Numéro National de Thèse : 2018LYSEN040

THÈSE de DOCTORAT DE L'UNIVERSITÉ DE LYON
opérée par
l'École Normale Supérieure de Lyon

École Doctorale N°52
École Doctorale de Physique et Astrophysique de Lyon (PHAST)

Spécialité de doctorat : Traitement du signal
Discipline : Physique

Soutenue publiquement le 26/09/2018, par :

Carlos Granero Belinchón

Multiscale Information Transfer in Turbulence

Transfert de l'information en turbulence : une vision multi-échelles

---

Devant le jury composé de :

Lesne, Annick	Directrice de recherche	CNRS, LPTMC	Rapporteure
Arneodo, Alain	Directeur de recherche	CNRS, LOMA	Rapporteur
Dubrulle, Bérengère	Directrice de recherche	CNRS, SPHYNX	Examinatrice
Tuckerman, Laurette	Directrice de recherche	CNRS, LPMMH	Présidente
Chevillard, Laurent	Chargé de recherche	CNRS, LPENSL	Examinateur
Garnier, Nicolas	Chargé de recherche	CNRS, LPENSL	Co-encadrant
Roux, Stéphane	Maître de conférences	ENS de Lyon, LPENSL	Directeur de thèse

Acknowledgements

First of all, I want to thank my PhD advisors Stèphane G. Roux and Nicolas B. Garnier for their inestimable help during the development of my thesis. They were always available to guide me with their great and innovative ideas, and to show me the research career. Without their help throughout these three years, this dissertation wouldn't have reached its final form.

I wish to thank, Alain Arneodo and Annick Lesne for being referees of my PhD dissertation, and Laurette Tuckerman, Bérengère Dubrulle and Laurent Chevillard for being members of the jury of my thesis.

Special thanks to Laurent Chevillard for helping me everytime I had questions on turbulence. His knowledge on the field enlightened some shadows on the theoretical aspects of turbulence, and his suggestions and remarks were very useful to understand them.

I want to thank Roberto Leonarduzzi for answering all my questions about multifractal analysis and for helping me to understand multifractality.

Finally, I wish to thank the two people who bring me in science research, when I was only an undergraduate student at the *Universidad Autónoma de Madrid*: Carlos Sánchez and José Ramón Ares. They, together with all the members of the *Materials of Interest in Renewable Energies* laboratory, show me what research was.

Abstract

This thesis aims at the statistical study of complex systems through the analysis of signals obtained from: experiments, synthetic generation and theoretical models. We have choosen turbulence as the paradigm of study because of its properties: non-linear dynamics, multi-scale behavior, energy cascade, intermittency ...

In order to perform a statistical characterization of a complex system, we study the distribution (probability density function), the correlations and dependences, and the Wiener causality relationships, of signals describing the system. Information theory appears as the ideal framework to perform this kind of analysis.

In this information theory framework, we developed two different, but related, methodologies to analyse the self-similarity properties of complex systems, and more precisely, of turbulence. The first methodology is based on the analysis of the increments of the studied process using Shannon entropy and Kullback-Leibler divergence. The second method, which allows to analyse non-stationnary processes, is based on the analysis of processes *via* Shannon entropy rate. We studied the relationship between these two methodologies which are able to characterize the distribution of information across scales and the deformation of the probability density function of the increments across scales.

Contents

Notation and Terminology **11**

Introduction **13**
Objectives 13
Framework 14
Contributions 15
Structure of this Manuscript 16

I Information Theory **19**

1 Theoretical aspects **21**
1.1 Stochastic and Dynamical processes 22
1.2 Multidimensional random variables: embedding 23
1.3 Static measures 23
1.3.1 Shannon Entropy 24
1.3.2 Kullback-Leibler Divergence 27
1.3.3 Mutual Information 28
1.4 Dynamic measures 28
1.4.1 Entropy Rate 28
1.4.2 Auto-Mutual Information 30
1.5 Information dynamics across scales 31
1.6 Causality measures 32
1.6.1 Delayed Mutual Information 33
1.6.2 Transfer Entropy 33
1.6.3 Directed Information 34
1.7 Specific distributions 35
1.7.1 Gaussian distribution 35
1.7.2 Other distributions 35

2 Non-parametric estimators **37**
2.1 Probability Density Function Estimates 37
2.1.1 Histogram algorithms 38
2.1.2 Kernel algorithms 39
2.1.3 Nearest Neighbors Search algorithms 40
2.1.4 Combinations of KDE and NNS 40
2.2 Kozachenko-Leonenko estimate for Shannon Entropy 41
2.3 Kraskov-Stögbauer-Grassberger estimate for Mutual Information 43
2.4 m-order entropy rate estimate 45
2.5 Dynamics at a given scale: Adapted Theiler prescription 48
2.6 Kraskov-Stögbauer-Grassberger estimate for Conditional Mutual Information 51

3 High order Statistics on non-Gaussian noises **53**
3.1 Data and procedure 54
3.1.1 Synthetic processes 54
3.1.2 Estimation Setup 56
3.2 Characterization of the estimates 56
3.2.1 Dependence on the data size N. 57
3.2.2 Dependence on the number of neighbors k. 58
3.2.3 Influence of the correlation strength 58
3.3 Dependence structure across scales 59
3.3.1 Dependence on the scale τ for fixed embedding dimensions. 59
3.3.2 Dependence on embedding dimensions for fixed τ. 62
3.4 Conclusion 63

4 Information theory of non-stationary processes **65**
4.1 Information Theory for Non-Stationary processes 66
4.1.1 General framework 66
4.1.2 T-length window framework 69
4.2 Data and procedure 71
4.3 Numerical analysis: T-length window framework 71
4.3.1 Characterization of the estimates 71
4.3.2 Dependence of $\bar{h}_T^{(m,\tau)}$ and $\bar{H}_T^{(m,\tau)}$ on T and the scale τ 72
4.3.3 Relation between the entropy rate of the motion and the entropy of its increments 74
4.4 Application: Discrimination between processes 76
4.5 Discussion and Conclusions 77

II Turbulence **79**

5 Fully developed turbulence **81**
5.1 Kolmogorov's theory 83
5.2 Multifractal approach of Kolmogorov's theory 87
5.2.1 Monofractal model 88
5.2.2 Multifractal models 88
5.2.3 Conclusion 90

6 An information theory viewpoint of K41 theory 91
6.1 Turbulent experimental systems and signals 91
6.2 Scaling of information in turbulence 92
  6.2.1 Grid turbulence 92
  6.2.2 Jet turbulence in Helium 96
6.3 Conditioned Entropy rate 97
6.4 Skewness of turbulent velocity increments 98
6.5 Conclusions 101

7 Rediscovering intermittency 103
7.1 Definitions 104
  7.1.1 KL divergence from Gaussianity 104
  7.1.2 Distance from Gaussianity across scales 105
  7.1.3 Methodology 105
7.2 Application to turbulent velocity signals 106
7.3 Modeling 108
  7.3.1 Synthetic processes 109
  7.3.2 Results 109
  7.3.3 Phenomenological model : the propagator formalism 112
7.4 Discussion and Conclusions 116

8 Wiener Causality across scales of turbulence 119
8.1 Definitions of the scale 119
  8.1.1 Increments 120
  8.1.2 Filtering 121
  8.1.3 Embedding 121
8.2 Results for different scale functions 122
  8.2.1 Increments 122
  8.2.2 Scales obtained with filtering 123
  8.2.3 Scales obtained by embedding 125
8.3 Results for different observables of turbulence 126
8.4 Conclusions and perspectives 126

III Biomedical Signal Processing 129

9 Information Theory to Probe Intrapartum Fetal Heart Rate Dynamics 131
9.1 Datasets: Intrapartum Fetal Heart Rate Times Series and Labor Stages . 133
9.2 Methods 134
  9.2.1 Features from Complexity Theory 134
  9.2.2 Connecting Complexity Theory and Information Theory 136
9.3 Results: Acidosis Detection Performance 137
  9.3.1 Comparison of Features' Performance, Using a Single Time Window, Just Before Delivery 137
  9.3.2 Effect of Window Size on Performance 140
  9.3.3 Effect of the Embedding Dimensions on the (Fetal Acidosis Detection) Performance of AMI 140

9.3.4 Dynamical Analysis 141
9.4 Discussion, Conclusions and Perspectives 144
  9.4.1 Acidosis Detection in the First Stage 145
  9.4.2 Acidosis Detection in Second Stage 145
  9.4.3 Probing the Dynamics 146

Discussion, Conclusions and Perspectives **149**

Notation and Terminology

General notation

$X^{(m,\lambda)}$ $\equiv$ process of dimension m and sampling distance λ (distance between consecutive samples).

$X_i^{(m,\lambda)}$ $\equiv$ random variable of $X^{(m,\lambda)}$ with index i, embedding dimension m and sampling distance λ.

$\mathbf{x}_i^{(m,\lambda)}$ $\equiv$ possible value of $X_i^{(m,\lambda)}$.

$\mathbf{x}^{(m,\lambda)}$ $\equiv$ possible value of $X^{(m,\lambda)}$.

$$\text{if } m = 1 : \begin{cases} X^{(m,\lambda)} \equiv X \\ X_i^{(m,\lambda)} \equiv X_i \\ \mathbf{x}_i^{(m,\lambda)} \equiv \mathbf{x}_i \\ \mathbf{x}^{(m,\lambda)} \equiv \mathbf{x} \end{cases}$$

The notation above is completely generic and can be applied to any kind of processes. Consequently, it will be used for general developments. Nevertheless, throughout this dissertation we mainly focus on temporal processes, and then we use the specific notation presented below when analysing them.

Time processes notation

$X^{(m,\tau)}$ $\equiv$ process of dimension m and sampling τ.

$X_t^{(m,\tau)}$ $\equiv$ random variable of $X^{(m,\tau)}$ at time t, embedding dimension m and sampling time τ.

$\mathbf{x}_t^{(m,\tau)}$ $\equiv$ possible value of $X_t^{(m,\tau)}$.

$\mathbf{x}^{(m,\tau)}$ $\equiv$ possible value of $X^{(m,\tau)}$.

$$\text{if } m = 1 : \begin{cases} X^{(m,\tau)} \equiv X \\ X_t^{(m,\tau)} \equiv X_t \\ \mathbf{x}_t^{(m,\tau)} \equiv \mathbf{x}_t \\ \mathbf{x}^{(m,\tau)} \equiv \mathbf{x} \end{cases}$$

Introduction

Signal processing is omnipresent in almost any domain of science. From physics to chemistry and throughout engineering, experimental measures and records have to be processed in order to be useful and informative. What do we mean by *signals*? and by *processing*? We define a *signal* as a function bringing information from the properties or from the behaviour of a given phenomenon. So, the temperature of a nuclear reactor, population of a city, the luminosity of a star, a voice record or a photograph, are signals. By *processing*, we consider the analysis, modification, recording, transport or synthesis of a signal [162]. This processing can have multiple goals, such as the extraction of information from the signal or the reduction of its size, and it can be as basic as the conversion of volts into Kelvin for thermocouple measures, but also as challenging as the analysis of image textures using multifractal analysis and wavelets decomposition.

The correct processing and analysis of the high-quality (not always as high) measures obtained from experimental set-ups, is one of the most important parts of experimental research.

