) the **Hankel** matrices, or close to it, of the red signal by using the expression (1.2) and rigorously computing the expression of the trajectory matrix  $\mathcal{S}$  in order to adhere to the framework and be able to discuss eigenvectors and eigenvalues. The size of the analysis window is  $5/6$  of the length of the red signal.



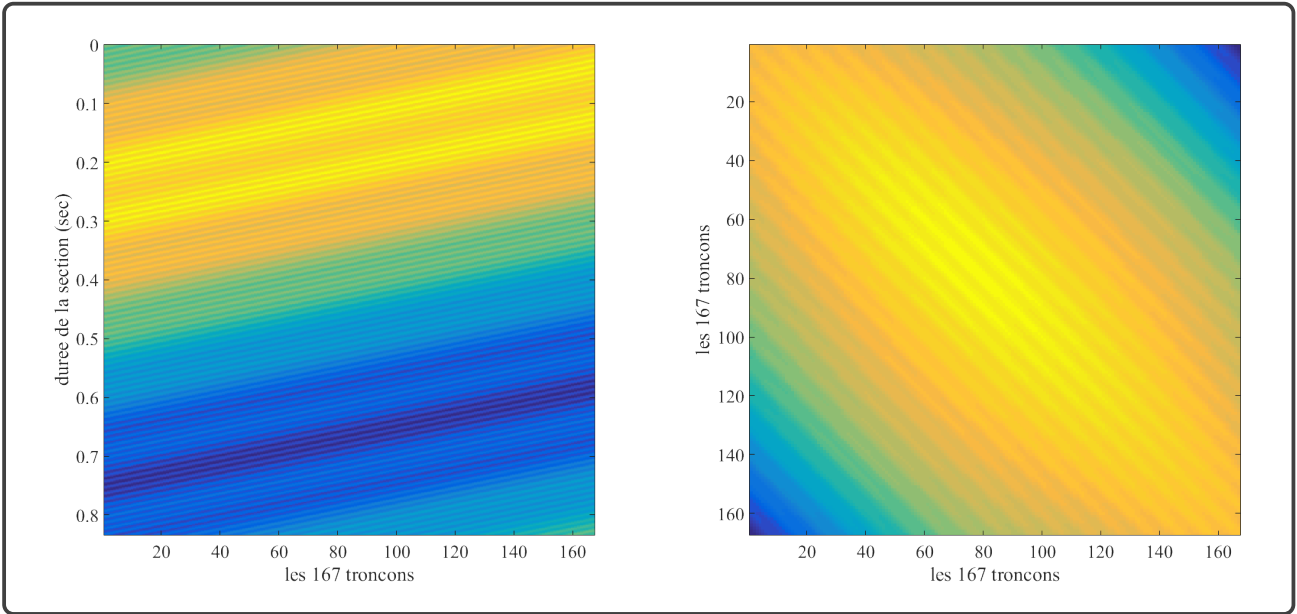


Figure 13.2: On the left is the rectangular matrix  $X$ , obtained using the `hankel.m` function in Matlab®. On the right is the product of  $X$  with its transpose, giving a square matrix.

We can proceed to step n°2 and apply SVD processing to these two matrices. Only the first 10 singularities and eigenvalues are shown here (Figure 13.3).

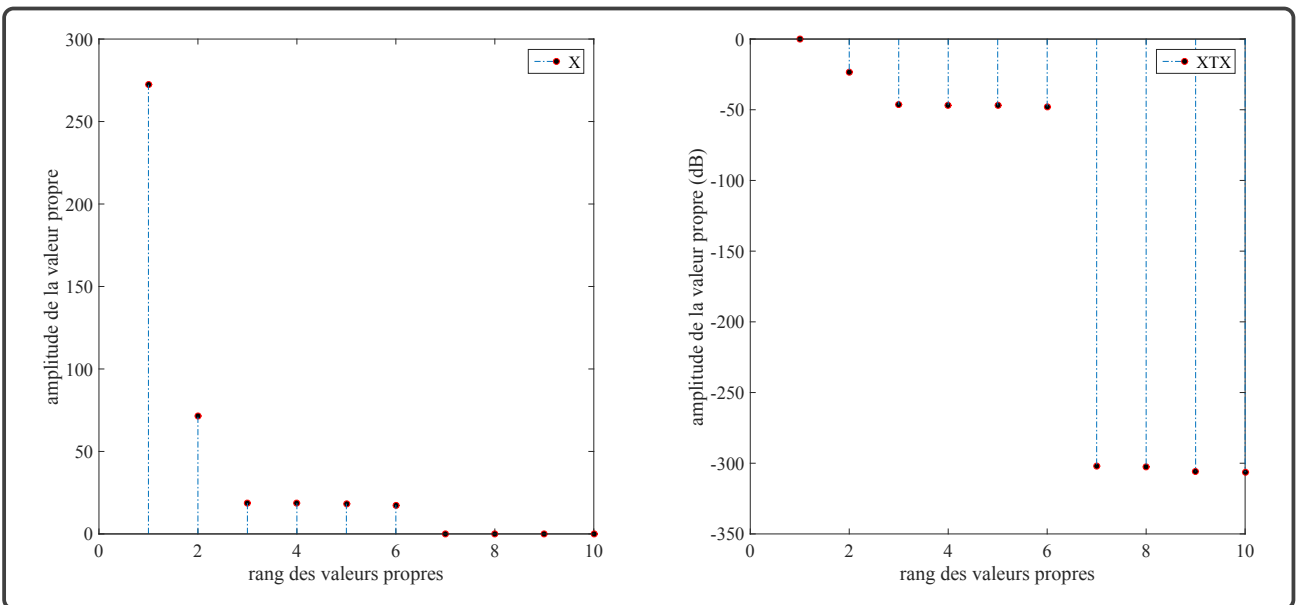


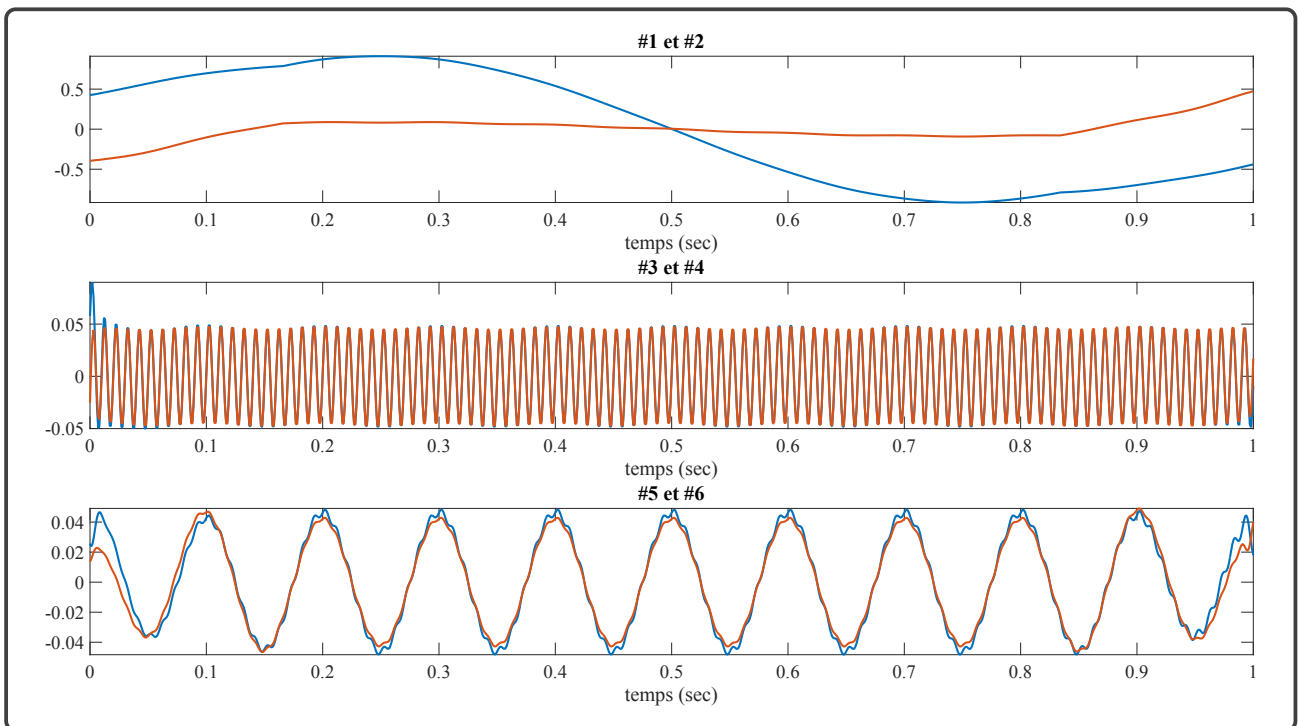
Figure 13.3: On the left, the first 10 singular values of  $X$ . On the right, the logarithm (in dB) of the first 10 eigenvalues of  $X^T X$ .

The logarithm was necessary because the square of an eigenvalue can be quite large. Nevertheless, we observe that in both cases there seems to be no significant energy above the 7<sup>th</sup> eigenvalue (or singular value). For this example we have chosen the limiting case where  $m = d$ , i.e. we will

## 1. SINGULAR SPECTUM ANALYSIS (SSA)

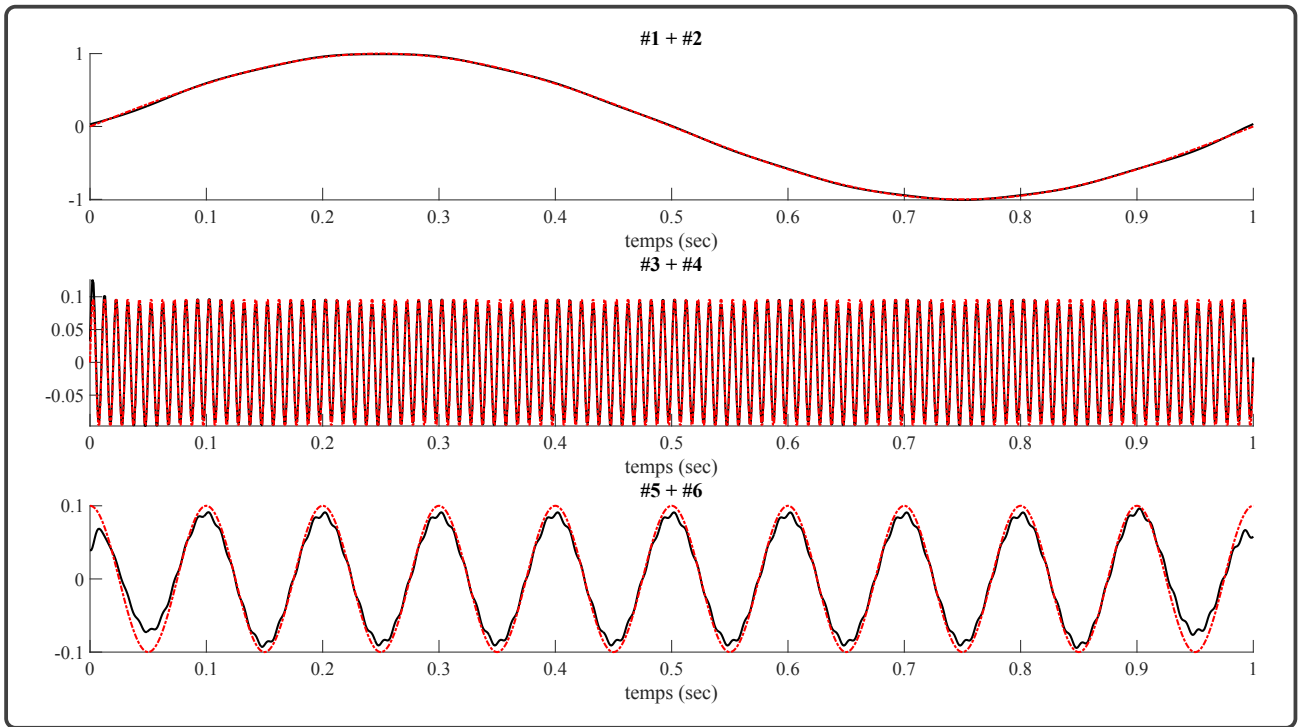
reconstruct the 6 first unitary matrices and thus the 6 first elementary signals. One last point of clarification: since it seems that using  $\mathbf{X}$  instead of  $\mathbf{X}'\mathbf{X}$  does not change the result, except for having to adjust the square of the  $\lambda$ s in the reconstruction formula and the sign of the original signal, we will use  $\mathbf{X}$  exclusively from now on.

We are left with the final step, sometimes referred to in the literature as "*hankelization*", which is diagonal averaging. Figure (13.4) shows the 6 elementary signals reconstructed by SSA. We have paired them for an obvious reason: the similarity of the patterns.



**Figure 13.4:** From top to bottom, the first 2 elementary signals appear to represent the 1 Hz sinusoid, followed by signals number 3 and 4, which correspond to the 10 Hz sinusoid, and finally, the last 2 signals can be attributed to the 100 Hz oscillation.

As noted at the beginning of this chapter, unlike *Fourier* analysis, we are working in real space (Re) for both the original signal and the grouping matrices. This allows us to sum the contributions of interest at each step of the operation, focusing here on shape similarity. Figure (13.5) shows these sums.



**Figure 13.5:** The red and blue curves for each pair in Figure (13.4) have been simply summed (black curves). They are compared to their respective original signals (red curves).

The result is quite remarkable; the SSA analysis has successfully detected and separated each contribution in terms of both phase and amplitude. However, the reconstruction is not perfect for several reasons, the most important of which is the size of the analysis window  $L$ . Here we have chosen it somewhat arbitrarily, but as with wavelets, its detection capability depends significantly on its length. Figure (13.6) shows the first 10 singular values computed by SSA for the three individual sinusoids and for the combined signal.

Two things become clear: first, it seems that two singular values, and thus two singular vectors, are needed to reconstruct a pure oscillation. These are called **Hilbert** pairs. The second observation is that theoretically these two pairs should have equal amplitudes, which is clearly not the case here. The **Hilbert** pairs for the 1 Hz oscillation are quite problematic, while those for the fastest oscillation are almost perfect. It is time to analyse these problems.

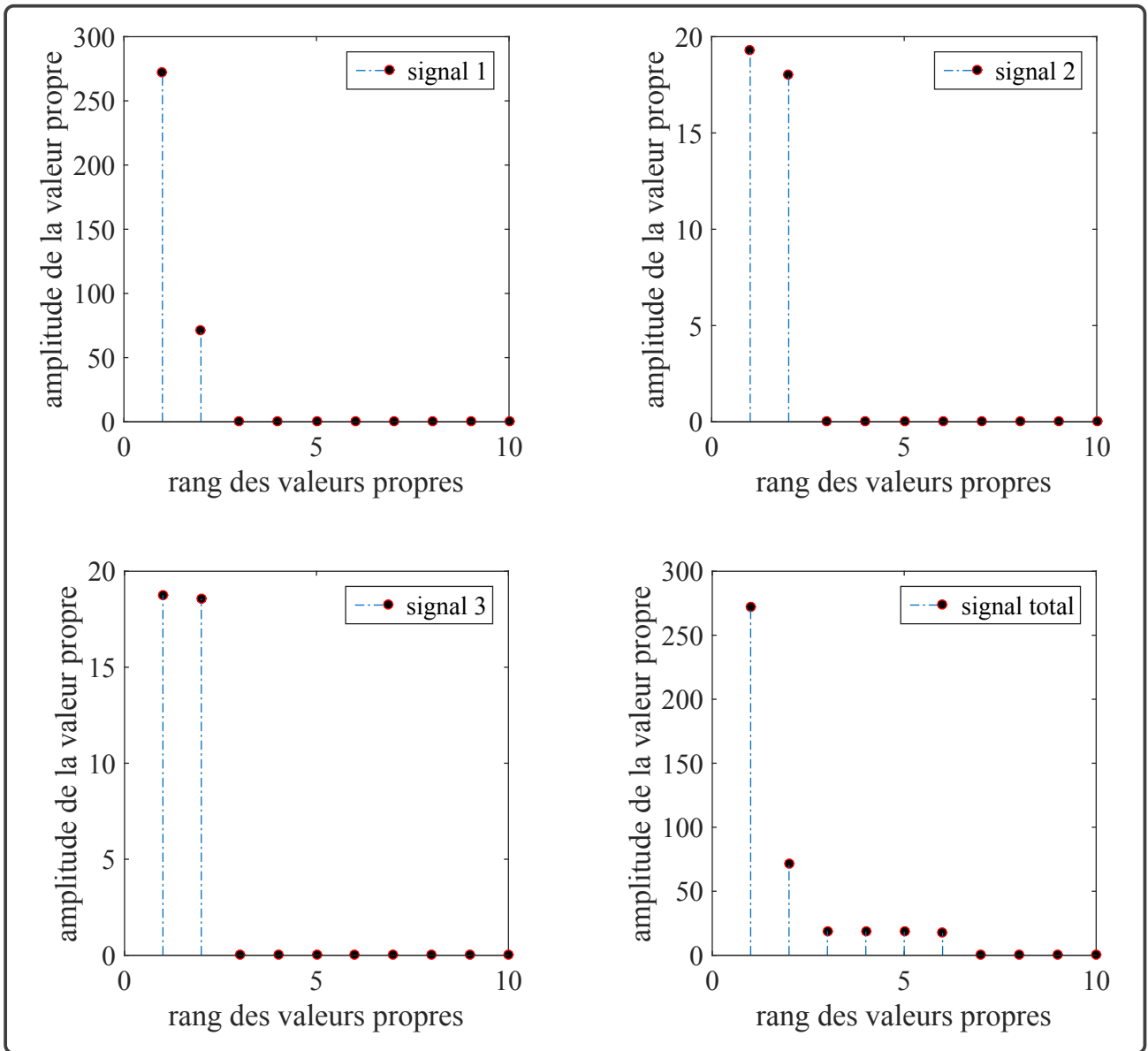


Figure 13.6: The top left shows the signal with the lowest period (1 Hz), the top right shows the singular values associated with the 10 Hz oscillation, the bottom left shows those associated with the 100 Hz oscillation, and finally the bottom right shows the singular values we have already presented for the total signal (Figure 13.3).

## 2 Analysis of the 4 stages of SSA

### 2.1 Embedding

The first stage of SSA analysis, embedding, involves projecting the one-dimensional time series  $\mathcal{X}_N = (x_1, \dots, x_N)$  into a multidimensional series space  $(X_1, \dots, X_k)$  such that the vectors  $X_i = (x_i, \dots, x_{i+L-1})^t$  belong to the space  $\mathcal{R}^L$ , where  $K = N - L + 1$ . This somewhat succinct definition was proposed and demonstrated in the early 1980s by [Mañé \(1981\)](#) and [Takens \(1981\)](#), with the aim

---

of constructing a space that accurately describes strange attractors, often a Banach space. A strange attractor is an object whose dynamical properties can evolve into chaos and are therefore non-linear in nature. The parameter controlling the embedding is  $L$ , the size of the analysis window;  $L$  is an integer between 2 and  $N - 1$ . The Hankel matrix (eg 1.2) has symmetry properties; its transpose  $\mathbf{X}^t$ , known as the trajectory matrix, has dimension  $K$ . Embedding is a mandatory step in the analysis of nonlinear series; formally, it involves empirically evaluating all pairs of distances between two shifted, lagged vectors to calculate the correlation dimension of the series under analysis. This dimension is quite close to the fractal dimension of the strange attractors that could generate such series, and in this particular case it is advisable to choose very small window sizes  $L$  (i.e. very large  $K$ ). *A contrario* for SSA,  $L$  must be sufficiently large so that each vector contains a significant part of the information contained in the original time series ( $\mathcal{X}_N$ ); from a mathematical point of view, one must consider the framework of *Structural Total Least Squares* (STLS) for a Hankel matrix (Lemmerling et Van Huffel, 2001), which contrasts with the fractal dimension discussed above. A second advantage of using very large values for  $L$  is the ability to consider the sub-vectors ( $\mathbf{X}_i$ ) as independent sub-series with different dynamics, thus allowing the identification of common features within collections of these sub-series.

## 2.2 Singular Value Decomposition

The SVD of the non-zero trajectory matrix ( $\mathbf{X}$ ), which has dimensions  $L \times K$ , is a decomposition of the form,

$$\mathbf{X} = \sum_{i=1}^d \sqrt{\lambda_i} U_i V_i^t \quad (2.1)$$

relation (2.1), in which we find the eigenvalues  $\lambda_i$  ( $i = 1, \dots, L$ ) of the matrix  $\mathbf{S} = \mathbf{X}\mathbf{X}^T$ , arranged in descending order of magnitude, the corresponding (left) eigenvectors  $U_i$ , and finally the (right) eigenvectors  $V_i$  given by the following relation,

$$V_i = \mathbf{X}^T U_i / \sqrt{\lambda_i}. \quad (2.2)$$

The equality (2.1) shows that the SVD has special symmetry properties, which leads to the fact that the (right) eigenvectors  $V_1, \dots, V_2$ , which also form an orthonormal basis, are arranged in the same order as the eigenvalues ( $\lambda_i$ ). Let  $\mathbf{X}_i$  be a submatrix of  $\mathbf{X}$ ,

$$\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^t, \quad (2.3)$$

## 2. ANALYSIS OF THE 4 STAGES OF SSA

---

then the embedding matrix  $\mathbf{X}$  can be represented as a simple linear sum of elementary matrices  $\mathbf{X}_j$ . If all the eigenvalues are equal to one, then (1.3) is uniquely defined.

Now to the nature and characteristics of the embedding matrix: Note that its rows and columns are subsets of the original time signal. Consequently, the eigenvectors  $(U_i, V_i)$  have a temporal structure and can therefore be considered as a representation of the time series data. Let  $\mathbf{X}$  be a sequence of  $L$  delayed parts of  $\mathcal{X}$  and  $(X_1, \dots, X_K)$  the linear basis of these eigenvectors. If we set,

$$Z_i = \sum_{i=1}^d \sqrt{\lambda_i} V_i, \quad (2.4)$$

with  $i = 1, \dots, d$ , then (2.1) can be expressed in the form,

$$\mathbf{X} = \sum_{i=1}^d U_i Z_i^t \quad (2.5)$$

ie for the elementary matrix  $j^{\text{th}}$ ,

$$X_j = \sum_{i=1}^d z_{ji} U_i \quad (2.6)$$

where  $z_{ji}$  is a component of the vector  $Z_i$ . This means that the vector  $Z_i$  is composed of the  $i^{\text{th}}$  components of the vector  $X_j$ . In the same way, if we introduce,

$$Y_i = \sum_{i=1}^d \sqrt{\lambda_i} U_i \quad (2.7)$$

we obtain for the transposed trajectory matrix,

$$X_j^t = \sum_{i=1}^d U_i Y_i^t \quad (2.8)$$

which corresponds to a representation of the  $K$  lagged vectors in the orthogonal basis  $(V_1, \dots, V_d)$ . This illustrates why the SVD is an excellent choice for analyzing the embedding matrix, as it provides us with two geometric descriptions.

---

**Please note 1** There are strong similarities between performing an SVD of the trajectory matrix, as in the case of SSA, and multivariate analyses such as Principal Component Analysis (PCA) or Karhunen-Loève (KL) decompositions commonly used in time series analysis. However, SSA differs in the nature of its trajectory matrix; it is a Hankel matrix with a particular structure, where its rows and columns are subsets of the signal being analysed and thus have a meaningful temporal and physical sense relative to each other. This is not the case for PCA and KL.

**Please note 2** In general, the orthonormal basis ( $U_i$ ) associated with the trajectory matrix and obtained by SVD can be replaced by any orthonormal basis ( $P_i$ ). In this case, the relation (1.3) becomes  $X_i = P_i Q_i^t$  with  $Q_i = X^t P_i$ . A classic example of an alternative basis are the eigenvectors of an autocovariance matrix (Toeplitz SSA).

## 2.3 Grouping of SVD components

The topic here is the separation of additive components of a time series, which involves addressing the critically important question: the concept of "separability".

Let  $\mathcal{X}$  be the sum of two time series  $\mathcal{X}^{(1)}$  and  $\mathcal{X}^{(2)}$  such that  $x_i = x_i^{(1)} + x_i^{(2)}$  for all  $i \in [1, N]$ . Let  $L$  be the fixed-length analysis window, and let  $X$ ,  $X^{(1)}$  and  $X^{(2)}$  be the embedding matrices for the series  $\mathcal{X}$ ,  $\mathcal{X}^{(1)}$  and  $\mathcal{X}^{(2)}$ . These two subsets are separable (even weakly) in relation (1.3) if there exists a collection of indices  $\mathcal{I} \subset 1, \dots, d$  such that  $\mathbf{X}^{(1)} = \sum_{i \in \mathcal{I}} \mathbf{X}_i$ , or if there exists a collection of indices such that  $\mathbf{X}^{(2)} = \sum_{i \notin \mathcal{I}} \mathbf{X}_i$ .

For example, in the case of separability, the contribution of  $\mathbf{X}^{(1)}$  corresponds to the simple ratio of its eigenvalues ( $\sum_{i \in \mathcal{I}} \lambda_i$ ) to the total eigenvalues ( $\sum_{i=1}^d \lambda_i$ ). We have illustrated this case with figure (13.6).

Still in the context of the relation (1.3), let  $\mathcal{I} = \mathcal{I}_1$  be the set of indices corresponding to the first signal, with the corresponding matrix denoted by  $X_{\mathcal{I}_1}$ . If this matrix, as well as the matrix corresponding to the second signal ( $\mathbf{X}_{\mathcal{I}_2} = \mathbf{X} - \mathbf{X}_{\mathcal{I}_1}$ ), are close to a Hankel matrix, or are Hankel matrices themselves, then the signals are separable or approximately separable. It is therefore clear that the concept of grouping SVD components can be summarised\* as the decomposition of the initial trajectory matrix into several elementary matrices†.

---

\*only theoretically, as the actual problem is much more complex

†whose structures should be as close as possible to that of a Hankel matrix

### 3. WHAT SSA CAN DO

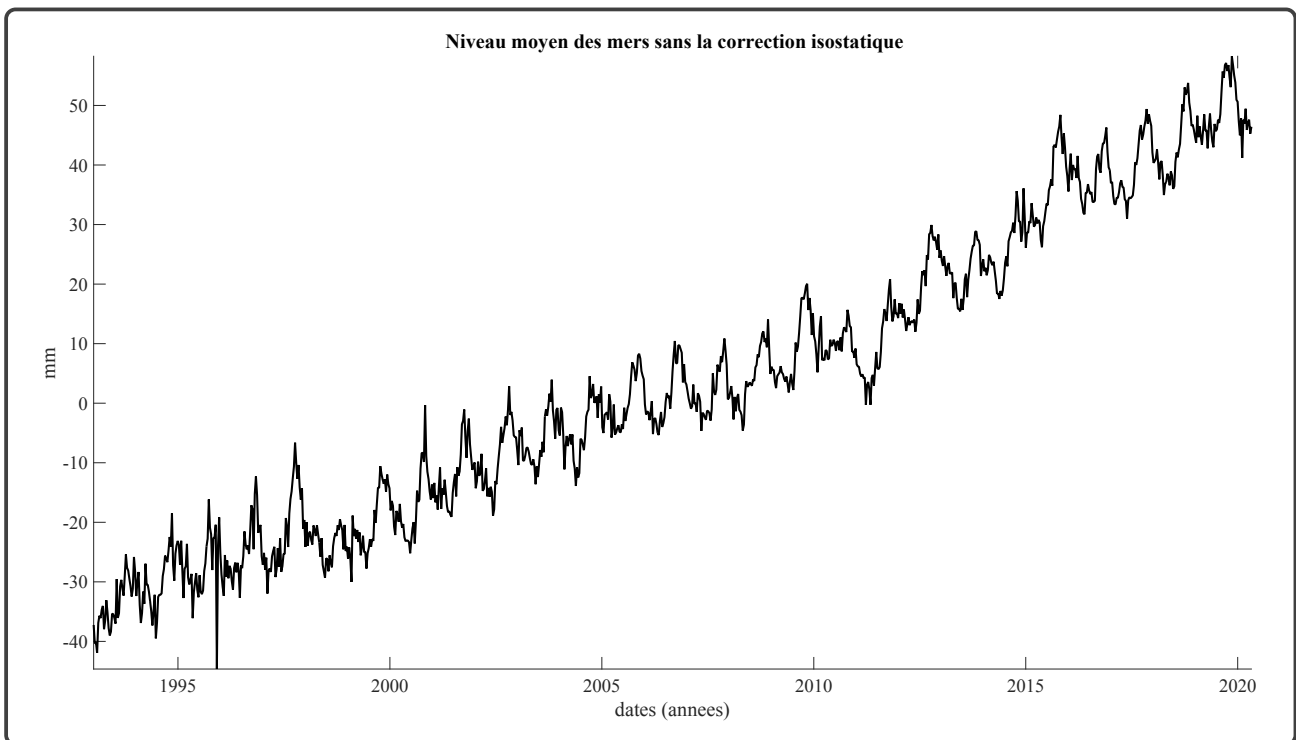
---

We will stop here, as the problem is very complex. Although the idea is simple, several procedures are available to us; these will be discussed later in this chapter..

## 3 What SSA can do

### 3.1 Trend extraction

To illustrate our point, we will apply SSA to physical data. Since the early 1990s, NASA has been measuring and providing\* mean sea level (*cf* Figure (13.7)). One of the questions for geodesists is the effect of isostasy on tectonics in general and on the axis of rotation of the poles in particular (*eg* Courtillot *et al.*, 2022). For example, how does the melting of ice and the redistribution of surface masses affect the Earth's axis of rotation? One way to understand this phenomenon is to study the evolution of global mean sea level from satellite measurements (Poseidon/Topex, Jason I, II and III).



**Figure 13.7: Mean sea level from 1993 up to the present day**

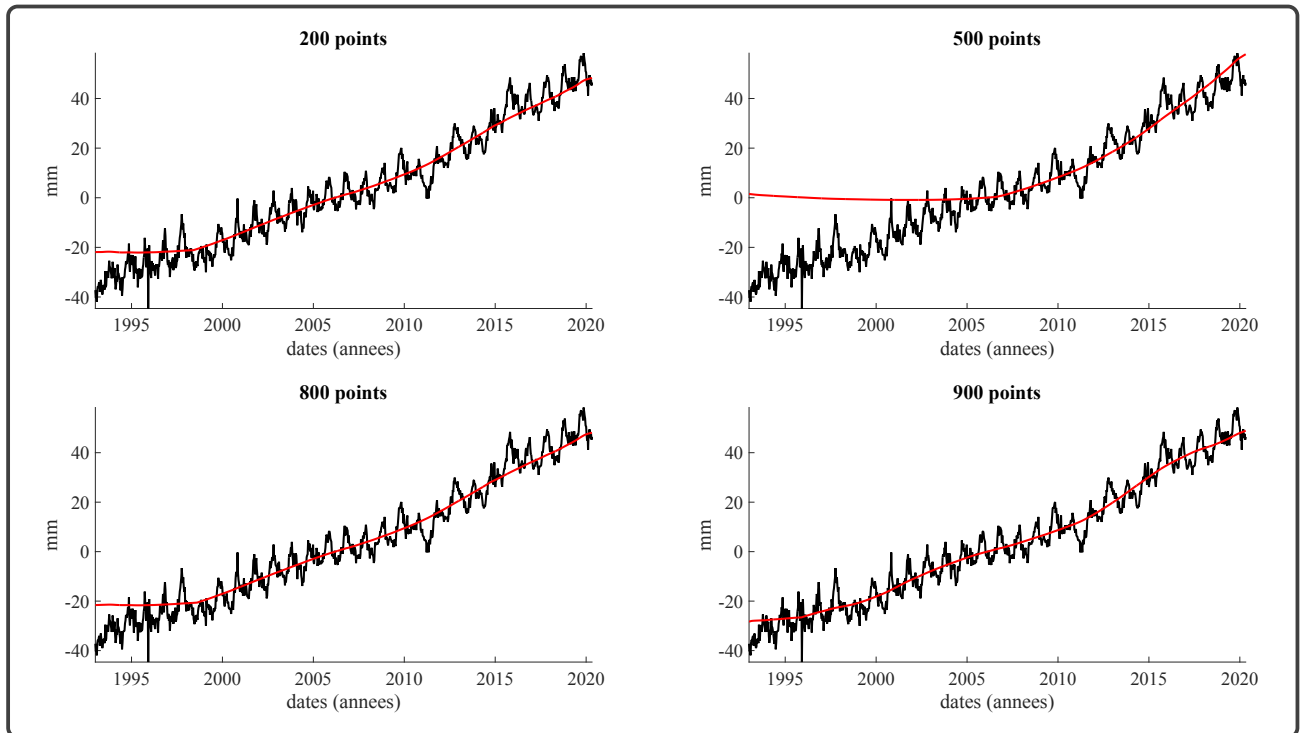
In figure (13.7) we have plotted this mean sea level curve, which obviously shows a superposition of a more or less non-linear trend and an annual oscillation due to the Earth's rotation. The data start in 1993 and extend to September 2020, with a temporal sampling of about one point every 10

---

\*<https://climate.nasa.gov/vital-signs/sea-level/>



days. We will perform the SSA without any precautions and represent the first computed component, the trend, using the first elementary matrix  $\mathbf{X}_{\mathcal{I}_1}$  with  $\mathcal{I}_1 = 1$ . We have chosen different values of  $L$ : 200 points ( $\approx 5.5$  years), 500 points ( $\approx 13.7$  years), 800 points ( $\approx 21.9$  years) and 900 points ( $\approx 24.6$  years). The shift between two consecutive vectors in  $\mathbf{X}$  is a sample point. The trends obtained with the script `ex_ssa02.m` are shown in figure (13.8).



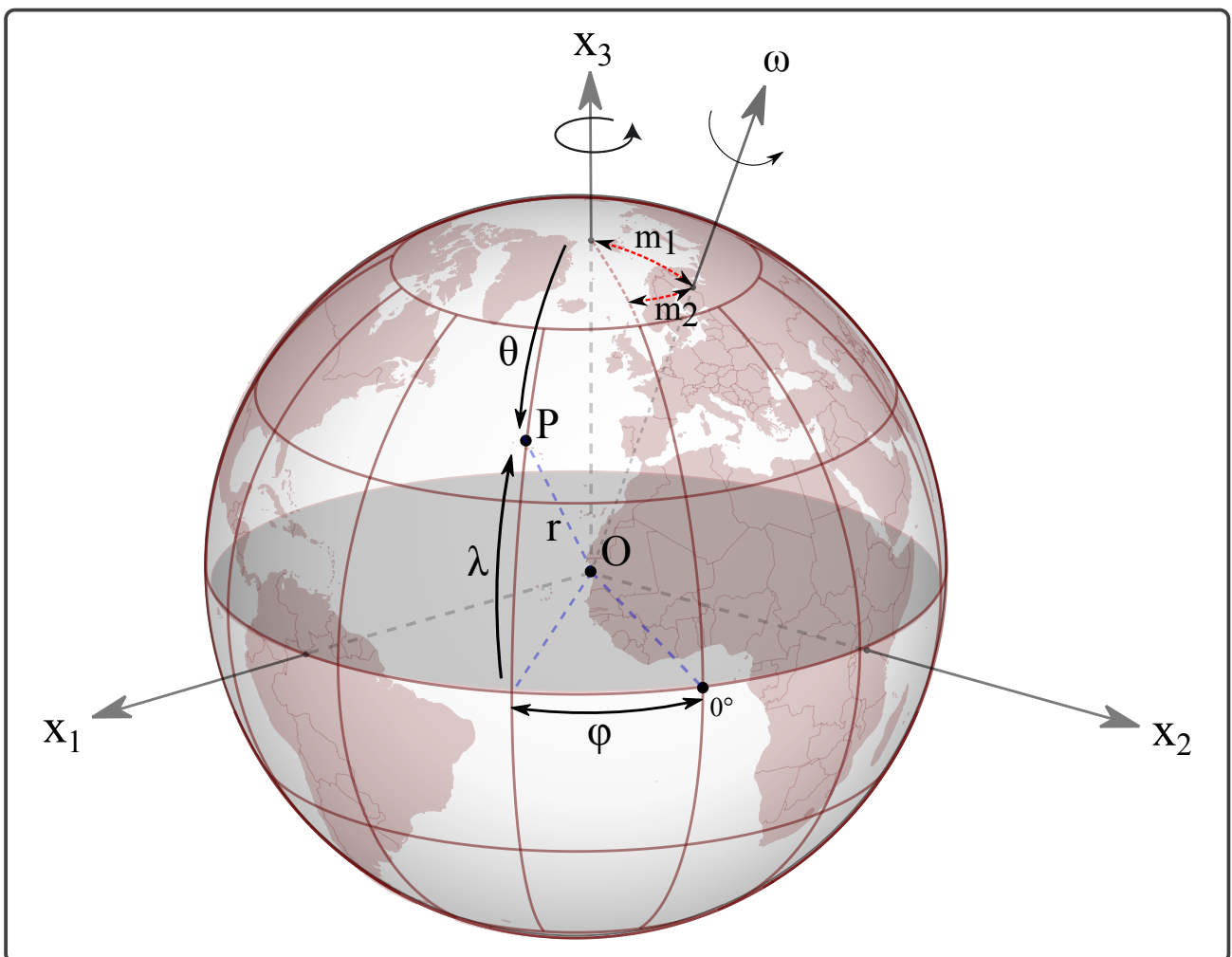
**Figure 13.8:** Superposées sur la courbe du niveau moyen des océans brute, les tendances extraites par SSA pour différentes valeurs de  $L$  (courbes rouges).