Throughout this thesis, which combines *signal processing* and *physics*, we use signal processing to progress in the understanding of physical systems and, at the same time, we use a physical approach to develop signal processing methods.

Objectives

The objectives of this PhD can be summarized as developing methodologies for the characterization of complex systems.

The definition of **complex systems** still remains vague, and so its frontiers are imprecise [144, 98]. We understand a complex system as being *composed of several parts interacting together in a non-linear way.* Most of the time they are multi-scale, *i.e.* they present a continuum of coupled scales, in which the different scales exhibit different behaviours. This definition of complex systems includes a huge number of systems such as: financial markets [24, 26, 140], ecosystems [141], the brain [179], the Earth [54], colonies of ants [143], geographycal systems [2] ...

A prototype of complex system is **turbulence**, which is highly non-linear and multiscale [66].

Most people when hearing the word "turbulence" think of the violent mouvements shaking an aeroplane, and indeed, these motions are due to turbulence, more precisely to the turbulent behaviour of the air around the plane. When a fluid flow is turbulent, many eddies of different sizes appear. But turbulence is not only present in the air around planes, we can find turbulence in the flows of rivers, the smoke from cigars, the storm clouds and even in breakfast coffee (when we stir the coffee to cool it, we are modifying the state of the fluid "coffee", and more precisely we are inducing turbulence). A turbulent flow is characterized by chaotic trajectories of its fluid particles, and we say that a flow is turbulent when two very close particles of the fluid at a given time t will be arbitrarily far at another time $t + \Delta t$, for a small Δt. Understanding turbulence is very important for many different fields such as meteorology and climate sciences, aeronautic, naval or vehicle engineering, or, mixing reactions in chemistry. Because of the easy generalization to any complex system of methodologies capable of characterizing turbulence, and also because of the very interesting topic it represents, we use fully developed turbulence as our object of study almost all along the PhD.

In order to accomplish the description of a complex system we are interested in the study of the distribution, the statistical dependences and the (Wiener) causality relationships of experimental measures of different observables of the system. **Information theory** [176, 49] provides the perfect framework to perform this kind of analysis [163, 132].

Framework

Throughout this thesis, we propose to characterize **complex systems** and in particular **turbulence**, in terms of Shannon information [176]. We thus take a statistical viewpoint of complex systems. Experimental measures can be used to study the properties and behaviours of systems and phenomena, for example, the changes in temperature of a fluid during a chemical reaction. Once the experimental measures have been obtained, it is necessary to analyse them in order to grasp all the information they contain. We aim to characterize complex systems by analysing experimental measures with information theory tools.

Information theory characterizes a signal or process in terms of complexity, the amount of information needed to characterize the signal, and in terms of dependences, the amount of information shared by different parts of the signal. Furthermore, information theoretic measures are able to grasp non-linear effects, and high order dependences. This together with its applicability to stochastic as well as to deterministic processes [117, 118, 62, 128], and the consistency of its interpretation, makes information theory a very interesting approach in the analysis of complex systems [152, 209, 217] and more precisely, in the study of turbulence, where information theory has been very rarely used [29, 96, 33].

A very large number of systems in different domains such as physics, sociology, ecology *etc*, present **multi-scale behaviour** [17, 130, 19, 18, 152, 141]. Therefore, in these kind of systems the analysis at different scales (how the system behaviour evolves with the scale), and the analysis of possible interactions between scales (how the behaviour at a given scale can forecast the behaviour at another scale), become of utmost significance.

For this reason, during the PhD we developed an information theory approach to analyse the complexity and dependences of a process across scales. We also studied the possible (Wiener) causality relationships between different scales.

Contributions

Along the PhD we:

- Developed a multi-scale framework to analyse the distribution and dependences of a process and to characterize it in terms of complexity. We showed the ability of information theoretic measures to grasp high order statistics and the whole dependence structure of the studied processes.
    - To illustrate this ability we performed the discrimination between two processes with identical marginal distribution and correlation function but with differences at high order statistics.
    - We used this framework to analyse fully developed turbulence and recover Kolmogorov's 1941 theory [108, 109, 107]. We characterized fully developed turbulence in terms of complexity by studying experimental velocity signals and theoretical models, as well as synthetic processes modelling turbulence. Finally, we found the energy cascade of turbulence by performing a very precise analysis, also based on information theory, of the skewness of the velocity increment distributions.
    - We used this framework to analyse fetal heart rate signals during delivery in order to detect acidosis and characterize the dynamics of its development.
- Developed a methodology which allows to analyse the scale invariant properties, mono and multi fractality, of a process across scales. This methodology characterizes the deformation of the distribution of a process across scales. This methodology is applicable to any multi-scale system.
    - We applied this methodology to fully developed turbulence and found the multifractal nature of it. We compared the results for a set of experimental turbulent signals at different Reynolds and theoretical models, in order to corroborate the validity of the models.
- We presented a general framework to analyse non-stationary processes at a given time t using information theory. We also developed a new framework in which we compute information theory quantities on time windows of length T of the processes.
    - We applied the analysis of non-stationnary processes to study their self-similarity properties.

Structure of this thesis

This thesis is composed of three main parts:

Information theory:

- In chapter 1 we present the main information quantities used during the PhD, we also show their properties and interpretations. We classify them in three different groups: static measures, dynamic measures and causality measures.
- In chapter 2 we enumerate and characterize the main non-parametric algorithms used to compute information theory quantities. Then, we choose the most robust estimators in order to use them along the PhD.
- In chapter 3 we show the ability of information theory to measure dependences and to probe high order statistics of a process. We present some information theory tools as more general versions of correlation function and power spectrum.
- In chapter 4, we generalize information theory to the analysis of non-stationary processes. In order to do that, we develop two different frameworks to study non-stationary processes. The obtained analytical results are supported by numerical computations.

Turbulence:

- Chapter 5 is a summary of the state of the art in three dimensional fully developed turbulence. We present Kolmogorov theories and the multifractal theory of turbulence first introduced by Frisch and Parisi.
- In chapter 6, we study turbulence using information theory and develop a new framework capable to explain Kolmogorov 1941 theory. This chapter is based on the article "*Scaling of information in turbulence*" published in "*Europhysics letters*" [80].
- In chapter 7 we go one step further and develop an information theory framework capable of quantifying multifractality in turbulence across scales by measuring the deformation of the probability density functions with the scale. This framework can be generally applied to characterize the multifractal nature of any multi-scale system. This chapter is based on the article "*Kullback Leibler divergence measure of intermittency: Application to turbulence*" published in "*Physical Review E*" [81].
- In chapter 8 we try to measure flows of information between scales in turbulence, understood as Wiener causality relations between the different scales. To do that we analyse interactions between scales on several different systems. The interactions are quantified using transfer entropy and directed information.

Biomedical applications:

- In chapter 9 we apply information theory to analyse fetal heart rate dynamics during delivery. We perform this analysis in order to discriminate between healthy babies and acidotic ones. This chapter is based on the article "*Information Theory to Probe Intrapartum Fetal Heart Rate Dynamics*" published in the journal "*Entropy*" [77].

A **conclusion** chapter with the summary of contributions and perspectives closes the dissertation.

Part I

Information Theory

CHAPTER 1

Theoretical aspects

Information Theory is a theoretical viewpoint aiming at describing the mathematical laws governing the transmission, storage, processing, measurement and representation of information. Claude E. Shannon developped the proper basis of information theory in 1948, in his article “A Mathematical theory of Communication” [176] and expanded it later together with Weaver in a book [177]. As the title of the first article (of Shannon) points out, Information Theory appeared initially as a theory of communication. Nevertheless since the fifties, applications started to generalize to other fields such as dynamical systems [110, 111, 183, 113]. This leap from stochastic processes, the natural object of study of information theory, to dynamical processes suppose a very important advance in the development of information theory, allowing its spread to a very wide range of fields.

Nowadays Information theory is used in an extraordinary large array of different fields: thermodynamics [100, 101, 206, 156], quantum thermodynamics [93], neuroscience [209, 217, 221], etc. In addition, from the first article of Shannon until now, the number of different information theoretical objects increased significantly. Amongst these theoretical objects we can find: Entropy rate [177], Kullback Leibler divergence [121, 120], Transfer Entropy [172], Directed Information [115] ...

Recently, access to increasingly powerful computers together with the broader fields of application of Information Theory render it a very interesting research topic both in theoretical aspects and in applied ones.

Here, we present the Information Theory framework used during the PhD for the analysis of stationnary processes, with Shannon entropy as the elementary brick. In addition, as information theory has been proved to be able to describe both stochastic and dynamical systems [117, 118, 62, 128], throughout this manuscript information theory is applied over both kinds of processes.

1.1 Stochastic and Dynamical processes

Random variables

A measurable space $(\Omega, \mathcal{B})$ consists of a sample space Ω together with a σ-algebra $\mathcal{B}$. This σ-algebra $\mathcal{B}$ is a collection of subsets of Ω that includes the empty subset and is closed under any countable sequence of set theoretic operations such as union, intersection, complement ...

A random variable, f, defined on a measurable space $(\Omega, \mathcal{B})$ and taking values on another measurable space $(\mathbb{A}, \mathcal{B}_{\mathbb{A}})$ is a function $f : (\Omega, \mathcal{B}) \to (\mathbb{A}, \mathcal{B}_{\mathbb{A}})$ such that:

$$\text{if } F \in \mathcal{B}_{\mathbb{A}}, \text{ then } f^{-1}(F) = \{\omega : f(\omega) \in F\} \in \mathcal{B}$$

One then defines a probability space as $(\mathbb{A}, \mathcal{B}_{\mathbb{A}}, P_f)$, where the measurable space $(\mathbb{A}, \mathcal{B}_{\mathbb{A}})$ is called the alphabet of the random variable f, and P_f is the probability measure, which assigns a real number to every member F of $\mathcal{B}_{\mathbb{A}}$. The probability measure should satisfy the following conditions:

- Nonnegativity: $P_f(F) \geq 0 \;\; \forall \;\; F \in \mathcal{B}_{\mathbb{A}}$.
- Normalization: $P_f(\mathbb{A}) = 1$
- Countable additivity: If $F_i \in \mathcal{B}_{\mathbb{A}}$, $i = 1, 2, \cdots$ are disjoint, then $P_f(\cup_{i=1}^{\infty} F_i) = \sum_{i=1}^{\infty} P_f(F_i)$

The given definition of random variable ensures that the output of the random variable inherits its own probability measure, also called the probability density distribution of the random variable [85, 155]:

$$P_f(F) = P(f^{-1}(F))$$

Two random variables defined on a common probability space are considered to be the same if their distributions are the same [85, 155].

Random processes

A random process X is a family of random variables $\{X_i\}_{i \in \mathbb{I}}$ defined on a common probability space $(\Omega, \mathcal{B}, P)$, where $\mathbb{I}$ is a d-dimensional index set. In general $\mathbb{I}$ is only required to be closed under addition. It is assumed that all the random variables X_i share a common alphabet $\mathbb{A}$ [85, 155].

Dynamical processes

On the other hand, a dynamical system consists of a probability space $(\Omega, \mathcal{B}, P)$ together with a measurable transformation $T : \Omega \to \Omega$ [85, 193]. By measurable, one means that if $F \in \mathcal{B}$, then also $T^{-1}F \in \mathcal{B}$. $(\Omega, \mathcal{B}, P, T)$ is called a dynamical system in ergodic theory. Roughly speaking, in a dynamical system a function describes the time dependence of a point in a geometrical space.

This definition of a dynamical system leads to an alternative description of random processes. This description, based on dynamical systems, consists on considering a random variable together with a transformation defined on the underlying probability space [85]. With inverse reasoning, a chaotic (deterministic) dynamical system defined by a time function $f_t(x)$ can be seen as a random process where the input/state (x) of the dynamical system is a random variable.

The instability of chaotic systems with respect to initial conditions implies that time sequences generated by chaotic systems behave, from the point of view of information theory, as authentic stochastic processes [62, 128]. This equivalence between time sequences of dynamical systems and time sequences of stochastic processes is shown by Jewett-Krieger theorem [117, 118, 62], see section 1.4.1.

1.2 Multidimensional random variables: embedding

It is possible to show that collections of random variables, $X^{\mathbb{I}} = \{X_i; i \in \mathbb{I}\}$ such as $X^n = X_1, X_2, \cdots, X_n$, are also random variables [11, 85]. The measurable space $(\Omega, \mathcal{B})$ and the probability measure P of the new constructed random variables $X^{\mathbb{I}}$ are different from those of the separate single random variables X_i.

In order to properly characterize a stochastic process X taking into account both its static and dynamic properties, we need families or collections of random variables formed by sampling the process. With this purpose one uses a delay-embedding procedure [197] to construct a m dimensional stochastic process $X^{(m,\lambda)}$, sampled by λ, from the initial stochastic process X by defining m-dimensional random variables from the 1-dimensional ones:

$$X_i^{(m,\lambda)} = (X_i, X_{i-\lambda}, X_{i-2\lambda}, ..., X_{i-(m-1)\lambda}) \tag{1.1}$$

where m is the embedding dimension, λ is the sampling distance between random variables of X, and $X_i^{(m,\lambda)}$ is a random variable of $X^{(m,\lambda)}$.

This development of embedding is independent of the base dimension d of the initial random variables X_i. For a generic process X of dimension $d \neq 1$ the total dimension obtained for the embedded process $X^{(m,\lambda)}$, is $m \times d$. Throughout this dissertation, for simplicity and because of the perfect equivalence between dimensions d and embedding dimensions m, we fix $d = 1$ unless noted otherwise. Consequently the total dimension of the embedded process $X^{(m,\lambda)}$ is given by the embedding dimension m.

1.3 Static measures

In this section, we present several information theory objects capable of studying static properties of processes, *i.e.* properties that take into account neither interactions nor dependences between previous and next samples of the process. In the case of a one-dimensional process X, these objects depend on the marginal distribution of X. For n-dimensional processes, they depend on the n-dimensional joint distribution of the process.

The embedding procedure (section 1.2) allows the static measures to take also into account the dynamics of a process X: in the case of embedded processes $X^{(m,\tau)}$, static measures are functions of the m-dimensional joint PDF, and consequently, they take into account interactions between samples inside the m-dimensional embedded vector.

1.3.1 Shannon Entropy

Discrete framework

Information theory started with the definition of Shannon Entropy [176]. The Shannon entropy of a discrete process X is defined as the uncertainty that one has of the process X *i.e.* the amount of information that characterizes the process. It is a function of the probability density function (pdf) of the process:

$$H(X) = -\sum_{x \in \mathbb{A}} p_X(x) \log(p_X(x)) \tag{1.2}$$

where the sum is over all the possible values x of X and p_X is the probability distribution of X.

Information and thus Shannon entropy should satisfy some important properties [176, 128]:

- The entropy should be continuous in the probabilities: small changes in the PDF should imply small changes in the entropy.
- For a given number n of possible events defining the alphabet $\mathbb{A}$, the entropy is a maximum and equal to $\log(n)$ when all the $p_X(x)$ are equal ($p_X(x) = 1/n \ \ \forall x \in \mathbb{A}$). This is the most uncertain situation. With equally likely events there is more uncertainty when there are more possible events n.
- Additivity: The entropy of a system Y can be expressed in function of the entropies of its subsystems X if the relations between the system and the subsystems and between the subsystems themselves are known. We define a system $Y = \{Y_i\}$ as:

$$\begin{aligned} Y_1 &= \{X_1, \cdots, X_{k_1}\} \\ Y_2 &= \{X_{k_1+1}, \cdots, X_{k_1+k_2}\} \\ &\vdots \\ Y_m &= \{X_{n-k_m+1}, \cdots, X_n\} \end{aligned}$$

  with the probability of Y_i as:

$$p_{Y_i} = \sum_{l=k_1+\cdots+k_{(i-1)}}^{k_1+\cdots+k_i-1} p_{X_l}$$

  where p_{X_l} is the probability of X_l. We then define the entropy of X as:

$$H(p_{X_1}, \cdots, p_{X_n}) = H(p_{Y_1}, \cdots, p_{Y_m}) + \sum_{i=1}^{m} p_{Y_i} H\left(\frac{p_{X_{k_1+\cdots+k_{(i-1)}}}}{p_{Y_i}}, \cdots, \frac{p_{X_{k_1+\cdots+k_i-1}}}{p_{Y_i}}\right)$$

- Concavity: Shannon entropy is concave in the probability density function.

  $$H(ap_{X_1} + (1-a)p_{X_2}) \geq aH(p_{X_1}) + (1-a)H(p_{X_2})$$

  for any probability density functions and $0 \leq a \leq 1$.

- The entropy is positive and vanishes if and only if all the $p_X(x)$ but one are zero, thus when there is no uncertainty.

- Given two different processes X and Y, let $p_{X,Y}(x,y)$ be the probability of the joint occurrence of $x \in \mathbb{A}_X$ and $y \in \mathbb{A}_Y$. The joint entropy is the uncertainty of the joint occurrence:

  $$H(X,Y) = H(Y,X) = -\sum_{x\in\mathbb{A}_X, y\in\mathbb{A}_Y} p_{X,Y}(x,y)\log(p_{X,Y}(x,y)). \tag{1.3}$$

  It should respect:

  $$H(X,Y) \leq H(X) + H(Y) \tag{1.4}$$

  where the equality is obtained if and only if the two processes are independent.

- Given two processes X and Y, for any particular value x that X can take with probability $p_X(x) > 0$, we can define the conditional probability, $p_{Y|X}(y|x)$, that Y has the value y:

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

  We can define the conditional entropy of Y, $H(Y|X)$ as the average of the entropy of Y for each value of X weighted according to the probability of getting that particular value of X:

  $$H(Y|X) = -\sum_{x\in\mathbb{A}_X, y\in\mathbb{A}_Y} p_{X,Y}(x,y)\log(p_{Y|X}(y|x)). \tag{1.5}$$

  The conditional entropy measures how uncertain we are of Y when X is known. It satisfies:

  $$H(Y|X) = H(X,Y) - H(X). \tag{1.6}$$

  From eq.(1.4) and eq.(1.6) we see that:

  $$H(Y) \geq H(Y|X)$$

  which can be intuitively inferred by noting that knowing X can only help in the characterization of Y. In the worst-case scenario, if X and Y are independent, knowing X will not alter how uncertain we are of Y, and then $H(Y|X) = H(Y)$.

Continuum framework

Generalizing the information measure to continuous processes is not straightforward. Differential entropy defined by Shannon [177] as:

$$H(X) = -\int_{\mathbb{A}} p_X(x)\log(p_X(x))dx \tag{1.7}$$

is not invariant under change of variables and can be negative. In addition, differential entropy is maximal for the multivariate Gaussian distribution instead of the uniform distribution. Thus differential entropy is not a good measure of information [102]. Instead of the straightforward definition in eq.(1.7) intuited by Shannon, the correct measure of information in the continuous case is [102]:

$$H(X) = -\int_{\mathbb{A}} p_X(x) \log\left(\frac{p_X(x)}{m(x)}\right) dx \tag{1.8}$$

where $m(x)$ is the measure function defined by:

$$\int_a^b m(x)dx = \lim_{N_e\to\infty} \frac{(\#\text{ of events in } a < x < b)}{N_e} \tag{1.9}$$

In eq.(1.9) N_e is the number of discrete events in the whole domain of $\mathbb{A}$, and a and b define an infinitely small segment of $\mathbb{A}$.

In the following we use the discrete formalism unless said otherwise.

Interpretation of Shannon entropy and relation with thermodynamical entropy

- **Shannon entropy**, as presented above, is a measure of the total uncertainty that one has of a process. Shannon entropy measures the amount of information that completely characterizes a processs . The more complex the process the larger the Shannon entropy.

- **Thermodynamical entropy**, was defined by Clausius [31, 43] as:

  $$\Delta S_C(1 \to 2) = \int_1^2 \frac{dQ}{T} \tag{1.10}$$

  where T is the temperature of the system and dQ is the heat.

  In his article, "*Gibbs vs Boltzmann entropies*" [103], Jaynes showed that Gibbs entropy defined as:

  $$H_G = k_B \int_{\Omega} W_N \log(W_N) d\Omega, \tag{1.11}$$

  where k_B is the Boltzmann constant and W_N is the N-particle distribution function, which gives the probability density in the full phase space Ω of the system, verifies in the case of thermal equilibrium (canonical distribution) and reversible path 1 → 2:

  $$(H_G)_2 - (H_G)_1 = \Delta H_G(1 \to 2) = \int_1^2 \frac{dQ}{T}. \tag{1.12}$$

The fact that Shannon entropy and Gibbs entropy share the same mathematical expression, does not in itself establish any connection between this two entropies.

- **Maximum-Entropy Principle**

  In order to properly relate Shannon entropy to thermodynamic entropy, Jaynes formulated the maximum-entropy principle [100, 101]. Jaynes built maximum-entropy principle in the framework of *subjectivist* or Bayesian statistics, which understands probabilities as expressions of human ignorance. Thus, the probability of an event is a formal expression of our expectation that the event occur, based on the available information. The problem that Jaynes formulated was to find "... *a probability assignment which avoids bias, while agreeing with whatever information is given*" [100].

  Information theory provides a measure of the amount of uncertainty characterizing a discrete probability distribution. The only unbiased statement one can conclude with a given amount of information, is that one must use the probability distribution which maximizes the entropy subject to whatever is known. This is known as the maximum-entropy principle.

  Then Jaynes interpreted the probability density in the full phase space of the system at time zero, $W_N(t=0)$, and consequently its Gibbs entropy as measuring "*our degree of ignorance as to the true unknown microstate, when the only information we have consists of the macroscopic thermodynamic parameters*" [103].