First important observation: SSA has a significant smoothing power; the trends obtained (red curves), although non-linear, are all smooth. Next, we observe quite different behaviour, especially for a value of  $L$  of 500 points. Formally, for the first eigentriplet\*, there are only minor differences between a moving average and the first component extracted by SSA. A priori, for a window length of 13.7 years (500 points), shifted one sample point at a time, the conditions seem to be met to characterise two behaviours in the data: a plateau from 1993 to around 2007, followed by an affine trend from 2007 to the present. As mentioned in the previous section, the length of the analysis window should be as large as possible; this is not a mathematical criterion. Figure (13.8) shows that the trends obtained for 200, 800 and 900 points belong to the same family. SSA, like other tools presented in this paper, is by no means a magic tool. Depending on the question, only the geophysicist should have the final say.

\*a single and unique triplet, *eg* for  $i = 1$  ( $\lambda_1, U_1, V_1'$ )

### 3.2 Pseudo cycle separation

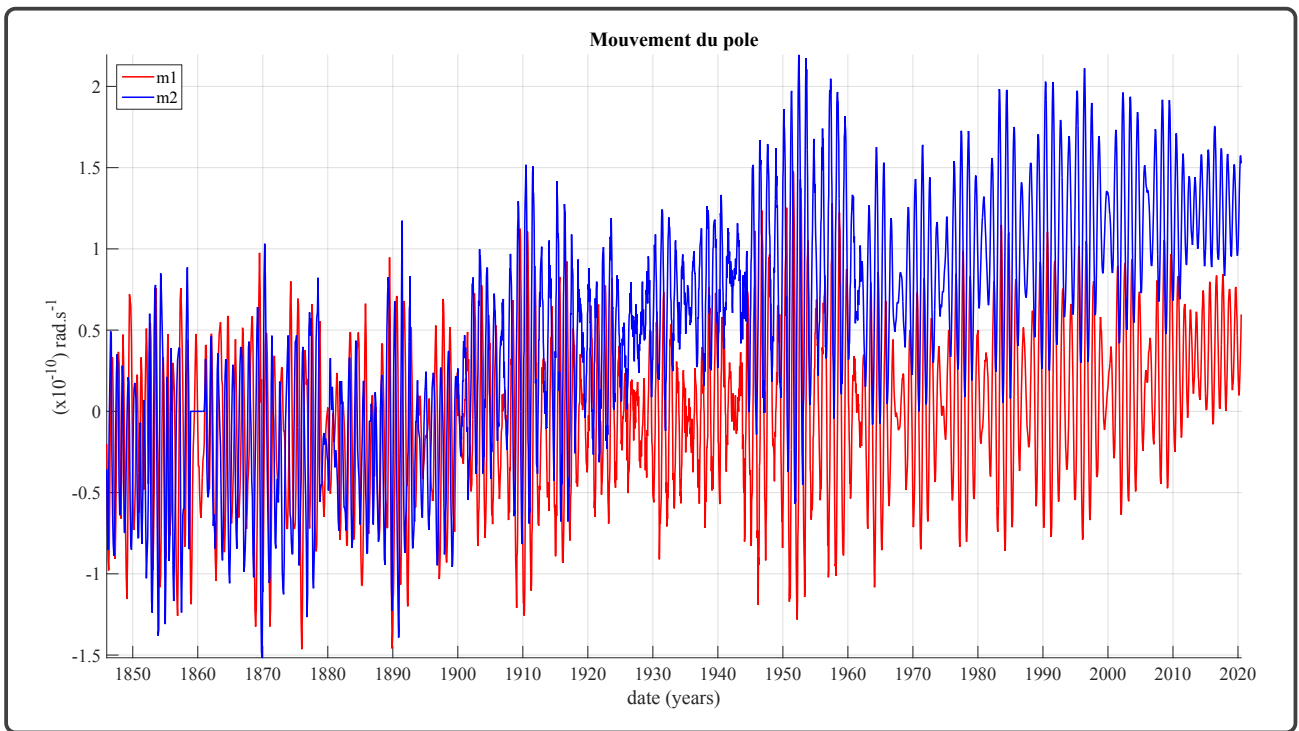
Continuing with the trend extraction just discussed, we will now analyse a new real signal to extract its main pseudo-cycles\*, namely the movement of the Earth's rotation pole (eg Lopes *et al.*, 2021). The movement of the Earth's rotation pole has been measured since 1846, initially using stars and now using laser measurements from satellites. The *International Earth Rotation and Reference Systems Service*† provides us with the time series of the pair  $(m_1, m_2)$ , the coordinates of the rotation pole (cf Figure (13.9)). The time series of this pair is shown in Figure (13.10).



**Figure 13.9: Geodetic Reference for the Movement of the Pole.**  $m_1$  is the North-South distance from the geographic North Pole, and  $m_2$  is the East-West distance, with reference to the Greenwich Meridian

\*cycles whose periods and amplitudes vary over time

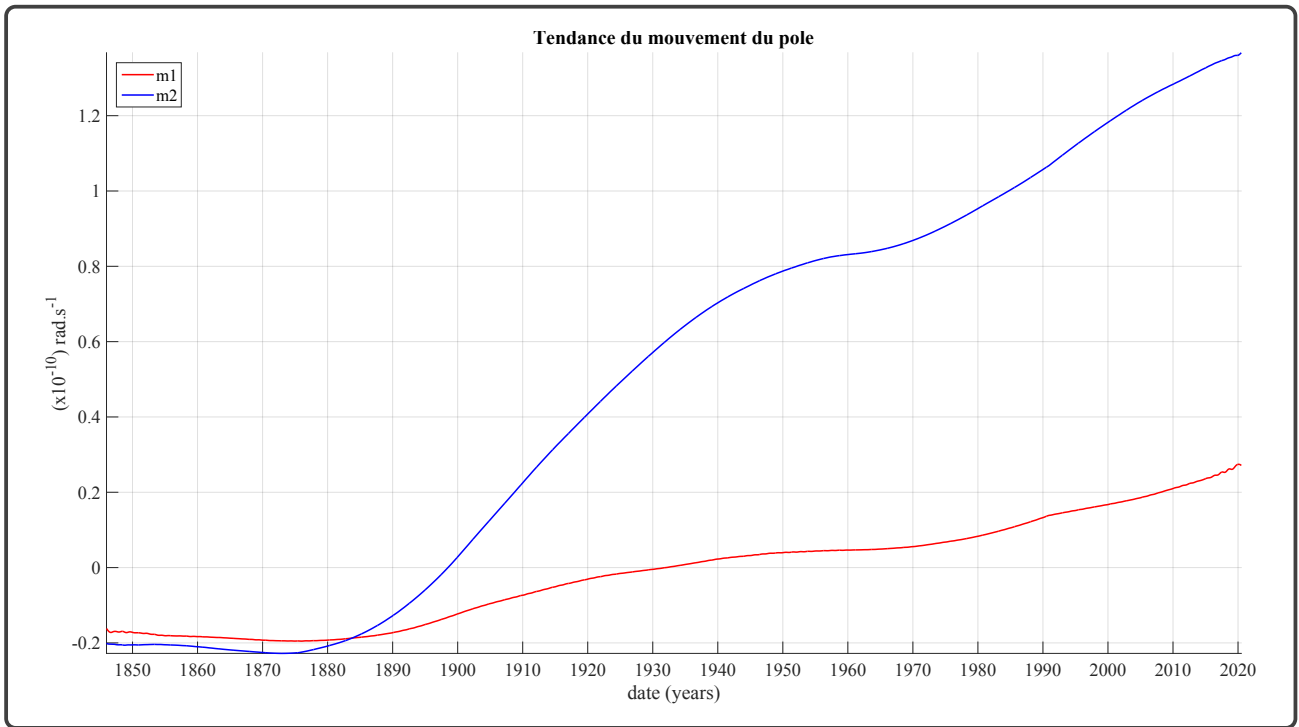
†<https://www.iers.org/IERS/EN/DataProducts/EarthOrientationData/eop.html>



**Figure 13.10: Temporal evolution of the components ( $m_1$ ,  $m_2$ ) from 1846 to the present.**

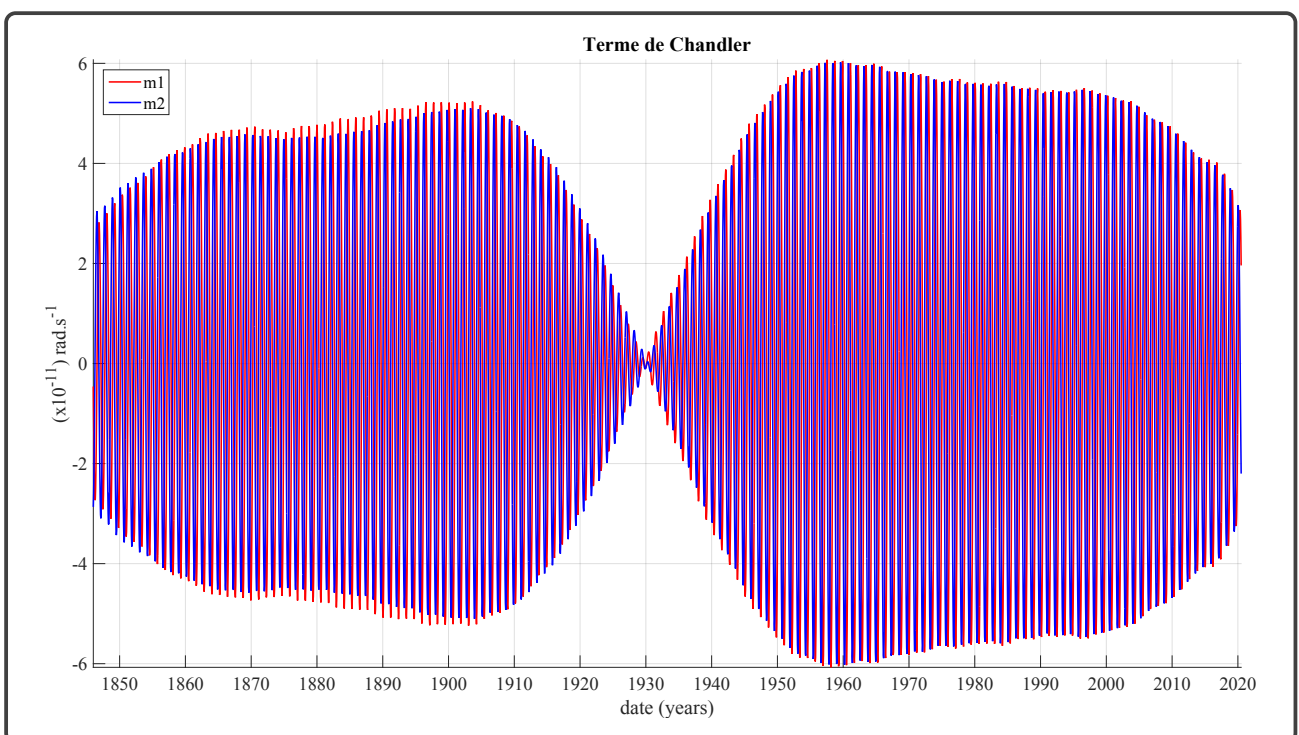
It has been known since the late 18th century that the movement of the pole follows the first-order linear partial differential equations of [Liouville-Euler](#). This system exhibits a forced oscillation, traditionally called the annual oscillation, resulting from the Earth's revolution around the Sun; and a free oscillation, known as the Chandler oscillation ([Chandler \(1891a\)](#); [Chandler \(1891b\)](#)), characterised by a dramatic phase jump during the 1920s and 1940s. These two pseudo-oscillations are superimposed on a pole drift discussed in the previous section, with time constants corresponding to the drift of the plates ( $\approx 10$  cm/year). This drift was first identified in the 1960s by [Markowitz](#) ([Markowitz et Guinot, 1968](#)). The following figures have been produced using the script [ssa\\_03.m](#). Figure (13.11) shows the SSA analysis components 1 ( $m_1$ ) and 5 ( $m_2$ ). As can be seen in figure (13.10), since component  $m_2$  drifts more than its longitudinal counterpart, it is normal for its trend to appear before that of  $m_1$ .

### 3. WHAT SSA CAN DO



*Figure 13.11: Trends of the Pair  $(m_1, m_2)$  Extracted by SSA*

Next, in the same order, are 1) Chandler oscillations (components 1 and 2 for  $m_1$ , 2 and 3 for  $m_2$ ; cf Figure (13.12)), and 2) forced oscillations (components 3 and 4 for  $m_1$ , 4 and 5 for  $m_2$ ; cf Figure (??)).



*Figure 13.12: Chandler pseudo-cycles of the pair  $(m_1, m_2)$  extracted by SSA.*

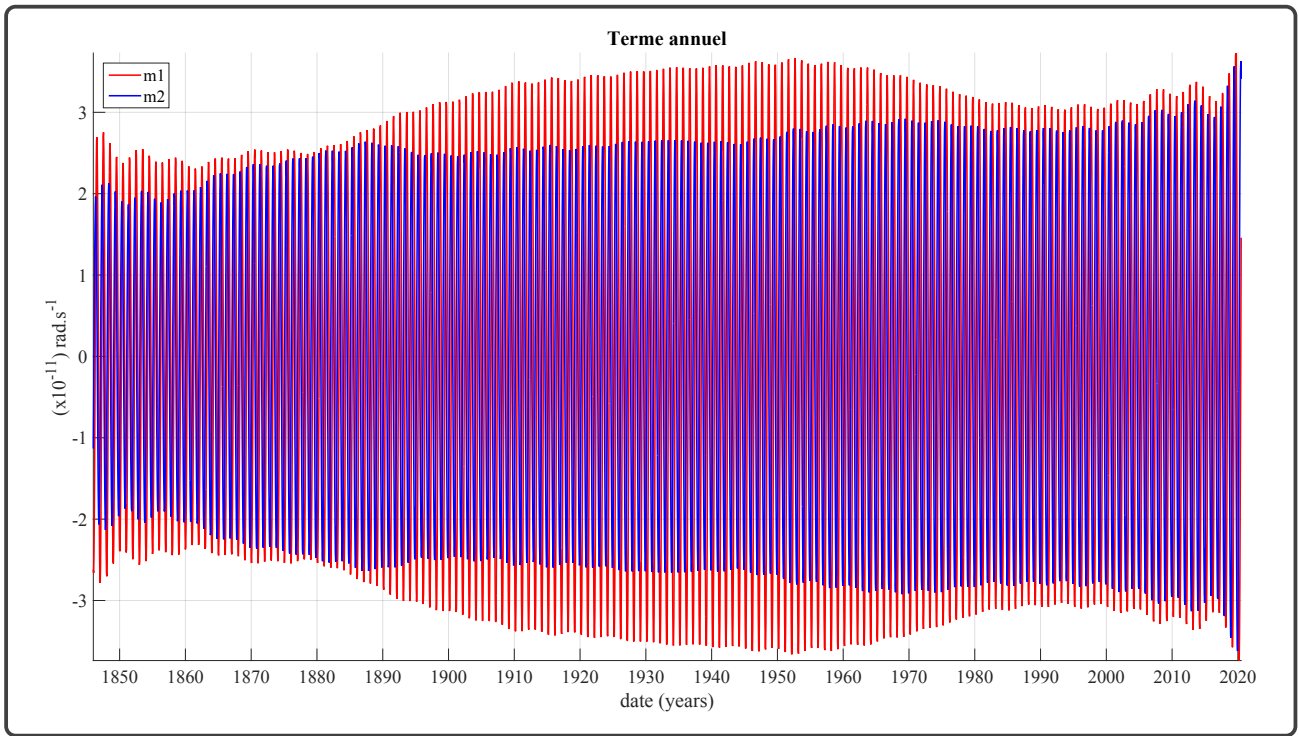


Figure 13.13: Forced pseudo-cycles of the pair  $(m_1, m_2)$  extracted by SSA

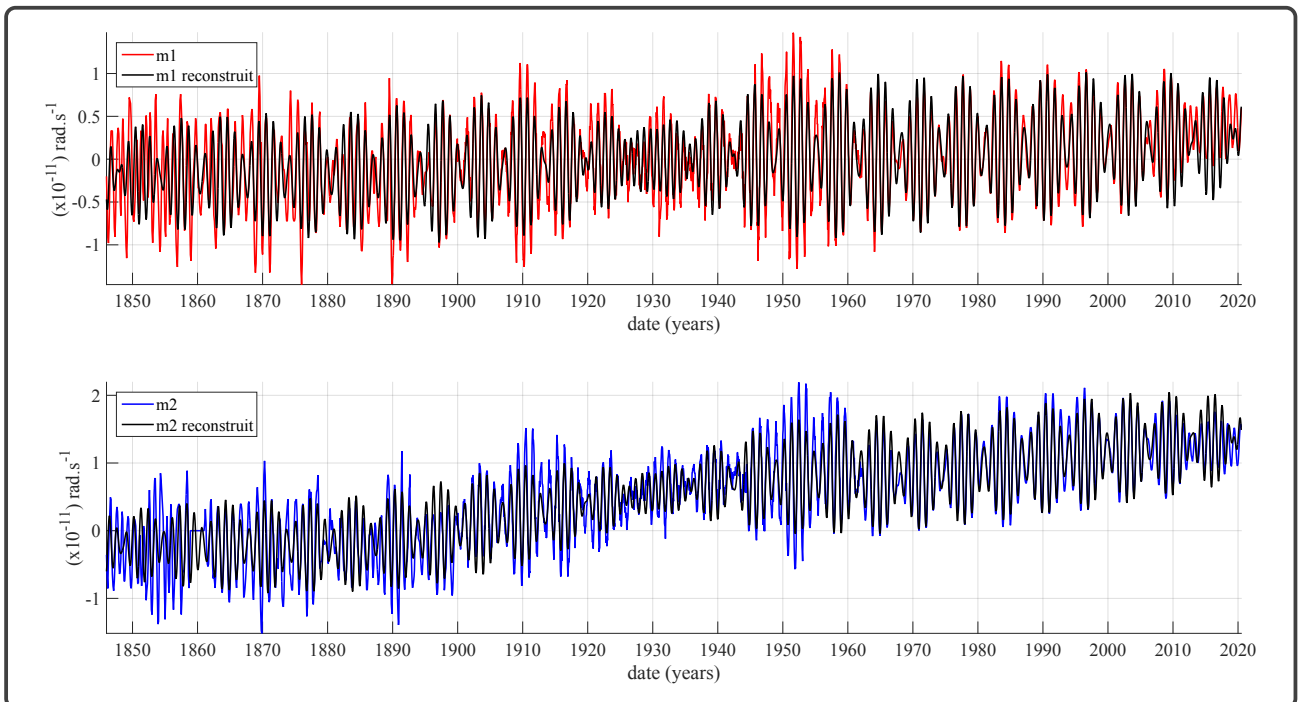


Figure 13.14: Comparison between the raw component  $m_1$  (red curve) and the sum of the extracted components (black curve). Below, the same comparison for  $m_2$ .

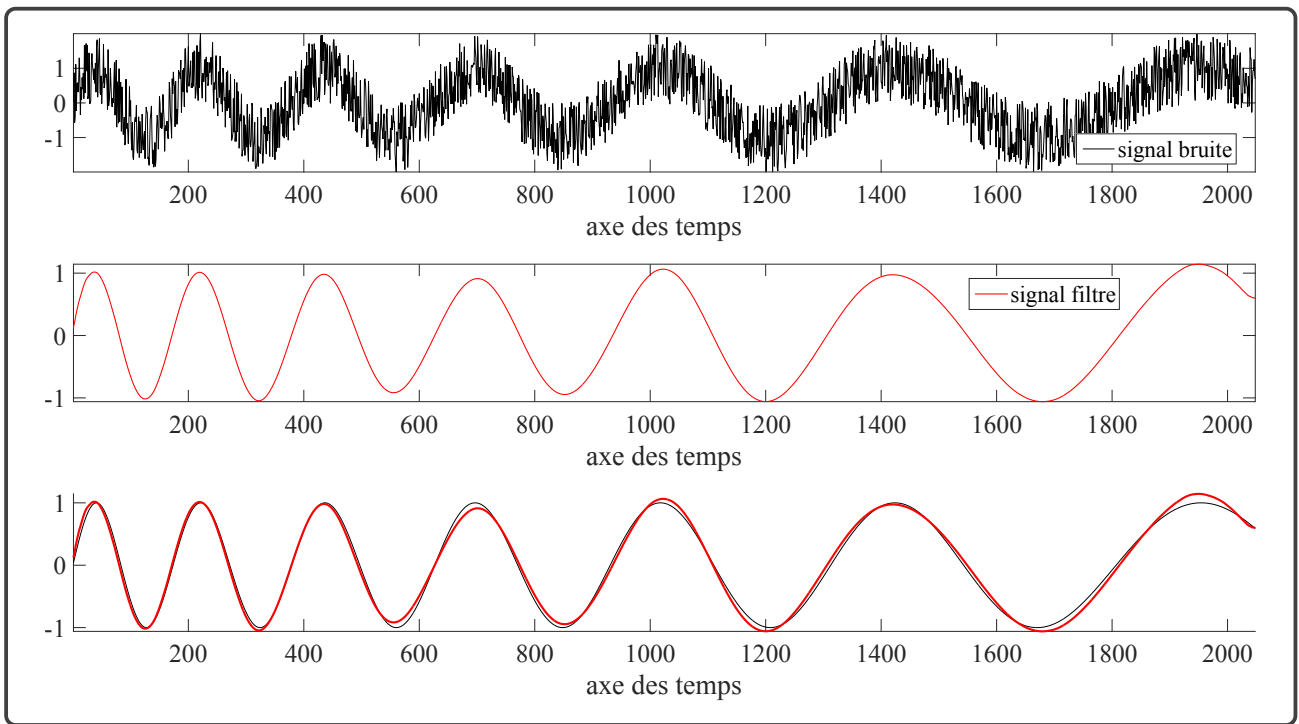
Since SSA operates only in the data space and not in the transform space, unlike [Fourier](#) or [Wavelets](#), it is possible to reconstruct a signal from the extracted cycles and trends that best fits a theory such as [Liouville-Euler](#). SSA thus allows us to discard any information that is not accounted

for by a system of equations and that could complicate its resolution or inversion. The SSA acts as a non-linear physical filter. In the problem of interest here, namely how much the pole drift and the free and forced oscillations contribute to the original signal, we simply need to sum them up and compare them with the originals (*cf* . Figure (13.14)).

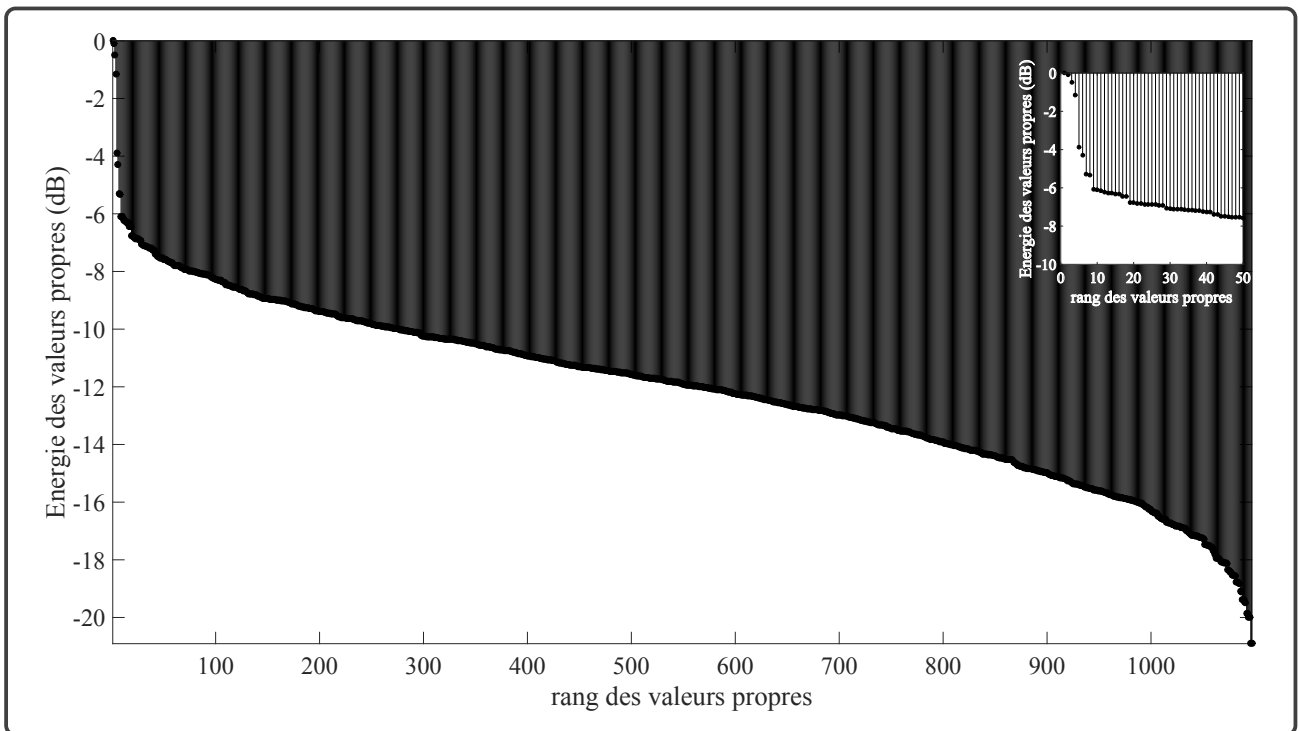
We can see here that considering only the first 5 components obtained from the decomposition of  $\mathbf{X}$ , grouped in 3 sets, largely explains this polar motion signal.

### 3.3 Nonlinear Filtering

We will revisit the example from Figure (12.7), where we filtered a noisy signal with a phase that increased over time using wavelets. The aim here is not to determine whether wavelets or SSA provide better filtering, but rather to compare the implementation of the two approaches on the same signal. This signal has two very interesting advantages for us. Firstly, this sine wave has a phase that varies linearly with time, and secondly, the additive noise is of a magnitude significantly greater than that of the sine wave itself. Using the [ssa\\_04.m](#) script, we obtain the results shown in the figure (13.15). First, at the top, we see the signal to be analysed, followed in the middle by its filtered version obtained by SSA. This almost perfect filter was expected. The Hankel matrix, unlike the EOF autocorrelation matrix, is composed of segments of the signal to be analysed; therefore, if the signal is only the sum of a first-order predictable signal with white noise, then the pattern of the corresponding Hankel matrix will be that of the first-order signal alone (i.e. without the noise, *cf* Figure (13.17)). It will then be easy for the SVD to isolate this noise into low energy eigenvalues (see Figure (13.16)). It is clear that after the eighth eigenvalue, the energy drops below -6 dB, which is less than 50% of the signal amplitude.



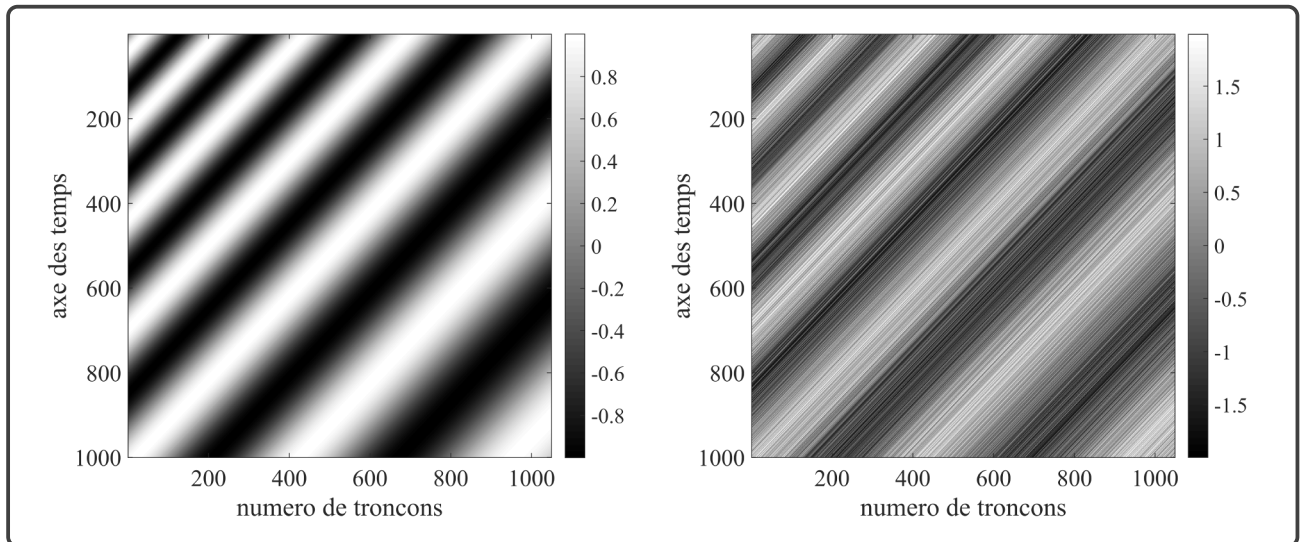
**Figure 13.15:** At the top, in black, is a sine wave with an increasing phase over time, to which we have added noise. In the middle, in red, is the result of the SSA filtering. Below, superimposed, are the original signal (unnoised, in black) and the filtered signal (in red).



**Figure 13.16:** Eigenvalues in dB of the noisy signal from figure (13.15). On the top right is a zoom of these eigenvalues between ranks 1 and 50.

### 3. WHAT SSA CAN DO

---



**Figure 13.17: Hankel matrices: on the left, the signal before adding noise; on the right, the noisy signal.**





---

---

# CHAPTER 14

---

## INVERSE PROBLEM

<b>1</b>	<b>Introduction</b> . . . . .	<b>226</b>
<b>2</b>	<b>An inverse problem example</b> . . . . .	<b>226</b>
<b>3</b>	<b>General structure of inverse problems</b> . . . . .	<b>228</b>
<b>4</b>	<b>A little bit of history</b> . . . . .	<b>232</b>
4.1	The 1960s . . . . .	232
4.2	The 1970s . . . . .	232
4.3	The 1980s . . . . .	233
4.4	The 1990s . . . . .	233
4.5	The 2000s . . . . .	234
<b>5</b>	<b>Our philosophy</b> . . . . .	<b>234</b>

---

# 1 Introduction

The theory of inverse problems often has a bad reputation. Among other things, it is considered to be too mathematical, detached from reality, and impractical. For these reasons, it is seen by many as the preserve of a community admired for its intellectual achievements but not taken seriously when it comes to practical applications with real data collected from the field. While it is true that some geophysicists working on inverse problems deserve such criticism, it is unfair to generalise this negative impression. In fact, the last decade has seen numerous successes of inverse methods. In geophysics, seismic tomography inversions have provided images of the Earth's mantle. In seismology, the most advanced 3D migration methods are based on nonlinear inversion techniques. In geomagnetism, magnetotelluric inversions and those reconstructing the flow of liquid iron at the surface of the outer core can be cited. Meteorology and oceanography have also seen significant progress in inverse problems, the specificity of which (large volumes of data, spatio-temporal variability of models) requires the development of methods such as data assimilation, which is gaining increasing interest in geophysics. In medicine, inverse problems are increasingly used to analyse electroencephalographic and electrocardiographic data. They are also present in more traditional imaging algorithms such as ultrasound or electrical tomography.

The applications mentioned above show that the theory of inverse problems now constitutes a *corpus* of considerable volume, resulting from an evolution of research over the last 40 years. In addition to theoretical advances, computational innovations, which have also progressed spectacularly, now allow the implementation of methods that were considered inapplicable only 20 years ago. Before delving into the history, it is useful to define what an inverse problem is by describing a very simple case that will allow us to illustrate the various concepts we will be exploring throughout this course.

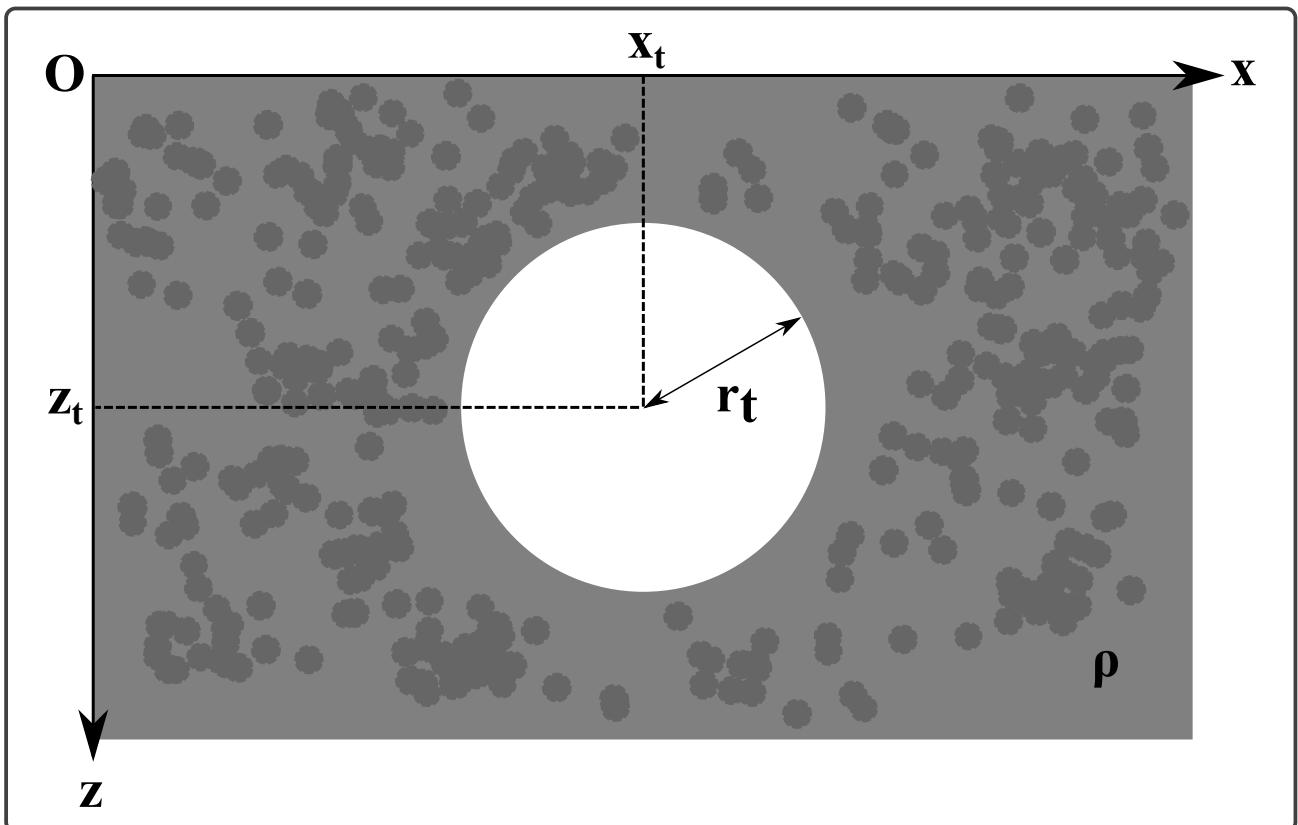
## 2 An inverse problem example

The task is to locate a tunnel by measuring the gravity field along a profile perpendicular to the tunnel axis. We will therefore work in the two-dimensional approximation  $(x, z)$  and assume that the tunnel has a circular cross-section. The vertical component of the gravitational anomaly generated by the tunnel and calculated along the profile  $(x, z = 0)$  is given by

$$g_z(x) = \frac{2\pi \cdot G \cdot \rho \cdot r_t^2 \cdot z_t}{(x - x_t)^2 + z_t^2}, \quad (2.1)$$

## 2. AN INVERSE PROBLEM EXAMPLE

$\rho$  is the density of the rock in which the tunnel of radius  $r_t$  is located. The coordinates of the tunnel axis are  $x_t$  and  $z_t < 0$ . The left-hand side of the equation represents the gravimetric anomaly, which is comparable to the data that will allow us to determine the parameters. This data is the primary information for the inverse problem, as it will allow us to improve our understanding of the tunnel model (see figure 14.1). However, this primary information is not the only information available to us, as some symbols on the right hand side of the equation can be considered as more or less known. This is another source of information, known as *a priori* information.



**Figure 14.1: Diagram of a classic subsurface geophysical problem: detection of a gravimetric anomaly due to the presence of a tunnel.**

Depending on our level of knowledge, the symbols on the right can be either data or unknowns - parameters of the problem. For example, we might assume that the density, radius and horizontal position are known and that the only parameter of the inverse problem is the depth  $z_t$ . In this case we face a non-linear problem because, for example, if  $z_t$  is multiplied by 2, the gravimetric anomaly is certainly not multiplied by 2. It is also possible that the only unknown parameter is  $\rho$ , and then the problem is linear, because when  $\rho$  is multiplied by 2, the gravimetric anomaly is doubled. In the most general case, we can assume that we are solving for the four parameters  $\{\rho, r_t, x_t, z_t\}$ .

Depending on the *a priori* information available - or believed to be available ! - the inverse

---

problem will take different analytical forms. It is clear that the initial parameterisation of an inverse problem implicitly contains a lot of information. The last symbol we have not yet discussed is the position  $x$  at which the gravitational measurements are made. It is usually assumed that this position, known as the independent variable, is perfectly known. We will see that even this variable can be considered as imperfectly known when working within the most general framework of inverse problem theory. An ultimate complication can be added if the accuracy of the equation itself is questioned by considering that the tunnel may not necessarily have a circular cross section, but rather a "potato-like" shape.

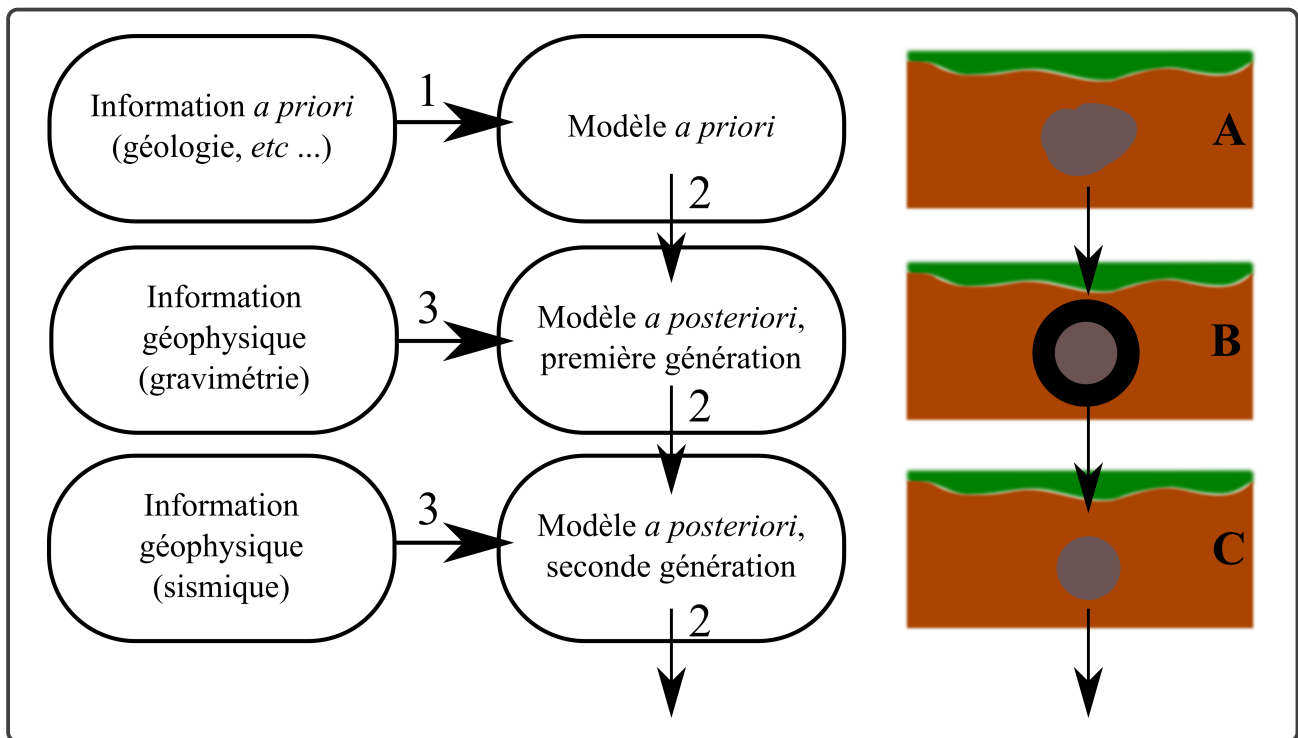
As we can see, the inverse problem of locating a tunnel can be made as complex as desired to illustrate many aspects of inverse problem theory. We will reformulate it throughout the course, gradually increasing its complexity until we reach its most general form.

### 3 General structure of inverse problems

The example presented earlier has characteristics found in most inverse problems, the general structure of which is summarised in Figure (14.2). The goal of inverse problems is to improve our understanding of an object or phenomenon by using more and more information about it. For example, general geological knowledge about a region (top left case) might indicate the presence of a fossil magma chamber underground. The same general knowledge indicates that a magma chamber is a more or less spherical structure composed of rocks with densities within a certain range. This leads to a set of *a priori* models - top, center - an example of which is shown in box A. These models are infinite in number and are often described in vague and non-numerical terms, which means that they are difficult to manipulate on a computer. However, the geologist's expertise allows the design of a gravimetric experiment based on these models, which will provide information-middle left box-that will refine our knowledge of the magmatic chamber in such a way that the set of acceptable models-centre box-is smaller than the *a priori* models. The links 2 and 3 leading to these sets are an inverse problem. The resulting *a posteriori* models are more accurate. For example, the models in box B are spherical, with a possible radius within a relatively narrow range and a depth that is fairly well defined. One could then carry out a second geophysical experiment, such as a seismic test, to provide new information - lower left box - that would allow the positioning of the roof of the magmatic chamber - box C - and thus significantly reduce the set of acceptable models - lower middle box. The links 2 and 3 leading to this new set of *a posteriori* models form a second inverse problem, where the *a priori* models are the *a posteriori* models from the first inverse problem. In this way, inverse

### 3. GENERAL STRUCTURE OF INVERSE PROBLEMS

problems can be linked sequentially to improve our understanding of the magmatic chamber.



**Figure 14.2: General structure of inverse problems**

Figure (14.3) details the structure of a specific inverse problem, specifically the links 1, 2, and 3 from Figure (14.2). Link 1 is a model generator that produces *a priori* models compatible with the initial information available before acquiring geophysical data. The model generation step is crucial in this general framework and represents one of the significant challenges in inverse problem theory. This difficulty arises because geological information is often vague and non-numeric, making it challenging to generate *a priori* models in a computer that adequately cover the wide range of models envisioned by the expert geologist. In this course, we will explore partial solutions to this problem (geostatistics, projection onto convex sets, ...). The next step is arrow 2, which represents the forward problem. This involves selecting *a priori* models and calculating their geophysical response to compare with the data collected in the field *via* arrows 3 and 4, which lead to the decision box. The forward problem lies at the heart of the inverse problem and often needs to be solved many times. Therefore, it is crucial that the forward problem can be solved as quickly as possible on the computer, which sometimes necessitates the use of approximate solutions. For example, in seismics, asymptotic methods (ray tracing) are often faster than wave equation methods (finite differences or finite elements). In the case of the magma chamber, a simplified forward problem might be to assume that the chamber is spherical. It is up to the expert geologist to decide whether such an approximation is acceptable given his *a priori* knowledge; if the magma chamber could be oblong, it

---

is clear that the spherical approximation does not allow proper exploration of the range of models the geologist has in mind. In such cases, the forward problem must be adapted to use, for example, ellipsoidal shapes, which may better suit the geologist's ideas. The decision step allows the selection of *a priori* models that are acceptable and will belong to the set of *a posteriori* models. This set is the solution to the inverse problem, and it can sometimes be a challenge to present it in a way that is simple and easy for the user to understand. It is sometimes possible to create a visual representation, such as a film showing the *a posteriori* models in proportion to their likelihood, but this is not always very meaningful.

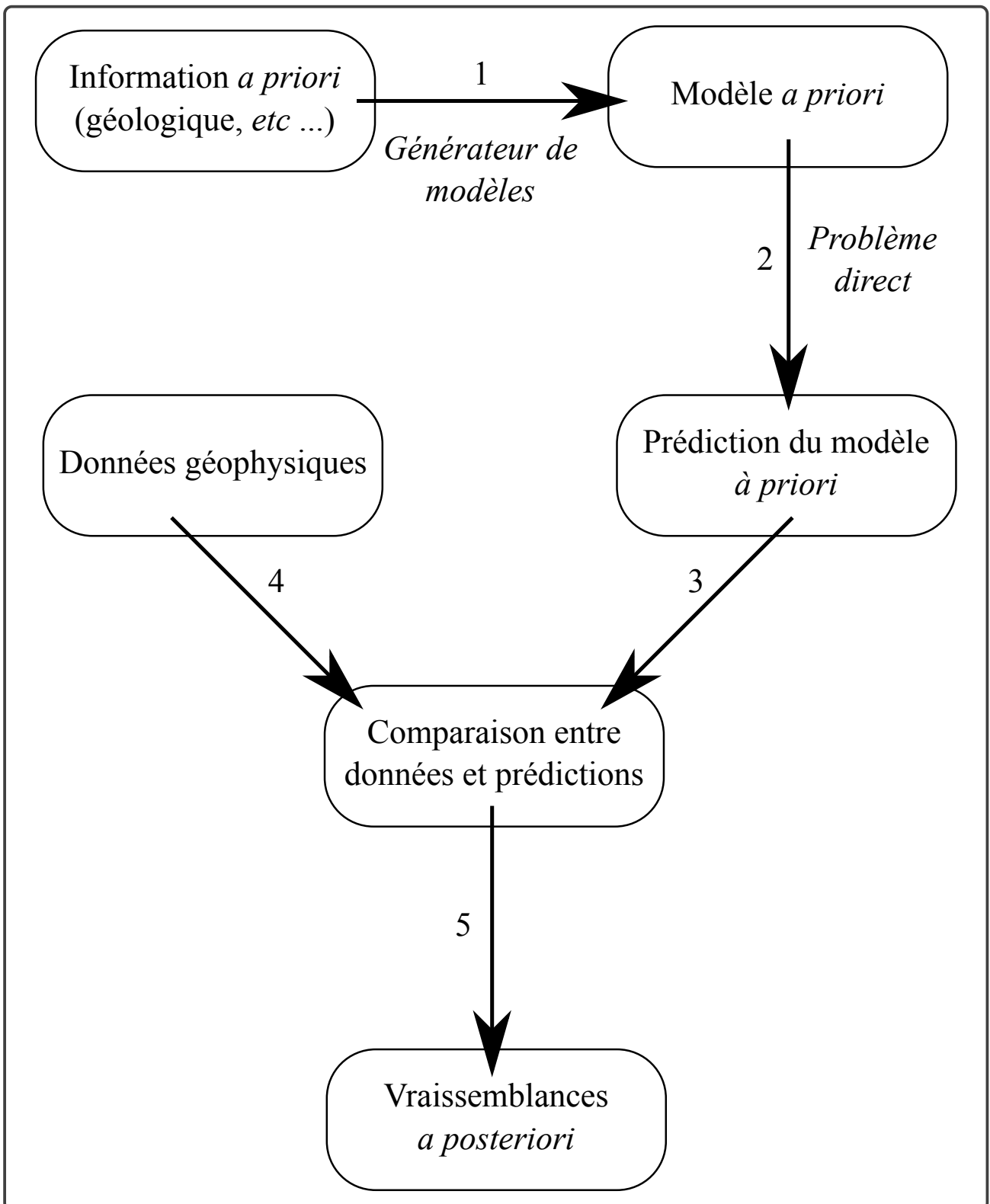


Figure 14.3: General Structure of Inverse Problems



---

## 4 A little bit of history

### 4.1 The 1960s

Inverse problems were introduced into geophysics towards the end of the 1960s when [Backus et Gilbert \(1967\)](#), [Backus et Gilbert \(1968\)](#), [Backus et Gilbert \(1970\)](#) published a series of theoretical papers laying the foundations of the theory. Numerous papers followed, either within the same theoretical framework or focusing on specific applications. The 1970s was thus a period of considerable development in the theory of inverse problems. This was particularly true for the theory of linear and linearised problems. Several fundamental principles were established, such as that the statistical uncertainty in the parameters of a model decreases as the resolution of the model - its "fineness" - increases. Although the notion of an ill-posed problem dates back to the early twentieth century ([Hadamard](#)), it became commonplace, and it was recognised that geophysical data alone are generally not sufficient to produce an unambiguous model. Basic algorithms were proposed to deal with these challenges as efficiently as possible, such as inversion by singular value decomposition and singular vectors [Jackson \(1972\)](#), which had been proposed much earlier in applied mathematics [Levenberg \(1944\)](#), [Penrose \(1955\)](#).

### 4.2 The 1970s

It was in the early 1970s that a different perspective was proposed by [Franklin \(1970\)](#), who published a paper explaining that certain ill-posed linear inverse problems become well-posed when formulated in probabilistic terms. The title of this paper, "*Well-posed stochastic extension of ill-posed linear problems*", implies that if the inverse problem is formulated in terms of finding the probability density of various models within the *a priori* model space, then the solution sought (i.e. the probability density) is unique. In this sense, the problem is well-posed. At the same time, there has been a growing recognition of the importance of *a priori* information that helps to reduce or even eliminate the ill-posed nature of an inverse problem. This information enhances the data provided by geophysical measurements, as if the data were more abundant, of a different nature and less noisy. *a priori* information is also used to introduce constraints on the parameters being sought. For example, in gravimetry the constraint that density must be positive can be applied. Unfortunately, *a priori* information proved difficult to incorporate into the formalisms of the time [Jackson \(1979\)](#).

### 4.3 The 1980s

The 1980s saw many developments in theory. Applications were also plentiful, but many remained unconvincing, mainly because they produced results that were difficult to integrate into broader frameworks. For example, certain electromagnetic inversion results were difficult to reconcile with geological interpretations, while other seismic inversion results provided little guidance for decisions such as whether to drill an oil well. The main problems in the early 1980s were that inverse methods often paid too little attention to *a priori* information, the geological nature of which did not fit easily into highly mathematical formalisms. Even the joint inversion of different geophysical data (seismic + gravimetry, *etc.*) remained rare (Vozoff et Jupp (1975), (Lines, Schultz, et Treitel 1988)). Another major drawback was that inversion methods often failed to account for the multiplicity of possible solutions resulting from insufficient and noisy data. This was a significant handicap when the inversion was intended to inform decision making. The decade of the 1980s is important because inverse problems began to be developed in other scientific fields such as astrophysics, meteorology and medical imaging. Each of these scientific fields contributed to the improvement of the techniques. For example, medicine made significant innovations in imaging dynamic media and developed methods suitable for inverse problems where parameters vary over time. Meteorology invented data assimilation methods, useful when new data are constantly arriving and need to be incorporated into an inversion.

### 4.4 The 1990s

The article by Franklin Franklin (1970) laid the foundations for a stochastic approach to inverse problems, but it took about twenty years for this approach to become commonplace. Among the foundational papers in the probabilistic approach to inverse problems is that of Tarantola et Valette (1982), published in 1982, where the authors establish the basis for inversion in terms of probability densities of parameter values. This perspective has its roots in the work of Bayes (1702-1761) (Barnard et Bayes, 1958) and has been the subject of numerous publications. In this course we will see that the *Bayesian* approach to inverse problems is very flexible and allows *a priori* information to be explicitly considered. However, it is only recently that this approach has become popular, largely because we now have sufficiently powerful computers to take full advantage of its benefits. Advances in computing have revived algorithms published in the 1950s that were impractical for intensive use at the time. This is the case of the Metropolis algorithm (Metropolis et al., 1953), proposed in 1953, shortly after the advent of the first computers.

---

## 4.5 The 2000s

Paradoxically, after a long period of heavy mathematisation accompanying the development of approximate methods (gradient methods, perturbations, *etc*), work on inverse problems has become more refined and is currently focused on the challenging problem of incorporating *a priori* information and solving highly nonlinear problems. Several global solution search algorithms, such as simulated annealing, which we will explore later, can now be implemented for inverse problems of realistic complexity (i.e. combinatorial complexity). These methods have been applied in geophysics to solve inverse problems in seismology, seismic imaging, electrical tomography, *etc*. Considerable effort is devoted to improving these algorithms (simulated annealing, genetic algorithms, neural networks, *etc*), which are based on intensive computations and require immense computing power. Continuing advances in computing now make it possible to access affordable computing power via PC *clusters* or the Internet, allowing inverse problems to be formulated realistically, i.e. using realistic models. Within the next decade, computers will be powerful enough to solve many inverse problems, and the next decade should also see inversion methods becoming more widely used and integrated into the geophysicist's toolbox alongside signal processing techniques.

## 5 Our philosophy

The series of examples we will see is merely an introduction. We have chosen to focus more specifically on the underlying philosophy of inverse problems, and to describe only a few techniques that are both easy to implement and general enough to be applicable in a wide range of cases. We have chosen to frame the inverse problem in terms of information theory because we believe this is the most general way to approach the subject. Indeed, one could say that solving an inverse problem involves transporting information. The transport of information can be subtle and may not only rely on physics and mathematics, but also require considerable expertise. An example of this is the inverse problems in palaeontology, which involve reconstructing the life history of an animal from an incomplete skeleton. The approach we will take is useful for understanding the difficulty of palaeontologists' tasks, but the mathematics we develop will certainly not be of much help. A major problem with inverse problem theory is that it relies on a mathematical formulation that is difficult to apply to the natural sciences. As a result, many geological inverse problems still defy rigorous theoretical approaches. Some attempts have been made through geostatistics, which has gained prominence for its ability to 'mathematise' geological information. However, a true theory of inverse problems applicable to geology remains to be established.

---

---

# CHAPTER 15

---

## INFORMATION & INVERSE PROBLEMS

<b>1</b>	<b>The definition of information</b> . . . . .	<b>237</b>
1.1	Information and Complexity . . . . .	237
1.2	Information and Probabilities . . . . .	238
1.3	Equally likely answers = maximum information . . . . .	240
1.4	About the Tunnel . . . . .	242
<b>2</b>	<b>Mutual Information</b> . . . . .	<b>242</b>
2.1	Coupling Information . . . . .	242
2.2	Conditional information . . . . .	245
2.3	About the Tunnel . . . . .	246
<b>3</b>	<b>Case of continuous distributions</b> . . . . .	<b>247</b>
3.1	There is a problem ! . . . . .	247
3.2	A new definition of Information . . . . .	248
3.3	Information conjunction . . . . .	249
<b>4</b>	<b>Direct problem = information</b> . . . . .	<b>250</b>
4.1	Still in the tunnel . . . . .	250
4.2	Direct problem = conditional probability . . . . .	251

---

<b>5</b>	<b>Inverse problem = information transfer</b>	<b>251</b>
5.1	<i>a posteriori</i> conditional information	251
5.2	The Bayes formula (discrete events)	252
5.3	The generalised Bayes formula (continuous case)	252

# 1 The definition of information

## 1.1 Information and Complexity

We will concentrate on defining information quantitatively, so that it can be treated as a measurable quantity. The definition of information that we will adopt is the one proposed by [Léon Brillouin](#) in 1959 ([Brillouin, 1959](#)), which is based on statistical considerations. Consider a problem with an a priori number of possible answers equal to  $N$ , for which we have no information. Under these conditions, all possible answers are equally probable, and we will say that the information  $I$  needed to uniquely determine the number of a posteriori answers is defined by

$$I = \ln N \tag{1.1}$$

The greater the number of *a priori* answers, the more information is needed to obtain a unique *a posteriori* answer. This is intuitive. The unit of information is the *nep* when the natural logarithm is used in the definition above; it becomes the *digit* for the decimal logarithm and the *bit* for the base-2 logarithm. Consider the example of a problem where the number of a priori answers is limited to  $N = 2$ . The information needed to solve this problem is  $I = \ln 2 \simeq 0.693$  nep.

The choice of a logarithmic function is due to the desire for information to have the property of additivity. For example, consider two independent problems with a priori numbers of answers  $N_1$  and  $N_2$  respectively. The number of answers to the combined problems is therefore,

$$N_{1,2} = N_1 \times N_2, \tag{1.2}$$

which gives,

$$\begin{aligned} I_{1,2} &= \ln(N_1 \times N_2) \\ &= \ln(N_1) + \ln(N_2) \\ &= I_1 + I_2 \end{aligned} \tag{1.3}$$

The information needed to solve both problems simultaneously is simply the sum of the individual pieces of information. This property also corresponds to our intuition. If the number of *a*

*posteriori* answers is no longer 1, but  $N'$ , then the information gained is given by,

$$\begin{aligned}
 I' &= \ln\left(\frac{N}{N'}\right) \\
 &= \ln N - \ln N' \\
 &< I.
 \end{aligned}
 \tag{1.4}$$

We can verify that this expression correctly reduces to the one previously discussed when the *a posteriori* answer is unique. It also shows that the information needed to partially solve a problem is less than the information needed to fully solve it.

## 1.2 Information and Probabilities

Let us now consider the case where the possible *a priori* answers are no longer equally probable. Each answer  $R_i$  is associated with a probability - a likelihood -  $p_i$ . Of course we do,

$$\sum_i p_i = 1.
 \tag{1.5}$$

Let us return to the simple problem. The set of *a priori* answers contains only two elements,

$$\mathcal{R} = \{R_1, R_2\}.
 \tag{1.6}$$

We know that the information needed to solve this problem is about 0.693 nep if the two *a priori* answers are equally probable. Let us express the probabilities as,

$$p_1 = \frac{N_1}{N_1 + N_2}, \quad p_2 = \frac{N_2}{N_1 + N_2},
 \tag{1.7}$$

where  $N_1$  and  $N_2$  are positive integers. The complexity  $N$  of the problem, whose *a priori* answers have probabilities  $p_i$ , is equal to the number - divided by  $N_1 + N_2$  - of ways in which a sequence of  $N_1 + N_2$  symbols  $R_i$  can be formed, knowing that there are  $N_1$  equal to  $R_1$  and, obviously,  $N_2$  equal to  $R_2$ . A simple counting calculation shows that the complexity,

$$\begin{aligned}
 N &= \frac{(N_1 + N_2)(N_1 + N_2 - 1)(N_1 + N_2 - 2) \times \cdots \times (N_2 + 1)}{2 \times 3 \times \cdots \times N_1} \\
 &= \frac{(N_1 + N_2)!}{N_1! \times N_2!}
 \end{aligned}
 \tag{1.8}$$

## 1. THE DEFINITION OF INFORMATION

---

where the division by  $N_1!$  is due to the fact that the  $N_1$  symbols  $R_1$  are interchangeable. If we compute the information from  $N$ , we get,

$$\begin{aligned} I &= \ln N \\ &= [\ln(N_1 + N_2)! - \ln N_1! - \ln N_2!]. \end{aligned} \quad (1.9)$$

If  $N_1$  and  $N_2$  are chosen large enough - that is,  $> 100$  - we can use [Stirling's](#) formula,

$$\ln Q! \simeq Q(\ln Q - 1), \quad (1.10)$$

to find,

$$\begin{aligned} I &\simeq (N_1 + N_2) \ln(N_1 + N_2) - N_1 \ln N_1 - N_2 \ln N_2 \\ &= -(N_1 + N_2) \left[ \frac{N_1}{N_1 + N_2} \ln \frac{N_1}{N_1 + N_2} + \frac{N_2}{N_1 + N_2} \ln \frac{N_2}{N_1 + N_2} \right] \\ &= -(N_1 + N_2) [p_1 \ln p_1 + p_2 \ln p_2]. \end{aligned} \quad (1.11)$$

The last expression still depends on  $N_1$  and  $N_2$ , which is problematic because these numbers are not uniquely determined. For example, the probabilities  $1/3$  and  $2/3$  can be represented either by  $N_1 = 1$  and  $N_2 = 2$  or by  $N_1 = 2000$  and  $N_2 = 4000$ . However, the information should not depend on any particular choice. Therefore the complexity  $N$  has to be normalised to get an acceptable expression. To do this, it is sufficient to divide the above information by the number of realisations, so that,

$$I = \frac{1}{N_1 + N_2} [\ln(N_1 + N_2)! - \ln N_1! - \ln N_2!], \quad (1.12)$$

which results in a measure of information that depends only on the laws of probability,

$$I = -[p_1 \ln p_1 + p_2 \ln p_2]. \quad (1.13)$$

At the level of complexity, this renormalisation amounts to a choice,

$$N = \left[ \frac{(N_1 + N_2)!}{N_1! \times N_2!} \right]^{1/(N_1 + N_2)}. \quad (1.14)$$

Although obtained using [Stirling's](#) approximation, the expression for the information can be



considered exact in the sense that  $N_1$  and  $N_2$  can always be chosen to be as large as desired. If you consider one of the answers to be certain,

$$p_1 = 1 \text{ et } p_2 = 0 \Rightarrow I = 0 \text{ nep} \Rightarrow N = 1, \quad (1.15)$$

we find that the information needed to solve the problem is zero, which is obvious since the answer to the problem is known *a priori*. The complexity is then 1. Other examples,

$$p_1 = \frac{4}{10} \text{ et } p_2 = \frac{6}{10} \Rightarrow I \simeq 0.673 \text{ nep} \Rightarrow N = 1.96, \quad (1.16)$$

$$p_1 = \frac{2}{10} \text{ et } p_2 = \frac{8}{10} \Rightarrow I \simeq 0.500 \text{ nep} \Rightarrow N = 1.65, \quad (1.17)$$

$$p_1 = \frac{1}{10} \text{ et } p_2 = \frac{9}{10} \Rightarrow I \simeq 0.325 \text{ nep} \Rightarrow N = 1.38, \quad (1.18)$$

show that the information required decreases as the probability of one of the answers decreases. The associated complexity then varies from 2 to 1.

The calculations just performed can be generalised to any number of *a priori* answers  $R_i$  associated with probabilities  $p_i$ . We then obtain the definition originally proposed by [Claude Shannon](#),

$$I = - \sum_i p_i \ln p_i \quad (1.19)$$

### 1.3 Equally likely answers = maximum information

We will now show that the information required to answer a question is maximized when the  $N$  *a priori* answers  $R_i$  are equally probable. Referring to equation (1.19), we seek the probabilities  $p_i$  such that,

$$\frac{\partial I}{\partial p_i} = 0, \quad (i = 1, \dots, N), \quad (1.20)$$

which returns to the definition of information,

$$\frac{\partial}{\partial p_i} \sum_k p_k \ln p_k = 0, \quad (i = 1, \dots, N), \quad (1.21)$$

## 1. THE DEFINITION OF INFORMATION

---

which, when expanded, yields,

$$\ln p_i + 1 + \sum_{k \neq i} \frac{\partial p_k}{\partial p_i} \ln p_k + \sum_{k \neq i} \frac{\partial p_k}{\partial p_i} = 0, \quad (i = 1, \dots, N). \quad (1.22)$$

Note that the probabilities are normalised, *ie*

$$p_k = 1 - \sum_{i \neq k} p_i, \quad (1.23)$$

which implies,

$$\frac{\partial p_{k \neq i}}{\partial p_i} = -1 \quad (1.24)$$

and consequently the condition (1.22) becomes,

$$\ln p_i - \sum_{k \neq i} \ln p_k - N + 2 = 0, \quad (i = 1, \dots, N). \quad (1.25)$$

By evaluating the equation (1.25) for two different indices  $i$  and  $j$  and performing the subtraction, we obtain

$$\ln p_i - \sum_{k \neq i} \ln p_k - \ln p_j + \sum_{k \neq j} \ln p_k = 0, \quad (i = 1, \dots, N).$$

Finally, by cancelling the opposing terms, we find

$$2 \ln p_i - 2 \ln p_j = 0$$

which is satisfied when the two probabilities are equal,

$$p_i = p_j \quad (1.26)$$

Strictly speaking, the proof should be completed with an analysis of the signs of the second partial derivatives to establish that the *extremum* identified is indeed a *maximum*.

## 1.4 About the Tunnel

Let's illustrate equation (1.19) using our tunnel-finding problem. Suppose the only unknown parameter is the depth  $z_t$  of the centre of the tunnel. One way of framing the problem is to say that we need to find  $z_t$  within a set of possible values,

$$\{R_i\} = \{5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}. \quad (1.27)$$

If all depths are *a priori* equally likely, the information needed to find the correct depth is given by,

$$I = 2.773 \text{ nep}. \quad (1.28)$$

Geological information may lead us to believe that depths below 10 metres or above 18 metres are unlikely. We can express this using the following probability table,

$$\left\{ \frac{1}{34}, \frac{1}{34}, \frac{1}{34}, \frac{1}{34}, \frac{1}{34}, \frac{3}{34}, \frac{3}{34}, \frac{3}{34}, \frac{3}{34}, \frac{3}{34}, \frac{3}{34}, \frac{3}{34}, \frac{3}{34}, \frac{1}{34}, \frac{1}{34} \right\}. \quad (1.29)$$

The information required is now given by,

$$I = 2.654 \text{ nep} \quad (1.30)$$

and we can therefore say that the geological information provided is equivalent to

$$I_{\text{géol}} = 2.773 - 2.654 = 0.119 \text{ nep}. \quad (1.31)$$

## 2 Mutual Information

### 2.1 Coupling Information

Suppose the problem to be solved involves finding two answers from two *a priori* sets of answers,  $A_i$  and  $B_j$ . Let  $p(A_i, B_j)$  denote the probabilities of all possible *a priori* pairs  $(A_i, B_j)$ . We have,

$$\sum_i \sum_j p(A_i, B_j) = 1 \quad (2.1)$$

and we have the marginal probabilities,

$$\begin{aligned} p(A_i) &= \sum_j p(A_i, B_j), \\ p(B_j) &= \sum_i p(A_i, B_j). \end{aligned} \tag{2.2}$$

It can be shown in a direct way that,

$$\sum_i p(A_i) = \sum_j p(B_j) = \sum_i \sum_j p(A_i) \times p(B_j) = 1. \tag{2.3}$$

The coupling information is given by,

$$I(A, B) = - \sum_i \sum_j p(A_i, B_j) \ln p(A_i, B_j), \tag{2.4}$$

and the marginal information,

$$\begin{aligned} I(A) &= - \sum_i p(A_i) \ln p(A_i) \\ &= - \sum_i \sum_j p(A_i, B_j) \ln p(A_i), \end{aligned} \tag{2.5}$$

$$\begin{aligned} I(B) &= - \sum_j p(B_j) \ln p(B_j) \\ &= - \sum_j \sum_i p(A_i, B_j) \ln p(B_j). \end{aligned} \tag{2.6}$$

Note that,

$$\begin{aligned} I(A) + I(B) &= - \sum_i \sum_j p(A_i, B_j) \ln [p(A_i) \times p(B_j)] \\ &= - \sum_i \sum_j p(A_i, B_j) \ln [p(A_i, B_j) + q(A_i, B_j)] \\ &= - \sum_i \sum_j p(A_i, B_j) \ln \left[ p(A_i, B_j) \left( 1 + \frac{q(A_i, B_j)}{p(A_i, B_j)} \right) \right] \\ &= I(A, B) - \sum_i \sum_j p(A_i, B_j) \ln \left[ 1 + \frac{q(A_i, B_j)}{p(A_i, B_j)} \right]. \end{aligned} \tag{2.7}$$

To proceed further, it is necessary to prove a useful result. Let us consider the function,

$$f(x) = x - \ln(1+x) \quad (2.8)$$

defined in the interval  $] -1, +\infty[$ . It is easily verified that  $f(0) = 0$ ,  $f'(0) = 0$  et  $f'(x) = x/(1+x)$ , which gives,

$$\begin{aligned} f'(x > 0) > 0 &\implies f(x > 0) > f(0) \\ f'(-1 < x < 0) < 0 &\implies f(-1 < x < 0) > f(0). \end{aligned} \quad (2.9)$$

The function  $f$  is therefore minimal at  $x = 0$  and strictly positive everywhere else. Hence we have,

$$x \geq \ln(1+x) \quad (2.10)$$

in the whole domain of  $f$ . Now set

$$x \equiv \frac{q(A_i, B_j)}{p(A_i, B_j)}, \quad (2.11)$$

we have,

$$-q(A_i, B_j) \leq -p(A_i, B_j) \ln \left[ 1 + \frac{q(A_i, B_j)}{p(A_i, B_j)} \right], \quad (2.12)$$

and it follows,

$$I(A, B) - \sum_i \sum_j q(A_i, B_j) \leq I(A, B) - \sum_i \sum_j p(A_i, B_j) \ln \left[ 1 + \frac{q(A_i, B_j)}{p(A_i, B_j)} \right]. \quad (2.13)$$

Furthermore, the very definition of  $q(A_i, B_j)$  implies that,

$$\sum_i \sum_j q(A_i, B_j) = 0. \quad (2.14)$$

## 2. MUTUAL INFORMATION

---

Combining the various results obtained, we then find that,

$$I(A, B) \leq I(A) + I(B), \quad (2.15)$$

*ie* the coupling information is less than or equal to the sum of the marginal information. If the two answers to be found are independent, the probability is

$$p(A_i, B_j) = p(A_i) \times p(B_j) \quad (2.16)$$

and equality holds,

$$I(A, B) = I(A) + I(B). \quad (2.17)$$

### 2.2 Conditional information

The probability of the pairs can be expressed in the form,

$$\begin{aligned} p(A_i, B_j) &= p(A_i) \times p(B_j|A_i) \\ &= p(B_j) \times p(A_i|B_j) \end{aligned} \quad (2.18)$$

where  $p(B_j|A_i)$  is the conditional probability of  $B_j$  given that  $A_i$  has occurred. The coupling information then takes the form,

$$\begin{aligned} I(A, B) &= - \sum_i \sum_j p(A_i, B_j) \ln p(A_i, B_j) \\ &= - \sum_i \sum_j p(A_i, B_j) \ln [p(A_i) \times p(B_j|A_i)] \\ &= - \sum_i \sum_j p(A_i, B_j) \ln p(A_i) - \sum_i \sum_j p(A_i, B_j) \ln p(B_j|A_i) \\ &= - \sum_i p(A_i) \ln p(A_i) - \sum_i \sum_j p(A_i, B_j) \ln p(B_j|A_i) \\ &= I(A) + I(B|A), \end{aligned} \quad (2.19)$$

where the conditional information has been introduced

$$\begin{aligned}
 I(B|A) &= -\sum_i \sum_j p(A_i, B_j) \ln p(B_j|A_i) \\
 &= -\sum_i \sum_j p(A_i, B_j) \ln \frac{p(A_i, B_j)}{p(A_i)}.
 \end{aligned} \tag{2.20}$$

This information is furthermore such that,

$$I(A) + I(B) \geq I(A, B) = I(A) + I(B|A), \tag{2.21}$$

which implies that,

$$I(B|A) \leq I(B). \tag{2.22}$$

This expression means that the information about the responses  $B_j$  given the response  $A$  is reduced compared to the information about the responses  $B_j$  alone. If the responses  $A_i$  and  $B_j$  are independent – that is, if knowing  $A$  does not provide any additional information about  $B$  – the conditional information is equal to the marginal information.

## 2.3 About the Tunnel

Let us illustrate equation (2.22) using our tunnel search problem, assuming that we want to determine the depth  $z_t$  of the centre of the tunnel at two locations separated by several tens of metres. So the problem is to find the two depths  $z_t$ ,

$$\begin{aligned}
 \{A_i\} &= \{5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\} \\
 \{B_j\} &= \{5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}.
 \end{aligned} \tag{2.23}$$

As we have shown, if all depths are equally likely, the information is

$$I(A) = I(B) = 2.773 \text{ nep.}$$

If the depth estimates are independent, then

$$I(A, B) = I(A) + I(B) = 5.546 \text{ nep.}$$

### 3. CASE OF CONTINUOUS DISTRIBUTIONS

---

On the other hand, if technical information indicates that the difference between the two depths should not exceed 3 metres, because it is known that the tunnel does not have a slope greater than a certain value, then it becomes clear that knowing the first depth provides information about the second, still unknown, depth. This information limits the number of a priori possible depths, reducing the number of *a priori* answers from  $N = 16^2$  to only  $N = 100$ . This reduces the coupling information

$$I(A, B) = \ln 100 = 4.605 \text{ nep},$$

and the conditional information in this case is

$$I(B|A) = I(A, B) - I(A) = 4.605 - 2.773 = 1.832 \text{ nep}.$$

The information related to the tunnel slope constraint is equal to the difference between the coupling information calculated with and without the constraint,

$$I_{\text{pente}} = 5.546 - 4.605 = 0.941 \text{ nep}.$$

## 3 Case of continuous distributions

### 3.1 There is a problem !

The definition of information given by [Shannon](#) and that we have seen so far can be generalised to the case of probability densities  $\rho(x)$  where  $x$  can vary continuously. This is, for example, the case of the tunnel depth, which we initially assumed to take discrete values, whereas in reality it can take any value within a given interval *a priori*. The [Shannon](#) information for a probability density  $\rho$  is given by

$$I_{\text{Shannon}} \equiv - \int \rho(x) \ln \rho(x) dx. \quad (3.1)$$

Note by the way that the density  $\rho$  is such that,

$$\int \rho(x) dx = 1, \quad (3.2)$$

and that we can have  $\rho(x) > 1$  for certain values of  $x$ .



Now consider the case of determining the tunnel depth when the depth is *a priori* contained within the interval  $S = [z_{inf}, z_{sup}]$ . If we assume that the depths are equally likely, then,

$$\rho(z_t \in S) = \frac{1}{z_{sup} - z_{inf}}, \quad \rho(z_t \notin S) = 0. \quad (3.3)$$

The information associated with this probability density is given by,

$$I_{Shannon} = -\ln \frac{1}{z_{sup} - z_{inf}}. \quad (3.4)$$

If  $z_{sup} - z_{inf} = 1$  metres, we find that  $I_{Shannon} = 0$ , which, according to what we have seen so far, implies that we have the answer to the question of determining the depth of the tunnel. However, this is not the case, since the depth is contained within an interval of one metre in width. Worse still, if we now express the distances in centimetres, we find that the associated information is  $I_{Shannon} = \ln 100!$  This means that the quantification of information depends on the choice of units, which means that information loses the absolute character we had previously ascribed to it.

In his 1948 paper, [Shannon \(1948\)](#) notes this problem and points out that it is not serious, since what really matters is the variation of information for a fixed choice of units.

### 3.2 A new definition of Information

In their paper, [Albert Tarantola](#) and [Bernard Valette](#) ([Tarantola et Valette 1982](#)) propose a definition of information that is invariant under changes in coordinate systems or units. They propose,

$$I_{TarVal} \equiv \int \rho(x) \ln \frac{\rho(x)}{\mu(x)} dx, \quad (3.5)$$

where the probability density  $\mu$  represents the *maximum* state of ignorance about the variable  $x$ . Note that with this definition, the information obtained no longer represents the information needed to answer the question posed, but rather the information available to answer the question. It is, in a sense, the complementary information to that considered previously.