1.3.2 Kullback-Leibler Divergence

Kullback-Leibler divergence, also called Relative entropy or Kullback-Leibler distance, is defined as the measure of the distinguishability between two probability density functions, p_X and q_X, of the same process X. Kullback-Leibler divergence measures the inefficiency of assuming probability density function q_X when the true one is p_X [120, 49]:

$$\mathcal{K}_{p_X||q_X} = \sum_{x\in\mathbb{A}} p_X(x)\log\left(\frac{p_X(x)}{q_X(x)}\right) \tag{1.13}$$

Kullback-Leibler divergence measures how "similar" the distribution q_X is to the distribution p_X. Some Kullback-Leibler divergence properties are:

- It's always non-negative and only zero for $p_X = q_X$.
- It's not a distance, since it is not symmetric:

  $$\mathcal{K}_{p_X||q_X} \neq \mathcal{K}_{q_X||p_X} \tag{1.14}$$

- It remains well-defined for continuous distributions:

  $$\mathcal{K}_{p_X||q_X} = \int_{-\infty}^{+\infty} p_X(x)\log\left(\frac{p_X(x)}{q_X(x)}\right)dx \tag{1.15}$$

Chernoff-Stein's lemma gives an easy interpretation of Kullback-Leibler divergence in terms of "surprise" [41, 49]. After n realizations of X, with X distributed with $p_X(x)$, the probability to mistake $\hat{p}_{X,n}(x)$ (observed distribution after n realizations) for $q_X(x)$ decays exponentially with $\mathcal{K}_{p_X||q_X}$:

$$P(q_X(x)|\hat{p}_{X,n}(x)) \approx e^{-n\mathcal{K}_{p_X||q_X}} \tag{1.16}$$

1.3.3 Mutual Information

A specific but famous case of Kullback-Leibler divergence is mutual information. Given two different processes X and Y, mutual information measures the shared (common) information between both processes, *i.e.* the "distance" between independence and dependence of the processes. It is defined as the Kullback-Leibler divergence between the joint probability of both processes $p_{X,Y}(x,y)$ (general case) and the product of the marginals $p_X(x)p_Y(y)$ (thus assuming the independence of X and Y) [177]:

$$I(X,Y) = \sum_{x\in\mathbb{A}_X, y\in\mathbb{A}_Y} p_{X,Y}(x,y)\log\left(\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\right) \tag{1.17}$$

Mutual Information can be expressed in function of Shannon entropies:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{1.18}$$

From eq.(1.18) it is easy to verify the symmetry of mutual information, $I(X,Y) = I(Y,X)$.

As for Shannon entropy it is possible to define a conditional version of Mutual information [222]:

$$I(X,Y|Z) = \sum_{z\in\mathbb{A}_Z} p_Z(z) \sum_{x\in\mathbb{A}_X, y\in\mathbb{A}_Y} p_{X,Y|Z}(x,y|z)\log\left(\frac{p_{X,Y|Z}(x,y|z)}{p_{X|Z}(x|z)p_{Y|Z}(y|z)}\right) \tag{1.19}$$

It can also be expressed in function of conditional entropies:

$$I(X,Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z) \tag{1.20}$$

Conditional Mutual information in eq.(1.19) measures the shared information between X and Y that is not contained in Z.

1.4 Dynamic measures

Dynamic measures are able to characterize interactions and dependences between previous and next samples of a process X, and then, connections between the different samples of the process.

1.4.1 Entropy Rate

In order to measure how the entropy of a sequence of m random variables grows with m, Shannon presented the entropy rate [177]. Entropy rate of a process $X = \{X_i\}_{i\in\mathbb{I}}$ is defined as [49]:

$$h(X) = \lim_{m\to\infty} \frac{1}{m} H(X_1, X_2, X_3, \cdots, X_m) \tag{1.21}$$

Entropy rate defined in eq.(1.21) is the entropy per symbol of the m-dimensional random variable characterizing the process. Another possible, definition of entropy rate is [49]:

$$h'(X) = \lim_{m\to\infty} H(X_{m+1}|X_m, X_{m-1}, X_{m-2}, \cdots, X_1) \tag{1.22}$$

In this case (eq.(1.22)) entropy rate is defined as the entropy of the $m+1$th random variable, knowing the previous ones, $X_i \; \forall\, i \in [1:m]$. For a stationary process, these two definitions of entropy rate (eq.(1.21) and eq.(1.22)) are equivalent, both limits exist and are equal [49].

Entropy rate probes the dynamical evolution of the process. It measures the effect of the previous samples of the process on the next one (eq.(1.22)), or the contribution of each one-dimensional random variable X_i to the total entropy of the process characterized by a m-dimensional random variable, eq.(1.21).

m-order entropy rate

m-order entropy rate is an approximation of entropy rate from eq.(1.22), where m doesn't go to infinity but takes a finite value m, indicating the dimension of the analysed random variables (section 1.2) and hence the order of the entropy rate [128]:

$$\begin{aligned} h^{(m)}(X) &= H(X_{m+1}|X_m, X_{m-1}, X_{m-2}, \cdots, X_1) \qquad (1.23) \\ &= H(X_{m+1}, X_m, \cdots, X_1) - H(X_m, X_{m-1}, \cdots, X_1) \end{aligned}$$

Due to the impracticability of computing entropy rate as defined in eq.(1.21) or eq.(1.22), the approximation of entropy rate to its mth order (eq.(1.23)) has been extensively used [58, 92, 23, 129]. In definitions (1.21),(1.22) and (1.23) of entropy rate the distance between consecutive samples of process X is fixed to $\lambda = 1$. In order to generalize the definition of m-order entropy rate the use of embedding notation presented in section 1.2 can be useful:

$$h^{(m,\lambda)}(X) = H(X_{i+\lambda}|X_i^{(m,\lambda)}) = H(X^{(m+1,\lambda)}) - H(X^{(m,\lambda)}) \tag{1.24}$$

This abreviated notation allows to easily indicate variations on the index i and the sampling distance λ.

m-order entropy rate can be understood as a kind of derivative of the entropy, measuring how much the entropy varies between two consecutive dimensions m and $m+1$. It can be seen also as a measure of the information of $X_{i+\lambda}$ not contained in the m previous random variables $X_i^{(m,\lambda)}$, or, the new information about the process X brought by $X_{i+\lambda}$ if $X_i^{(m,\lambda)}$ is known.

A very important characteristic of entropy rate is that it can be applied in the analysis of data from both deterministic and stochastic sources [62, 128].

Kolmogorov-Sinai Entropy and relation with entropy rate

In the 1960s, Kolmogorov and Sinai adapted Shannon's information theory to the study of dynamical systems. The divergence of trajectories starting from different, but undistinguishable initial conditions can be pictured as creating uncertainty, so creating information. Metric entropy, also known as the Kolmogorov–Sinai entropy $h_{\mathrm{KS}}(\rho)$, measures the mean rate of creation of information by a dynamical system with ergodic probability measure ρ. Metric entropy is constructed exactly as the Shannon entropy rate from information theory, but using the density ρ of trajectories in phase space instead of the probability density p. Metric entropy is related to Shannon entropy rate by $h_{\mathrm{KS}} = h$ if Shannon entropy rate is defined using natural logarithm [62, 128]. A very important result relating dynamical system and stochastic processes is Jewett-Krieger theorem [117, 118]: *A continuous valued deterministic system in discrete time and with finite entropy is equivalent to a stochastic process with a finite number of states. The minimum number of states, m_s, required for the theorem to be valid, is given by $e^{h_{\mathrm{KS}}} < m_s < 1 + e^{h_{\mathrm{KS}}}$.*

The exact analogy between the Shannon entropy rate and the metric entropy supports the validation and meaning of Information Theory analysis over dynamical systems [62, 128].

1.4.2 Auto-Mutual Information

Another way to express Shannon m-order entropy rate is using mutual information between two consecutive chunks of a process $X_{i+\lambda}$ and $X_i^{(m,\lambda)}$:

$$h^{(m,\lambda)}(X) = H(X) - I(X_{i+\lambda}, X_i^{(m,\lambda)}) \tag{1.25}$$

where $H(X)$ is the total Shannon entropy of the process X.

From this expression of Shannon entropy rate, we can separate two different contributions: $H(X)$, which only depends on the one-point distribution of X and is hence a static property; Conversely, $I^{(m,1,\lambda)}(X)$ gathers all information conveyed by linear and non linear dynamics, irrespective of the variance of X, but dependent on the shape of its distribution [80].

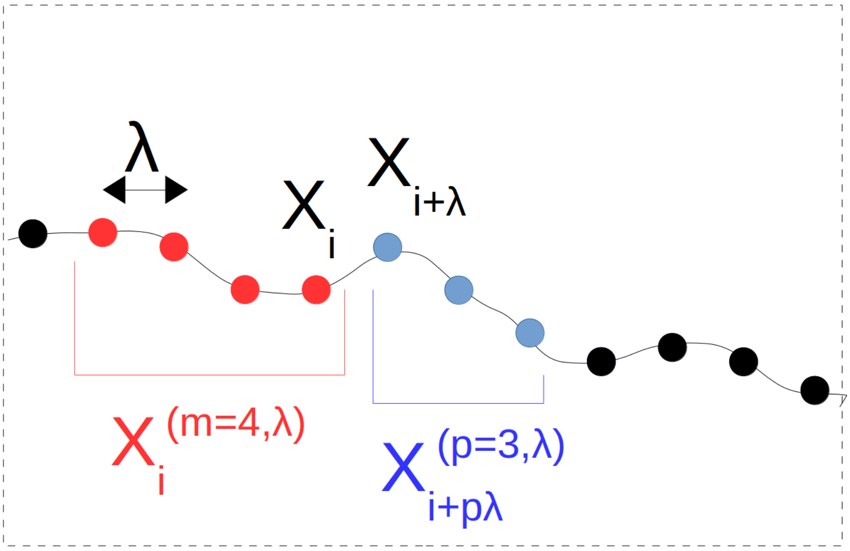

Figure 1.1: Embedded vectors $X_{i+p\lambda}^{(p,\lambda)}$ and $X_i^{(m,\lambda)}$ of process X.

We define auto-mutual information [49, 73, 61, 223], also called information storage, as the information shared between $X_{i+\lambda}$ and $X_i^{(m,\lambda)}$, *i.e.* the mutual information between them, $I(X_{i+\lambda}, X_i^{(m,\lambda)})$. Information storage can be understood as the part of information contained in sample $i+\lambda$ of process X ($X_{i+\lambda}$) that can be predicted by the knowledge of previous samples ($X_i^{(m,\lambda)}$).

Information storage definition can be generalized to "the mutual information between $X_{i+p\lambda}^{(p,\lambda)}$ and $X_i^{(m,\lambda)}$" (figure 1.1):

$$I(X_{i+p\lambda}^{(p,\lambda)}, X_i^{(m,\lambda)}) = H(X_{i+p\lambda}^{(p,\lambda)}) + H(X_i^{(m,\lambda)}) - H(X_{i+p\lambda}^{(m+p,\lambda)}) \tag{1.26}$$

In order to simplify the expression of auto mutual information we denote $I(X_{i+p\lambda}^{(p,\lambda)}, X_i^{(m,\lambda)}) = I^{(m,p,\lambda)}(X)$

Auto-mutual information Symmetry

Auto-mutual information of stationnary processes is symmetrical with respect to the embedding dimension:

$$I^{(m,p,\lambda)}(X) = I^{(p,m,\lambda)}(X) \tag{1.27}$$

Proof. If we develop the MI expression in function of the entropies, for both cases, left and right sides of the above equation:

$$I(X_{i+p\lambda}^{(p,\lambda)}, X_i^{(m,\lambda)}) = H(X_{i+p\lambda}^{(p,\lambda)}) + H(_i^{(m,\lambda)}) - H(X_{i+p\lambda}^{(p,\lambda)}, X_i^{(m,\lambda)})$$

As we work with the same process X delayed:

$$H(X_{i+p\lambda}^{(p,\lambda)}, X_i^{(m,\lambda)}) = H(X_{i+p\lambda}^{(p+m,\lambda)}) \tag{1.28}$$

and invoking stationarity:

$$\begin{aligned} H(X_{i+p\lambda}^{(p,\lambda)}) &= H(X^{(p,\lambda)}) \\ H(X_i^{(m,\lambda)}) &= H(X^{(m,\lambda)}) \\ H(X_{i+p\lambda}^{(p+m,\lambda)}) &= H(X^{(p+m,\lambda)}) \end{aligned}$$

□

1.5 Information dynamics across scales

In order to analyse the dynamics across scales of a 1-dimensional process X, we build m-embedded processes $X^{(m,\lambda)}$, where λ indicates the scale at which the process is studied (section 1.2). Then, we estimate the entropy rate (section 1.4.1) and the auto mutual information (section 1.4.2) for different λ values.