How should the *maximum* state of ignorance be chosen? The idea is that this state should be the one that provides the least information about the answer to the question posed. A natural choice is to use a uniform distribution over the *a priori* interval, since we have seen that the case of equally probable outcomes corresponds to the state that requires the most information to answer the question. However, this choice is not always appropriate. Consider, for example, the problem of

### 3. CASE OF CONTINUOUS DISTRIBUTIONS

---

locating an earthquake on the Earth's surface. If we work in Cartesian coordinates  $(x, y)$ , the natural choice is,

$$\mu(x, y) = \text{constant}. \quad (3.6)$$

However, if we work in spherical coordinates  $(\theta, \phi)$ , where the surface element is  $ds = R \sin \theta, d\theta, d\phi$ , the probability density corresponding to an equally probable distribution with respect to the surface is given by,

$$\mu(\theta, \phi) = \text{constant} \cdot R \cdot \sin \theta. \quad (3.7)$$

### 3.3 Information conjunction

The introduction of the *maximum* state of ignorance requires an adaptation of the formula for the conjunction of information seen in the case of discrete events. Let  $\sigma$  be the probability density associated with the information  $I_\sigma$  corresponding to the conjunction of two pieces of information  $I_1$  and  $I_2$ , whose respective probability densities are  $\rho_1$  and  $\rho_2$ . We then have,

$$\begin{aligned} I_\sigma &= \int \sigma(x) \ln \frac{\sigma(x)}{\mu(x)} dx \\ I_1 &= \int \rho_1(x) \ln \frac{\rho_1(x)}{\mu(x)} dx \\ I_2 &= \int \rho_2(x) \ln \frac{\rho_2(x)}{\mu(x)} dx. \end{aligned} \quad (3.8)$$

These different pieces of information must be combined according to logical rules that take into account the existence of the *maximum* state of ignorance. These rules are,

$$I_\sigma = (I_1 \& I_2) = (I_2 \& I_1) \quad \text{commutation} \quad (3.9)$$

$$\rho_1(x) = 0 \Rightarrow \sigma(x) = 0 \quad \text{absorption} \quad (3.10)$$

$$\rho_1(x) = \mu(x) \Rightarrow \sigma(x) = \rho_2(x) \quad \text{non-information} \quad (3.11)$$

---

The condition 3.9 simply states that the conjunction of information must be commutative. This requires a symmetric form of  $\sigma$  with respect to the densities  $\rho_1$  and  $\rho_2$ . The condition 3.10 corresponds to the fact that if one of the probability densities is zero for certain values of  $x$ , then the density  $\sigma$  must also be zero for those values. This absorption property is analogous to multiplication, which implies that  $\sigma$  must be a function of the product  $\rho_1 \times \rho_2$ , which automatically satisfies the commutativity imposed by the first condition. The third condition 3.11 takes into account the maximum ignorance  $\mu$ . Finally, considering the form of  $\sigma$  dictated by the first two conditions, we find that,

$$\sigma(x) = \frac{\rho_1(x)\rho_2(x)}{\mu(x)}. \quad (3.12)$$

## 4 Direct problem = information

### 4.1 Still in the tunnel

We will start with our favourite example to illustrate and intuitively grasp the developments that will follow. To do this, we will rephrase it slightly to introduce the concept of a direct problem. We have seen that knowing one depth can provide information to determine a second depth. That is, providing a piece of data - the first depth - can improve the information we have about an unknown - the second depth. Building on this observation, it is easy to modify the formulation of the problem slightly and assume that the data is no longer the first depth, but a measurement of the gravitational field. Similarly, the constraint on the slope of the tunnel - which allowed us to 'connect' the two depths - can be replaced by [Newton's](#) law, which relates the tunnel depth to the gravitational anomaly. You might think that [Newton's](#) law is perfectly known and, unlike the slope constraint, leaves no room for tolerance. This is incorrect; there are many reasons why [Newton's](#) law is 'fuzzy' when applied to our tunnel! For example, we do not know the exact density of the surrounding rock, we are not sure if the tunnel is perfectly cylindrical, etc. In short, the direct problem of calculating the gravitational anomaly as a function of depth is an imprecise law that can be described by a distribution of conditional probabilities, which we will denote by,

$$p(g|z_t). \quad (4.1)$$

## 4.2 Direct problem = conditional probability

The perspective we have just illustrated with the tunnel example is extremely powerful because it allows not only to relax the rigidity of the mathematical relations describing the direct problem, but also to take measurement uncertainties into account. This is certainly what makes the information-theoretic approach to inverse problems so attractive. From the most general point of view, the direct problem, which relates the data to the parameters that are the unknowns of the inverse problem, is thus expressed in terms of a conditional probability density,

$$\text{DIRECT PROBLEM} = \text{PROBABILITY}(\text{DATA}|\text{PARAMETERS}). \quad (4.2)$$

## 5 Inverse problem = information transfer

### 5.1 *a posteriori* conditional information

We have seen that the coupling information is given by,

$$I(A, B) = I(A) + I(B|A) = I(B) + I(A|B), \quad (5.1)$$

which implies,

$$I(A|B) = I(A) - I(B) + I(B|A). \quad (5.2)$$

This equation provides the solution to an inverse information transfer problem: the *a posteriori* conditional information we can obtain about the answer  $A$  is equal to the *a posteriori* information about  $A$  minus the *a posteriori* information about  $B$  and plus the conditional information about  $B$  given  $A$ . As we have already shown,

$$I(A|B) \leq I(A), \quad (5.3)$$

*ie* the *a posteriori* information needed to know  $A$  is less than the *a priori* information we had. In other words, the *a posteriori* knowledge we have is greater than the *a priori* knowledge. Seen in this way, solving the inverse problem involves increasing our knowledge about the answer  $A$ .

---

## 5.2 The Bayes formula (discrete events)

We will now bridge to the next chapter concerning probabilities. Let's begin with the case of discrete events by explaining solution 5.2,

$$-\sum_i \sum_j p(A_i, B_j) \ln p(A_i|B_j) = -\sum_i \sum_j p(A_i, B_j) \ln \frac{p(A_i) p(B_j|A_i)}{p(B_j)}, \quad (5.4)$$

or, in equivalent terms,

$$\sum_i \sum_j p(A_i, B_j) \ln \left[ \frac{p(A_i|B_j) p(B_j)}{p(A_i) p(B_j|A_i)} \right] = 0. \quad (5.5)$$

If we want this relationship to hold in general, the logarithmic term must be identically zero, so we have,

$$p(A_i|B_j) = \frac{p(A_i) p(B_j|A_i)}{p(B_j)}. \quad (5.6)$$

This relation is known as the Bayes formula. It plays a very important role in probability theory. In the following chapters we will see how this formula can be used to solve inverse problems.

By recalling this, a slightly different form of the Bayes formula can be obtained,

$$p(B_j) = \sum_k p(A_k, B_j) = \sum_k p(A_k) p(B_j|A_k). \quad (5.7)$$

We then find that,

$$p(A_i|B_j) = \frac{p(A_i) p(B_j|A_i)}{\sum_k p(A_k) p(B_j|A_k)}. \quad (5.8)$$

We have just established that the manipulation of information can be reduced to the manipulation of probability laws.

## 5.3 The generalised Bayes formula (continuous case)

Although the Bayes formula (5.8) is indeed used to solve many inverse problems, it is important to remember that its derivation is within the framework of discrete event theory. In the continuous case, one must use the information conjugate seen earlier, for which the resulting probability density

## 5. INVERSE PROBLEM = INFORMATION TRANSFER

---

is given by the equation 3.12, repeated here,

$$\sigma(x) = \frac{\rho_1(x)\rho_2(x)}{\mu(x)}, \quad (5.9)$$

which is the continuous case equivalent of the Bayes formula.



---

---

# CHAPTER 16

---

## BAYESIAN INVERSION

<b>1</b>	<b>Probabilities &amp; Inverse Problems</b>	<b>257</b>
1.1	Probabilities, Frequencies, and Information	257
1.2	Probability Densities	257
1.3	Mathematical expectation value of a function	260
1.4	Multivariate probabilities	260
<b>2</b>	<b>A few common probability distributions</b>	<b>261</b>
2.1	The normal distribution (Gauss)	261
2.2	Generalised Gaussian distributions	261
2.3	The <i>log-normal</i> distribution	261
2.4	The Poisson distribution	262
2.5	The <i>gamma</i> ( $\Gamma$ -) distribution	262
2.6	The beta ( $\beta$ -) distribution	262
2.7	The Pareto distribution	263
2.8	The <i>binomial</i> distribution	263
2.9	The Cauchy distribution	264
2.10	The Weibull distribution	264



---

<b>3</b>	<b>Bayes' formula and inversion"</b>	<b>264</b>
3.1	General solution	264
3.2	Solution for <i>a priori</i> independent data and parameters	265
3.3	Solution for an exact physical law	266
3.4	Solution using Bayes' formula	266
<b>4</b>	<b>The tunnel again</b>	<b>267</b>
4.1	Example 1: one data and one parameter	267
4.2	Example 2: two data and one parameter	269
4.3	Example 3: One data and two parameters	272
4.4	Example 4: Two data and two parameters	276
<b>5</b>	<b>Summary of examples 1, 2, 3 and 4</b>	<b>277</b>

# 1 Probabilities & Inverse Problems

## 1.1 Probabilities, Frequencies, and Information

It is useful to begin with some thoughts on the concept of probability. In its purest sense, the concept of probability is associated with the idea of repeating an experiment in which the outcome is not identical but, on the contrary, varies from trial to trial. The most common example is throwing a dice. The result of a single throw is an integer between 1 and 6. The number obtained from one roll to the next is not necessarily the same. In signal theory, a stochastic process refers to the system under consideration in the experiments. In our example, the stochastic process is the system consisting of the dice, the receiving surface and the thrower. Each throw is a realisation of the stochastic process. The characterisation of a process is done in terms of statistics and in particular probabilities. In the case of dice, we calculate the frequency of occurrence of each possible number. If this frequency of occurrence is calculated from a very large number of throws, to the point where the number can be considered infinite, the frequency of occurrence is called a probability. In this case, the notion of probability is clearly defined and is based on counting within a set of realisations of a stochastic process with a finite number of possible outcomes.

Inverse problem theory uses a notion of probability that is sometimes different from what we have just discussed. Here, probabilities are used to quantify the likelihood of an event. For example, a certain possible depth of the tunnel might be considered unlikely if engineers or geologists consider it unlikely. This is rarely a probability calculated in the same way as a die, i.e. by running statistics on a large number of tunnel depths. Probability is a more ambiguous concept that can, of course, include objective statistical data, but also subjective and more difficult to define information. In fact, many of the probabilities dealt with in inverse problem theory are actually likelihoods. This creates a gap in the theory because the foundations on which our rigorous calculations are based can be questioned. For example, one could move away from traditional probabilities in favour of fuzzy logic, which combines information differently.

## 1.2 Probability Densities

Until now, we have only discussed discrete probabilities calculated for a finite number of possible outcomes. For example, in the case of dice, where the number of outcomes is limited to 6. In this context, probability is a measure that involves counting the elements of the sets under consideration. A measure must satisfy the following basic properties,

- the measure is always positive,,  $M(\cdot) \geq 0$ ;
- the measure of the empty set is zero,  $M(\emptyset) = 0$ ;
- the measure of the entire space is 1,  $M(\Omega) = 1$ ;
- the measure satisfies the additivity property for a collection of disjoint sets,  $M(\Omega_1 \cup \Omega_2) = M(\Omega_1) + M(\Omega_2)$  if  $\Omega_1 \cap \Omega_2 = \emptyset$ .

When working with continuous random variables, a different measure must be adopted, which we will define over an interval  $\mathcal{I}$ , so that for any interval  $\mathcal{A} \subset \mathcal{I}$  we have,

$$M(\mathcal{A}) \equiv \int_{\mathcal{A}} \rho(x) dx \quad (1.1)$$

We have a valid measure if  $\rho(x) \geq 0$ . We will say that  $\rho(x)$  is a probability density function if,

$$\int_{\mathcal{I}} \rho(x) dx = 1 \quad (1.2)$$

The probability density function allows you to calculate the probability that a realisation  $x'$  of the random variable  $x$  lies within a given interval  $\mathcal{A} \subset \mathcal{I}$ ,

$$P(x' \in \mathcal{A}) = \int_{\mathcal{A}} \rho(x) dx \quad (1.3)$$

We can calculate the mathematical expectation value – that is, the mean – of the random variable  $x$ ,

$$\bar{x} \equiv E[x] = \int_{\mathcal{I}} x \rho(x) dx \quad (1.4)$$

and the variance,

$$\sigma_x^2 \equiv E[(x - \bar{x})^2] = \int_{\mathcal{I}} (x - \bar{x})^2 \rho(x) dx \quad (1.5)$$

By generalisation we will define the  $n$ -th central moment as,

$$E[(x - \bar{x})^n] = \int_{\mathcal{I}} (x - \bar{x})^n \rho(x) dx \quad (1.6)$$

## 1. PROBABILITIES & INVERSE PROBLEMS

---

An example of a probability density is given by  $\rho(x) = \pi^{-1/2} \exp(-x^2)$  where the interval  $\mathcal{I} = \mathbb{R}$ .

Indeed, one can verify that:

† the normalisation condition is satisfied,

$$M(\mathcal{I}) = \frac{1}{\sqrt{\pi}} \int_{\mathbb{R}} \exp(-x^2) dx = 1 \quad (1.7)$$

† the positivity of the measure,

$$M(\mathcal{A} \subset \mathcal{I}) = \frac{1}{\sqrt{\pi}} \int_{\mathcal{A}} \exp(-x^2) dx \geq 0, \quad (1.8)$$

† the additivity of the measure,

$$M(\mathcal{A} \cup \mathcal{A}') = M(\mathcal{A}) + M(\mathcal{A}') \quad (1.9)$$

† when  $\mathcal{A} \cap \mathcal{A}' = \emptyset$  we finally have,

$$M(\emptyset) = 0. \quad (1.10)$$

When probability densities are used, Bayes' formula (5.8) takes the form,

$$\rho(y|x) = \frac{\rho(y) p(x|y)}{\int_{\mathcal{I}} \rho(y) p(x|y) dy} \quad (1.11)$$

The function  $\rho(y)$  is the *a priori* probability density, and  $\rho(y|x)$  is the *a posteriori* probability density.

---

### 1.3 Mathematical expectation value of a function

Bayes' formula for probability densities involves the integral,

$$\int_{\mathcal{S}} \rho(y) p(x|y) dy \quad (1.12)$$

More generally, you will often encounter integrals of the form,

$$E[f(x)] \equiv \int_{\mathcal{S}} f(x) \rho(x) dx \quad (1.13)$$

which, as an extension of what we saw in the previous section, we will define as the mathematical expectation value of the function  $f(x)$  with respect to the probability density  $\rho(x)$ .

### 1.4 Multivariate probabilities

The generalisation to the case of multivariate probability densities is immediate by introducing the vector random variable  $\mathbf{x}$  and the function  $\rho(\mathbf{x})$ . As before, we define the mean by,

$$\bar{\mathbf{x}} \equiv \int_{\mathcal{S}} \mathbf{x} \rho(\mathbf{x}) d\mathbf{x} \quad (1.14)$$

and the covariance matrix by,

$$C_{ij}(\bar{\mathbf{x}}) \equiv \int_{\mathcal{S}} (x_i - \bar{x}_i) (x_j - \bar{x}_j) \rho(\mathbf{x}) d\mathbf{x} \quad (1.15)$$

The marginal probability allows us to determine the probability of finding a realisation of a component  $x_i$  of the random variable within an interval  $\mathcal{A}_i$ ,

$$P(x_i \in \mathcal{A}_i) = \iint_{\mathcal{A}_i} \rho(\mathbf{x}) d\mathbf{x} \quad (1.16)$$

where inner integration is performed over the complete intervals corresponding to the components of  $\mathbf{x}$  except  $x_i$ .

An example of a bivariate probability density defined on  $\mathcal{S} = \mathbb{R} \times \mathbb{R}$  is,

$$\rho(x, y) = \frac{1}{\pi} \exp[-(x^2 + y^2)] \quad (1.17)$$

One verifies that the marginal probability density for  $x$  actually gives a univariate probability density,

$$\begin{aligned}\rho(x) &= \frac{1}{\pi} \int_{\mathbb{R}} \exp[-(x^2 + y^2)] dy \\ &= \frac{1}{\sqrt{\pi}} \exp(-x^2)\end{aligned}\tag{1.18}$$

## 2 A few common probability distributions

### 2.1 The normal distribution (Gauss)

The normal distribution is given by,

$$\rho(\mathbf{x}) = \frac{1}{(2\pi)^{N/2} \sqrt{\det \mathbf{C}}} \exp\left[-\frac{1}{2} (\mathbf{x} - \bar{\mathbf{x}})^t \mathbf{C}^{-1} (\mathbf{x} - \bar{\mathbf{x}})\right]\tag{2.1}$$

where  $N$  is the dimension of the vector  $\mathbf{x}$  and  $\mathbf{C}$  is the covariance matrix, which is symmetric and positive definite. If the components of  $\mathbf{x}$  are independent variables, this matrix is diagonal, and its elements are the variances associated with each component of  $\mathbf{x}$ .

### 2.2 Generalised Gaussian distributions

Generalised Gaussian distributions are defined by the family,

$$\rho_p(x) \equiv \frac{p^{1-1/p}}{2\sigma_p \Gamma(1/p)} \exp\left(-\frac{|x - \bar{x}|^p}{p(\sigma_p)^p}\right) \quad p \geq 1,\tag{2.2}$$

where,

$$\sigma_p \equiv \left(\int_{\mathcal{I}} |x - \bar{x}|^p \rho(x) dx\right)^{1/p},\tag{2.3}$$

is a generalised measure of the dispersion of a probability density  $\rho(x)$ .

### 2.3 The *log-normal* distribution

The log-normal distribution is defined for  $x \geq 0$ ,

$$\frac{1}{\sqrt{2\pi x} \sigma} \exp\left[-\frac{\ln(x/m)^2}{2\sigma^2}\right]\tag{2.4}$$

and has a mean of  $m \exp(\sigma^2/2)$ , a median of  $m$ , and a variance of  $m^2 \exp(\sigma^2) \exp(\sigma^2 - 1)$ . Each log-normally distributed variable  $x$  is associated with a variable  $\ln(x)$  that follows a normal distribution. Similar to how the normal distribution is often obtained by adding random variables, the log-normal distribution is often obtained by multiplying random variables. As a result, the log-normal distribution is often useful for representing fluctuations due to multiplicative effects. More formally, the log-normal distribution is used to represent variables that are subject to proportional changes, where the resulting value is obtained by applying a random factor to the previous value.

## 2.4 The **Poisson** distribution

The **Poisson** distribution is defined for positive integer variables  $x$ ,

$$\frac{\lambda^x}{x!} \exp(-\lambda). \quad (2.5)$$

The mean and the variance are both equal to  $\lambda$ , which must be positive. The **Poisson** distribution is often used to represent rare random events, such as earthquakes in intra-plate zones. Surprisingly, the **Poisson** distribution also accurately represents the sequence of fatal accidents caused by horse kicks in the Prussian army in the 19th century!

## 2.5 The **gamma** ( $\Gamma$ -) distribution

The  $\Gamma$ -distribution is defined for  $x \geq 0$  and is given by

$$\frac{\lambda^k}{\Gamma(k)} x^{k-1} \exp(-\lambda x) \quad (2.6)$$

where  $\lambda$  and  $k$  are two positive parameters representing the scale and shape of the distribution, respectively. The name of the distribution comes from its denominator  $\Gamma(k)$ , which ensures the normalization of the distribution. The mean is  $k/\lambda$  and the variance is  $k/\lambda^2$ . When  $k = 1$ , the distribution simplifies to the exponential distribution. The Gamma distribution is similar to the **Poisson** distribution but has a lighter tail, resulting in lower probabilities for extreme values.

## 2.6 The **beta** ( $\beta$ -) distribution

The  $\beta$ -distribution is defined for  $0 \leq x \leq 1$ ,

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1} \quad (2.7)$$

## 2. A FEW COMMON PROBABILITY DISTRIBUTIONS

---

The mean is  $a/(a+b)$ , and the variance is  $ab/[(a+b)^2(a+b+1)]$ . The two shape parameters,  $a$  and  $b$ , must be positive.

### 2.7 The Pareto distribution

This distribution, also known as the hyperbolic or power law, is named after the Italian economist [Vilfredo Pareto](#), who used it in the late 19th century to describe personal wealth in certain societies.

It is defined for  $x \geq a$  with positive shape parameters  $a$  and  $b$ ,

$$\frac{ba^b}{x^{1+b}}. \quad (2.8)$$

The mean is given by  $ab/(b-1)$  for  $b > 1$ . For  $b > 2$ , the variance is,

$$\frac{a^2b}{[(b-1)^2(b-2)]} \quad (2.9)$$

and is infinite for  $b \leq 2$ . The [Pareto](#) distribution is often used to represent scale laws found in nature. In this distribution, the probability that the variable  $x > u > a$  is given by  $(a/u)^b$ . A particular application of the [Pareto](#) distribution is in modelling flood peaks.

### 2.8 The binomial distribution

The binomial distribution is defined for positive integer values of  $x$ ,

$$\binom{n}{x} p^x (1-p)^{n-x} \quad (2.10)$$

The mean of this distribution is  $np$  and the variance is given by  $np(1-p)$ . This distribution gives the probability of  $x$  events occurring in a series of length  $n$ , given that the probability of an event occurring is  $p$ . The binomial distribution can be used to calculate the probabilities of events occurring that do not respond systematically to a given cause. For example, what is the probability of a seismological station being struck by lightning in a year with 45 thunderstorms?



---

## 2.9 The Cauchy distribution

The Cauchy distribution is defined by,

$$\frac{1}{\pi b \left[ 1 + \left( \frac{x-a}{b} \right)^2 \right]} \quad (2.11)$$

where the parameter  $b > 0$ . The Cauchy distribution has a slow-decaying tail, which assigns a relatively high probability to extreme values. As a result, the mean and variance are not defined. However, the median is equal to  $a$ . The Cauchy distribution is a Lévy-stable distribution, meaning that the sum of variables drawn from a Cauchy distribution will also follow a Cauchy distribution.

## 2.10 The Weibull distribution

This distribution is defined for positive integers of  $x$ ,

$$\left( \frac{a}{b^a} \right) x^{a-1} \exp \left[ - \left( \frac{x}{b} \right)^a \right], \quad (2.12)$$

where  $a$  is the shape parameter of the distribution and  $b$  is the scale parameter. The mean is given by  $b\Gamma(1 + 1/a)$  and the variance by  $b^2 [\Gamma(1 + 2/a) - \Gamma^2(1 + 1/a)]$ . When  $a = 1$ , the distribution reduces to the exponential distribution, and to the Rayleigh distribution when  $a = 2$ .

# 3 Bayes' formula and inversion"

## 3.1 General solution

Let us revisit the information conjunction formula 3.12 established for continuous variables, which is equivalent to Bayes' formula,

$$\sigma(z) = \frac{\rho(z) \theta(z)}{\mu(z)} \quad (3.1)$$

and examine the meaning of the different probability densities that make it up. First, it is important to note that, in an inverse problem, we traditionally have data  $x$  and parameters  $y$  that form the random variable  $z$  in the formula above, and so the formula can be rewritten in a more

explicit form as follows,

$$\sigma(x,y) = \frac{\rho(x,y)\theta(x,y)}{\mu(x,y)}. \quad (3.2)$$

The probability density  $\rho$  can be considered as the *a priori* probability on the parameters and data, while  $\theta$  represents the probabilistic version of the forward problem, *ie* the probability density relating the data to the parameters within the framework of a physical law or, in the absence of a law, *via* statistical relationships. By integrating 3.2 with respect to  $x$ , one obtains the marginal probability density for  $y$ ,

$$\sigma(y) = \int \frac{\rho(x,y)\theta(x,y)}{\mu(x,y)} dx \quad (3.3)$$

which is the most general solution to an inverse problem (?). Note that the marginal probability over the data  $x$  can also be evaluated to obtain the *a posteriori* probability over the measured values of the data,

$$\sigma(x) = \int \frac{\rho(x,y)\theta(x,y)}{\mu(x,y)} dy \quad (3.4)$$

### 3.2 Solution for *a priori* independent data and parameters

If the data and parameters are *a priori* independent, the probability densities take the form,

$$\rho(x,y) = \rho_x(x) \cdot \rho_y(y) \quad (3.5)$$

$$\theta(x,y) = \theta(x|y) \cdot \mu_y(y). \quad (3.6)$$

The probability  $\theta(x,y)$  given by equation 3.6 is a conditional probability that contains no information about the parameters since their marginal probability represents the maximum ignorance  $\mu_y$ . Substituting these expressions into the general solution 3.3, and assuming that  $\mu(x,y) = \mu_x(x) \cdot \mu_y(y)$ , gives the solution to the inverse problem when the data and parameters are *a priori* independent,

$$\sigma(y) = \rho_y(y) \int \frac{\rho_x(x)\theta(x|y)}{\mu_x(x)} dx \quad (3.7)$$

The very existence of the marginal probability  $\sigma(y)$  depends on whether the various probabilities that make up the equation 3.7 are consistent with each other. If the *a priori* probabilities are

inconsistent with the measured data and the forward problem, then it is possible that the marginal probability at  $y$  is zero everywhere. The marginal probability over the data  $x$  can be obtained in a similar way by substituting 3.5 and 3.6 in 3.4,

$$\sigma(x) = \frac{\rho_x(x)}{\mu_x(x)} \int \rho_y(y) \theta(x|y) dy \quad (3.8)$$

### 3.3 Solution for an exact physical law

If there is an exact physical law for predicting the data from the parameters, then there exists a function  $g$  such that,

$$x = g(y), \quad (3.9)$$

and the conditional probability associated with the forward problem can be written as,

$$\theta(x|y) = \delta(x - g(y)) \quad (3.10)$$

where  $\delta$  is the Dirac distribution. In this case, equation 3.7 becomes,

$$\sigma(y) = \rho_y(y) \int \frac{\rho_x(x) \delta(x - g(y))}{\mu_x(x)} dx \quad (3.11)$$

$$= \rho_y(y) \frac{\rho_x(g(y))}{\mu_x(g(y))}. \quad (3.12)$$

### 3.4 Solution using Bayes' formula

Let us recall Bayes' formula, from which we will derive the developments that follow,

$$\rho(y|x) = \frac{\rho(y) p(x|y)}{\int_{\mathcal{Y}} \rho(y) p(x|y) dy}. \quad (3.13)$$

It is important to clearly define the role of each term in this equation. The probability density  $\rho(y|x)$  is what we are looking for - it is the *a posteriori* conditional probability of having  $y$  given that  $x$  has occurred, and it is the most general answer to an inverse problem. The probability density  $\rho(y)$  is called the *a priori* because it is assumed to contain all the information available about  $y$  before the experiment was performed, *ie* before  $x$  was known. Finally, the conditional probability  $p(x|y)$  - also called the likelihood - takes into account the fact that the data  $x$  is uncertain and that models  $y$  that do not perfectly reproduce the data (i.e. the particular realisation of the random variable  $x$ )

are acceptable within limits defined by  $p(x|y)$ . Bayesian inversion depends critically on how the likelihood  $p(x|y)$  is defined, and a significant part of the expertise in physics lies in determining the likelihood accurately. This requires calibration of the method used, as well as the most credible simulations where the true response  $y$  is known, and so on. The establishment of a scientific fact can be considered achieved when the relevant community is convinced. In scientific debates, new or surprising results are typically challenged for their reliability, which in the language of [Bayesian](#) inversion amounts to debating the choice of probability  $p(x|y)$ .

## 4 The tunnel again

### 4.1 Example 1: one data and one parameter

Let us return to our favourite example and illustrate the use of Bayes' formula (3.13) to estimate the depth  $z_t$  from a measurement  $g_1 = -62; \mu Gal$  of the gravitational field taken at  $x_1 = x_t$ . In this particular case, we assume that all other parameters in equation (2.1) are sufficiently well known and do not need to be determined in the inverse problem. So, let us assume that  $x_t$  is known, and that  $r_t = 3; m$  and  $\rho = 2700; kg/m^3$ . The forward problem is then reduced to,

$$g'_1(z_t) = \frac{\alpha}{z_t} (\mu Gal) \quad (4.1)$$

where  $\alpha \simeq -1018.84$ . Suppose the measurement is accompanied by a Gaussian uncertainty with a standard deviation of  $\sigma_g = 5; \mu Gal$ . The probability of the measurement with respect to the true value – which is unknown to us, but in this example where we have taken  $z_t = 16; m$  is  $58.96; \mu Gal$  – of the gravity  $g'_1$  is given by,

$$p(g_1|g'_1) = \frac{1}{\sqrt{2\pi}\sigma_g} \exp\left[-\frac{(g_1 - g'_1(z_t))^2}{2\sigma_g^2}\right] \quad (4.2)$$

La dépendance de  $g'_1$  par rapport à  $z_t$  permet d'obtenir la vraisemblance nécessaire pour la formule de [Bayes](#),

$$p(g_1|z_t) = p[g_1|g'_1(z_t)] \quad (4.3)$$

Implementing Bayes' formula requires defining the *a priori* probability density on  $z_t$ . If we assume that all depths between  $z_{\min} = 5; m$  and  $z_{\max} = 25; m$  are equally probable, then we have,

$$\rho(z_t) = \frac{1}{\Delta z} \Pi\left(\frac{z_t - z_{moy}}{\Delta z}\right) \quad (4.4)$$

where  $\Delta z = z_{\max} - z_{\min}$  and  $z_{moy} = (z_{\max} + z_{\min})/2$ . The *a posteriori* probability density then becomes,

$$\begin{aligned} \rho(z_t|g_1) &= \frac{\rho(z_t) p(g_1|z_t)}{\int_{z_{\min}}^{z_{\max}} \rho(z_t) p(g_1|z_t) dz_t} \\ &= \frac{\Pi\left(\frac{z_t - z_{moy}}{\Delta z}\right) \exp\left[-(g_1 - \alpha/z_t)^2 / (2\sigma_g^2)\right]}{\int_{z_{\min}}^{z_{\max}} \exp\left[-(g_1 - \alpha/z_t)^2 / (2\sigma_g^2)\right] dz_t} \end{aligned} \quad (4.5)$$

It can be observed that the *a posteriori* probability density is no longer a uniform distribution and has a maximum relatively localised within the interval  $[z_{\min}, z_{\max}]$  (Figure 16.1, generated using the script [ex\\_tunnel\\_01.m](#)). We say that the parameter  $z_t$  is resolved, which means that the information provided by the data is useful in determining the unknown parameter. It is possible to compute the *a priori* and *a posteriori* information on  $z_t$  to see the effect of the data  $g_1$ . The information needed to determine the depth before using the gravimetric measurement is given by

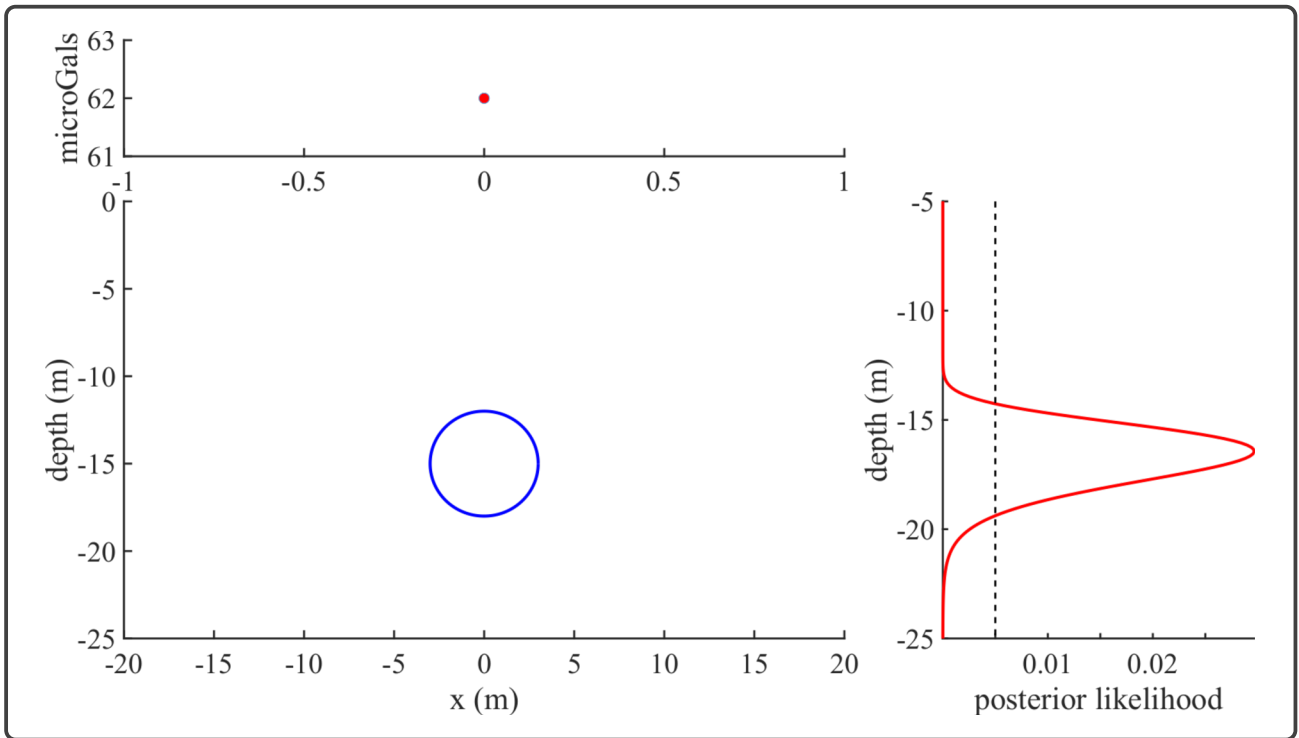
$$\begin{aligned} I_{priori}(z_t) &= - \int_{z_{\min}}^{z_{\max}} \rho(z_t) \ln \rho(z_t) dz_t \\ &= \frac{1}{\Delta z} \int_{z_{\min}}^{z_{\max}} \ln \Delta z dz_t \\ &= \ln \Delta z \\ &= 3.00 \text{ nep} \end{aligned} \quad (4.6)$$

The *a posteriori* information is given by,

$$\begin{aligned} I_{posteriori}(z_t|g_1) &= - \int_{z_{\min}}^{z_{\max}} \rho(z_t|g_1) \ln \rho(z_t|g_1) dz_t \\ &\simeq 1.74 \text{ nep} \end{aligned} \quad (4.7)$$

Thus, one can calculate the information provided by the gravimetric measurement,

$$I_{gravi}(g_1) = I_{priori} - I_{posteriori} \simeq 1.26 \text{ nep}. \quad (4.8)$$



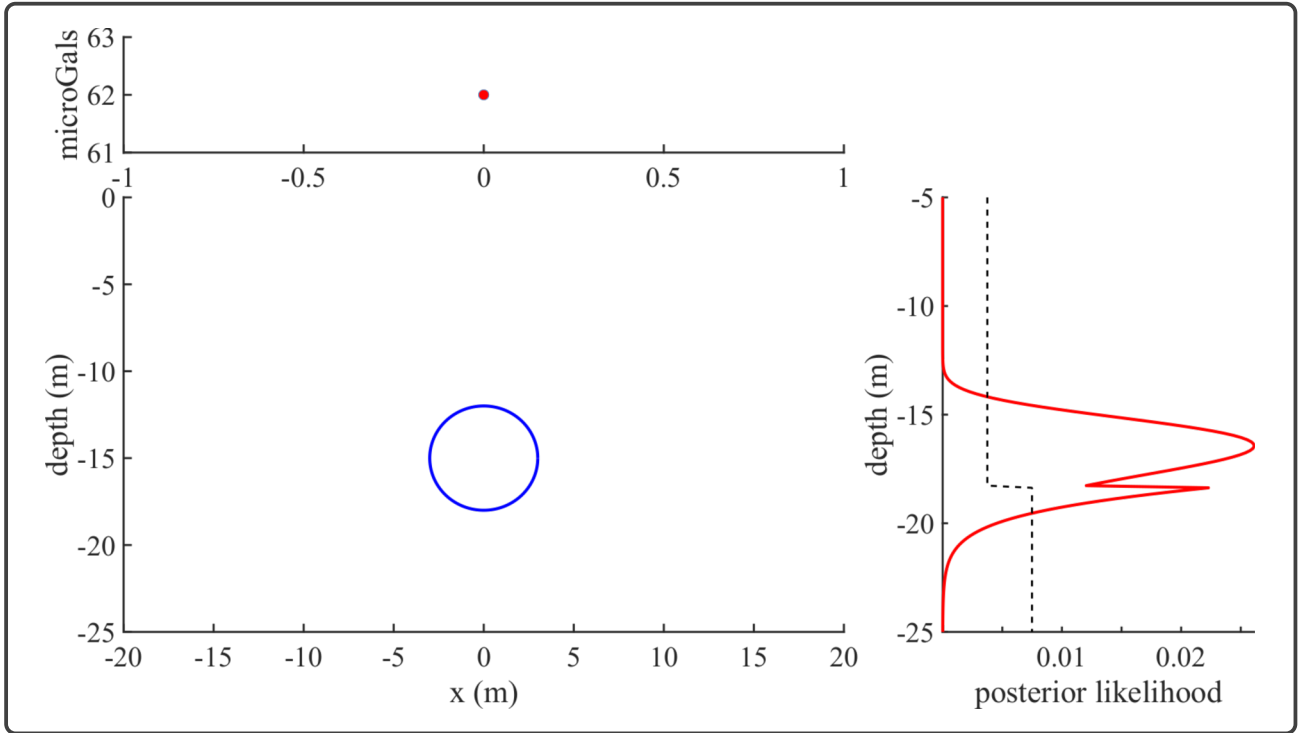
**Figure 16.1:** *a priori* probabilities (dashed line, uniform distribution) and *a posteriori* probabilities (solid curve) of the tunnel depth when a single gravimetric measurement is taken directly above the tunnel.

This information is not zero, which means that our knowledge of the depth  $z_t$  has increased. We say that the parameter is resolved. Consistent with intuition, the previous expressions show that the *a posteriori* information continues to decrease as the standard deviation of the measurement uncertainty decreases. There is a threshold beyond which the *a posteriori* information is almost equal to the *a priori* information, at which point the gravimetric measurement becomes essentially useless.

It is of course possible to start with a non-uniform prior probability, as in the example shown in figure 16.2. In this case, the *a posteriori* probability changes significantly, highlighting the importance of *a priori* information in solving inverse problems,

## 4.2 Example 2: two data and one parameter

Let us revisit Example 1 by adding a second gravimetric measurement and see what this means for our knowledge of the tunnel depth. Suppose the data are,



**Figure 16.2: a priori (dashed line) and a posteriori (solid line) probabilities of the tunnel depth when a single gravimetric measurement is taken directly above the tunnel. In this example, the a priori probability is not uniform, resulting in a significant change in the a posteriori probability (see figure 16.1).**

$x_i - x_t$ (m)	$g_i$ ( $\mu Gal$ )	$\sigma_g$ ( $\mu Gal$ )
0	-62.	5.
10.	-44	5.

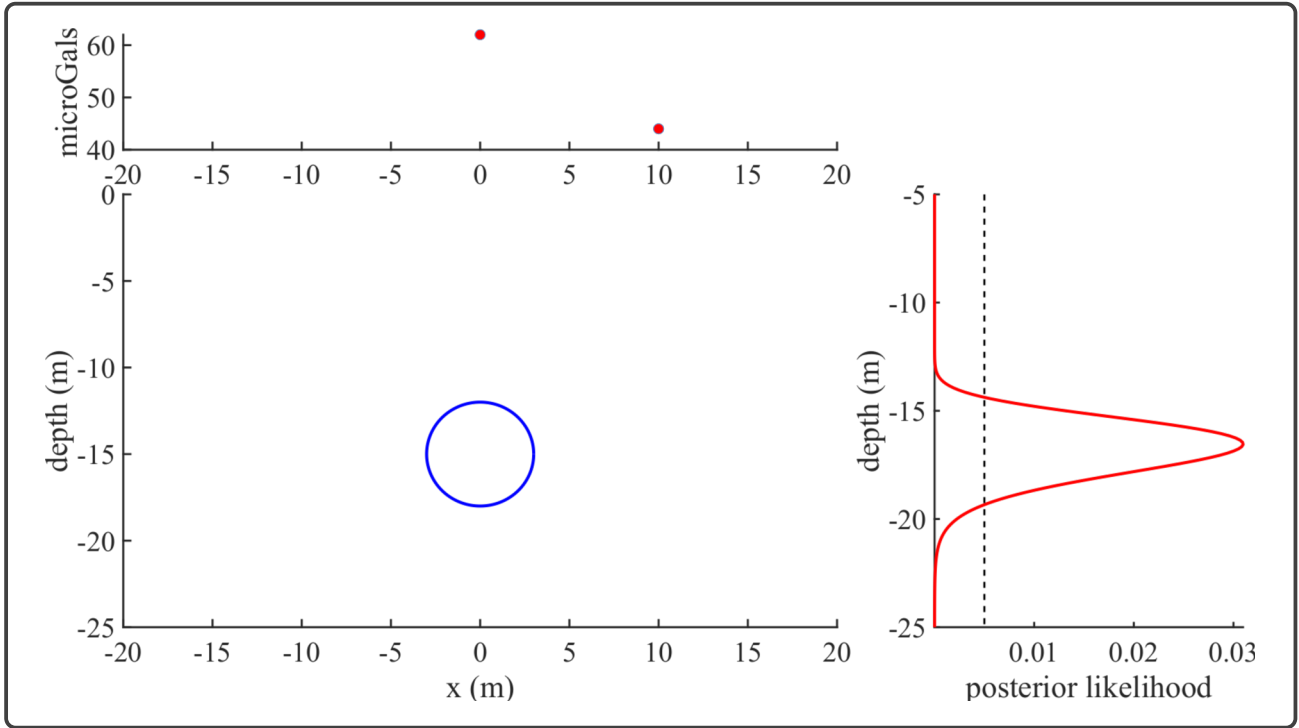
Using the vector notation,  $\mathbf{g}$ , to represent the data, the probability is then given by

$$p(\mathbf{g}|z_t) = \frac{1}{(\sqrt{2\pi}\sigma_g)^2} \exp\left[-\frac{\|\mathbf{g} - \mathbf{g}'(z_t)\|^2}{2\sigma_g^2}\right], \quad (4.9)$$

where  $\mathbf{g}'(z_t)$  represents the forward problem, that is, the calculation of the theoretical gravity as a function of the depth  $z_t$  that we wish to test. We have,

$$g'_i(z_t) = \frac{\alpha z_t}{(x_i - x_t)^2 + z_t^2}. \quad (4.10)$$

#### 4. THE TUNNEL AGAIN



**Figure 16.3:** *a priori* probabilities (dashed line, smooth curve) and *a posteriori* probabilities (solid line) from two gravimetric measurements, one directly above the tunnel and the other 10 metres above.

Using the same *a priori* probability density (4.4) as in Example 1, we find that (Figure 16.3)

$$\begin{aligned}
 \rho(z_t | \mathbf{g}) &= \frac{\rho(z_t) p(\mathbf{g} | z_t)}{\int_{z_{\min}}^{z_{\max}} \rho(z_t) p(\mathbf{g} | z_t) dz_t} \\
 &= \frac{\Pi\left(\frac{z_t - z_{moy}}{\Delta z}\right) \exp\left[-[\mathbf{g} - \mathbf{g}'(z_t)]^2 / (2\sigma_g^2)\right]}{\int_{z_{\min}}^{z_{\max}} \exp\left[-[\mathbf{g} - \mathbf{g}'(z_t)]^2 / (2\sigma_g^2)\right] dz_t}
 \end{aligned} \tag{4.11}$$



Of course, the prior information remains unchanged compared to Example 1 and is given by equation (4.6). However, the posterior information is given by,

$$I_{posteriori}(z_t|\mathbf{g}) = - \int_{z_{\min}}^{z_{\max}} \rho(z_t|\mathbf{g}) \ln \rho(z_t|\mathbf{g}) dz_t \simeq 1.69 nep \quad (4.12)$$

This allows us to calculate the information provided by the gravimetric measurements,

$$I_{gravi}(\mathbf{g}) = I_{priori} - I_{posteriori} \simeq 1.31 nep \quad (4.13)$$

This information is only slightly less than that obtained in the previous example, indicating that the parameter  $z_t$  is not better resolved and that the gravimetric data  $g_2$  has contributed negligible additional information. Let's examine this situation more closely by calculating the solution to the inverse problem using only the measurement  $g_2$  taken at  $x_2 - x_t = 10; m$ . The *a posteriori* probability density (Figure 16.4) is, in contrast to the previous case, poorly localised and has two *maxima*. The posterior information associated with this probability density is given by,

$$I_{posteriori}(z_t|g_2) = 2.93 nep \quad (4.14)$$

and so we have,

$$I_{gravi}(g_2) = 0.07 nep \quad (4.15)$$

which confirms that the data provide little additional information. The depth parameter is poorly resolved in this case. This can be understood by noting that the function  $g'_2(z_t)$  is equal to  $44; \mu Gal$  at two relatively different depths. This explains the presence of two maxima in the *a posteriori* probability density.

### 4.3 Example 3: One data and two parameters

We can complicate the inverse problem by assuming that the horizontal position  $x_t$  of the tunnel is poorly determined and is included as one of the parameters. In this example we will only use the

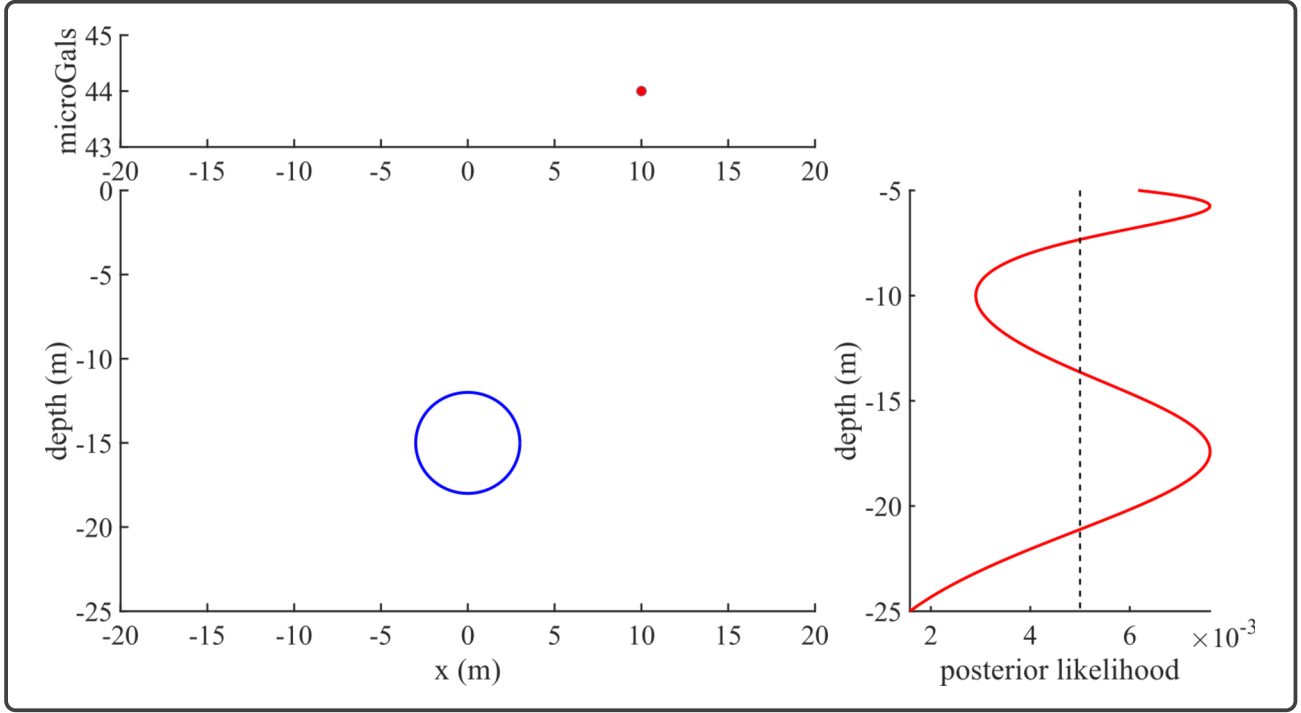


Figure 16.4: *a priori* probabilities (uniform distribution, dashed line) and *a posteriori* probabilities (solid curve) corresponding to the inverse problem solved with only the gravimetric data located 10 metres from the tunnel axis. It can be seen that the posterior probability is not well localised, indicating that the data do not effectively resolve the tunnel depth (cf [ex\\_tunnel\\_02.m](#)).

gravimetric data  $g_1$  from example 1. Under these conditions,

$$p(g_1|x_t, z_t) = \frac{1}{\sqrt{2\pi}\sigma_g} \exp \left[ -\frac{1}{2\sigma_g^2} (g_1 - g'_1(x_t, z_t))^2 \right] \quad (4.16)$$

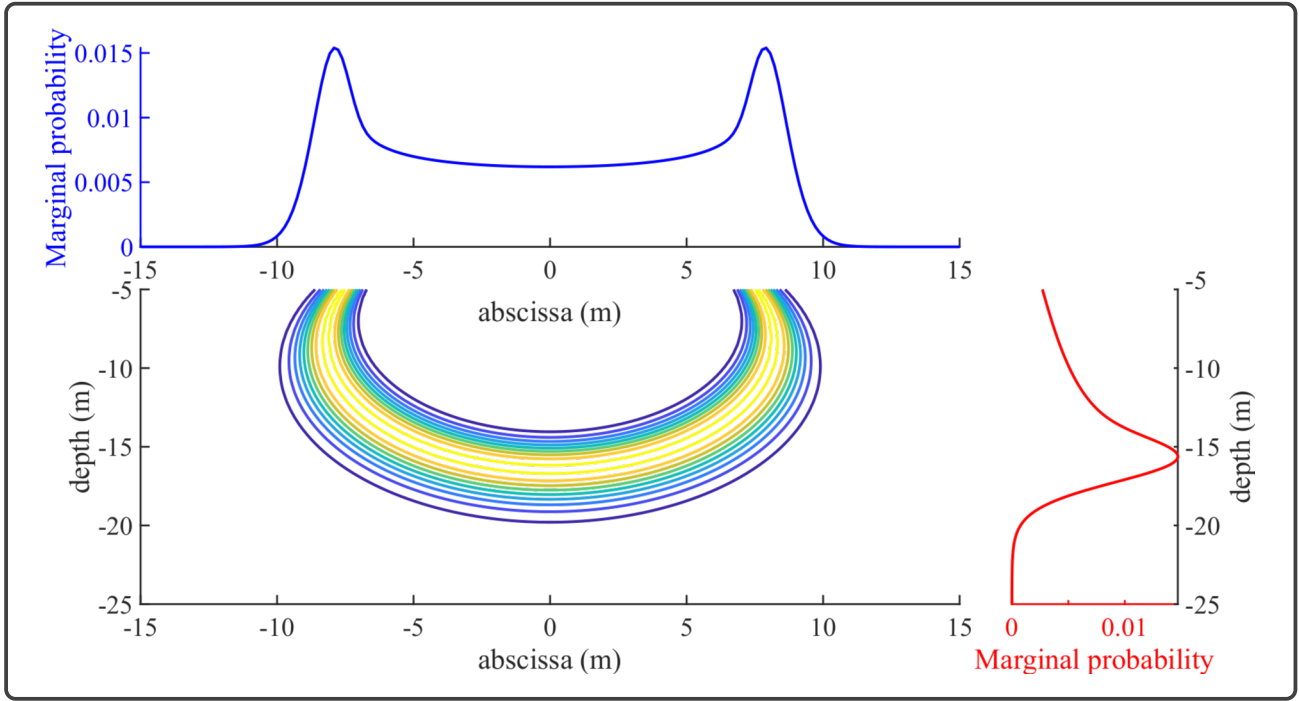
where the forward problem is given by,

$$g'_i(x_t, z_t) = \frac{\alpha z_t}{(x_i - x_t)^2 + z_t^2}. \quad (4.17)$$

For example, we can set the *a priori* probability density as,

$$\rho(x_t, z_t) = \frac{1}{\Delta x \Delta z} \Pi \left( \frac{x_t}{\Delta x} \right) \Pi \left( \frac{z_t - z_{moy}}{\Delta z} \right) \quad (4.18)$$

which indicates that the horizontal position is *a priori* within an interval of length  $\Delta x = 50; m$



**Figure 16.5: a posteriori** probability (contour plots, bottom left) of the horizontal position and depth of the tunnel obtained from a single gravimetric measurement (taken directly above the tunnel, although this is not known). The marginal probability for the depth is significantly less localised than when the data were used to determine the depth alone (see figure 16.1), indicating that the addition of parameters in an inverse problem affects the determination of the OTHER parameters (cf `ex_tunnel_03.m`).

centered on the measurement location. Thus, we have,

$$\rho(x_t, z_t | g_1) = \frac{\Pi\left(\frac{x_t}{\Delta x}\right) \Pi\left(\frac{z_t - z_{moy}}{\Delta z}\right) \exp\left[-\frac{1}{2\sigma_g^2} (g_1 - g'_1(x_t, z_t))^2\right]}{\int_{x_{\min}}^{x_{\max}} \int_{z_{\min}}^{z_{\max}} \exp\left[-\frac{1}{2\sigma_g^2} (g_1 - g'_1(x_t, z_t))^2\right] dx_t dz_t} \quad (4.19)$$

This *a posteriori* probability density is relatively complex (Figure 16.5) and has a horseshoe shape, indicating the correlation between the two parameters  $x_t$  and  $z_t$ . The *a priori* information is given by,

$$\begin{aligned} I_{\text{priori}}(x_t, z_t) &= - \int_{x_{\min}}^{x_{\max}} \int_{z_{\min}}^{z_{\max}} \rho(x_t, z_t) \ln \rho(x_t, z_t) dx_t dz_t \\ &= \ln(\Delta x \Delta z) \simeq 6.91 \text{ nep} \end{aligned} \quad (4.20)$$

and the *a posteriori* information is given by,

$$\begin{aligned} I_{\text{posteriori}}(x_t, z_t | g_1) &= - \int_{x_{\min}}^{x_{\max}} \int_{z_{\min}}^{z_{\max}} \rho(x_t, z_t | g_1) \ln \rho(x_t, z_t | g_1) dx_t dz_t \\ &\simeq 1.89 \text{ nep} \end{aligned} \quad (4.21)$$

#### 4. THE TUNNEL AGAIN

The information provided by the data  $g_1$  is therefore  $I_{gravi}(g_1) \simeq 5.02 \text{ nep}$ . The marginal probability densities are respectively,

$$\rho(x_t|g_1) = \frac{\Pi\left(\frac{x_t}{\Delta x}\right) \int_{z_{\min}}^{z_{\max}} \exp\left[-\frac{1}{2\sigma_g^2} (g_1 - g'_1(x_t, z_t))^2\right] dz_t}{\int_{x_{\min}}^{x_{\max}} \int_{z_{\min}}^{z_{\max}} \exp\left[-\frac{1}{2\sigma_g^2} (g_1 - g'_1(x_t, z_t))^2\right] dx_t dz_t} \quad (4.22)$$

and,

$$\rho(z_t|g_1) = \frac{\Pi\left(\frac{z_t - z_{moy}}{\Delta z}\right) \int_{x_{\min}}^{x_{\max}} \exp\left[-\frac{1}{2\sigma_g^2} (g_1 - g'_1(x_t, z_t))^2\right] dx_t}{\int_{x_{\min}}^{x_{\max}} \int_{z_{\min}}^{z_{\max}} \exp\left[-\frac{1}{2\sigma_g^2} (g_1 - g'_1(x_t, z_t))^2\right] dx_t dz_t} \quad (4.23)$$

The marginal informations are respectively,

$$I_{priori}(x_t) = \ln \Delta x \simeq 3.91 \text{ nep} \quad (4.24)$$

$$I_{priori}(z_t) = \ln \Delta z \simeq 3.00 \text{ nep}$$

$$I_{posteriori}(x_t|g_1) = - \int_{x_{\min}}^{x_{\max}} \rho(x_t|g_1) \ln \rho(x_t|g_1) dx_t \simeq 2.93 \text{ nep} \quad (4.25)$$

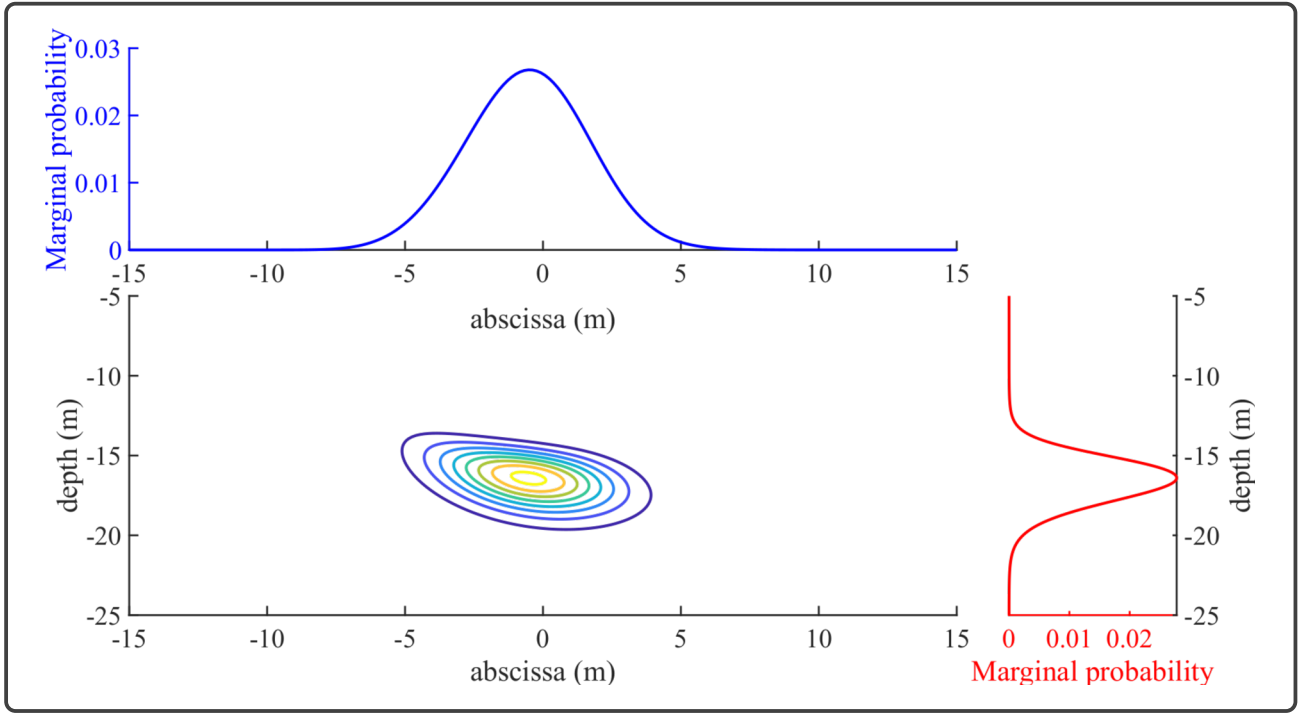
$$I_{posteriori}(z_t|g_1) = - \int_{z_{\min}}^{z_{\max}} \rho(z_t|g_1) \ln \rho(z_t|g_1) dz_t \simeq 2.60 \text{ nep}$$

It can be seen that the  $g_1$  data did not provide the same amount of information about the two parameters,

$$I_{gravi}(g_1 \rightsquigarrow x_t) \equiv I_{priori}(x_t) - I_{posteriori}(x_t|g_1) \simeq 0.98 \text{ nep} \quad (4.26)$$

$$I_{gravi}(g_1 \rightsquigarrow z_t) \equiv I_{priori}(z_t) - I_{posteriori}(z_t|g_1) \simeq 0.40 \text{ nep}$$

An important observation can already be made by comparing these results with those from the first example. It can be seen that in example 1 the data  $g_1$  contributed an information value of  $1.26; \text{nep}$  to our knowledge of the depth  $z_t$ . In contrast, in example 3, the same data contributes only  $0.40; \text{nep}$ , which is three times less. This illustrates a universal principle in inverse problem theory, which contrasts with the popular notion that data, often referred to as 'information', provides unchanging knowledge about a parameter.



**Figure 16.6:** *a posteriori* probability (contour plots, bottom left) for the horizontal position and depth of the tunnel using two gravimetric measurements. The marginal probabilities indicate that the two parameters are fairly well resolved (*cf ex\_tunnel\_04.m*).

#### 4.4 Example 4: Two data and two parameters

Let us add the second gravimetric measurement  $g_2$  to the inverse problem introduced in Example 3. The likelihood is given by the formula,

$$p(\mathbf{g}|x_t, z_t) = \frac{1}{(\sqrt{2\pi}\sigma_g)^2} \exp \left[ -\frac{\|\mathbf{g} - \mathbf{g}'(x_t, z_t)\|^2}{2\sigma_g^2} \right] \quad (4.27)$$

where  $\mathbf{g}'(x_t, z_t)$  represents the direct problem,

$$g'_i(x_t, z_t) = \frac{\alpha z_t}{(x_i - x_t)^2 + z_t^2}. \quad (4.28)$$

Using the same *a priori* probability density (4.18) as in example 3, we find (Figure 16.6),

$$\rho(x_t, z_t | \mathbf{g}) = \frac{\rho(x_t, z_t) p(\mathbf{g}|x_t, z_t)}{\int_{x_{\min}}^{x_{\max}} \int_{z_{\min}}^{z_{\max}} \rho(x_t, z_t) p(\mathbf{g}|x_t, z_t) dx_t dz_t} \quad (4.29)$$

The prior information remains unchanged from Example 3 and is given by equation (4.20), while

the *a posteriori* information is,

$$I_{posteriori}(x_t, z_t | \mathbf{g}) = - \int_{x_{\min}}^{x_{\max}} \int_{z_{\min}}^{z_{\max}} \rho(x_t, z_t | \mathbf{g}) \ln \rho(x_t, z_t | \mathbf{g}) dx_t dz_t \simeq 0.01 \text{ nep} \quad (4.30)$$

The information provided by the gravimetric measurements,

$$I_{gravi}(\mathbf{g}) = I_{priori} - I_{posteriori} \simeq 6.90 \text{ nep} \quad (4.31)$$

The prior marginal information is the same as in example 3, and the *a posteriori* marginal information is

$$\begin{aligned} I_{posteriori}(x_t | \mathbf{g}) &\simeq 2.22 \text{ nep}, \\ I_{posteriori}(z_t | \mathbf{g}) &\simeq 1.82 \text{ nep} \end{aligned} \quad (4.32)$$

The information provided about the two parameters is,

$$\begin{aligned} I_{gravi}(\mathbf{g} \rightsquigarrow x_t) &\equiv I_{priori}(x_t) - I_{posteriori}(x_t | \mathbf{g}) \simeq 1.69 \text{ nep}, \\ I_{gravi}(\mathbf{g} \rightsquigarrow z_t) &\equiv I_{priori}(z_t) - I_{posteriori}(z_t | \mathbf{g}) \simeq 1.18 \text{ nep} \end{aligned} \quad (4.33)$$

## 5 Summary of examples 1, 2, 3 and 4

It is time to make some summary remarks on the examples we have just discussed and to draw some conclusions that will guide the following sections. The main observations we can make are the following,

1. the addition of an extra data point may not improve our knowledge of a parameter (example 2),
2. the addition of a parameter can significantly reduce the knowledge of another parameter that was previously well resolved (example 3),
3. a data may improve our knowledge of one parameter, but not another (example 4),
4. the *a posteriori* probability density often contains several relative maxima that do not necessarily correspond to the true solution, which may, in contrast, correspond to a relative minimum (example 3).

In a more general sense, it was observed that the information provided by the data  $g_2$  was

---

used in very different ways from one inverse problem to another. In example 2 this information was used very sparingly and it can be said that the data  $g_2$  was practically useless. In contrast, in example 4 this data proved to be important, where it contributed significantly to the knowledge of the parameter  $x_i$ . This reflects a very classic behaviour of information in human contexts, where, for example, a message may be revealing to one person but meaningless to another. In the realm of inverse problems, different parameters have different sensitivities, or 'resolutions', to different pieces of data. Perhaps even more surprisingly, two gravimetric measurements that might initially be expected to play similar roles can have such different degrees of importance *a posteriori*.

---

---

# CHAPTER 17

---

## MONTE CARLO METHODS

<b>1</b>	<b>Introduction</b> . . . . .	<b>280</b>
<b>2</b>	<b>Integration by the Monte Carlo method</b> . . . . .	<b>281</b>
<b>3</b>	<b>Metropolis algorithm</b> . . . . .	<b>284</b>
3.1	Importance sampling . . . . .	284
3.2	Markov chain . . . . .	285
3.3	The Metropolis algorithm . . . . .	286
3.4	Example . . . . .	288
3.5	Example of the tunnel . . . . .	289



---

# 1 Introduction

The *Bayesian* solution to an inverse problem is the *a posteriori* probability density

$$\rho(\mathbf{y}|\mathbf{x}) = \frac{\rho(\mathbf{y})p(\mathbf{x}|\mathbf{y})}{\int_{\mathcal{S}} \rho(\mathbf{y})p(\mathbf{x}|\mathbf{y}) d\mathbf{y}} \quad (1.1)$$

where  $\mathbf{x}$  and  $\mathbf{y}$  represent the data and parameter vectors of the problem, respectively. In general, the dimensions  $D_x$  and  $D_y$  of these vectors are very large, and as soon as  $D_y > 3$  one encounters difficulties in visualising the function  $\rho(\mathbf{y}|\mathbf{x})$ . Moreover, the exhaustive and systematic exploration of the *a priori* solution space  $\mathcal{S}$ , which was feasible for the examples concerning the tunnel, is no longer possible because the number of computations required is immense.

The visualisation problem can be partially solved in several ways. For example, several authors, including [Albert Tarantola](#), advocate the creation of films in which the images consist of several acceptable *a posteriori* solutions. The more an *a posteriori* solution is probable, the more often its image appears in the film. However, this approach is rarely used because the practical realisation of these films must take into account the fact that the frequency of appearance is not necessarily a linear function of the probability, if one wants to take into account physiological factors such as retinal persistence and mental factors such as memory retention. For example, the images may need to be sorted in a certain way to help the viewer better grasp the different classes of solutions. We have already used this technique, which has proved very useful in certain cases, and we have found that a random appearance of the images makes the film very difficult to use. Although the film technique remains experimental and unusual for now, we believe it may become more important in the future as visualisation methods continue to improve.

Another solution to the visualisation problem is to visualise only the marginal probability densities,

$$\rho(y_i|\mathbf{x}) = \int \rho(\mathbf{y}|\mathbf{x}) d\mathbf{y}_{\neq i} \quad (1.2)$$

where  $y_i$  represents the parameter for which the marginal probability is computed, and  $d\mathbf{y}_{\neq i}$  is the vector of dimension  $D_y - 1$ , excluding the dimension corresponding to  $y_i$ . The representation of marginal probabilities is, of course, very simple since we are dealing with functions that depend only on the single variable  $y_i$

## 2 Integration by the Monte Carlo method

The calculation of the marginal probabilities 1.2 requires the integration of the *a posteriori* probability density  $\rho(\mathbf{y}|\mathbf{x})$ . However, we have found that it is practically impossible to evaluate this probability density systematically and uniformly over the entire *a priori* solution space. Therefore, the integral 1.2 cannot be evaluated by numerical methods that require systematic knowledge of  $\rho(\mathbf{y}|\mathbf{x})$ , but it is possible to use integration by the *Monte Carlo* method based on random sampling of the *a priori* space. Let  $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n, \dots, \mathbf{y}_N\}$  be a collection of  $N$  models, all with the same component  $y_i$  and with the other  $D_y - 1$  components randomly drawn from the *a priori* space of volume  $V$ . Then we have,

$$\rho(y_i|\mathbf{x}) = \int \rho(\mathbf{y}|\mathbf{x}) d\mathbf{y}_{\neq i} \approx V \langle \rho \rangle \pm V \sqrt{\frac{\langle \rho^2 \rangle - \langle \rho \rangle^2}{N}} \quad (2.1)$$

where,

$$\langle \rho \rangle \equiv \frac{1}{N} \sum_{n=1}^N \rho(\mathbf{y}_n|\mathbf{x}) \quad \langle \rho^2 \rangle \equiv \frac{1}{N} \sum_{n=1}^N \rho^2(\mathbf{y}_n|\mathbf{x}) \quad (2.2)$$

The term  $\pm$  in the equation 2.1 is an estimate of the uncertainty in the value of the integral.

Figures 17.1, 17.2, 17.3 and 17.4 show the marginal probabilities obtained by *Monte Carlo* integration. These probabilities differ quite significantly from the curves (grey lines) obtained by regular sampling of the *a priori* model space. Only a large number of samples, greater than that used for regular sampling (50 for figures 17.2 and 17.4), allows the recovery of curves that appear correct. Note that for figures 17.3 and 17.4, the error is significant in the region of maximum probability because most of the integral is contributed by a small region of the integration domain that is not properly sampled by the randomly drawn models.

Figures 17.1 and 17.3 were created with the script `ex_tunnel_05.m`; images 17.2, 17.4 and 17.5 were created with the script `ex_tunnel_06.m`.

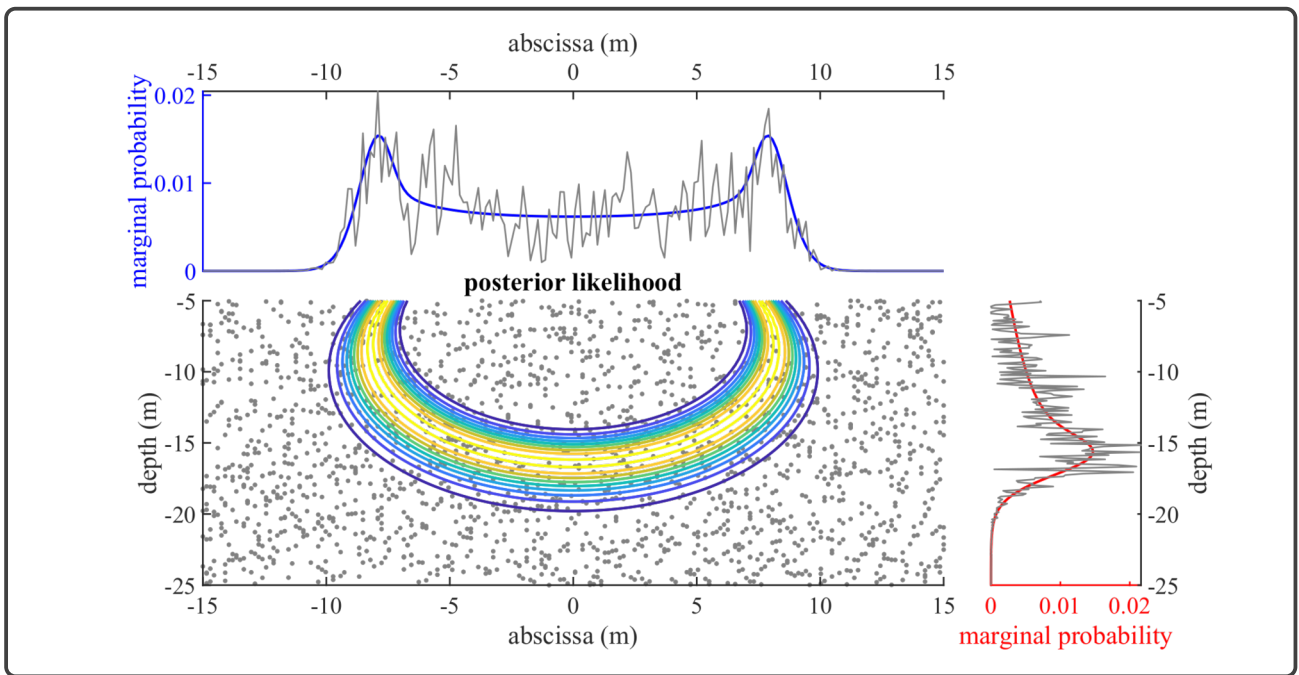


Figure 17.1: *a posteriori* probability (contour plots at bottom left) for the horizontal position and depth of the tunnel when only a single gravimetric measurement is used. The marginal probabilities obtained by the Monte Carlo method are shown with dashed lines. Ten samples were taken for each value of  $x_t$  (top curve) or  $z_t$  (right curve).

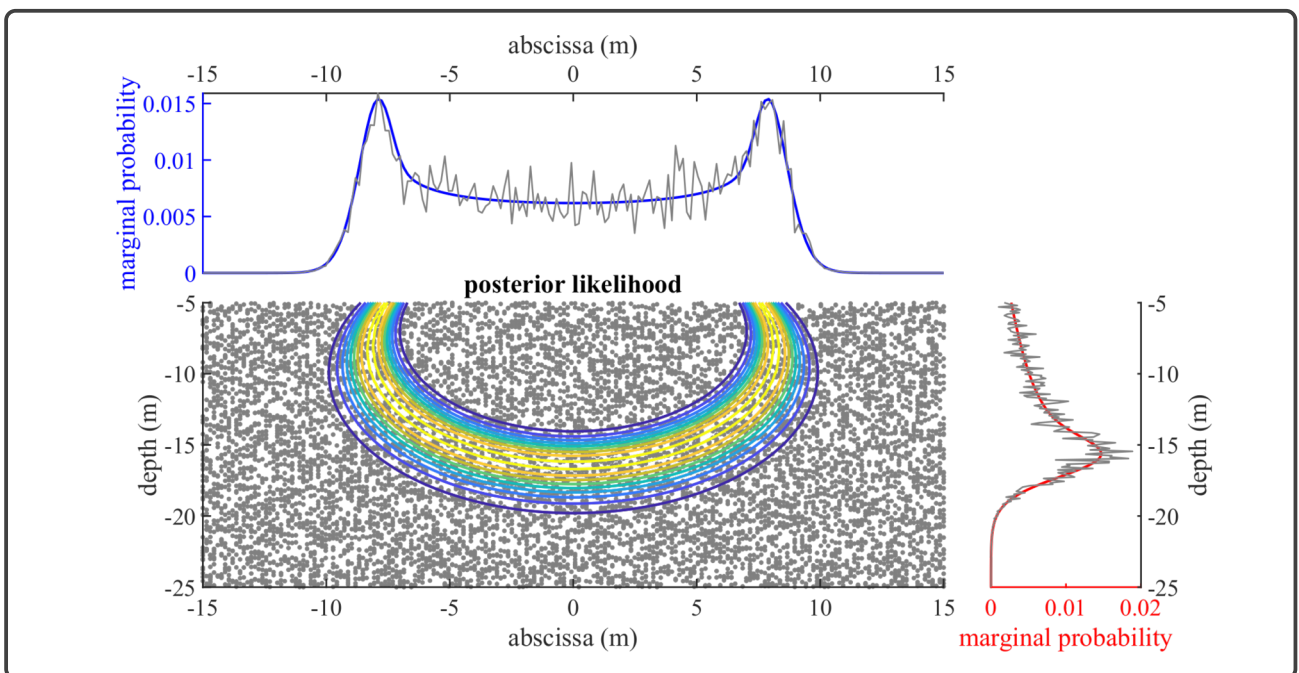


Figure 17.2: Similar to Figure 17.1, but for 50 Monte Carlo samples.