The scale λ of the statistics:

At the scale λ, consecutive elements of the m-dimensional vectors are samples of the initial process X with a distance λ between them, eq.(1.1). Consequently we focus on the dynamics of the process at this scale λ. When $\lambda \to 0$, so at the small scales, the dependences between two consecutive coordinates of the embedded vector increase. Conversely, when $\lambda \to +\infty$, then at the large scales of the process, all dependences disappear.

Of course the entropy of $X^{(m,\lambda)}$ ($H(X^{(m,\lambda)})$) and then, the entropy rate ($h^{(m,\lambda)}(X)$) and auto mutual information ($I^{(m,1,\lambda)}(X)$) probe the dynamics of X at this scale and behave accordingly to the dependences. When the dependences are null $H(X^{(m,\lambda)}) = mH(X)$, $I^{(m,1,\lambda)}(X) = 0$ and $h^{(m,\lambda)}(X) = H(X)$. However when the dependences increase, $H(X^{(m,\lambda)}) < mH(X)$, $I^{(m,1,\lambda)}(X)$ increases and $h^{(m,\lambda)}(X)$ decreases.

$I^{(m,1,\lambda)}(X)$ measures the shared information between current m-points dynamics and future 1-point dynamics at the given scale λ.

The order m of the statistics:

The Shannon entropy of a 1-dimensional stationnary process X involves arbitrarily high order moments of the PDF, and therefore high order statistics, although it importantly does not depend on the first moment of the PDF and one can work with centered processes. In the case of no embedding, there is no dependence of the entropy on λ, so we need at least two samples of the 1-dimensional process to define a scale λ.

Theoretically, the embedding dimension m indicates the number of points directly involved in the statistics, so, for embedding dimension equals m, m-points statistics will be directly caught by IT tools, even if all IT tools work over the complete PDF, and then indirectly they work always on all order statistics. In this way, for the m-embedded version of the stationnary process $X^{(m,\lambda)}$, the Shannon entropy focuses on the m-order statistics, and depends on the scale parameter λ.

The entropy rate of order m and the auto mutual information of order m, as the entropy, probe statistics of any order, but eq.(1.24) and eq.(1.25) suggest that for a given m, they focus on the extra information contained in order $m+1$ statistics. Therefore, when the embedding dimension m increases, higher and higher order statistics are taken into account. As the Shannon entropy of the m-embedded process $X^{(m,\lambda)}$, both Shannon entropy rate and auto mutual information depend on the scale λ of the embedding procedure.

The Gaussian and log-normal processes presented in section 1.7 will be used as illustration.

1.6 Causality measures

In this section we present three measures capable of quantifying causal relations between processes. The second and the third measures are based on Wiener causality [218, 82, 72, 170, 28, 7, 132], *i.e.* a process X causes another process Y if the prediction of Y_i using the previous samples of both processes X and Y is more succesful than the prediction of Y_i using only previous samples of Y. The first measure that we present is more basic and causality is here understood in a simpler way: a process X causes another process Y if Y_i and previous samples of X share some information.

1.6.1 Delayed Mutual Information

As we show in section 1.3.3, Mutual information measures the shared information between two processes X and Y. By adding a lag δ, MI can measure the directional exchanged information from X to Y and vice versa. Delayed mutual information from X to Y is defined as:

$$I(X \to Y)_\delta = I(X_i, Y_{i+\lambda}) \tag{1.29}$$

We can define equivalently, the delayed mutual information from Y to X:

$$I(Y \to X)_\delta = I(X_{i+\lambda}, Y_i) \tag{1.30}$$

Unfortunately delayed mutual information presents at least three main flaws [172, 153]:

1 Common history of both processes can influence the measure.

2 Common input for both processes can influence the measure. Even if neither X causes Y nor Y causes X some causality relationships can be obtained between X and Y, if a third process Z causes both of them.

3 Exchanged information measured in one direction can be a reminiscent residue of exchanged information in the other direction. A positive measure of $I(X \to Y)_\delta$ can be due to a residue of the Wiener causality from Y to X.

1.6.2 Transfer Entropy

In order to fix flaws 1 and 3, Schreiber developed Transfer Entropy (TE) [172, 153, 209]:

$$\begin{aligned} TE(X^{(p,\lambda)} \to Y^{(m,\lambda)})_\delta &= H(Y_i|Y^{(m,\lambda)}_{i-\delta}) - H(Y_i|Y^{(m,\lambda)}_{i-\delta}, X^{(p,\lambda)}_{i-\delta}) && (1.31)\\ &= MI(Y_i, X^{(p,\lambda)}_{i-\delta}|Y^{(m,\lambda)}_{i-\delta}) && (1.32)\end{aligned}$$

Transfer Entropy measures the causal relations, in Wiener-Granger terms, that X has over Y at a distance equals δ. It can be seen as the information contained in Y_i which is not contained in $Y^{(m,\lambda)}_{i-\delta}$ but in $X^{(p,\lambda)}_{i-\delta}$. Transfer entropy can be expressed as a conditional mutual information, see eq.(1.32). In order to give a dynamical meaning to CMI, the conditioning is done on the past of the target process, e.g. in eq.(1.32) on the past of Y.

As we can see in eq.(1.31), conditioning only by previous samples of Y in the first conditional entropy and previous samples of X and Y in the second conditional entropy makes Transfer Entropy not symmetrical, $TE(X^{(p,\lambda)} \to Y^{(m,\lambda)})_\delta \neq TE(Y^{(m,\lambda)} \to X^{(p,\lambda)})_\delta$. This non-symmetry allows to measure directionality and dynamics. In order for Transfer Entropy to be causal, δ must be larger than zero, in this way it does not take into account instantaneous interactions.

Later Wibral et al. state that, in order to be Granger-Wiener causal, another definition of TE is needed [217, 132]:

$$\begin{aligned} TE(X^{(p,\lambda)} \rightarrow Y^{(m,\lambda)})_{\delta} &= H(Y_i|Y_{i-1}^{(m,\lambda)}) - H(Y_i|Y_{i-1}^{(m,\lambda)}, X_{i-\delta}^{(p,\lambda)}) \\ &= MI(Y_i, X_{i-\delta}^{(p,\lambda)}|Y_{i-1}^{(m,\lambda)}) \end{aligned} \tag{1.33}$$

In this definition the knowledge of previous samples of Y come until the ith sample, and the distance δ only remains in the causation process X.

1.6.3 Directed Information

Another measure of influence of a system X to another one Y is Directed Information [115] (DI):

$$DI(X \rightarrow Y)_M = \sum_{m=1}^{M} \Big(H(Y_m|Y_{m-\lambda}^{(m-1,\lambda)}) - H(Y_m|Y_{m-\lambda}^{(m-1,\lambda)}, X^{(m,\lambda)}) \Big) \tag{1.34}$$

$$= \sum_{m=1}^{M} MI(X^{(m,\lambda)}, Y_m|Y_{m-\lambda}^{(m-1,\lambda)}) \tag{1.35}$$

where M is called the order and indicates the maximum embedding dimension of the measure.

We can compare Transfer Entropy to Directed Information, finding that the structure of the measures is very similar (compare eq.(1.31) to eq.(1.34)), both of them can be expressed as a conditional mutual information (see eq.(1.32) and eq.(1.35)). But contrary to Transfer entropy, Directed Information takes into account instantaneous interactions. These instantaneous interactions measured by DI are symmetric under directionality, *i.e.* are identical in both directions.

We present transfer entropy and directed information as measures of Wiener causality. Initially, both transfer entropy and directed information were presented as measures of directional information flows between two systems. Since their definitions they have been often used as measures of exchanged information, or measures of sent/received information. Some recent works have questioned this interpretation, pointing out some possible misconceptions and some possible problems in TE and DI [219, 220, 99, 40]. The most interesting example is the possibility of having a third system Z influencing both X and Y. A conditioning, not only on the past of X and the past of Y, but also on the rest of systems interacting, would solve, or at least reduce (It's imposible to take into account all the possible sources of influence), the problem. Some other objections, such as the possibility of adding synergetic information with the conditioning procedure (the conditioning on two, or more variables may provide information that is not available from either alone), or the possibility of being measuring influences not related to information exchanges, have to be studied. The interpretation of TE and DI as measures of influence and causality seems well accepted, and we adopt this point of view. Nevertheless we let the door open to other possible quantitities to measure information flows.

1.7 Specific distributions

In this section we express the Shannon entropy, the entropy rate and the auto mutual information for some specific distributions.

1.7.1 Gaussian distribution

If X is a stationary Gaussian process, hence fully defined by its variance σ^2 and normalized correlation function $c(\lambda)$, its entropy is [226]:

$$H(X^{(p,\lambda)}) = \frac{p}{2}\log\left(2\pi e\sigma^2\right) + \frac{1}{2}\log\left(|\Sigma^{(p)}|\right) \tag{1.36}$$

where $\Sigma^{(m)}$ is the $m \times m$ correlation matrix of the process X; $\Sigma_{i,j} = c(|i-j|\lambda)$ and $|\Sigma^{(1)}| = 1$.

Using eq.(1.36) we can obtain the information storage of X:

$$I(X_{i+p\lambda}^{(p,\lambda)}, X_i^{(m,\lambda)}) = \frac{1}{2}\log\left(\frac{|\Sigma^{(m)}||\Sigma^{(p)}|}{|\Sigma^{(m+p)}|}\right) \tag{1.37}$$

For the particular case $p = 1$ we have:

$$\begin{aligned} H(X) &= \frac{1}{2}\log\left(2\pi e\sigma^2\right) & (1.38)\\ I(X_{i+\lambda}, X_i^{(m,\lambda)}) &= \frac{1}{2}\log\left(\frac{|\Sigma^{(m)}|}{|\Sigma^{(m+1)}|}\right) & (1.39)\end{aligned}$$

so we directly obtain the m-order entropy rate of a Gaussian process:

$$h^{(m,\lambda)}(X) = \frac{1}{2}\log\left(2\pi e\sigma^2\right) - \frac{1}{2}\log\left(\frac{|\Sigma^{(m)}|}{|\Sigma^{(m+1)}|}\right) \tag{1.40}$$

which clearly illustrates the decomposition of the entropy rate according to eq.(1.25): the first term $H(X)$ depends only on the static (one-point) statistics (via σ^2), and the second term $I(X_{i+\lambda}, X_i^{(m,\lambda)})$ depends on the temporal dynamics (and in this simple case, only on the dynamics, via the auto-correlation function $c(\lambda)$).