## 2. INTEGRATION BY THE MONTE CARLO METHOD

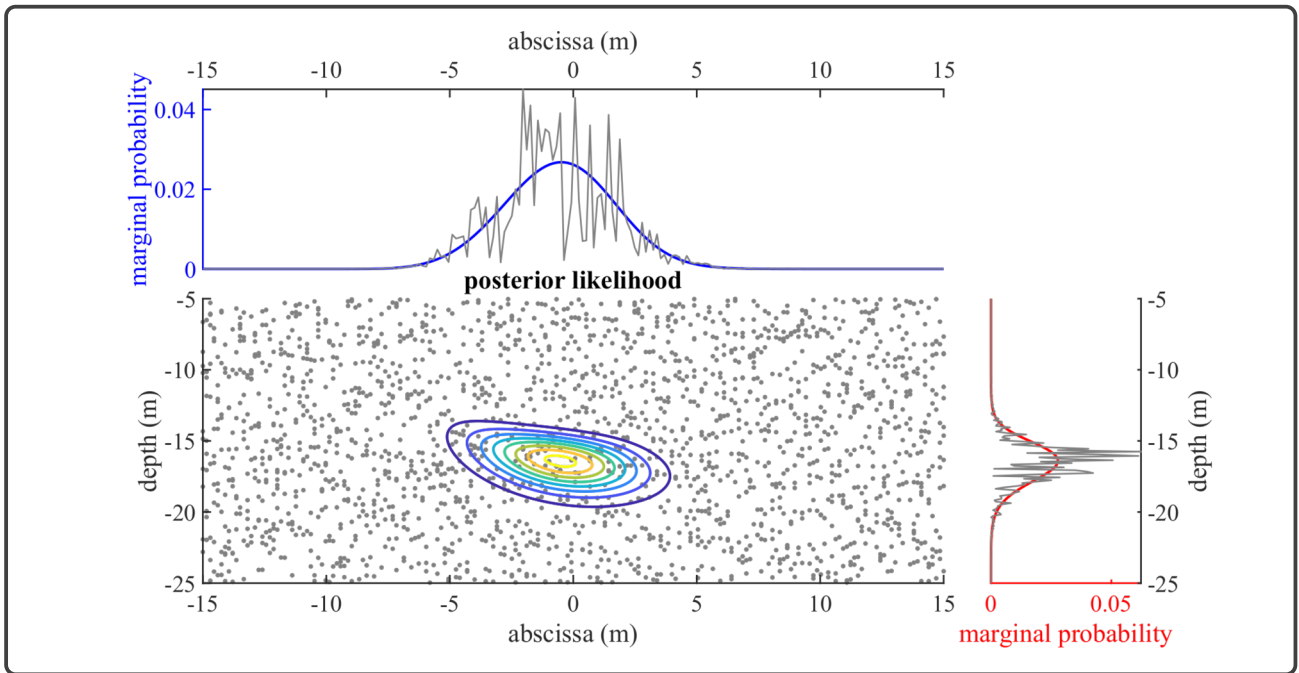


Figure 17.3: Similar to Figure 17.1, but for two gravimetric measurements.

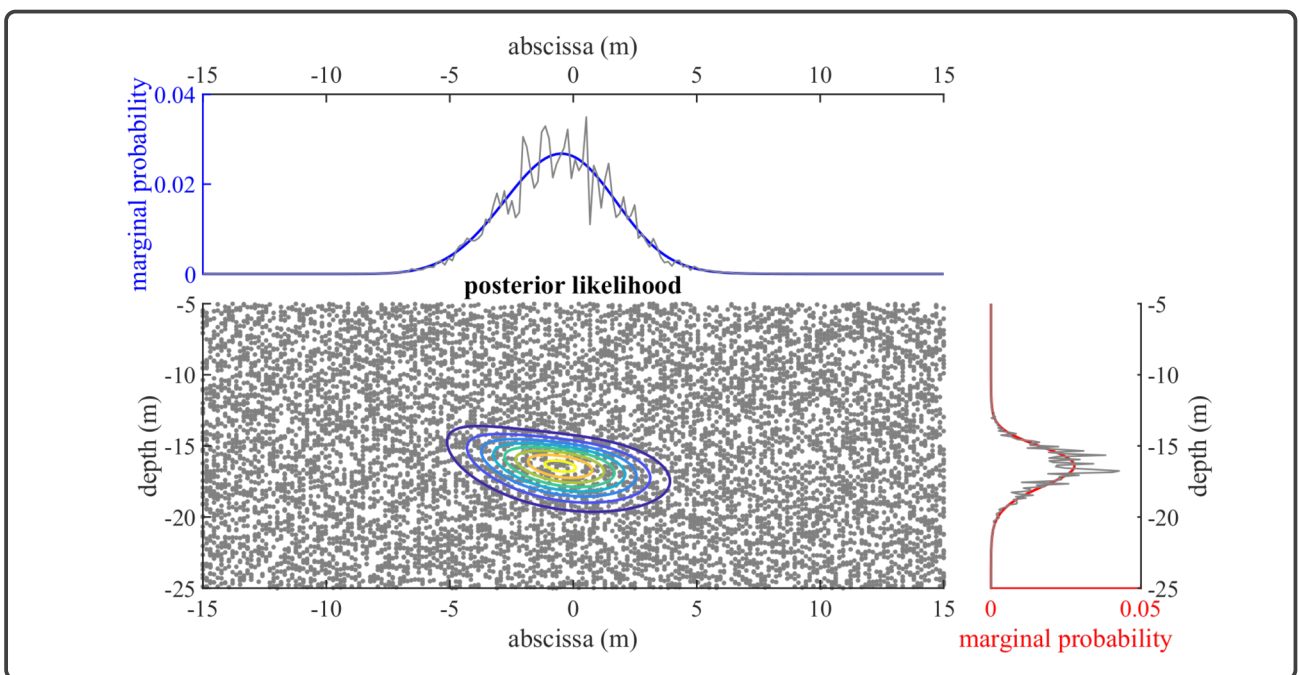


Figure 17.4: Similar to Figure 17.3, but for 50 Monte Carlo samples..

The integration error caused by random sampling decreases as  $N^{-1/2}$ , whereas the error caused by regular sampling decreases as  $N^{-1}$ . It is therefore tempting to perform a random sampling that has the advantage of regular sampling, *ie* one that is random but distributes the points relatively evenly. This can be achieved using quasi-random sequences, such as those of Sobol, which produce values with a quasi-uniform density that improves as the sequence lengthens (see figures 17.3 and

17.5). This type of sampling gives better numerical integrations, but is limited to a small number of parameters (typically less than 10) and is not significantly more efficient than regular and systematic sampling.

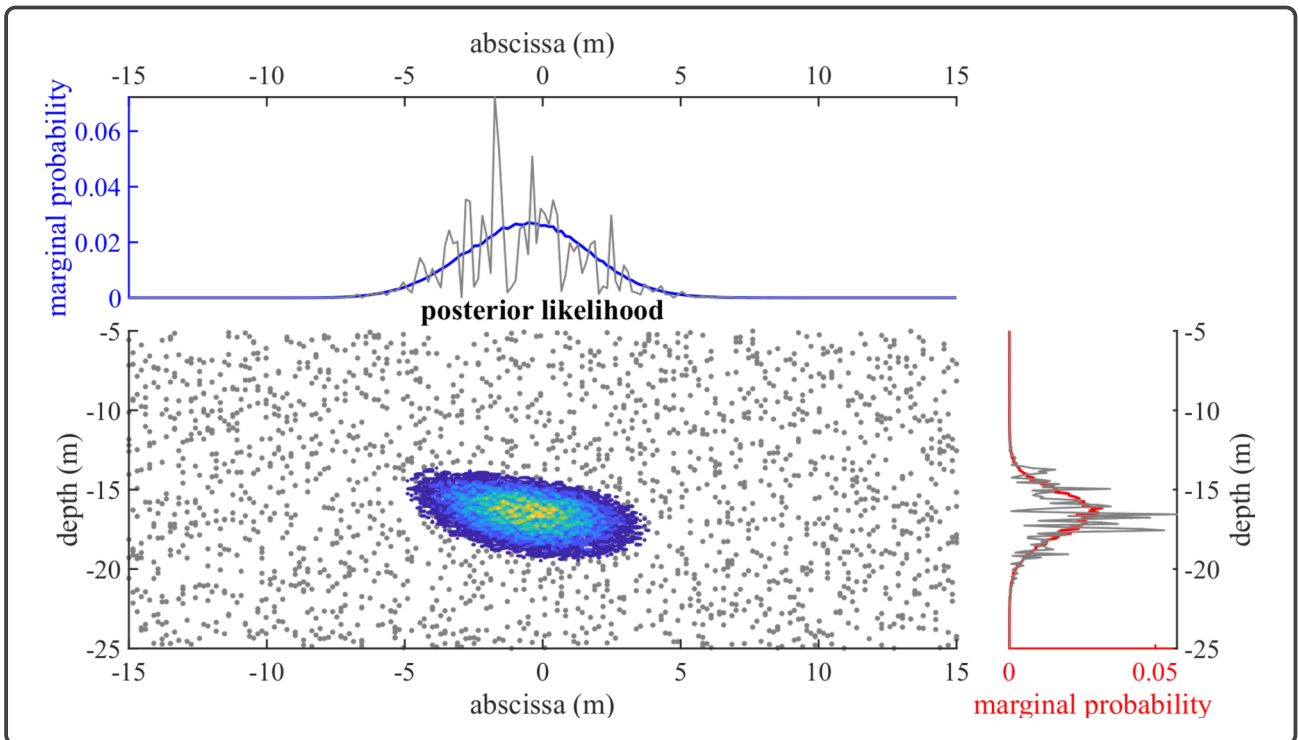


Figure 17.5: Similar to Figure 17.3, but for 10 samples from a Sobol sequence. The more regular sampling reduces the integration error.

### 3 Metropolis algorithm

#### 3.1 Importance sampling

Integration using the Monte Carlo method does not correctly integrate the marginal probabilities because the random sampling does not give sufficient weight to small regions where the probability density is significant. One way to overcome this is to generate a sequence of random models whose distribution is  $\rho(y|x)$  to reduce the error in the mean of equation 2.2

The Metropolis algorithm, invented in 1953 (Metropolis *et al.*, 1953) at the dawn of the computer age, enables this particular type of random sampling, known in the Anglo-Saxon literature as "importance sampling". Basically, the Metropolis algorithm is a Markov chain in which a model is replaced by a successor under the control of a process that is partly random and partly guided. It is this process that constrains the set of generated models to conform to the imposed probability density.

#### 3.2 Markov chain

A **Markov** chain is defined by a transition probability law,

$$P(\mathbf{y}_i^0 \rightarrow \mathbf{y}_j^1), \quad (3.1)$$

which generates a set  $\mathcal{M}1 = \{\mathbf{y}j^1\}$  from a set  $\mathcal{M}0 = \{\mathbf{y}i^0\}$ . That is to say, when the law  $P$  is applied to each element of  $\mathcal{M}0$ , it results in  $\mathcal{M}1$ . If we group the sets of solutions  $\mathbf{y}i^0$  and  $\mathbf{y}j^1$  into vectors  $\mathbf{Y}_0$  and matrix  $\mathbf{Y}_1$ ,

$$\mathbf{P} \cdot \mathbf{Y}_0 = \mathbf{Y}_1. \quad (3.2)$$

We want to repeat the transformation procedure by iteratively applying  $\mathbf{P}$  starting from the initial set  $\mathbf{Y}_0$ , so that, after a large number of iterations, the population of the final set satisfies the probability law  $\rho(\mathbf{y}|\mathbf{x})$ . This iterative process has the form

$$\lim_{n \rightarrow \infty} \mathbf{P}^n \cdot \mathbf{Y}_0 = \mathbf{Y}, \quad (3.3)$$

where the final set  $\mathbf{Y}$  consists of models  $\mathbf{y}_i$  with an appearance frequency of  $\rho(\mathbf{y}i|\mathbf{x})$ .

It is necessary for the algorithm to be stable, *ie* the point  $\mathbf{Y}$  must be the only fixed point of the flow,

$$\mathbf{P} \cdot \mathbf{Y} = \mathbf{Y}. \quad (3.4)$$

Three conditions are necessary to ensure the uniqueness of the fixed point. The first is to state that every initial model  $\mathbf{y}_i^0$  must have an image in the set  $\mathcal{M}1$ . This amounts to saying that the sum of the transformation probabilities is 1,

$$\sum_{\mathcal{M}1} P(\mathbf{y}_i^0 \rightarrow \mathbf{y}_j^1) = 1. \quad (3.5)$$

This equation simply means that it is certain that an element of the target set can be obtained by applying the transformation rule to the initial set. It also requires that any initial model  $\mathbf{y}_i^0$  can be

transformed, even with a very small probability, into one of the models in the target set  $\mathcal{M}_1$ ,

$$P(\mathbf{y}_i^0 \rightarrow \mathbf{y}_j^1) > 0, \forall (\mathbf{y}_i^0, \mathbf{y}_j^1) \in \mathcal{M}_0 \times \mathcal{M}_1. \quad (3.6)$$

This condition is known as the strong *ergodicity* condition. The third condition is sufficient, but not necessary, to ensure that the transformation  $P$  satisfies the desired properties. This is the microscopic equilibrium condition (detailed balance condition),

$$P(\mathbf{y}_i^0 \rightarrow \mathbf{y}_j^1) \rho(\mathbf{y}_i^0 | \mathbf{x}) = P(\mathbf{y}_j^1 \rightarrow \mathbf{y}_i^0) \rho(\mathbf{y}_j^1 | \mathbf{x}), \forall (\mathbf{y}_i^0, \mathbf{y}_j^1) \in \mathcal{M}_0 \times \mathcal{M}_1, \quad (3.7)$$

which we will see is satisfied by the [Metropolis](#) algorithm.

If conditions (3.5), (3.6), and (3.7) are satisfied, then we have,

$$\begin{aligned} \sum_{\mathcal{M}_0} P(\mathbf{y}_i^0 \rightarrow \mathbf{y}_j^1) \rho(\mathbf{y}_i^0 | \mathbf{x}) &= \sum_{\mathcal{M}_0} P(\mathbf{y}_j^1 \rightarrow \mathbf{y}_i^0) \rho(\mathbf{y}_j^1 | \mathbf{x}) \\ &= \rho(\mathbf{y}_j^1 | \mathbf{x}), \end{aligned} \quad (3.8)$$

where the properties (3.7) and (3.5) have been used successively. The relation (3.8) shows that the probability density  $\rho(\mathbf{y} | \mathbf{x})$  is indeed a fixed point of the *Markov* chain.

### 3.3 The [Metropolis](#) algorithm

Relation (3.8) can be satisfied in several ways, among which the [Metropolis](#) algorithm uses the transformation law defined by,

$$P(\mathbf{y}_i^0 \rightarrow \mathbf{y}_j^1) = 1 \text{ if } \rho(\mathbf{y}_i^0 | \mathbf{x}) < \rho(\mathbf{y}_j^1 | \mathbf{x}) \quad (3.9)$$

$$P(\mathbf{y}_i^0 \rightarrow \mathbf{y}_j^1) = \frac{\rho(\mathbf{y}_j^1 | \mathbf{x})}{\rho(\mathbf{y}_i^0 | \mathbf{x})} \text{ if } \rho(\mathbf{y}_i^0 | \mathbf{x}) \geq \rho(\mathbf{y}_j^1 | \mathbf{x}). \quad (3.10)$$

Equation (3.9) shows that the transformation  $\mathbf{y}_i^0 \rightarrow \mathbf{y}_j^1$  is accepted whenever the proposed image model has a higher probability than the previous model. In contrast, equation (3.10) shows that the transformation is possible, but not certain, if the image model is less probable than the previous model. Clearly then,

$$P[\rho(\mathbf{y}_i^0 | \mathbf{x}) < \rho(\mathbf{y}_j^1 | \mathbf{x})] + P[\rho(\mathbf{y}_i^0 | \mathbf{x}) \geq \rho(\mathbf{y}_j^1 | \mathbf{x})] = 1, \quad (3.11)$$

### 3. METROPOLIS ALGORITHM

---

which is the particular event, the algorithm defined by (3.9) and (3.10) satisfies condition (3.5).

Relation (3.9) implies,

$$P(\mathbf{y}_i^0 \rightarrow \mathbf{y}_j^1) \rho(\mathbf{y}_i^0 | \mathbf{x}) = \rho(\mathbf{y}_i^0 | \mathbf{x}) \quad \text{si } \rho(\mathbf{y}_i^0 | \mathbf{x}) < \rho(\mathbf{y}_j^1 | \mathbf{x}), \quad (3.12)$$

and relation (3.10) provides,

$$P(\mathbf{y}_j^1 \rightarrow \mathbf{y}_i^0) \rho(\mathbf{y}_j^1 | \mathbf{x}) = \rho(\mathbf{y}_i^0 | \mathbf{x}) \quad \text{si } \rho(\mathbf{y}_i^0 | \mathbf{x}) < \rho(\mathbf{y}_j^1 | \mathbf{x}) \quad (3.13)$$

Combining these two results, we find that

$$P(\mathbf{y}_i^0 \rightarrow \mathbf{y}_j^1) \rho(\mathbf{y}_i^0 | \mathbf{x}) = P(\mathbf{y}_j^1 \rightarrow \mathbf{y}_i^0) \rho(\mathbf{y}_j^1 | \mathbf{x}) \quad \text{si } \rho(\mathbf{y}_i^0 | \mathbf{x}) < \rho(\mathbf{y}_j^1 | \mathbf{x}) \quad (3.14)$$

Furthermore, the relation (3.10) allows us to write that,

$$P(\mathbf{y}_i^0 \rightarrow \mathbf{y}_j^1) \rho(\mathbf{y}_i^0 | \mathbf{x}) = \rho(\mathbf{y}_j^1 | \mathbf{x}) \quad \text{si } \rho(\mathbf{y}_i^0 | \mathbf{x}) \geq \rho(\mathbf{y}_j^1 | \mathbf{x}) \quad (3.15)$$

while relation (3.9) implies that,

$$P(\mathbf{y}_j^1 \rightarrow \mathbf{y}_i^0) \rho(\mathbf{y}_j^1 | \mathbf{x}) = \rho(\mathbf{y}_i^0 | \mathbf{x}) \quad \text{si } \rho(\mathbf{y}_i^0 | \mathbf{x}) \geq \rho(\mathbf{y}_j^1 | \mathbf{x}) \quad (3.16)$$

These two formulas provide the condition,

$$P(\mathbf{y}_i^0 \rightarrow \mathbf{y}_j^1) \rho(\mathbf{y}_i^0 | \mathbf{x}) = P(\mathbf{y}_j^1 \rightarrow \mathbf{y}_i^0) \rho(\mathbf{y}_j^1 | \mathbf{x}) \quad \text{si } \rho(\mathbf{y}_i^0 | \mathbf{x}) \geq \rho(\mathbf{y}_j^1 | \mathbf{x}) \quad (3.17)$$

Combining the conditional results (3.14) and (3.17), it follows that,

$$P(\mathbf{y}_i^0 \rightarrow \mathbf{y}_j^1) \rho(\mathbf{y}_i^0 | \mathbf{x}) = P(\mathbf{y}_j^1 \rightarrow \mathbf{y}_i^0) \rho(\mathbf{y}_j^1 | \mathbf{x}) \quad (3.18)$$

which is nothing other than the microscopic balance condition (3.7) seen earlier.



### 3.4 Example

Let's examine how the (3.9) and (3.10) algorithms work on a binary example where there are two possible models. In this case, the set of allowed models is,

$$\mathcal{I} = \{0, 1\}. \quad (3.19)$$

Suppose we want to generate a set of models with the respective probabilities given by,

$$\rho(0) = \frac{1}{3} \text{ et } \rho(1) = \frac{2}{3}. \quad (3.20)$$

Let the initial set be,

$$\mathcal{M}_0 = \{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\}. \quad (3.21)$$

The first iteration of the algorithm yields the set,

$$\mathcal{M}_1 = \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1\} \quad (3.22)$$

since only transformation (3.9) was successful, since model 0 is less likely than model 1. The second iteration tests the transformation  $1 \rightarrow 0$ , which will include the equation (3.10),

$$P(1 \rightarrow 0) = \frac{\rho(0)}{\rho(1)} = \frac{1}{2} \quad (3.23)$$

This transition is therefore random and indicates that the probability of performing the transformation is 0.5. The practical implementation of this transition is to use a random number generator to produce a number  $r \in [0, 1]$  drawn from a uniform distribution. If,

$$r \leq \frac{\rho(0)}{\rho(1)} \Rightarrow P(1 \rightarrow 0) \text{ accepts,} \quad (3.24)$$

$$r > \frac{\rho(0)}{\rho(1)} \Rightarrow P(1 \rightarrow 0) \text{ rejects.} \quad (3.25)$$

Using the random number generator on our calculator, we could find that,

$$\mathcal{M}_2 = \{1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 1\} \quad (3.26)$$

### 3. METROPOLIS ALGORITHM

The next iteration involves testing both transformations  $1 \rightarrow 0$  and  $0 \rightarrow 1$ . The first is random, as we observed in the second iteration, while the second is certain, as in the first iteration. Still using our pocket calculator, we obtained,

$$\mathcal{M}_3 = \{0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0\} \quad (3.27)$$

A similar calculation provides,

$$\begin{aligned} \mathcal{M}_4 &= \{1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1\}, \\ \mathcal{M}_5 &= \{0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1\}, \\ \mathcal{M}_6 &= \{1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0\}, \\ \mathcal{M}_7 &= \{0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1\}, \\ \mathcal{M}_8 &= \{1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0\}, \\ \mathcal{M}_9 &= \{1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1\}, \\ \mathcal{M}_{10} &= \{1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 1, 1\}. \end{aligned} \quad (3.28)$$

Given the fixed probabilities  $\rho(0)$  and  $\rho(1)$ , the sets generated should ideally contain seven 0s and fourteen 1s. The counts obtained are,

	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	$\mathcal{M}_5$	$\mathcal{M}_6$	$\mathcal{M}_7$	$\mathcal{M}_8$	$\mathcal{M}_9$	$\mathcal{M}_{10}$
0	21	0	9	5	7	8	7	5	11	4	7
1	0	21	12	16	14	13	14	16	10	17	14

Aggregating the results for the last 9 sets, the average is as follows,

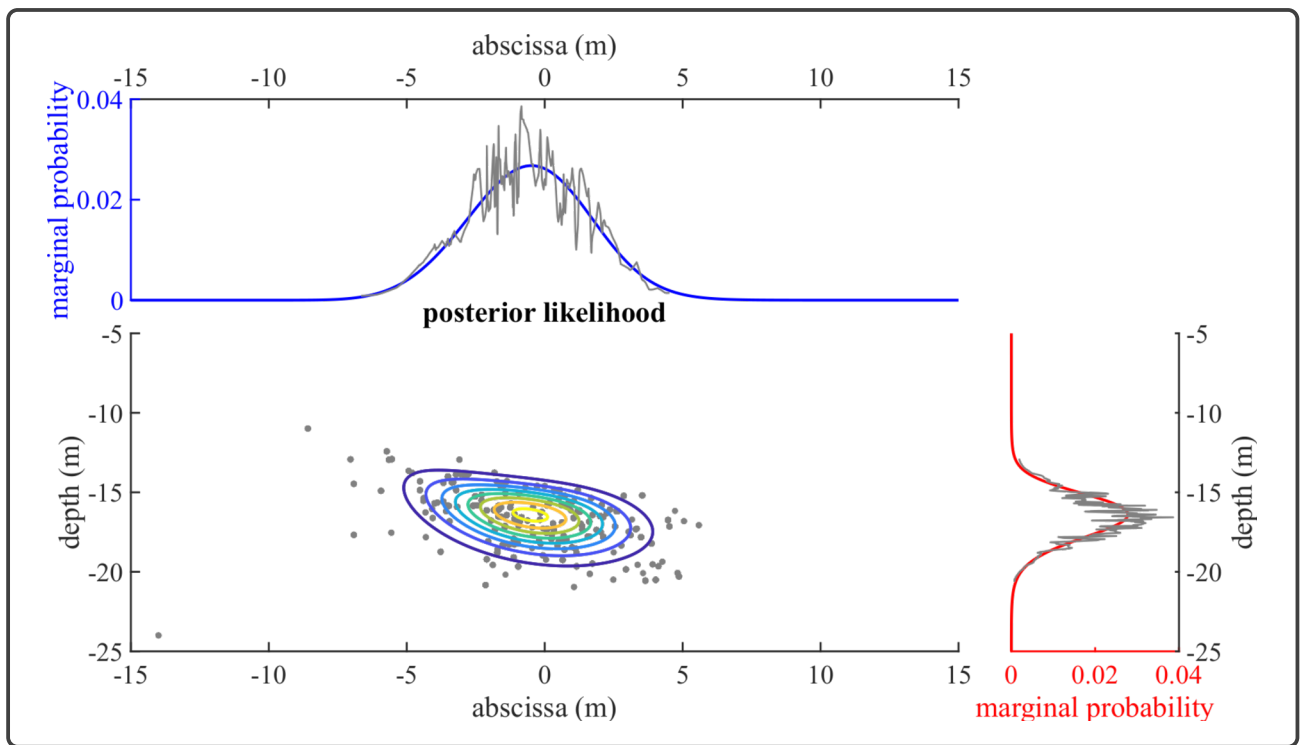
$$N(0) = \frac{63}{9} \Rightarrow \rho(0) = \frac{1}{3} \quad (3.29)$$

which is excellent. Of course, in practice, it is necessary to work with a sufficient number of iterations and with sets containing many elements, but it is remarkable to observe that even with small samples like those in our example, the [Metropolis](#) algorithm already yields good results.

### 3.5 Example of the tunnel

Let us return to the tunnel example used previously for integration with the [Monte Carlo](#) method. Figure 17.6, generated with the script [ex\\_tunnel\\_07.m](#), shows the result of the integration obtained with [Metropolis](#) sequences of 50 terms. A reduction of the error is observed, especially in the marginal

probability concerning  $x_t$  (see figure 17.4).



**Figure 17.6:** Similar to Figure 17.4, but for 50 samples from a **Metropolis** chain. Importance sampling helps to reduce the integration error. The dots represent the 2500 models from a **Metropolis** chain. Their distribution density follows the probability density. The small number of points compared to the 2500 models indicates that many models are duplicated many times, which means that the acceptance of new models was very rare. This is explained by the fact that each new model was randomly generated in the *a priori* space, independently of the previous model.

---

---

# CHAPTER 18

---

## SIMULATED ANNEALING

1	Aim of the method . . . . .	292
2	Control temperature . . . . .	292
3	Perturbing the models . . . . .	293
4	The Simulated Annealing algorithm . . . . .	294
5	Example: the traveling salesman problem . . . . .	294
5.1	Introduction . . . . .	294
5.2	Generation of models . . . . .	295
5.3	Example of how to operate . . . . .	295

---

# 1 Aim of the method

We have seen that the [Metropolis](#) algorithm can, in principle, generate a sequence of models  $\mathbf{y}$  according to a probability density  $\rho(\mathbf{y}|\mathbf{x})$ . However, this algorithm alone does not quickly yield good estimates of marginal probabilities when the probability density consists of local lobes in model space. In such cases, many tested models are rejected if they lie within one of the lobes, because most new models fall into areas of very low probability and therefore have little chance of being accepted. This leads to significant waste in the computation of direct problems, making the method inefficient. If the probability density is multimodal, the chances of exploring all the high-probability lobes are very low, leading to poor assessment of marginal probabilities.

The idea of simulated annealing ([Kirkpatrick et al., 1983](#)) is to guide the models towards the lobes of maximum probability density using the [Metropolis](#) algorithm, with two important modifications, which are,

1. the use of a model generation process that has a 'memory' so that the new models are in some sense close to the previous models,
2. a deformation of the probability density that gradually reveals the lobes during the [Metropolis](#) process.

The conjunction of these two aspects allows the generation of models that are confined to the vicinity of the probability density lobes, resulting in greater efficiency.

## 2 Control temperature

Fundamentally, Simulated Annealing is the [Metropolis](#) algorithm implemented with a probability density whose topology evolves over the course of iterations. This evolution allows, as mentioned above, a gradual transition from an almost uniform distribution to the *a posteriori* density  $\rho(\mathbf{y}|\mathbf{x})$ , which can potentially be multimodal. The evolution law assumes the following obvious relationship,

$$\rho(\mathbf{y}|\mathbf{x}) = \exp[\ln(\rho(\mathbf{y}|\mathbf{x}))]. \quad (2.1)$$

Let's rewrite this formula by including a parameter  $T$ , which we will call the temperature

$$\rho_T(\mathbf{y}|\mathbf{x}) = k_T \exp\left[\frac{\ln(\rho(\mathbf{y}|\mathbf{x}))}{T}\right], \quad T \geq 0, \quad (2.2)$$

where  $k_T$  is a normalization constant. Evidently,

$$\rho_1(\mathbf{y}|\mathbf{x}) = \rho(\mathbf{y}|\mathbf{x}), \quad (2.3)$$

and also,

$$\rho_\infty(\mathbf{y}|\mathbf{x}) = k_\infty, \quad (2.4)$$

that is, at infinite temperature,  $\rho_T(\mathbf{y}|\mathbf{x})$  approaches a uniform probability density. Thus, as  $T$  varies from 1 to infinity, the probability density  $\rho_T(\mathbf{y}|\mathbf{x})$  gradually deforms, providing a means to control the topology of the probability density that guides the [Metropolis](#) algorithm.

## 3 Perturbing the models

The second crucial aspect of Simulated Annealing is the memory of the process, *ie* the fact that the models generated retain certain parameters from previous models while modifying others. There is no precise mathematical rule to describe this process, but rather principles that should be followed by defining rules specific to the particular inverse problem at hand.

The primary principle is that the transition from one model to the next should not disrupt the guidance towards the modes of the probability density provided by the [Metropolis](#) process. For this reason, a completely random generation of models is not suitable, as it would result in a path through model space without memory. However, it is also essential that the path allows the exploration of large "territories" within this space in relatively few iterations to avoid algorithmic stagnation and the confinement of the series of models to a very limited volume. It is therefore clear that the model generation process must have somewhat contradictory properties: a substantial degree of movement similar to [Monte-Carlo](#) methods and a perturbative memory similar to gradient-based methods.

Depending on whether the inverse problem involves discrete variables, as in the case of the travelling salesman problem that we will discuss later, or continuous variables, as in the tunnel example, the model generation process may differ significantly. Indeed, even for a discrete problem like the travelling salesman problem, which can have a very large combinatorial space (*eg*  $2^{32}$ ), it remains finite and it is possible to design model generation processes where the distance, measured in terms of the number of random draws to move from one model to another, remains small (*eg* 100). In contrast, when dealing with continuous variables, the distance between two models becomes

---

infinite, even if they vary over a finite interval. In such cases, the model generation process may need to adapt as the temperature decreases during the iterations of [Metropolis](#).

## 4 The Simulated Annealing algorithm

Considering the above, the main steps of Simulated Annealing are as follows

1. loop  $j$  over the temperature
  - (a) function defining the temperature  $T_j$
  - (b) [Metropolis](#) loop  $i$ 
    - i. generation of the model to be tested  $\mathbf{y}_{i+1}^{T_j}$
    - ii. evaluation of the *a posteriori* probability  $\rho(\mathbf{y}_{i+1}^{T_j} | \mathbf{x})$
    - iii. acceptance or rejection of the transition  $\mathbf{y}_i^{T_j} \rightarrow \mathbf{y}_{i+1}^{T_j}$
  - (c) end of the  $i$ -th [Metropolis](#) loop
  - (d) convergence test
2. end of the  $j$ -th loop over the temperature

As we have already noted, the temperature control and model generation steps are particularly crucial and determine the success or failure of the method. Unfortunately, there are no precise and universal rules for the development of these steps, as their form depends on the specific inverse problem at hand. It is also worth noting that the sequential nature of the above algorithm poses challenges that we will address by proposing a modified algorithm in which several [Metropolis](#) loops operate in parallel. To the best of our knowledge, this new algorithm is very similar to genetic algorithms.

## 5 Example: the traveling salesman problem

### 5.1 Introduction

The travelling salesman problem is famous as a typical case of an optimisation problem with extremely high combinatorial complexity, and its efficient solution has been one of the reasons for the success of simulated annealing. The problem is to determine the order in which a travelling salesman should visit a given number of cities, exactly once, in order to minimise his travel distance. If  $N$  is the

number of cities to visit, then the number of possible *a priori* solutions is  $N!$ , which quickly leads to an extremely large combinatorial space. For example, if  $N = 32$ , the combinatorial space already exceeds  $10^{35}$ . Therefore, an exhaustive exploration of the solution space to find optimal solutions is out of the question.

### 5.2 Generation of models

For this problem, a model is an ordered list of cities, and the model generation process produces a list from another list. To satisfy both the memory constraints of the algorithm and the ability to explore the *a priori* model space quickly, the processes typically used for this problem involve randomly selecting a small number of cities from the list and permuting them, which can be deterministic. Often only 2 cities are selected and swapped, resulting in *eg*,

$$\{\text{Nice, Rennes, Brest, Lille, Caen}\} \mapsto \{\text{Nice, Caen, Brest, Lille, Rennes}\}. \quad (5.1)$$

It is observed that such a process has a clear memory effect, as a new list differs only slightly from the previous one. At the same time, it allows for rapid movement through the model space, since at most  $N - 1$  permutations are required to move from one list to any other.

### 5.3 Example of how to operate

The following example involves 32 cities whose geographical distribution follows a hierarchy of 'countries,' 'regions,' and 'municipalities.' For this example, the temperature was controlled *via* a geometric sequence,

$$T_{j+1} = 0.995T_j, \quad (5.2)$$

and 100 iterations were performed for each *Metropolis* loop (*ie* for a given temperature  $T_j$ ). Figures 18.1, 18.2, 18.3 and 18.4 illustrate the evolution of the path during the cooling process. It can be observed (curves at the top of each figure) that the cost decreases very rapidly when the temperature is around  $10^{-1.5}$  (figure 18.2) and stabilises when the temperature drops below  $10^{-2}$ . At the end of the run (Figure 18.4), the total number of models generated is only 137,900, which is very small compared to the combinatorial complexity of the problem, which is  $32!$ .



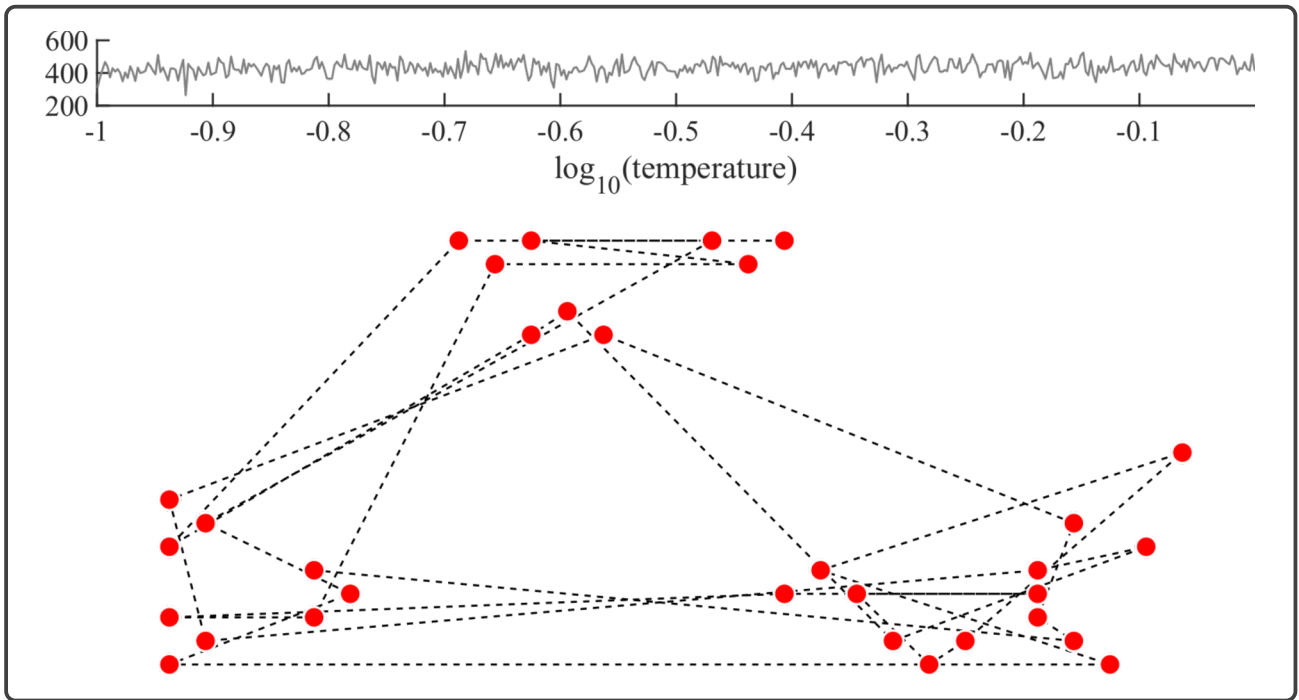


Figure 18.1: Path of the traveling salesman at the end of the *Metropolis* loop for  $T = 10^{-1}$ . The total distance is 309.