1.7.2 Other distributions

Given a bijective transformation F that maps the Gaussian signal X into a signal $Y = F(X)$, formulae (1.36,1.37) can be rewritten for the signal Y:

$$\begin{aligned} H(Y) &= H(X) + \langle \ln F\rangle_X \\ &= \frac{1}{2}\ln(2\pi e\sigma_x^2) + \langle \ln F\rangle_X & (1.41)\\ I^{(m,p,\tau)}(Y) &= I^{(m,p,\tau)}(X) \\ &= \frac{1}{2}\ln\left(\frac{|\Sigma^{(m)}||\Sigma^{(p)}|}{|\Sigma^{(m+p)}|}\right) & (1.42)\end{aligned}$$

where $\langle \ln F \rangle_X$ is the average of $\ln F(x)$ using the probability density of X, and hence is a constant, σ_x^2 is the variance of the Gaussian variable X, and $\Sigma^{(m)}$ is the m-dimensional covariance square matrix of X. As previously observed in [116], the MI is invariant under re-parametrization of its arguments. So, knowing the underlying Gaussian process X allows to estimate the theoretical entropy, auto-mutual information and then entropy rate of process Y.

Example: Log-normal distribution

If we consider as a particular case a stationary log-normal process Y, defined as $Y = F(X) = e^X$ with X Gaussian of mean μ and standard deviation σ, its entropy is [226]:

$$H(Y^{(p,\lambda)}) = \frac{p}{2}\log\left(2\pi e\sigma^2\right) + \frac{1}{2}\log\left(|\Sigma^{(p)}|\right) + p\mu \tag{1.43}$$

where $\Sigma^{(m)}$ is the $m \times m$ correlation matrix of the process X; $\Sigma_{i,j} = c(|i-j|\lambda)$ and $|\Sigma^{(1)}| = 1$, and μ is the mean of the process X.

Using eq.(1.43) we can obtain the information storage of Y:

$$I(Y_{i+p\lambda}^{(p,\lambda)}, Y_i^{(m,\lambda)}) = \frac{1}{2}\log\left(\frac{|\Sigma^{(m)}||\Sigma^{(p)}|}{|\Sigma^{(m+p)}|}\right) \tag{1.44}$$

For the particular case $p = 1$ of eq.(1.43) and eq.(1.44), we have:

$$\begin{aligned} H(Y) &= \frac{1}{2}\log\left(2\pi e\sigma^2\right) + \mu && (1.45)\\ I(Y_{i+\lambda}, Y_i^{(m,\lambda)}) &= \frac{1}{2}\log\left(\frac{|\Sigma^{(m)}|}{|\Sigma^{(m+1)}|}\right) && (1.46)\end{aligned}$$

The m-order entropy rate of a log-normal process is:

$$h^{(m,\lambda)}(Y) = \frac{1}{2}\log\left(2\pi e\sigma^2\right) - \frac{1}{2}\log\left(\frac{|\Sigma^{(m)}|}{|\Sigma^{(m+1)}|}\right) + \mu \tag{1.47}$$

CHAPTER 2

Non-parametric estimators

In order to apply Information Theory in Signal Processing, it is mandatory to develop reliable numerical codes which estimate properly entropy measures. From our knowledge, there are three main kinds of non-parametric algorithms used to estimate information quantities: algorithms based on histograms, algorithms based on kernel density estimations (KDE), or algorithms based on nearest neighbors search (NNS). Each kind of algorithm follows different approaches to estimate the probability density function, or to estimate the entropy measures directly. Of course, other kind of estimators of information quantities exist, nevertheless, in the course of the PhD only non-parametric estimates have been used.

Along this chapter we use previous literature and new original results to choose the most reliable estimates of Information theory measures. All the already known conclusions are supported with their respectives original papers. The new results obtained during the PhD, are properly explained (and related with their corresponding article). The chapter is organized as follows: the first section enumerates different algorithms to estimate PDF's, the second one presents the Kozachenko-Leonenko algorithm to compute Shannon entropy [114, 184, 127], the third section is dedicated to Kraskov-Stögbauer-Grassberger estimate of Mutual Information [116], the fourth section presents a new method to calculate the m-order Shannon entropy rate [78] (based on those of Kozachenko-Leonenko and Kraskov-Stögbauer-Grassberger), the fifth section explains how to use Theiler prescription to probe dynamics and complexity at a given scale λ, and finally the last section presents the Kraskov-Stögbauer-Grassberger estimate for conditional mutual information.

2.1 Probability Density Function Estimates

Historically, the first entropy estimates were based on histogram estimation of probability density functions [173, 198]. These estimates present two free parameters: the origin, and the size of the bins. Changes in both of them, especially the size of the bins, affect seriously the estimation of the PDF [198]. The bias of these algorithms is significant and renders the measure difficult to interpret [198]. In order to enhance estimations several variations of histograms alorithms were developed [174, 182], which "eliminate" the origin as a parameter of the estimate and then "eliminate" this source of bias. Nevertheless, using bins still generates a high bias of the measure [173, 198].

A very important improvement is achieved when the shape of the bin is made smoother. We call this a kernel estimate [171, 157, 30]. The function defining the shape of the bins must obey some properties which make the estimation well behaving and reduce its bias [173, 198]. However, as the size of the kernels is still fixed a bias remains [173, 198] specially for long tailed PDFs [182]. Various improvements on kernel estimations have been done along the last fifty years [200, 1, 201].

Nearest Neighbor Search algorithms were then developed to deal with constant-length problems of kernels [133]. The idea is to vary the size of the kernel according to the density of samples in the area [133, 182]. Bias due to fixed size of the kernels disappears but we recover the bias due to their "square" shape [135, 146]. Some mixtures of kernel and NNS estimates of PDF have been later developed [27], improving the estimation properties. Finally, Walter and Blum, and, Terrel and Scott showed that many different nonparametric PDF estimators, among them NNS, can be understood as kernel estimators [212, 201].

Once the PDF has been estimated, it should be inserted in the specific formula of the corresponding information theory measure. Calculating the PDF to later compute the information theory measure presents an important bias. For this reason, algorithms performing straightforward estimations of information theory objects are usually better than those that first estimate the PDF. These "direct" algorithms are normally based on the same strategies: histograms, KDE or NNS, and they search for bias reductions or compensations.

2.1.1 Histogram algorithms

Histogram algorithms base the estimation of the probability density function of a random variable on a simple box counting method. The domain of possible values that the random variable can take is divided in boxes of a chosen size with a given origin, and then the number of times the random variable X_i takes a value in each box is counted. Given an origin x_0 and a bin width Δx we define the bins as $[x_0+m\Delta x, x_0+(m+1)\Delta x)$ for positive and negative integers m. The estimate of the distribution is defined by [53, 182, 97, 175]:

$$\hat{f}(x) = \frac{1}{N} \times \frac{\text{nb. of } x_i \text{ in same bin as } x}{\Delta x} \tag{2.1}$$

where $x_i \in \mathbb{A}_i$ represents the possible values of X_i, N is the number of realizations of the random variable, and the bin width Δx controls the amount of smoothing in the procedure.

One can generalize the histogram algorithm by allowing the bin width to vary [53, 182, 175]. Nevertheless, histogram estimators depend on both origin x_0 and width of the bins Δx. The existence of two parameters affecting the distribution estimation together with the discontinuity of the estimated PDF lead to the necessity of more accurate estimators [182, 97].

2.1.2 Kernel algorithms

Kernel estimator is a generalization of histogram estimator in which every observation x_i is the center of a bin. In this way, kernel estimator does not depend anymore on the origin x_0 that disappear as a parameter of the estimation. On the other hand, kernel estimator allows a wide range of different shapes K for the bins. Then, kernel estimator with kernel K is defined by [157, 171, 30, 53, 182, 175, 145]:

$$\hat{f}(x) = \frac{1}{N\Delta x} \sum_{x_i \in \mathbb{A}_i} K\left(\frac{x - x_i}{\Delta x}\right) \tag{2.2}$$

where Δx is the window width or band width. The kernel estimator is a sum of bump functions K, centered at the observations x_i.

For $\hat{f}$ defined in eq.(2.2) to tend to the real distribution f, the kernel function should be a Borel function satisfying [157, 171, 30, 97]:

$$\sup_{-\infty<x<\infty} |K(x)| < \infty \tag{2.3}$$

$$\int_{-\infty}^{\infty} |K(x)| \mathrm{dx} < \infty \tag{2.4}$$

$$\lim_{x\to\infty} |xK(x)| = 0 \tag{2.5}$$

$$\int_{-\infty}^{\infty} K(x)dx = 1 \tag{2.6}$$

and the window width Δx should satisfy:

$$\lim_{N\to\infty} \Delta x|_N = 0 \tag{2.7}$$

For the most basic kernel estimator, called the naive estimator, the kernel function K is a weight function, $K(x) = 1/2$ if $|x| < 1$ and zero otherwise [157, 30, 192]. The choice of bin width still remains and controls the amount by which the data are smoothed to produce the estimate. As the basic histogram estimator the naive one is not a continuous function of x [182, 97].

More complex kernel estimators can be obtained by replacing the weight function by a smoother kernel function K as for example Gaussian. A table containing some kernel functions satisfying the above properties was presented by Parzen [157]. From the definition of kernel estimates one obtains that if K is a probability density function, then $\hat{f}$ is a probability density function which inherits all the continuity and the differentiability properties of the kernel [182, 97].

The main drawback of this kind of estimators appears when applied to data from long-tailed distributions. The density of this kind of distributions strongly varies from high values in the region near the mode of the distribution to small densities in the tails. Nevertheless the band width is fixed across the entire sample, and then there is a tendency for spurious noise to appear in the tails of the estimates, where density is small [182, 138].

2.1.3 Nearest Neighbors Search algorithms

In NNS estimators the degree of smoothing is determined by the distance to the kth nearest neighbor $d_k(x)$. The kth nearest neighbor density estimate is defined by [135, 146, 53, 182, 175]:

$$\hat{f}(x) = \frac{k}{2Nd_k(x)} \tag{2.8}$$

The nearest neighbor estimate is inversely proportional to the size of the box needed to contain k number of observations. Then, in the tails of the distribution, the distance d_k will be larger than in the center, and the problem of undersmoothing in the tails should be reduced.

The nearest neighbor estimate $\hat{f}$, is continuous and positive everywhere but have discontinuous derivatives [182].

2.1.4 Combinations of KDE and NNS

Kernel generalization of nearest neighbor estimate

With $K(x)$ a normalized kernel function, the generalized nearest neighbor estimate reads [53, 182, 175]:

$$\hat{f}(x) = \frac{1}{Nd_k(x)} \sum_{x_i \in \mathbb{A}_i} K\left(\frac{x - x_i}{d_k(t)}\right) \tag{2.9}$$

where $\hat{f}(x)$ is precisely the kernel estimate with window width $d_k(x)$. However, the derivative of the generalized nearest neighbor estimate will remain discontinuous when $d_k(x)$ has discontinuous derivative [182].

The variable kernel method

This estimate is constructed similarly to the classical kernel estimate but the scale parameter of the bumps is allowed to vary from one data point to another. We define K as the kernel function as above, and $d_{i,k}$ as the distance from x_i to the kth nearest observation. The variable kernel estimate with smoothing parameter Δx is [53, 182, 175]:

$$\hat{f}(x) = \frac{1}{N} \sum_{x_i \in \mathbb{A}_i} \frac{1}{\Delta x d_{i,k}} K\left(\frac{x - x_i}{\Delta x d_{i,k}}\right) \tag{2.10}$$

where the band width of the estimate placed on the point X_i is proportional to $d_{i,k}$, i.e. data points in regions where the data are sparse will have flatter kernels associated with them. Finally for fixed k the degree of smoothing depends on Δx.