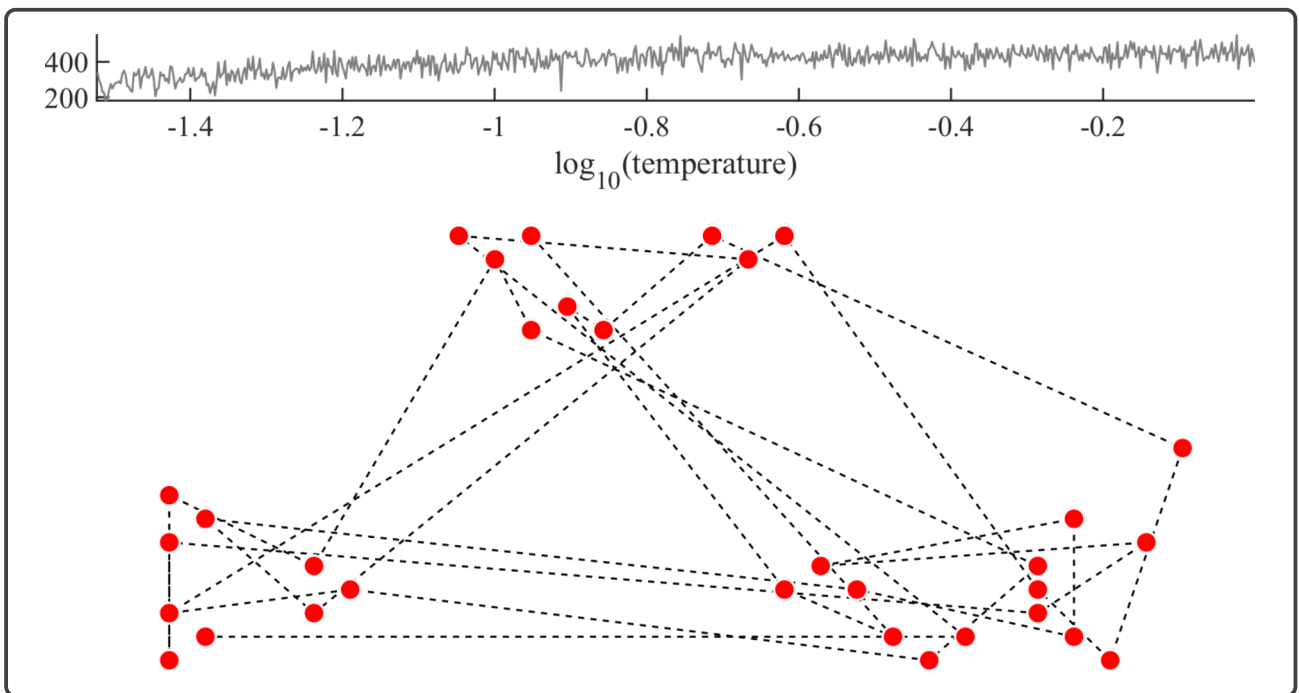
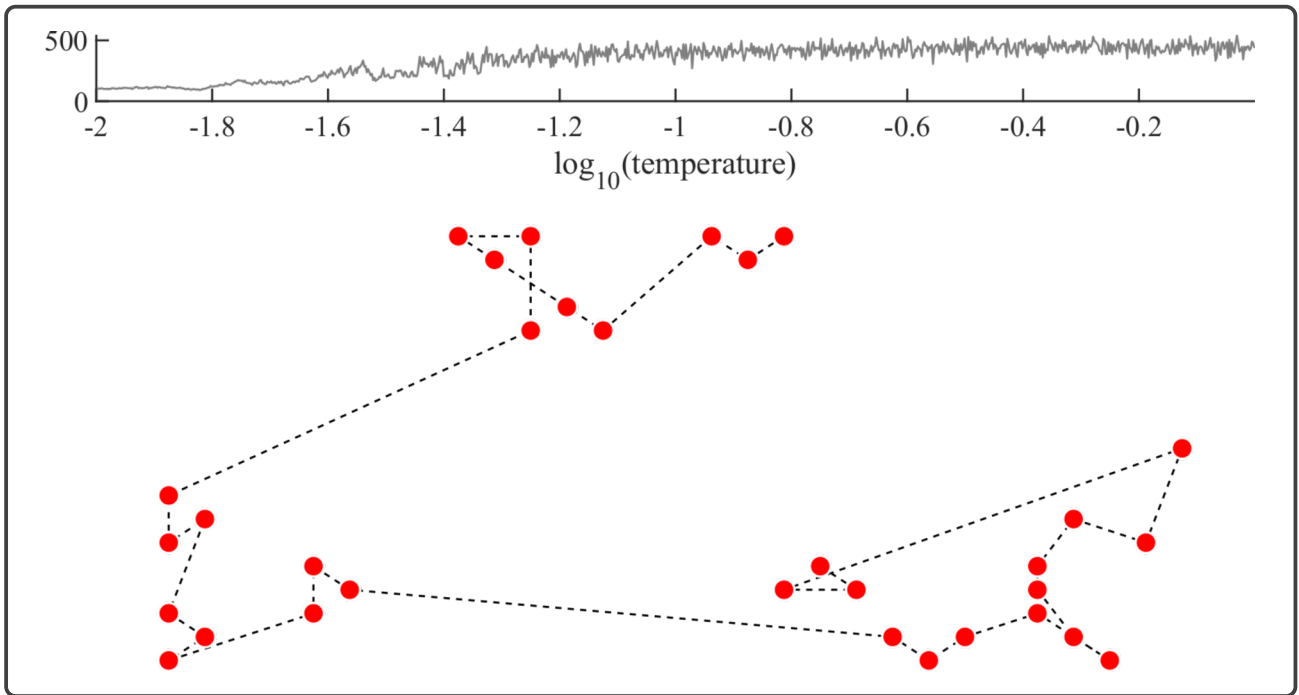
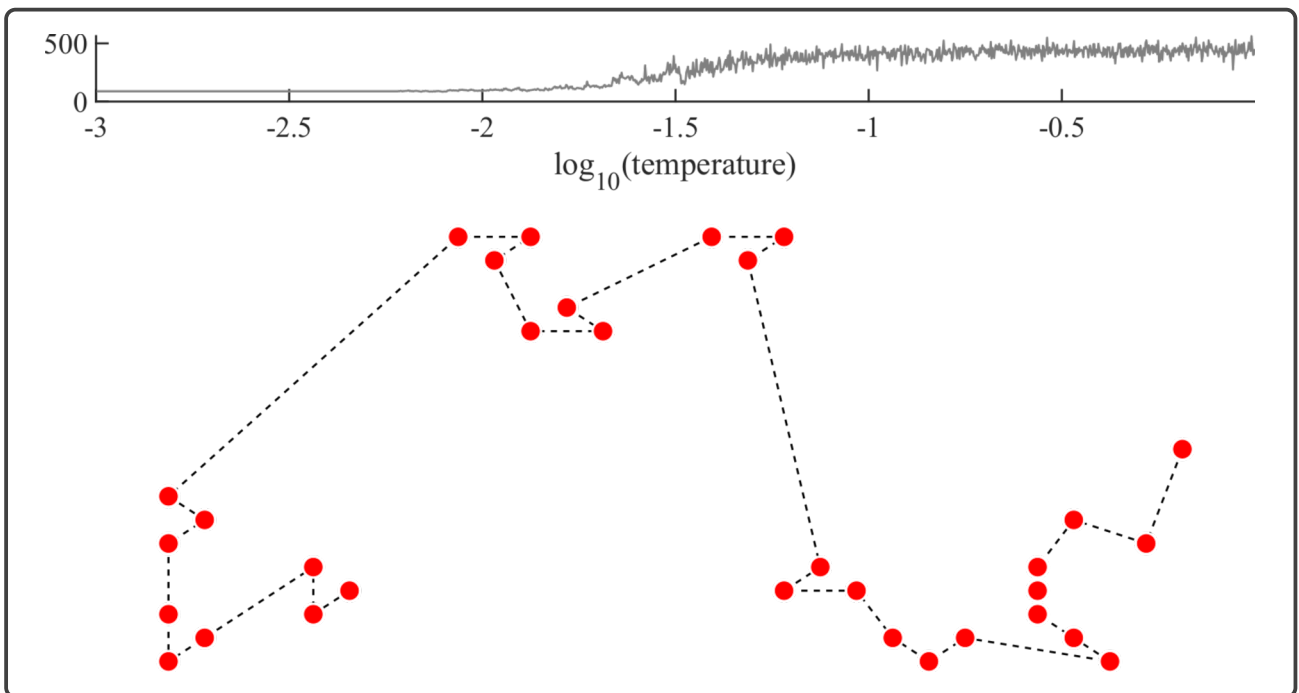


Figure 18.2: Path of the traveling salesman at the end of the *Metropolis* loop for  $T = 10^{-1.5}$ . The beginning of the cost reduction (path length) is noted. The total distance is 203.

## 5. EXAMPLE: THE TRAVELING SALESMAN PROBLEM



*Figure 18.3:* Path of the traveling salesman at the end of the *Metropolis* loop for  $T = 10^{-2}$ . The cost decreases in steps due to the hierarchical geography of the cities. The total distance is 104.



*Figure 18.4:* Path of the traveling salesman at the end of the *Metropolis* loop for  $T = 10^{-3}$ . This is the end of the cooling phase, and the cost no longer decreases significantly. The total distance is 88.



---

---

# CHAPTER 19

---

## METHODS OF LEAST SQUARES

1	Introduction . . . . .	300
2	Linear problem: the normal equations . . . . .	301
3	Singular Value Decomposition & Singular Vectors . . . . .	303
4	Solution provided by the spectral decomposition . . . . .	305
5	Obtained solution properties . . . . .	306
6	Example: Signal deconvolution . . . . .	307

---

# 1 Introduction

Let us revisit the *Bayesian* solution to an inverse problem,

$$\rho(\mathbf{y}|\mathbf{x}) = \frac{\rho(\mathbf{y})p(\mathbf{x}|\mathbf{y})}{\int_{\mathcal{J}} \rho(\mathbf{y})p(\mathbf{x}|\mathbf{y}) d\mathbf{y}}, \quad (1.1)$$

and develop it for the particular case where the *a priori* probability on the models is uniform,  $\rho(\mathbf{y}) = \rho_y$ , and where the errors on the data  $\mathbf{x}$  are distributed according to a centred normal (Gaussian) distribution with zero mean and covariance matrix  $\mathbf{C}$ . We then have,

$$p(\mathbf{x}|\mathbf{y}) = \frac{1}{(2\pi)^{N/2} \sqrt{\det \mathbf{C}}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \mathbf{x}_y)^t \mathbf{C}^{-1} (\mathbf{x} - \mathbf{x}_y) \right], \quad (1.2)$$

where  $\mathbf{x}_y$  represents the predictions (synthetic data) corresponding to the model  $\mathbf{y}$ , and  $N$  is the number of data points considered, which form the components of the vector  $\mathbf{x}$ . When the errors are uncorrelated, the covariance matrix is a diagonal matrix with elements  $\sigma_i$ , and by expanding equation 1.2 above, we obtain,

$$p(\mathbf{x}|\mathbf{y}) = \frac{1}{(2\pi)^{N/2} \prod_{i=1}^N \sigma_i} \exp \left[ -\sum_{i=1}^N \frac{(x_i - x_{y,i})^2}{2\sigma_i^2} \right]. \quad (1.3)$$

The *a posteriori* probability of the models  $\mathbf{y}$  given the data  $\mathbf{x}$  is thus such that,

$$\rho(\mathbf{y}|\mathbf{x}) = \frac{\rho_y}{(2\pi)^{N/2} \prod_{i=1}^N \sigma_i} \exp \left[ -\sum_{i=1}^N \frac{(x_i - x_{y,i})^2}{2\sigma_i^2} \right], \quad (1.4)$$

and we see that the model  $\mathbf{y}_{mc}$  corresponding to the maximum probability density is such that,

$$S(\mathbf{y} = \mathbf{y}_{mc}) \equiv \sum_{i=1}^N \frac{(x_i - x_{\mathbf{y}_{mc},i})^2}{\sigma_i^2} \text{ MINIMUM.} \quad (1.5)$$

The model  $\mathbf{y}_{mc}$  is therefore the one that minimises the sum of the squared differences between the data and the predictions. For this reason, this model is called the least squares solution to the inverse problem.

## 2 Linear problem: the normal equations

When the forward problem is linear, the relationship between the predictions  $\mathbf{x}_y$  and the model parameters  $\mathbf{y}$  is of the form,

$$\mathbf{x}_y = \mathbf{L} \cdot \mathbf{y} \quad (2.1)$$

where the matrix  $\mathbf{L}$  contains the coefficients  $L_{i,j}$  such that the predictions  $x_{y,i}$  are obtained as a linear combination of the  $M$  parameters  $y_j$

$$x_{y,i} = \sum_{j=1}^M L_{i,j} y_j. \quad (2.2)$$

Inserting this equation into the equation 1.5, we obtain,

$$S(\mathbf{y}) = \sum_{i=1}^N \frac{\left(x_i - \sum_{j=1}^M L_{i,j} y_j\right)^2}{\sigma_i^2}, \quad (2.3)$$

and, for  $S$  to be minimised, its partial derivatives with respect to the parameters  $y_k$  must be set to zero, which means that,

$$\frac{\partial S}{\partial y_k} = -2 \sum_{i=1}^N \frac{1}{\sigma_i^2} \left(x_i - \sum_{j=1}^M L_{i,j} y_{mc,j}\right) L_{i,k} = 0 \quad k = 1, \dots, M. \quad (2.4)$$

By changing the order of summation and eliminating some multiplicative factors, the equation 2.4 takes the following form,

$$\sum_{j=1}^M \left(\sum_{i=1}^N \frac{L_{i,k} L_{i,j}}{\sigma_i^2}\right) y_{mc,j} = \sum_{i=1}^N \frac{L_{i,k} x_i}{\sigma_i^2} \quad k = 1, \dots, M. \quad (2.5)$$

Notice that the  $M$  terms on the right side of the above equation are components of a vector such

that,

$$\begin{pmatrix} \vdots \\ \sum_{i=1}^N \frac{L_{i,k}x_i}{\sigma_i^2} \\ \vdots \end{pmatrix} = \begin{bmatrix} \vdots & & & \\ \frac{L_{1,k}}{\sigma_1} & \dots & \frac{L_{i,k}}{\sigma_i} & \dots & \frac{L_{N,k}}{\sigma_N} \\ \vdots & & & & \end{bmatrix} \begin{pmatrix} x_1/\sigma_1 \\ \vdots \\ x_i/\sigma_i \\ \vdots \\ x_N/\sigma_N \end{pmatrix}. \quad (2.6)$$

By introducing the matrix,

$$\mathbf{L}_\sigma \equiv \begin{bmatrix} \frac{L_{1,1}}{\sigma_1} & \dots & \frac{L_{1,k}}{\sigma_1} & \dots & \frac{L_{1,M}}{\sigma_1} \\ \vdots & & \vdots & & \vdots \\ \frac{L_{i,1}}{\sigma_i} & \dots & \frac{L_{i,k}}{\sigma_i} & \dots & \frac{L_{i,M}}{\sigma_i} \\ \vdots & & \vdots & & \vdots \\ \frac{L_{N,1}}{\sigma_N} & \dots & \frac{L_{N,k}}{\sigma_N} & \dots & \frac{L_{N,M}}{\sigma_N} \end{bmatrix}, \quad (2.7)$$

and the vector,

$$\mathbf{x}_\sigma \equiv \begin{pmatrix} x_1/\sigma_1 \\ \vdots \\ x_i/\sigma_i \\ \vdots \\ x_N/\sigma_N \end{pmatrix}, \quad (2.8)$$

equation 2.6 can be written in compact form as

$$\begin{pmatrix} \vdots \\ \sum_{i=1}^N \frac{L_{i,k}x_i}{\sigma_i^2} \\ \vdots \end{pmatrix} = \mathbf{L}_\sigma^T \cdot \mathbf{x}_\sigma. \quad (2.9)$$

Applying the same procedure to the left-hand side of the equation 2.5, this equation becomes,

$$(\mathbf{L}_\sigma^T \cdot \mathbf{L}_\sigma) \cdot \mathbf{y}_{mc} = \mathbf{L}_\sigma^T \cdot \mathbf{x}_\sigma \quad (2.10)$$

The least squares solution  $\mathbf{y}_{mc}$  is formally obtained by solving the equation 2.10,

$$\mathbf{y}_{mc} = (\mathbf{L}_\sigma^T \cdot \mathbf{L}_\sigma)^{-1} \cdot \mathbf{L}_\sigma^T \cdot \mathbf{x}_\sigma. \quad (2.11)$$

Unfortunately, the direct solution of this equation does not generally give an acceptable solution for various reasons that we will examine later, and it is preferable to work directly with the system 2.1.

$$\mathbf{x}_y = \mathbf{L} \cdot \mathbf{y}. \quad (2.12)$$

This system is generally rectangular, since the matrix  $\mathbf{L}$  has  $N$  rows and  $M$  columns, and its solution must be obtained formally by,

$$\mathbf{y}_{mc} = \mathbf{L}_\sigma^\dagger \cdot \mathbf{x}_\sigma, \quad (2.13)$$

where the matrix  $\mathbf{L}_\sigma^\dagger$  is an operator known as the generalised inverse of  $\mathbf{L}_\sigma$ .

## 3 Singular Value Decomposition & Singular Vectors

The decomposition of matrices into singular values and vectors (**SVD**) has its roots in the work of [Eugenio Beltrami](#) (1835-1899) published in 1873, which considered the decomposition of real square matrices. A year later, the mathematician [Camille Jordan](#) independently made the same discoveries. It was not until 1936 that [Eckart and Young](#) established the decomposition of complex rectangular matrices. The most widely used algorithm for performing the **SVD** decomposition of matrices is due to [Gene Golub](#) and [Christian Reinsch](#) (Golub et Reinsch, 1971).

The **SVD** decomposition theorem states that any matrix  $\mathbf{L} \in \mathbf{R}^{N \times M}$  can be factored in the form,

$$\mathbf{L} = \mathbf{U} \cdot \mathbf{\Lambda} \cdot \mathbf{V}^T, \quad (3.1)$$

where  $\mathbf{U} \in \mathbf{R}^{N \times K}$ ,  $\mathbf{V} \in \mathbf{R}^{M \times K}$ , and  $\mathbf{\Lambda} \in \mathbf{R}^{K \times K}$ . The  $K$  columns  $\mathbf{u}_i$  of the matrix  $\mathbf{U}$  are the singular vectors of the matrix  $\mathbf{L} \cdot \mathbf{L}^T$ , and those  $\mathbf{v}_i$  of the matrix  $\mathbf{V}$  are the singular vectors of the matrix  $\mathbf{L}^T \cdot \mathbf{L}$ .



---

So, we have,

$$\mathbf{U}.\mathbf{U}^T = \mathbf{U}^T.\mathbf{U} = \mathbf{I}_N, \quad (3.2)$$

and,

$$\mathbf{V}.\mathbf{V}^T = \mathbf{V}^T.\mathbf{V} = \mathbf{I}_M. \quad (3.3)$$

The matrix  $\Lambda$  is diagonal and its  $K$  elements are the square roots of the singular values of the matrices  $\mathbf{L}.\mathbf{L}^T$  and  $\mathbf{L}^T.\mathbf{L}$ . In general, there are  $K \leq \min(M, N)$  non-zero singular values, where  $K$  is the rank of the matrix  $\mathbf{L}$ . So, we have,

$$\mathbf{L}^T.\mathbf{u}_i = \lambda_i \mathbf{v}_i \quad i = 1, \dots, K, \quad (3.4)$$

and,

$$\mathbf{L}.\mathbf{v}_i = \lambda_i \mathbf{u}_i \quad i = 1, \dots, K. \quad (3.5)$$

The decomposition given by the equation 3.1 yields two sets of orthogonal and normalised vectors  $\mathbf{u}_i$  and  $\mathbf{v}_i$ . The vectors  $\mathbf{u}_i$ , of which there are  $N$  and of dimension  $N$ , form a basis for a subspace of the vector space containing the data vector  $\mathbf{x}$ . The vectors  $\mathbf{v}_i$ , of which there are  $M$  and of dimension  $M$ , form a basis for a subspace of the vector space of the parameters  $\mathbf{y}$ . It is important to note that the matrix  $\mathbf{L}$  is reconstructed using the  $K$  vectors  $\mathbf{u}_i$  and  $\mathbf{v}_i$  associated with the  $K$  non-zero singular values  $\lambda_i$ . To construct bases for the vector spaces of the data  $\mathbf{x}$  and the parameters  $\mathbf{y}$ , it is necessary to complete the bases formed by the vectors  $\mathbf{u}_i$  and  $\mathbf{v}_i$  by adding other orthonormal vectors  $\mathbf{u}_{0,i}$  and  $\mathbf{v}_{0,i}$ . These vectors, which can be considered as singular vectors corresponding to a zero eigenvalue of the matrices  $\mathbf{L}$  and  $\mathbf{L}^T$ , form bases for the zero subspaces of the vector spaces of dimensions  $N$  and  $M$  containing the data and the parameters, respectively. For the basis vectors of the null subspaces, the equations 3.4 and 3.5 are simplified to,

$$\mathbf{L}^T.\mathbf{u}_{0,i} = \mathbf{O} \quad i = K + 1, \dots, N, \quad (3.6)$$

and,

$$\mathbf{L} \cdot \mathbf{v}_{0,i} = \mathbf{O} \quad i = K + 1, \dots, M. \quad (3.7)$$

Similar to the matrices  $\mathbf{U}$  and  $\mathbf{V}$ , whose columns are the vectors  $\mathbf{u}_i$  and  $\mathbf{v}_i$ , the vectors  $\mathbf{u}_{0,i}$  and  $\mathbf{v}_{0,i}$  can be grouped to form matrices denoted  $\mathbf{U}_0$  and  $\mathbf{V}_0$  of dimensions  $N \times (N - K)$  and  $M \times (M - K)$  respectively.

## 4 Solution provided by the spectral decomposition

The spectral decomposition of the matrix  $\mathbf{L}$ , as discussed in the previous section, allows to represent the vectors  $\mathbf{x}$  and  $\mathbf{y}$  in the bases of the singular vectors,

$$\mathbf{x} = \sum_{i=1}^K b_i \mathbf{u}_i + \sum_{i=K+1}^N b_{0,i} \mathbf{u}_{0,i} \quad (4.1)$$

$$= \mathbf{U} \cdot \mathbf{b} + \mathbf{U}_0 \cdot \mathbf{b}_0, \quad (4.2)$$

$$\mathbf{y} = \sum_{i=1}^K a_i \mathbf{v}_i + \sum_{i=K+1}^M a_{0,i} \mathbf{v}_{0,i} \quad (4.3)$$

$$= \mathbf{V} \cdot \mathbf{a} + \mathbf{V}_0 \cdot \mathbf{a}_0, \quad (4.4)$$

where the vectors  $\mathbf{b}$ ,  $\mathbf{b}_0$ ,  $\mathbf{a}$  and  $\mathbf{a}_0$  are the components of the data  $\mathbf{x}$  and the parameters  $\mathbf{y}$ . Using these notations and introducing the decomposition 3.1, the system becomes 2.12,

$$\mathbf{U} \cdot \Lambda \cdot \mathbf{V}^T [\mathbf{V} \cdot \mathbf{a} + \mathbf{V}_0 \cdot \mathbf{a}_0] = \mathbf{U} \cdot \mathbf{b} + \mathbf{U}_0 \cdot \mathbf{b}_0, \quad (4.5)$$

and finding  $\mathbf{y}$  amounts to finding the components  $\mathbf{a}$  and  $\mathbf{a}_0$ . Rewriting the equation 4.5 and premultiplying each term by  $\Lambda^{-1} \cdot \mathbf{U}^T$ , we get,

$$\mathbf{a} = \Lambda^{-1} \cdot \mathbf{b} \quad (4.6)$$

$$= \Lambda^{-1} \cdot \mathbf{U}^T \cdot \mathbf{x} \quad (4.7)$$

The vector  $\mathbf{a}_0$  cannot be determined in the same way from the equation 4.5 and must be set arbitrarily or determined using additional information to that contained in  $\mathbf{x}$ . Therefore the solution  $\mathbf{y}$  is given by

$$\mathbf{y} = \mathbf{V} \cdot \Lambda^{-1} \cdot \mathbf{U}^T \cdot \mathbf{x} + \mathbf{V}_0 \cdot \mathbf{a}_0 \quad (4.8)$$

## 5 Obtained solution properties

The solution given by the equation 4.8 has certain properties which we will now examine. The first of these is that the components  $\mathbf{a}_0$  are arbitrary, which means that the solution is not unique. Uniqueness is only achieved when the number of singular values is equal to the dimension  $M$ , *ie* the number of parameters, because in this case the base  $\mathbf{V}_0$  is empty and  $\mathbf{y}$  is uniquely defined by the components  $\mathbf{a}$ , which are themselves determined by the data  $\mathbf{x}$  in equation 4.7.

The second important property is that the solution obtained via equation 4.8 is a least squares solution in the sense that the residual vector, which contains the discrepancies between the data  $\mathbf{x}$  and the model predictions  $\mathbf{x}_y$ , is such that,

$$\mathbf{e} \equiv \mathbf{L} \cdot \mathbf{y} - \mathbf{x} \quad (5.1)$$

$$= \mathbf{U} \cdot \Lambda \cdot \mathbf{V}^T \cdot [\mathbf{V} \cdot \mathbf{a} + \mathbf{V}_0 \cdot \mathbf{a}_0] - \mathbf{U} \cdot \mathbf{b} - \mathbf{U}_0 \cdot \mathbf{b}_0 \quad (5.2)$$

$$= \mathbf{U} \cdot [\Lambda \cdot \mathbf{a} - \mathbf{b}] - \mathbf{U}_0 \cdot \mathbf{b}_0. \quad (5.3)$$

The norm of this vector is,

$$\mathbf{e}^T \cdot \mathbf{e} = \|\Lambda \cdot \mathbf{a} - \mathbf{b}\|^2 + \|\mathbf{b}_0\|^2, \quad (5.4)$$

and is minimal if  $\mathbf{a} = \Lambda^{-1} \cdot \mathbf{b}$  (equation 4.6), *i.e.* if the solution  $\mathbf{y}$  is the one given by equation 4.8. In this case the quadratic error is,

$$\mathbf{e}^T \cdot \mathbf{e} = \|\mathbf{b}_0\|^2, \quad (5.5)$$

and is solely controlled by the projection of the data  $\mathbf{x}$  onto the vectors  $\mathbf{u}_{0,i}$  of the null subspace. Based on this result, the equation 4.8 can be rewritten using the notation denoting the least squares solution,

$$\mathbf{y}_{mc} = \mathbf{V} \cdot \Lambda^{-1} \cdot \mathbf{U}^T \cdot \mathbf{x} + \mathbf{V}_0 \cdot \mathbf{a}_0. \quad (5.6)$$

## 6 Example: Signal deconvolution

We will now illustrate the previous sections with an example commonly encountered in signal processing: deconvolution. It is indeed common, as in seismology, to try to recover the input signal  $y(t)$  of a system (assumed to be linear and stationary) from the output signal  $x(t)$  and the impulse response of the system  $l(t)$ . The relationship is given by,

$$x(t) = l(t) * y(t). \quad (6.1)$$

In practice, the convolution  $*$  is applied to discrete and truncated signals ( $\mathbf{x}$ ,  $\mathbf{y}$ , and  $\mathbf{l}$ ) *via* the Z-transform,

$$\sum_{n=1}^N x_n Z^n = \left( \sum_{j=1}^J l_j Z^j \right) \left( \sum_{m=1}^M y_m Z^m \right) \quad (6.2)$$

In this example, we will consider a system where the output  $\mathbf{x}$  is the second derivative of the input signal  $\mathbf{y}$ , taking,

$$\mathbf{l} = \begin{pmatrix} -1 \\ +2 \\ -1 \end{pmatrix} \quad (6.3)$$

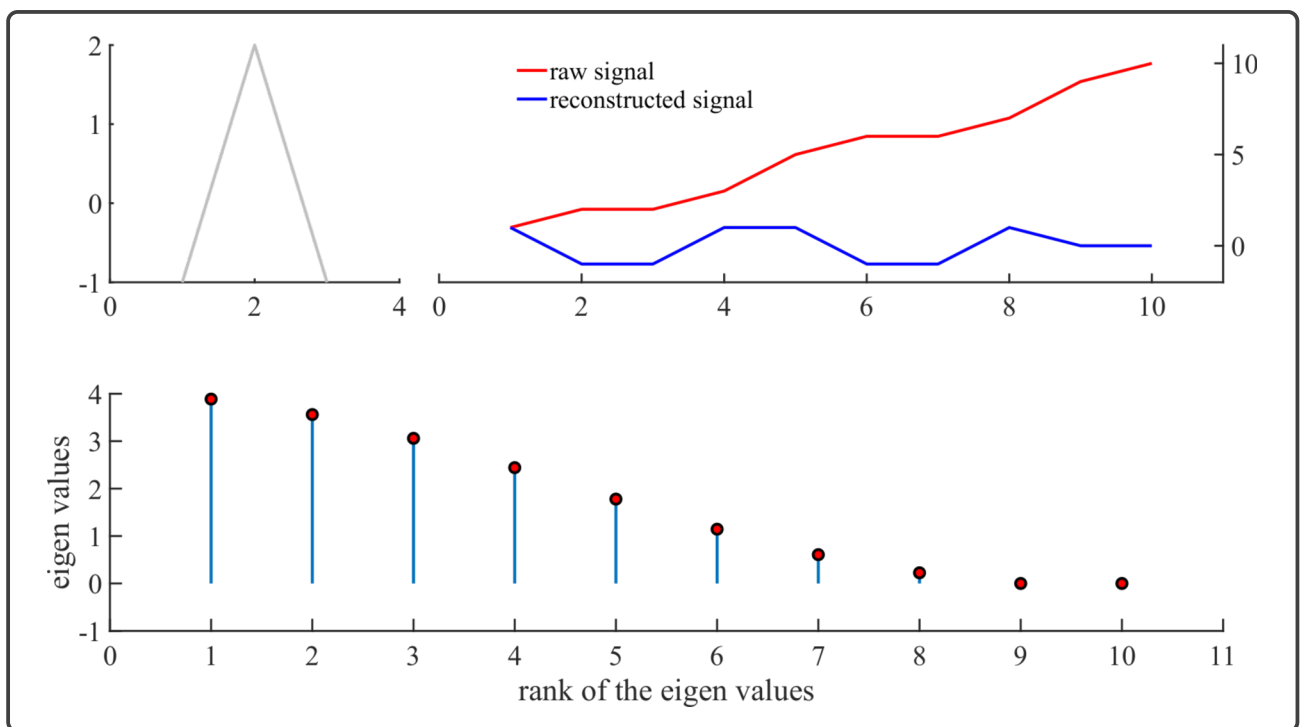
The application of this filter to an input signal  $\mathbf{y}$  can be written in matrix form, revealing the matrix  $\mathbf{L}$  as discussed in the previous sections.

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \end{pmatrix} = \begin{pmatrix} -1 & 2 & -1 & 0 & \cdots \\ 0 & -1 & 2 & -1 & \cdots \\ 0 & 0 & -1 & 2 & \cdots \\ 0 & 0 & 0 & -1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \vdots \end{pmatrix} \quad (6.4)$$

where the matrix, which has a very specific structure, is called a **Toeplitz** matrix and in this case has been written by removing the edge effects of the convolution given by the equation 6.2.

Figure 19.1 shows the results of the inversion by singular value decomposition (SVD) and singular vectors of the **Toeplitz** matrix for an input signal  $\mathbf{y}$  with  $M = 10$  values and an output

signal  $\mathbf{x}$  with  $N = 8$  values. This is an example of an underdetermined problem, since the number of unknowns exceeds the number of data points. The **Toeplitz** matrix therefore has at most  $K = 8$  non-zero singular values. This is illustrated in the lower part of figure 19.1, which shows the spectrum of the singular values of the **Toeplitz** matrix. Since the signal  $\mathbf{y}$  belongs to a vector space of dimension  $M = 10$ , there are two basis vectors corresponding to zero singular values, which form a basis for the zero subspace. The solution  $\mathbf{y}_{mc}$  obtained from equation 5.6 by setting  $\mathbf{a}_0 = 0$  is shown in the top right of figure 19.1 (solid line). It can be seen that this solution differs significantly from the theoretical solution, shown as a dashed line, and contains a significant trend that is not captured in the  $\mathbf{y}_{mc}$  solution. Figure 19.2 shows the singular vectors that form the basis of the  $\mathbf{y}$  solution space. The vectors corresponding to zero singular values are numbers 8 and 10 from the bottom, and it is evident that these vectors model a linear trend and a constant value. This explains why the trend is not found in the solution; it belongs to the zero subspace. This is logical since the filter is a second derivative operator that cancels constant or linear functions. It is of course possible to obtain a solution identical to the theoretical one, but this requires the choice of the correct vector  $\mathbf{a}_0$ , which can only be done using *a priori* information provided in addition to the data  $\mathbf{x}$ .



**Figure 19.1: Results of signal deconvolution using a second-order finite difference filter.**

## 6. EXAMPLE: SIGNAL DECONVOLUTION

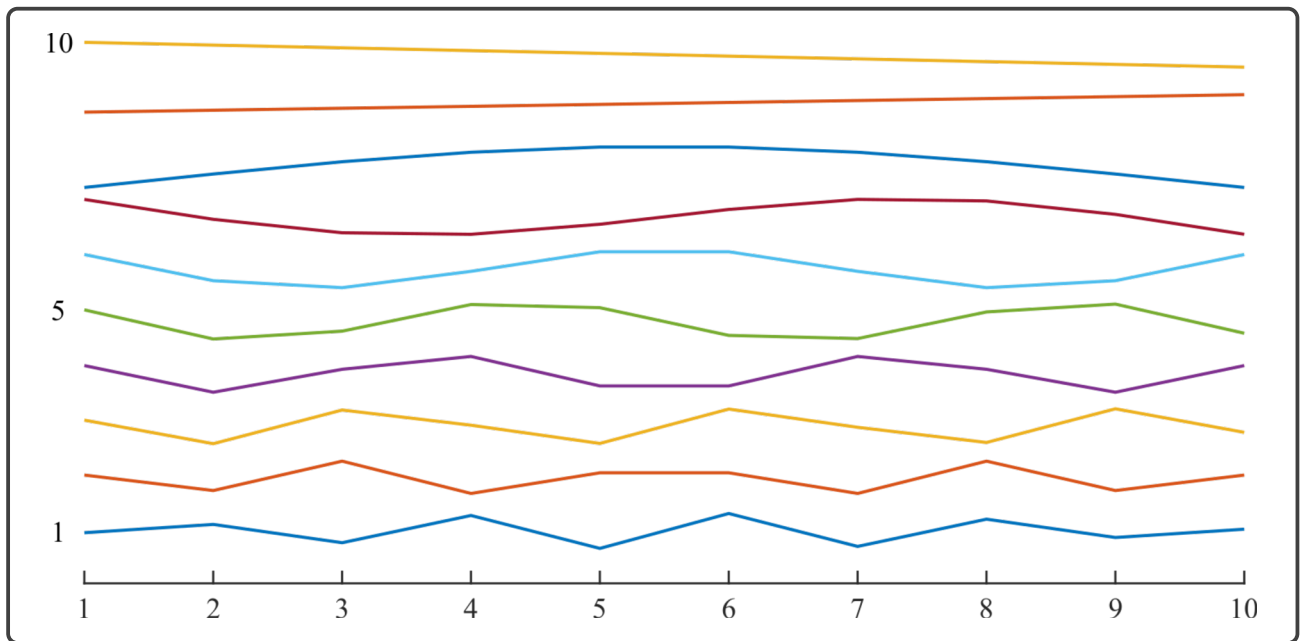


Figure 19.2: Singular vectors corresponding to the singular values shown at the bottom of Figure 19.1.



---

---

# CHAPTER 20

---

## GENERATION OF *A PRIORI* MODELS

<b>1</b>	<b>Introduction</b> . . . . .	<b>312</b>
<b>2</b>	<b>Convex sets: Definitions</b> . . . . .	<b>312</b>
<b>3</b>	<b>Projections onto convex sets</b> . . . . .	<b>313</b>
3.1	Imposed Values . . . . .	313
3.2	Valeurs bornées . . . . .	313
3.3	Discontinuity . . . . .	313
3.4	Sequencing . . . . .	314
3.5	Imposed Mean . . . . .	314
3.6	Maximum Energy . . . . .	315



---

# 1 Introduction

We have seen the benefit of being able to generate models according to the *a priori* probability density  $\rho(\mathbf{y})$ . This capability significantly improves the efficiency of simulated annealing in the sense that more models constructed in this way are retained. Furthermore, generating models according to the *a priori* probability allows for more easily escaping from local minima when dealing with a multimodal *a posteriori* density.

It is therefore interesting to have methods capable of producing models that immediately satisfy a set of constraints, which may be of a highly variable nature. Such techniques exist, and one of the most popular in the geosciences is undoubtedly geostatistics, whose success is based on its ability to incorporate qualitative and disparate geological information quantitatively. Another way of incorporating a priori constraints relatively easily is to use the method of projection onto convex subspaces. This technique allows an arbitrary model to be modified into one that comes close to satisfying the required constraints.

## 2 Convex sets: Definitions

Let us first establish the basic mathematical concepts. We say that a set of models  $\mathcal{E} = \{\mathbf{y}_i\}$  is convex if,

$$\forall (\mathbf{y}_i, \mathbf{y}_j) \in \mathcal{E} \times \mathcal{E} \text{ et } \lambda \in [0, 1], \text{ alors } \lambda \mathbf{y}_i + (1 - \lambda) \mathbf{y}_j \in \mathcal{E}. \quad (2.1)$$

It is easy to show that,

$$\mathcal{E} \text{ et } \mathcal{E}' \text{ convexes} \Rightarrow \mathcal{E} \cap \mathcal{E}' \text{ convexe.} \quad (2.2)$$

We will say that a convex set is a cone if,

$$\forall \mathbf{y} \in \mathcal{E} \text{ et } \mu > 0, \text{ alors } \mu \mathbf{y} \in \mathcal{E}. \quad (2.3)$$

The sets  $\mathbb{R}^n$  and  $\mathbb{R}_+^n$  are convex. A closed interval  $[a, b]$  in  $\mathbb{R}$  is convex, as well as the set of positive continuous functions.