As the basic kernel estimate, the variable kernel estimate is a probability density function by definition.

In the above sections we have presented three different strategies of computing the PDF of a process: histograms, KDE and NNS. In addition, for each strategy we presented different algorithms. Nevertheless, estimating the entropy of a process by computing its PDF and using eq.(1.2) is not the less biased approach. In sections 2.2, 2.3 and 2.4 we present a straightforward methodology based on NNS to compute respectively the entropy, the mutual information and the m-order entropy rate. These three straightforwards estimators based on NNS have been proved to be less biased than the strategies presented above and other straightforward strategies [210, 116, 154, 213, 203].

2.2 Kozachenko-Leonenko estimate for Shannon Entropy

The nearest neighbor search estimator described by Kozachenko and Leonenko [114, 184, 127] is based on the measure of the density $\mu(x)$ of the distribution, by calculating ϵ so that there are k points inside a ball of radius $r = \epsilon/2$ whose center is the point x_i of the process X. The expression of the entropy becomes:

$$H(X) = -\int_{\mathbb{A}} dx\mu(x)\log(\mu(x)) = -\langle\log(\mu(x))\rangle = -\frac{1}{N}\sum_{i=1}^{N}\log\mu(x_i) \tag{2.11}$$

where N is the number of realizations of the process X.

It's important to remark that in the Kozachenko estimator the number of neighbors k is fixed and ϵ varies in function of the density $\mu(x)$ around x_i.

The recipe to compute the Kozachenko-Leonenko estimate of entropy is:

i) For a process X, with values in a metric space, and whose density is $\mu(x)$, we define its Shannon entropy in eq.(2.11). Unbiased estimator of the density $\widehat{\log(\mu(x))} = \log(\mu(x))$ implies unbiased estimator of the entropy $\widehat{H}(X) = H(X)$.

ii) We want to calculate $\widehat{\log(\mu(x_i))}$. In order to do this, we consider:

1. $P_k(\epsilon) \equiv$ Probability distribution for the distance between x_i and its kth nearest neighbour.
2. $P_k(\epsilon)d\epsilon \equiv$ Probability of having one point with distance $r \in [\epsilon/2, \epsilon/2 + d(\epsilon/2)]$ from x_i, having $k-1$ points with smaller distances, and $N-k-1$ with larger distances.
3. $p_i \equiv$ Mass of the ϵ ball centered at x_i.

iii) We can obtain (using the trinomial formula):

$$P_k(\epsilon)d\epsilon = \frac{(N-1)!}{1!(k-1)!(N-k-1)!}\frac{dp_i(\epsilon)}{d\epsilon}d\epsilon \times p_i^{k-1}(1-p_i)^{N-k-1} \tag{2.12}$$

with

$$\int d\epsilon P_k(\epsilon) = 1 \tag{2.13}$$

iv) We can therefore calculate the expectation value of $\log(p_i(\epsilon))$:

$$\mathbb{E}(\log(p_i)) = \int_0^\infty d\epsilon P_k(\epsilon)\log(p_i(\epsilon)) = \psi(k) - \psi(N) \tag{2.14}$$

with $\psi(x) \equiv$ Digamma function and $\mathbb{E}$ the average over the positions of the other $N-1$ points, with x_i kept fixed.

v) We suppose $\mu(x) \equiv cte$ within the ϵ ball (This is the only approximation).

$$p_i(\epsilon) \approx c_d \epsilon^d \mu(x_i) \tag{2.15}$$

where d is the dimension of X, c_d is the volume of the d-dimensional unit ball (it depends on the space metric).

vi) We take the logarithm on both sides of eq.(2.15) and then its expectation value, so we arrive at:

$$\log\mu(x_i) \approx \psi(k) - \psi(N) - d\mathbb{E}(\log(\epsilon)) - \log(c_d) \tag{2.16}$$

vii) We finally obtain:

$$\widehat{H}(X) = -\psi(k) + \psi(N) + \log(c_d) + \frac{d}{N}\sum_i^N \log(\epsilon(i)) \tag{2.17}$$

with $\epsilon(i)$ twice the distance from x_i and its kth nearest neighbour.

The only approximation made in eq.(2.15) implies that $\widehat{\log(\mu(x)})$ and then $\widehat{H}(X)$ are unbiased only if the density $\mu(x)$ is strictly constant. In the following, unless said otherwise, we use the Kozachenko-Leonenko estimator for entropy calculations and use infinite norm when computing it. We reserve the notation $\widehat{H}(X)$ for the Kozachenko-Leonenko estimate of entropy.

Rate of Convergence: Bias and Variance

Gao et al. studied the rate of convergence of the Kozachenko-Leonenko estimator over probability density functions with bounded support [69]. They showed that under some assumptions for the PDF's and for finite $k = O(1)$ and dimension $m = O(1)$ the bias of the Kozachenko-Leonenko estimator is $\tilde{O}(N^{-\frac{1}{d}})$, its variance is $\tilde{O}(1/N)$ and hence the l_2 error of the Kozachenko-Leonenko estimator is $\tilde{O}(\frac{1}{\sqrt{N}} + N^{-\frac{1}{d}})$.

Comparisons between Kozachenko-Leonenko algorithm and kernel density estimation algorithms for Shannon entropy estimates have shown the advantages of the first one, especially for high dimensions [69, 223]. The Kozachenko-Leonenko Shannon entropy estimate presents also better properties, as smaller bias or faster convergence, than bin strategies for entropy estimation [210].

2.3 Kraskov-Stögbauer-Grassberger estimate for Mutual Information

As MI can be expressed as a combination of entropies, a direct way to calculate MI is to use the Kozachenko-Leonenko estimator for the three entropies involved in MI definition. Procceding this way, the bias of the measure is large because the different bias provided of the different entropy estimations have no reason to compensate and add up into a huge bias for the MI measure [116, 69].

To get a better measure of the Mutual Information, Kraskov et al. [116] developed two methods based on the Kozachenko-Leonenko estimate for Shannon entropy. The Kraskov-Stögbauer-Grassberger estimators fix a k neighbors value for the joint space $(X \cup Y)$, then, in this space they obtains the distance to the kth neighbor, ϵ. Once ϵ is obtained in the joint space, it remains fixed for the marginal spaces (X and Y), where the number of neighbors can vary. Fixing ϵ for all the entropy calculations produces a bias compensation between them. As such, we can understand the Kraskov-Stögbauer-Grassberger estimate as a mixture between NNS and KDE.

Next we detail the recipe to compute the Kraskov-Stögbauer-Grassberger estimates for mutual information.

First Mutual Information estimator. We consider maximum norm and the 2D random variable $Z(X,Y)$, defined as $Z = X \cup Y$. We take one point z_i and we consider the distance $\epsilon/2$ to its kth nearest neighbour. We find eq.(2.12) and eq.(2.14) and we can write eq.(2.15) in 2D:

$$p_i(\epsilon) \approx c_{d_x} c_{d_y} \epsilon^{d_z} \mu(z_i) \tag{2.18}$$

where $d_z = d_x + d_y$. And we obtain:

$$\widehat{H}(X,Y) = -\psi(k) + \psi(N) + \log(c_{d_x} c_{d_y}) + \frac{d_x + d_y}{N} \sum_{i}^{N} \log(\epsilon(i)) \tag{2.19}$$

With this method $\widehat{H}(X)$, $\widehat{H}(Y)$, $\widehat{H}(X,Y)$ could be given, and we could find the MI using eq.(1.18). But if we use the same number of neighbors k for $\widehat{H}(X)$, $\widehat{H}(Y)$ and $\widehat{H}(X,Y)$ their bias is so different that they do not cancel.

It's therefore better to have different values of k in the calculation of $\widehat{H}(X)$, $\widehat{H}(Y)$ and $\widehat{H}(X,Y)$. To obtain adequate values for the number of neighbors of each different distribution, we fix k for the joint distribution $Z(X,Y)$ and we obtain the distance to the kth nearest neighbour $\epsilon(i)$. Then using this distance $\epsilon(i)$, we calculate the number of neighbours $n_x(i) + 1$ lying on a segment $x = x_i \pm \frac{\epsilon(i)}{2}$ of the X dimension, *i.e.* lying on $-\infty \leq Y \leq \infty$, $x_i - \epsilon(i)/2 \leq X \leq x_i + \epsilon(i)/2$.

Consequently, $\epsilon(i)/2$ is the distance to the $n_x(i) + 1$st neighbour of x_i. We use the same developement for Y and we obtain that $\epsilon(i)/2$ is the distance to the $n_y(i) + 1$st neighbour of y_i. That's not really true, but is not a bad approximation. The errors in $\widehat{H}(X)$, $\widehat{H}(Y)$, $\widehat{H}(X,Y)$ will not cancel exactly in general, but the chances that they will do approximately are bigger with this last procedure. We arrive to:

$$\textbf{1) } \widehat{I}(X,Y) = \psi(k) + \psi(N) - \langle \psi(n_x) + \psi(n_y) \rangle \tag{2.20}$$

Second Mutual Information estimator. In this development we don't use hypercubes in the joint space but hyper-rectangles. Now we don't have a single ϵ but two, ϵ_x and ϵ_y. We therefore replace $P_k(\epsilon)$ by:

$$P_k(\epsilon_x,\epsilon_y) = \begin{pmatrix} N-1 \\ k \end{pmatrix} \frac{d^2[q_i^k]}{d\epsilon_x d\epsilon_y}(1-p_i)^{N-k-1} + (k-1) \begin{pmatrix} N-1 \\ k \end{pmatrix} \frac{d^2[q_i^k]}{d\epsilon_x d\epsilon_y}(1-p_i)^{N-k-1} \tag{2.21}$$

where $q_i \equiv q_i(\epsilon_x,\epsilon_y)$ is the mass of the rectangle of size $\epsilon_x \times \epsilon_y$ centered at (x_i, y_i) and p_i is the mass of the square of size $\epsilon = \max \epsilon_x, \epsilon_y$. Using the maximum norm we guarantee that there are no points in the square that are not in the rectangle.
We can calculate the expectation value of the logarithm of the mass of the hyper-rectangles:

$$\mathbb{E}(\log q_i) = \int\int_0^\infty d\epsilon_x d\epsilon_y P_k(\epsilon_x,\epsilon_y)\log(q_i(\epsilon_x,\epsilon_y)) = \psi(k) - \frac{1}{k} - \psi(N) \tag{2.22}$$

If we say now that $n_x(i)$ and $n_y(i)$ are the number of points with distance less than or equal to $\epsilon_x/2$ and $\epsilon_y/2$ respectively, and following the above procedure we arrive to:

$$\textbf{2) } \widehat{I}(X,Y) = \psi(k) + \psi(N) - \langle \psi(n_x) + \psi(n_y) \rangle - \frac{1}{k} \tag{2.23}$$

From here till the end of the PhD manuscript, even if both mutual information estimates are equally robust, when computing mutual information we will use the first mutual information estimator with infinite norm, unless said otherwise.