### 3 Projections onto convex sets

We will now consider some convex sets that are particularly interesting for generating models in the simulated annealing algorithm. For each of these sets, we will also show how to project an arbitrary model onto these convex sets.

#### 3.1 Imposed Values

We will denote by  $\mathcal{C}$  the set of models  $\mathbf{y}_i$  for which certain components have known and fixed values  $c_k$ , that is to say,

$$\mathcal{C} \equiv \{\mathbf{y}_i; y_{i,k} = c_k\}. \quad (3.1)$$

It is easy to show that this set is convex. The projection of any model  $\mathbf{y}$  onto  $\mathcal{C}$  is obtained by assigning the fixed values  $c_k$  to the corresponding components,

$$\mathbf{y} \rightarrow \mathbf{y}_{\mathcal{C}}; y_{\mathcal{C},k} = c_k. \quad (3.2)$$

#### 3.2 Valeurs bornées

The set  $\mathcal{B}$  denotes the class of models whose components are bounded,

$$\mathcal{B} \equiv \{\mathbf{y}_i; a_k \leq y_{i,k} \leq b_k\}. \quad (3.3)$$

Projection onto this convex set involves adjusting the components whose values are outside the allowed interval,

$$\mathbf{y} \rightarrow \mathbf{y}_{\mathcal{B}}; y_{\mathcal{B},k} = \max[\min(y_{i,k}, b_k), a_k]. \quad (3.4)$$

#### 3.3 Discontinuity

The set  $\mathcal{D}$  denotes the class of models with a discontinuity of amplitude  $\mathbf{d}(x)$  along a boundary defined by  $f(x) = 0$ . The set is thus defined by,

---


$$\mathcal{D} \equiv \{\mathbf{y}_i; \mathbf{y}_i(x^+) - \mathbf{y}_i(x^-) = \mathbf{d}(x)\}, \quad (3.5)$$

and the projection onto this convex set is such that,

$$\begin{aligned} \mathbf{y} \rightarrow \mathbf{y}_{\mathcal{D}} &= \mathbf{y} + \frac{1}{2} \mathbf{d}(x) ; f(x) > 0 \\ \mathbf{y} \rightarrow \mathbf{y}_{\mathcal{D}} &= \mathbf{y} - \frac{1}{2} \mathbf{d}(x) ; f(x) < 0. \end{aligned} \quad (3.6)$$

### 3.4 Sequencing

The set  $\mathcal{S}$  denotes the class of models with a specified ordering along. We have,

$$\mathcal{S} \equiv \{\mathbf{y}_i; \mathbf{y}_i(x_1) \geq \mathbf{y}_i(x_2)\} \quad (3.7)$$

and the projection onto this convex set is achieved by,

$$\mathbf{y}(x_1) = \mathbf{y}(x_2) = \frac{1}{2} (\mathbf{y}(x_1) + \mathbf{y}(x_2)) \quad (3.8)$$

This constraint allows, for example, for the imposition of rivers when generating fractal terrains.

### 3.5 Imposed Mean

The set  $\mathcal{M}$  denotes the class of models with an imposed mean  $y_{moy}$ . We have,

$$\mathcal{M} \equiv \{\mathbf{y}_i; \langle \mathbf{y}_i \rangle = y_{moy}\} \quad (3.9)$$

and the projection is performed by,

$$\mathbf{y} \rightarrow \mathbf{y}_{\mathcal{M}} = \mathbf{y} - \langle \mathbf{y} \rangle + y_{moy}. \quad (3.10)$$

#### 3.6 Maximum Energy

The set  $\mathcal{E}$  denotes the class of models with energy  $e$  less than or equal to a certain value  $e_0$ . We have,

$$E \equiv \left\{ \mathbf{y}_i; e = \int y^2(x) dx \leq e_0 \right\} \quad (3.11)$$

and the projection is performed by,

$$\mathbf{y} \rightarrow \mathbf{y}_E = \mathbf{y}^* \sqrt{e_0/e} \quad (3.12)$$



---

## REFERENCES

- Backus, G. et F. Gilbert (1968). The resolving power of gross earth data. *Geophysical Journal International* 16(2), 169–205.
- Backus, G. et F. Gilbert (1970). Uniqueness in the inversion of inaccurate gross earth data. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 266(1173), 123–192.
- Backus, G. E. et J. Gilbert (1967). Numerical applications of a formalism for geophysical inverse problems. *Geophysical Journal International* 13(1-3), 247–276.
- Barnard, G. A. et T. Bayes (1958). Studies in the history of probability and statistics: Ix. thomas bayes’s essay towards solving a problem in the doctrine of chances. *Biometrika* 45(3/4), 293–315. [233](#)
- Bernoulli, D. (1753). Réflexions et éclaircissemens sur les nouvelles vibrations des cordes. *Hist. Mém Acad. R. Sci. Lett. Berlin* 9, 147–172. [19](#), [21](#)
- Bracewell, R. N. et R. N. Bracewell (1986). *The Fourier transform and its applications*, Volume 31999. McGraw-Hill New York. [17](#)
- Brillouin, L. (1959). La science et la théorie de l’information. [237](#)
- Brown, R. (1828). Xxvii. a brief account of microscopical observations made in the months of june, july and august 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. *Philosophical Magazine Series* 2 4(21), 161–173. [106](#)
- Chandler, S. (1891a). On the variation of latitude, i. *The Astronomical Journal* 11, 59–61.
- Chandler, S. (1891b). On the variation of latitude, ii. *The Astronomical Journal* 11, 65–70.
- Cipra, B. A. (2000). The best of the 20th century: Editors name top 10 algorithms. *SIAM news* 33(4),

---

1–2. [95](#)

- Claerbout, J. F. (1985). Fundamentals of geophysical data processing. [91](#)
- Claerbout, J. F. (1992). *Earth soundings analysis: processing versus inversion*. Blackwell Scientific Publications. [95](#)
- Cohen-Tannoudji, C., B. Diu, F. Laloë, et B. Crasemann (1998). Quantum mechanics. [51](#)
- Cooley, J. W. et J. W. Tukey (1965). An algorithm for the machine calculation of complex fourier series. *Mathematics of computation* 19(90), 297–301. [95](#)
- Courtillot, V., J.-L. Le Mouël, F. Lopes, et D. Gibert (2022). On sea-level change in coastal areas. *Journal of Marine Science and Engineering* 10(12), 1871. [215](#)
- Danielson, G. C. et C. Lanczos (1942). Some improvements in practical fourier analysis and their application to x-ray scattering from liquids. *Journal of the Franklin Institute* 233(5), 435–452. [95](#)
- Dirac, P. A. (1925). The fundamental equations of quantum mechanics. Dans *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, Volume 109, pp. 642–653. The Royal Society. [68](#)
- Feynman, R., R. Leighton, et M. Sands (2013). *Le cours de physique de Feynman*. Dunod.
- Feynman, R. P. (1980). *La nature de la physique*. Le Seuil.
- Feynman, R. P., R. B. Leighton, et M. Sands (2011). *The Feynman lectures on physics, Vol. I: The new millennium edition: mainly mechanics, radiation, and heat*, Volume 1. Basic books. [121](#)
- Fourier, J. (1822). *Theorie analytique de la chaleur, par M. Fourier*. Chez Firmin Didot, père et fils. [21](#)
- Franklin, J. N. (1970). Well-posed stochastic extensions of ill-posed linear problems. *Journal of mathematical analysis and applications* 31(3), 682–716.
- Gabor, D. (1946). Theory of communication. part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering* 93(26), 429–441. [38](#)
- Golub, G. H. et C. Reinsch (1971). Singular value decomposition and least squares solutions. Dans *Linear Algebra*, pp. 134–151. Springer. [205](#), [303](#)
- Grossmann, A. et J. Morlet (1984). Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM journal on mathematical analysis* 15(4), 723–736. [177](#)
- Gubbins, D. (1971). Two dimensional digital filtering with haar and walsh transforms. Dans *Annales de Geophysique*, Volume 27, pp. 85–104. [185](#)
- Haar, A. (1909). *Zur theorie der orthogonalen funktionensysteme*. Georg-August-Universitat, Göttingen. [177](#)
- Hartley, R. V. (1942). A more symmetrical fourier analysis applied to transmission problems. *Proceedings of the IRE* 30(3), 144–150. [29](#)

- Hauer, J. F., C. Demeure, et L. Scharf (1990). Initial results in prony analysis of power system response signals. *IEEE Transactions on power systems* 5(1), 80–89. [12](#)
- Heil, C., D. Walnut, et I. Daubechies (2006). *Fundamental papers in wavelet theory*. Princeton University Press. [177](#)
- Heisenberg, W. (1927). Über den anschaulichen inhalt der quantentheoretischen kinematik und mechanik. [121](#)
- Hildebrand, F. B. (1956). *Introduction to numerical analysis*. McGraw-Hill Book, Co.
- Hudson, J. et J. Heritage (1981). The use of the born approximation in seismic scattering problems. *Geophysical Journal International* 66(1), 221–240. [51](#)
- Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Trans. Amer. Soc. Civil Eng.* 116, 770–808. [111](#)
- Jackson, D. D. (1972). Interpretation of inaccurate, insufficient and inconsistent data. *Geophysical Journal International* 28(2), 97–109.
- Jackson, D. D. (1979). The use of a priori data to resolve non-uniqueness in linear inversion. *Geophysical Journal International* 57(1), 137–157.
- Kac, M. (1966). Can one hear the shape of a drum? *The american mathematical monthly* 73(4), 1–23. [35](#)
- Kanasewich, E. (1981a). *Time sequence analysis in geophysics*. University of Alberta Press.
- Kanasewich, E. R. (1981b). *Time sequence analysis in geophysics*. University of Alberta. [132](#)
- Kasdin, N. J. (1995). Discrete simulation of colored noise and stochastic processes and  $1/f^\alpha$  power law noise generation. *Proceedings of the IEEE* 83(5), 802–827. [106](#)
- Kay, S. M. et S. L. Marple (1981). Spectrum analysis—a modern perspective. *Proceedings of the IEEE* 69(11), 1380–1419.
- Kirkpatrick, S., C. D. Gelatt Jr, et M. P. Vecchi (1983). Optimization by simulated annealing. *science* 220(4598), 671–680. [292](#)
- Kumazawa, M., Y. Imanishi, Y. Fukao, M. Furumoto, et A. Yamamoto (1990). A theory of spectral analysis based on the characteristic property of a linear dynamic system. *Geophysical Journal International* 101(3), 613–630. [12](#)
- Lemmerling, P. et S. Van Huffel (2001). Analysis of the structured total least squares problem for hankel/toeplitz matrices. *Numerical Algorithms* 27(1), 89–114. [212](#)
- Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics* 2(2), 164–168.
- Lines, L. R., A. K. Schultz, et S. Treitel (1988). Cooperative inversion of geophysical data. *Geophysics* 53(1), 8–20. [233](#)



- 
- Lopes, F., J. Le Mouél, V. Courtillot, et D. Gibert (2021). On the shoulders of laplace. *Physics of the Earth and Planetary Interiors* 316, 106693. [217](#)
- Mallat, S. (1999). *A wavelet tour of signal processing*. Academic press. [39](#)
- Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence* 11(7), 674–693.
- Mañé, R. (1981). On the dimension of the compact invariant sets of certain non-linear maps. Dans *Dynamical systems and turbulence, Warwick 1980*, pp. 230–242. Springer.
- Markowitz, W. et B. Guinot (1968). *Continental drift, secular motion of the pole, and rotation of the Earth*. Springer. [218](#)
- Menke, W. (1984). *Geophysical data analysis: Discrete inverse theory*. Academic Press. [100](#)
- Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, et E. Teller (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics* 21(6), 1087–1092. [233](#), [284](#)
- Morlet, J., G. Arens, E. Fourgeau, et D. Glard (1982). Wave propagation and sampling theory—part i: Complex signal and scattering in multilayered media. *Geophysics* 47(2), 203–221. [177](#)
- Morse, P. et H. Feshbach (1953). *Methods of theoretical physics*. 1953. McGraw-Hill.
- Oppenheim, A. V. (1965). Superposition in a class of nonlinear systems. [129](#)
- Oppenheim, A. V. et R. W. Schafer (2004). From frequency to quefrequency: A history of the cepstrum. *IEEE signal processing Magazine* 21(5), 95–106. [129](#)
- Papoulis, A. (1984). *Probability, random variables and stochastic processes*. McGraw-Hill. [104](#)
- Penrose, R. (1955). A generalized inverse for matrices. Dans *Mathematical proceedings of the Cambridge philosophical society*, Volume 51, pp. 406–413. Cambridge University Press.
- Pisarenko, V. F. (1973). The retrieval of harmonics from a covariance function. *Geophysical Journal International* 33(3), 347–366.
- Press, W., B. Flannery, S. Teutolsky, et W. Vetterling (1986). *Numerical recipes*. Cambridge University Press.
- Prony, R. (1795). Essai experimental–,–. *J. de l'Ecole Polytechnique* 2.
- Roach, G. F. (1982). *Green's functions*, Volume 239. Cambridge University Press Cambridge. [50](#)
- Schwartz, L. (1950). Théorie des distributions, vols. i–ii hermann. *Paris (1950–1951)*. [68](#)
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal* 27(3), 379–423.
- Shannon, C. E., W. Weaver, et A. W. Burks (1951). The mathematical theory of communication. [83](#)
- Takens, F. (1981). Detecting strange attractors in fluid turbulence, in rand, d. and young, l.-s., eds. *dynamical systems and turbulence*, springer-verlag, berlin.

## REFERENCES

---

- Talwani, M., J. L. Worzel, et M. Landisman (1959). Rapid gravity computations for two-dimensional bodies with application to the mendocino submarine fracture zone. *Journal of Geophysical Research* 64(1), 49–59.
- Tarantola, A. et B. Valette (1982). Generalized nonlinear inverse problems solved using the least squares criterion. *Reviews of Geophysics* 20(2), 219–232. [248](#)
- Thom, R. et E. Noël (1991). *Prédire n'est pas expliquer*. Eshel.
- Vautard, R. et M. Ghil (1989). Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series. *Physica D-Nonlinear Phenomena* 35, 395–424.
- Vautard, R., P. Yiou, et M. Ghil (1992). Singular-spectrum analysis: A toolkit for short, noisy chaotic signals. *Physica D: Nonlinear Phenomena* 58(1-4), 95–126.
- Vozoff, K. et D. Jupp (1975). Joint inversion of geophysical data. *Geophysical Journal International* 42(3), 977–991.



---

# LIST OF FIGURES

1.1	Examples of Fourier transforms of simple signals. On the left, the real part of the transform corresponds to the frequency carried by the cosine component of the signal, and the imaginary part corresponds to the sine component. On the right, phase shifts cause the two frequencies present in the signal to decompose into the real and imaginary parts of the Fourier transform. . . . .	23
1.2	Illustration of the upward extension of a gravity anomaly. The discrepancy between the theoretical anomaly and the extended anomaly is due to numerical inaccuracies resulting from sampling. . . . .	27
1.3	Polygon ABCDEF, of infinite dimension along Oy, used to calculate the theoretical magnetic or gravity anomaly produced by the perturbing body. . . . .	28
1.4	Sine function between $-\pi$ and $\pi$ . In blue and red, the areas (negative and positive) of the sine function illustrating the integration of the sine over a period. The sum of these two areas is zero. . . . .	37
1.5	Symmetries of the Fourier transform for different types of functions . . . . .	38
1.6	Heisenberg box schematizing the time-frequency duality of a Fourier atom . . . . .	41
2.1	Diagram of Earth's structure obtained from seismic wave tomography. The yellow triangles represent the seismometers used by geophysicists. . . . .	47

---

2.2	Temporal convolutions. This example is frequently encountered in seismics; the two signals at the top are the impulse responses of the subsurface. This is what would be obtained if we were capable of emitting a Dirac impulse using seismic sources. In practice, seismic sources are Ricker impulses (left center) or sweeps or chirps (right center). The resulting seismic traces (bottom) are the convolution of the impulse response (top) with the sources (middle). Note that the sweep, which has a long duration, completely masks the events present in the impulse response. How to retrieve them? See the section on correlation and Figure ??.	50
2.3	Example of cross-correlation. The convolution of an ideal seismic impulse response (top left) with a sweep (top right) emitted by a vibratory truck produces a trace (bottom right) in which arrivals are indistinguishable. The cross-correlation (bottom left) between this trace and the sweep helps to better discern the arrivals. This operation is routinely performed in seismic surveys when vibratory trucks are used as sources. Its effectiveness is due to the very specific shape (frequency sweep) of the sweep.	56
3.1	<b>Envelope Calculation by Hilbert Transform.</b> The initial signal (top, in gray) has an analytic function which is a complex function whose real part (bottom, in green) is equal to the initial signal itself, and whose imaginary part (bottom, in blue) is the Hilbert transform of the initial signal. The envelope (bottom, in red) is obtained by calculating the magnitude of the analytic signal.	59
4.1	Sine cardinal functions. In blue, the most commonly known unnormalized form, and in red, the normalized form.	63
4.2	From top to bottom, and from left to right, different stages of the self-convolution product of a rectangular function (blue curve) by itself (green curve). As can be seen, the result is the triangle function (black curve).	66
4.3	Top: Evolution of the window function as $T$ approaches 0. Bottom: Evolution of the Gaussian function as the standard deviation $\sigma$ approaches 0. As can be seen in both cases, the temporal support decreases, the functions thin out and tend to infinity, thus approaching Dirac distributions.	70

5.1 Illustration of the effects of truncation on frequency resolution  $\delta u$ . The two frequencies become indistinguishable when the truncation no longer reveals the presence of beats in the signal. At that point, the truncated signal can be interpreted as a single damped sinusoid instead of two sinusoids producing beats. We indeed find that the amplitude of the Fourier transform is also modified and corresponds to half the duration of the truncation window (*cf* equation 1.3). . . . . 82

5.2 Illustration of the effects of discretization and spectral aliasing on the apparent frequencies of three sinusoids with frequencies of 0.2 Hz, 0.067 Hz, and 0.033 Hz. The sampling periods, from bottom to top (from black to red), are 1 second (1 Hz), 2 seconds (0.5 Hz), 4 seconds (0.25 Hz), and 8 seconds (0.125 Hz). The initial signal (top left) is discretized with a one-second interval and results from the superposition of sinusoids, including two high-frequency ones, as shown by the magnitude of the Fourier transform of the signal (top right). Under-sampling with a two-second interval (blue curves) violates the Shannon condition and causes spectral aliasing. Under-sampling with a four-second interval (green curves) further distorts the spectral lines. . . . . 85

5.3 Example of spectral aliasing. At the top, the identical waveforms of four sinusoids with different frequencies, whose spectra are also identical because, in 3 out of 4 cases (red, green, and blue curves), we do not meet the Shannon sampling theorem. 87

5.4 Loss of information due to spectral aliasing. The signal in the middle left is constructed by under-sampling the initial signal (top left) with a two-second interval followed by Shannon interpolation at a one-second interval. Although the constructed signal is sampled as finely as the initial signal, the aliasing that occurred during under-sampling has not been eliminated, as shown by the amplitude spectrum in the middle right. The difference (bottom right) between this spectrum and that of the initial signal (top right) shows that the interpolated signal no longer contains the original high frequencies between 0.3 and 0.4 Hz, which have been aliased around 0.1 Hz. This is also evident in the difference (bottom left) between the two signals, which contains the high-frequency oscillations missing in the interpolated signal. . . . . 88

7.1 Comparison of computation times between the DFT and the FFT. The Fourier transform is performed on a signal consisting of  $2^N$  points in red and  $2^N - 1$  points in blue. It is observed that the so-called fast algorithm is nearly 5 times faster. . . . . 96

7.2 Direct and inverse Fourier transforms obtained by inverting the Vandermonde matrix. 102

---

8.1	<b>1/f Noise + Sine Wave.</b> These noises have identical phase spectra, which gives them correlated morphologies and allows us to observe that the black noise is a "smoother" version of the <i>brownian</i> noise, which in turn is a "smoother" version of the pink noise, <i>etc</i> . . . . .	106
8.2	<b>1/f Noise.</b> Representation of the energy spectra, in logarithmic scale, for white noise (blue curve), pink noise (red curve), brownian noise (brown curve), and black noise (black curve). . . . .	107
8.3	Quantization noise. The initial signal (top left). Its severe quantization (middle left) produces an amplitude spectrum (middle right) that is noisy compared to the initial spectrum (top right). Curves obtained with the program <code>ex_bruit_quantification.m</code> . . . . .	108
8.4	Quantization noise. The initial signal (top left) is still our "favorite" previously shown in several figures. Its severe quantization (middle left) yields an amplitude spectrum (middle right) that is noisy compared to the initial spectrum (top right). The difference between the two signals (bottom left) is the quantization noise, whose spectrum (bottom right) is close to that of white noise. . . . .	109
8.5	Pink noise generated by the superposition of simple processes with different time constants. In this example, the processes are six dice, with the first being rolled at each time increment, the second only every other time, the third every fourth time, <i>etc</i> The average of the values displayed by the dice is calculated at each time increment, producing the noise on the left, whose amplitude spectrum (on the right) is reasonably pink. Thus, it is seen that the superposition of a small number of processes with different time constants easily produces pink noise. Such situations likely occur frequently in Nature. . . . .	111
8.6	<b>Two stable laws.</b> The distributions of Gauss (red) and Cauchy (blue). Although these two distributions do not appear very different at first glance, the resulting statistical consequences are dramatically so, as you will see later in Figures ?? and ??. These two curves were produced using the program <code>ex_lois_stables.m</code> , which calls the sub-functions <code>fct_cauchy.m</code> and <code>fct_normal.m</code> . . . . .	112

8.7 Illustration of the central limit theorem. The initial distribution (top left) has a finite variance and differs markedly from the Gaussian distribution with the same variance. Convolution of this initial distribution with itself (top right) yields a distribution that is already closer to the Gaussian. The triple autocorrelation (bottom left) and quadruple autocorrelation (bottom right) show that convergence is very rapid. Redo this figure by changing the initial distribution using the program `ex_theoreme_central_limite.m`; you will see that the convergence is striking in almost all cases. . . . . 113

8.8 **Gaussian noise.** Addition of traces to increase the signal-to-noise ratio. In this example, the noise is white and Gaussian. The goal is to recover the initial signal (top left) by calculating the average of a certain number of noisy realizations. The average of 9 realizations (top right) allows for the identification of the first two events of the pure signal. Averages taken over more realizations (middle and bottom) improve the signal-to-noise ratio and allow for the recovery of the pure signal arrivals. An average taken over an infinite number of realizations converges stochastically to the pure signal. . . . . 115

8.9 **Cauchy noise.** Addition of traces to increase the signal-to-noise ratio. The disaster becomes apparent when summing the traces; the signal-to-noise ratio increases dramatically! . . . . . 116

9.1 Homomorphic Deconvolution. The first figure illustrates the signal to be analyzed. It simply results from the convolution of a Ricker wavelet, shown in black in the bottom figure, with a sequence of positive (+1) or negative (-1) reflectors. The cepstrum, obtained using the relations (7.2), is shown in the center in brown. By canceling out the part of this curve corresponding to the Green's function of the medium, it is possible to reconstruct the source through an inverse cepstral transform. The result of the deconvolution (red Ricker in the bottom figure) is superimposed on the original source (black Ricker). . . . . 130

10.1 Gain of the discrete filter (1,1,1,1,1)/5. Comparison between the gain given by expression (1.7), in red, and that of a perfect filter (black dashed lines) obtained by a 5-second duration rectangular pulse. (`ex_gain_transformeeZ.m`) . . . . . 134



---

10.2	Gains of finite difference operators. This example shows the gains (solid lines) of the two second derivative operators, $\{1; -2; 1\}$ and $\{-\frac{1}{12}; \frac{15}{12}; -\frac{28}{12}; \frac{15}{12}; -\frac{1}{12}\}$ . The yellow curve represents the gain of the ideal operator, $(2\pi u)^2$ , which is best approximated by the 5-term filter (red curve). Both filters are non-phase-shifting as they are centered on the origin. The calculations assume a unit sampling step, and you will notice that proper use of the filters is only possible if limited to low frequencies, approximately Nyquist/5 for the less efficient operator, which requires sampling the signals with a finer step than that dictated by the Shannon rule. . . . .	136
10.3	Gain and phase of the narrowband filter. The upper filter was constructed with $\varepsilon = 0.1$ , and the lower one with $\varepsilon = 0.01$ . (ex_bande_etroite.m) . . . . .	138
10.4	Applications of the narrowband filter. The initial signal (top left) consists of a pure frequency (0.20 Hz) and Gaussian white noise with unit variance. The amplitude spectrum (top right) clearly shows the spectral component and the noise level. The first filtering attempt (middle left) removes a significant portion of the noise (middle right). The filter used is the one at the top of Figure 10.3. The second attempt (bottom left) was performed with the filter closer to the ideal shown at the bottom of Figure 10.3. The spectral analysis of the filtered signal (bottom right) shows that the noise has indeed been further eliminated. However, the filtered signal (bottom left) shows a very disturbing boundary effect. The more or less significant nature of these boundary effects can be understood by examining the impulse response of the filters ( <i>cf</i> Figure 10.5). . . . .	140
10.5	Impulse responses of the narrowband filters shown in Figure 10.3. The filter constructed with $\varepsilon = 0.1$ has a much shorter impulse response (on the left) compared to the one (on the right) constructed with $\varepsilon = 0.01$ . These differences in duration explain the more or less significant boundary effects that occur when implementing recursive filters. . . . .	141
10.6	An unstable low-pass filter. This filter, designed to retain only the low frequencies of the initial signal (top left), is unstable and produces an unusable filtered signal (top right) showing exponential numerical divergence. Although the filter's gain (bottom left) is as expected, instability occurs because some poles are inside the unit circle (bottom right). . . . .	143

10.7 A stable low-pass filter. This filter is stable, as shown by the filtered signal (top right). It was obtained from the unstable filter in Figure 10.6 by making  $D(Z)$  a minimum-phase filter, which does not alter the gain (bottom left) but places all the poles outside the unit circle (bottom right) . . . . . 143

10.8 Examples of Butterworth low-pass filters (ex\_butter\_2\_6\_12.m) for orders of 2, 6, and 12. The higher the order, the closer the filter is to a window function. . . . . 145

10.9 The first four terms of the "bilinear approximation". The bilinear transformation, in the strict sense, has a validity range restricted to approximately Nyquist/4 (top left), as shown by the comparison with the line of slope 1 (black). The approximation using only the first two terms of the expansion (top right) has a broader validity range but is practically unusable due to spectral aliasing caused by the function not being bijective. The approximation with three terms (bottom left) is practically usable and has a validity range significantly larger than the classical approximation. *etc* . . . . . 147

10.10 Butterworth band-pass filters. The filter at the top is a first-order filter with a passband of 0.04-0.14 Hz, while the filter at the bottom is of the same passband but of fourth order. . . . . 149

10.11 Band-pass filtering. The initial signal (top left) contains spectral lines between 0.04 Hz and 0.14 Hz, which we will isolate using the filters shown in Figure 10.10. The filtering performed with the first-order filter is shown in the top right, and that with the fourth-order filter is shown in the bottom left. Both of these filtrations were performed in a forward-only manner and are therefore phase-shifting. A non-phase-shifting filtration using the fourth-order filter is shown in the bottom right. . . . . 150

10.12 Wiener Filtering . . . . . 153

11.1 Signal truncation limits the frequency resolution. The signal analysed in this example consists of two sinusoids (0.048 Hz and 0.058 Hz) sampled with  $\tau = 1;s$ . The frequency resolution, approximately equal to  $1/T$ , is sufficient to resolve the spectral lines at  $T = 512;s$  (top left),  $T = 256;s$  (top right), and  $T = 128;s$  (bottom left). However, a stronger truncation,  $T = 64;s$  (bottom right), no longer allows the resolution of the two spectral lines. Note that the amplitude of the peaks in these spectra decreases as they widen. Note also the increasing prominence of the secondary lobes associated with the main peaks as the duration of the signal analysed decreases. This phenomenon, known as *leakage*, can be reduced by using apodization windows. . . . . 164

---

11.2	Apodization windows. These three windows are the Dirichlet (top left), Bartlett (middle left) and Parzen (bottom left) windows. They are obtained by successive auto-convolutions of the function $\Pi(t)$ . As the number of auto-convolutions increases (from top to bottom), the window becomes smoother, and its Fourier transform (right) has attenuated secondary lobes and a wider central lobe, corresponding to a degradation in frequency resolution. From the central limit theorem illustrated in a previous chapter, you know that the limiting window obtained by this auto-convolution process is the Gaussian window, which is not very different from the Parzen window . . . . .	165
11.3	Apodization windows. These three windows are the Blackman (top left), Hamming (middle left) and Welch (bottom left) windows. Note that the Blackman window is very similar to the Parzen window. . . . .	166
11.4	Effects of the apodisation window. The analysed signal consists of two sinusoids with different amplitudes (1 and 0.05), sampled with $\tau = 1;s$ and $T = 512;s$ . The amplitude spectra obtained after apodising the signal with windows smoother than the Dirichlet window (top left) allow a better resolution of the low amplitude spectral line. . . . .	167
11.5	Effects of the presence of a trend in the analysed signal. When the signal contains a significant trend (middle left), its spectrum (middle right) is primarily representative of that of the trend alone (bottom left). Some details that are visible in the spectrum (top right) of the signal without the trend (top left) may then be obscured. . . . .	168
12.1	Illustration of the inadequacy of the Fourier transform for non-stationary signals: a signal consisting of two successive sinusoids (bottom left) has an amplitude spectrum (bottom right) very similar to that of a signal consisting of the superposition of the two sinusoids (top left), despite their different temporal structures (top right). . . .	178
12.2	Example of the calculation of a simple spectrogram using a Dirichlet window to extract segments of the signal. The amplitude of the spectrogram is plotted in the time-frequency plane. . . . .	179
12.3	Analysis function $g_T$ for three different frequencies. . . . .	180
12.4	Morlet wavelet for three different dilations $a$ . . . . .	182
12.5	The magnitude of the spectrogram (top) and the Morlet wavelet transform (middle) of a rectangular pulse. . . . .	183
12.6	Some functions of the Haar basis for $a = 1, 2, 4,$ and $8$ . . . . .	187

12.7	Example of filtering in the Haar basis. At the top, the desired signal (in blue) and its noisy version (in green). In the middle, the cumulative energy of the wavelet coefficients. At the bottom, the reconstructed signal retaining the 15% most energetic coefficients . . . . .	192
13.1	At the top is the 1 Hz sine wave, followed in descending order by the 10 Hz and 100 Hz sine waves. These three sinusoids are shown in black. The last one at the bottom, in red, is the sum of these sinusoids. . . . .	207
13.2	On the left is the rectangular matrix $\mathbf{X}$ , obtained using the <code>hankel.m</code> function in Matlab®. On the right is the product of $\mathbf{X}$ with its transpose, giving a square matrix. . . . .	208
13.3	On the left, the first 10 singular values of $\mathbf{X}$ . On the right, the logarithm (in dB) of the first 10 eigenvalues of $\mathbf{X}^t\mathbf{X}$ . . . . .	208
13.4	From top to bottom, the first 2 elementary signals appear to represent the 1 Hz sinusoid, followed by signals number 3 and 4, which correspond to the 10 Hz sinusoid, and finally, the last 2 signals can be attributed to the 100 Hz oscillation. . . . .	209
13.5	The red and blue curves for each pair in Figure (13.4) have been simply summed (black curves). They are compared to their respective original signals (red curves). . . . .	210
13.6	The top left shows the signal with the lowest period (1 Hz), the top right shows the singular values associated with the 10 Hz oscillation, the bottom left shows those associated with the 100 Hz oscillation, and finally the bottom right shows the singular values we have already presented for the total signal (Figure 13.3). . . . .	211
13.7	Mean sea level from 1993 up to the present day . . . . .	215
13.8	Superposées sur la courbe du niveau moyen des océans brute, les tendances extraites par SSA pour différentes valeurs de $L$ (courbes rouges). . . . .	216
13.9	Geodetic Reference for the Movement of the Pole. $m_1$ is the North-South distance from the geographic North Pole, and $m_2$ is the East-West distance, with reference to the Greenwich Meridian . . . . .	217
13.10	Temporal evolution of the components $(m_1, m_2)$ from 1846 to the present. . . . .	218
13.11	Trends of the Pair $(m_1, m_2)$ Extracted by SSA . . . . .	219
13.12	Chandler pseudo-cycles of the pair $(m_1, m_2)$ extracted by SSA. . . . .	219
13.13	Forced pseudo-cycles of the pair $(m_1, m_2)$ extracted by SSA . . . . .	220
13.14	Comparison between the raw component $m_1$ (red curve) and the sum of the extracted components (black curve). Below, the same comparison for $m_2$ . . . . .	220

---

13.15	At the top, in black, is a sine wave with an increasing phase over time, to which we have added noise. In the middle, in red, is the result of the SSA filtering. Below, superimposed, are the original signal (unnoised, in black) and the filtered signal (in red).	222
13.16	Eigenvalues in dB of the noisy signal from figure (13.15). On the top right is a zoom of these eigenvalues between ranks 1 and 50. . . . .	222
13.17	Hankel matrices: on the left, the signal before adding noise; on the right, the noisy signal. . . . .	223
14.1	Diagram of a classic subsurface geophysical problem: detection of a gravimetric anomaly due to the presence of a tunnel. . . . .	227
14.2	General structure of inverse problems . . . . .	229
14.3	General Structure of Inverse Problems . . . . .	231
16.1	<i>a priori</i> probabilities (dashed line, uniform distribution) and <i>a posteriori</i> probabilities (solid curve) of the tunnel depth when a single gravimetric measurement is taken directly above the tunnel. . . . .	269
16.2	<i>a priori</i> (dashed line) and <i>a posteriori</i> (solid line) probabilities of the tunnel depth when a single gravimetric measurement is taken directly above the tunnel. In this example, the <i>a priori</i> probability is not uniform, resulting in a significant change in the <i>a posteriori</i> probability (see figure 16.1). . . . .	270
16.3	<i>a priori</i> probabilities (dashed line, smooth curve) and <i>a posteriori</i> probabilities (solid line) from two gravimetric measurements, one directly above the tunnel and the other 10 metres above. . . . .	271
16.4	<i>a priori</i> probabilities (uniform distribution, dashed line) and <i>a posteriori</i> probabilities (solid curve) corresponding to the inverse problem solved with only the gravimetric data located 10 metres from the tunnel axis. It can be seen that the posterior probability is not well localised, indicating that the data do not effectively resolve the tunnel depth ( <i>cf</i> <code>ex_tunnel_02.m</code> ). . . . .	273
16.5	<i>a posteriori</i> probability (contour plots, bottom left) of the horizontal position and depth of the tunnel obtained from a single gravimetric measurement (taken directly above the tunnel, although this is not known). The marginal probability for the depth is significantly less localised than when the data were used to determine the depth alone (see figure 16.1), indicating that the addition of parameters in an inverse problem affects the determination of the <b>OTHER</b> parameters ( <i>cf</i> <code>ex_tunnel_03.m</code> ). . . . .	274

16.6 *a posteriori* probability (contour plots, bottom left) for the horizontal position and depth of the tunnel using two gravimetric measurements. The marginal probabilities indicate that the two parameters are fairly well resolved (*cf* `ex_tunnel_04.m`). . . . 276

17.1 *a posteriori* probability (contour plots at bottom left) for the horizontal position and depth of the tunnel when only a single gravimetric measurement is used. The marginal probabilities obtained by the Monte Carlo method are shown with dashed lines. Ten samples were taken for each value of  $x_t$  (top curve) or  $z_t$  (right curve). . . 282

17.2 Similar to Figure 17.1, but for 50 *Monte Carlo* samples. . . . . 282

17.3 Similar to Figure 17.1, but for two gravimetric measurements. . . . . 283

17.4 Similar to Figure 17.3, but for 50 *Monte Carlo* samples.. . . . . 283

17.5 Similar to Figure 17.3, but for 10 samples from a Sobol sequence. The more regular sampling reduces the integration error. . . . . 284

17.6 Similar to Figure 17.4, but for 50 samples from a Metropolis chain. Importance sampling helps to reduce the integration error. The dots represent the 2500 models from a Metropolis chain. Their distribution density follows the probability density. The small number of points compared to the 2500 models indicates that many models are duplicated many times, which means that the acceptance of new models was very rare. This is explained by the fact that each new model was randomly generated in the *a priori* space, independently of the previous model. . . . . 290

18.1 Path of the traveling salesman at the end of the *Metropolis* loop for  $T = 10^{-1}$ . The total distance is 309. . . . . 296

18.2 Path of the traveling salesman at the end of the *Metropolis* loop for  $T = 10^{-1.5}$ . The beginning of the cost reduction (path length) is noted. The total distance is 203. . . 296

18.3 Path of the traveling salesman at the end of the *Metropolis* loop for  $T = 10^{-2}$ . The cost decreases in steps due to the hierarchical geography of the cities. The total distance is 104. . . . . 297

18.4 Path of the traveling salesman at the end of the *Metropolis* loop for  $T = 10^{-3}$ . This is the end of the cooling phase, and the cost no longer decreases significantly. The total distance is 88. . . . . 297

19.1 Results of signal deconvolution using a second-order finite difference filter. . . . . 308

19.2 Singular vectors corresponding to the singular values shown at the bottom of Figure 19.1. . . . . 309