Rate of Convergence: Bias and Variance

In their initial article, Kraskov *et al.* derived and showed a bias behaviour of the Kraskov-Stögbauer-Grassberger estimate in $\frac{k}{N}$ [116]. This result was later studied and supported by Gao et al. [69] who showed that the bias of the estimator is $\tilde{O}((\frac{k}{N})^{\frac{1}{d_x+d_y}})$ and the variance $\tilde{O}(1/N)$. Its l_2 error is $\tilde{O}(\frac{1}{\sqrt{N}} + (\frac{k}{N})^{\frac{1}{d_x+d_y}})$.

Since the apparition of Kraskov-Stögbauer-Grassberger estimator of mutual information numerous comparisons between this NNS based estimator and others based on KDE or bins have been done [116, 69, 223, 203]. While Kraskov et al. [116], Gao et al. [69], and Xiong et al. [223], support the use of the Kraskov-Stögbauer-Grassberger estimator over KDE and bins, Tobin et al. [203] show that their KDE estimator produce very similar results. Kraskov-Stögbauer-Grassberger estimator appears always as better than bins estimators.

2.4 *m*-order entropy rate estimate

The simplest entropy rate estimation would use eq.(1.24) with two Kozachenko-Leonenko entropy estimators:

$$\hat{h}_1^{(m,\lambda)}(X) = \hat{H}(X^{(m+1,\lambda)}) - \hat{H}(X^{(m,\lambda)}). \tag{2.24}$$

We propose instead to use eq.1.25 with Kozachenko-Leonenko estimate for the entropy and the Kraskov-Stögbauer-Grassberger estimate for the mutual information $\widehat{I}$.

$$\hat{h}_2^{(m,\lambda)}(X) = \hat{H}(X) - \hat{I}(X_{i+\lambda}, X_i^{(m,\lambda)}) \tag{2.25}$$

As we showed above, both estimators $\hat{H}$ and $\hat{I}$ are consistent with a variance $\mathcal{O}(1/N)$ and a bias $\mathcal{O}((\frac{k}{N})^{\frac{1}{d_{total}}})$ where N is the number of samples of the analysed process and k the number of neighbors in the search.

From here till the end we estimate the m-order entropy rate by computing the Kozachenko-Leonenko entropy and the Kraskov-Stögbauer-Grassberger mutual information ($\hat{h}_2^{(m,\lambda)}(X)$), unless said otherwise.

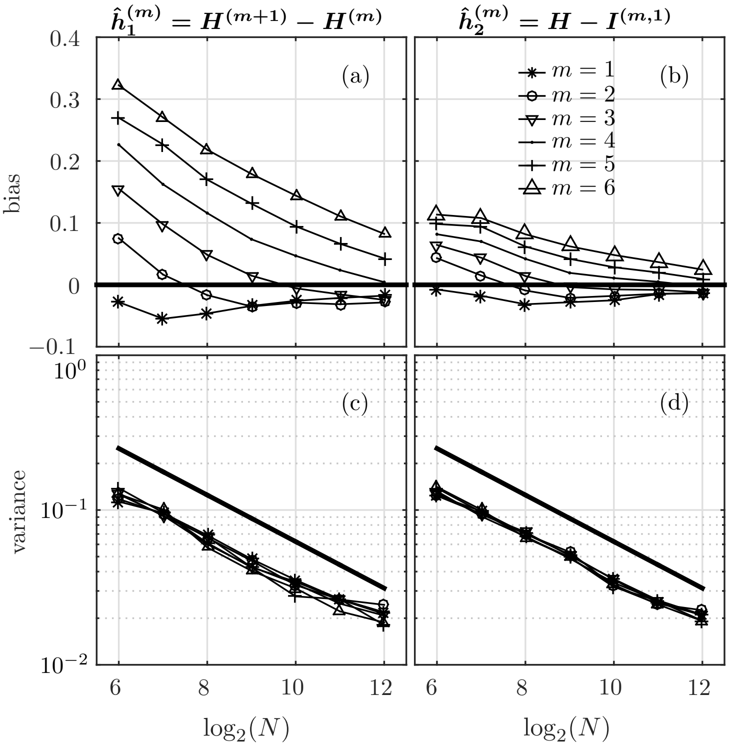

Figure 2.1: **Entropy rate performances (at fixed τ) on the Gaussian process.** Bias and standard variation of the estimates in function of the number of samples N for both entropy rate estimators. The scale τ is fixed ($\tau = 3$) and chosen such that the correlation coefficient between two succesive points is $c(\tau) = 0.94$. In c) and d) the straight line correspond to $f \propto \sqrt{\frac{1}{N}}$

Rate of Convergence: Bias and Variance

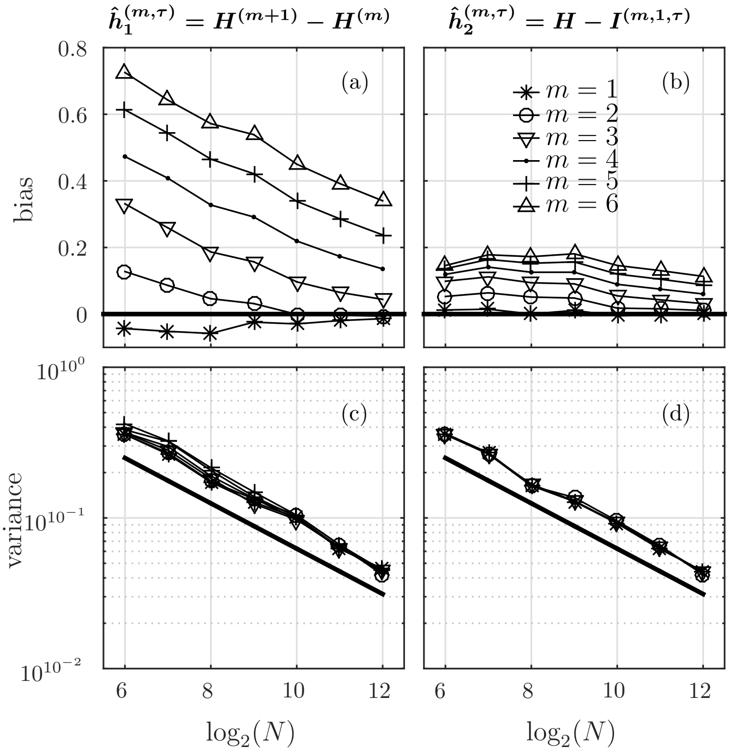

Figure 2.2: **Entropy rate performances (at fixed τ) on the log-normal process.** Bias and standard variation of the estimates in function of the number of samples N for both entropy rate estimators. The scale τ is fixed ($\tau = 3$) and chosen such that the correlation coefficient between two succesive points is $c(\tau) = 0.94$. In c) and d) the straight line correspond to $f \propto \sqrt{\frac{1}{N}}$

In order to compare both estimators of entropy rate, we study their bias and standard deviation [78] when computed for two kinds of temporal processes (for temporal processes the sampling distance λ is noted τ) with different marginals and correlation functions but with the same entropy rate [226]. The first process X is obtained by low-pass filtering a Gaussian white noise. The used filter is defined as:

$$F(\omega) = \frac{1}{1 + \mathrm{i}\frac{f}{f_c}} \tag{2.26}$$

where f_c represents the cut-off frequency of the filter and $i = \sqrt{-1}$. Next we normalize the signal by its variance, so that $\sigma^2 = 1$, and we obtain a Gaussian process whose correlation function is:

$$c(t - t') = e^{-2\pi f_c |t-t'|}. \tag{2.27}$$

The entropy rate of process X is deduced from eq.(1.40).

The second process is defined as $Y = e^X$ with X being the above defined Gaussian process. We define thus a process with log-normal marginal and correlation function:

$$c_{\mathrm{LN}}(t-t') = \frac{e^{c(t-t')}-1}{e^1-1}. \tag{2.28}$$

If the mean of the Gaussian process X is zero, the entropy and entropy rate of the log-normal process Y defined in eq.(1.43) and eq.(1.47) have the same expressions as the entropy and entropy rate of the Gaussian process X [226], see eq.(1.36) and eq.(1.40).

In figures 2.1 and 2.2 we analyse the evolution of the bias and standard deviation in function of the size N of the analysed process, the scale parameter is fixed to $\tau = 3$ and the number of neighbors k is fixed to 5. Indeed the different estimators depend very slightly on this parameter if $k \ll N$.

The new entropy rate estimator defined as the difference between the entropy and the mutual information appears as better than the estimator defined as the difference of entropies. Indeed, this new estimator is not only less biased and less dependent on the embedding dimension but also its bias is much less dependent on the process statistics (see figures 2.1 a) and b) and figures 2.2 a) and b)). This is an important feature as the statistics of the processes are *a priori* unknown in applications.

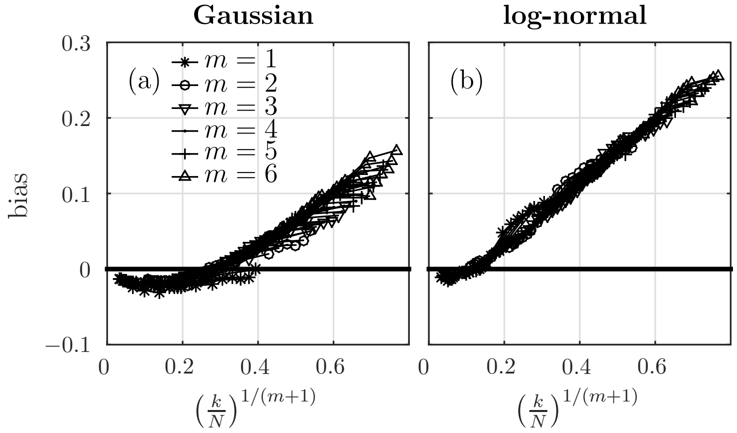

Figure 2.3: **Analysis of the Bias of the Gaussian and the log-normal processes** in function of $\left(\frac{k}{N}\right)^{\frac{1}{m+1}}$ for the new entropy rate estimator. The scale τ is fixed at ($\tau = 3$) and chosen in order for the correlation coefficient between two succesive points to be $c(\tau) = 0.94$. The number of neighbors is $5 \leq k \leq 10$, the sizes of the processes are comprised between $6 \leq \log_2(N) \leq 12$ and the embedding dimensions are $1 \leq m \leq 6$.

Figure 2.3 verifies that the bias of the new entropy rate estimator behaves as $\mathcal{O}((\frac{k}{N})^{\frac{1}{m+1}})$ and supports the results already shown by Kraskov *et al.* [116] and by Gao *et al.* [69], but this time for our estimator of entropy rate. In figure 2.3 the number of neighbors in the search varies between 5 and 10, $k = [5, 10]$. The number of samples increases from a minimum value of $N = 64$ to a largest value of $N = 4096$. The embedding dimension varies from $m = 1$ to $m = 6$. In both figures 2.3(a) and (b) the bias goes to zero when $k/N \to 0$. In addition there is another intermediate value of k/N for which the bias is zero, $k/N|_0$. This intermediate value depends on the analyzed process. When k/N

is between zero and $k/N|_0$, the theoretical value $h^{(m,\lambda)}(X)$ is underestimated. However when k/N is higher than $k/N|_0$ the theoretical value is overestimated. These results are in agreeement with the Kraskov et al. study [116].

We also checked that the variance of the new entropy rate estimator behaves as $\mathcal{O}(1/N)$ (see figures 2.1 d) and 2.2 d)).

2.5 Dynamics at a given scale: Adapted Theiler prescription

Takens embedding procedure (section 1.2) inherently defines a distance λ between samples of the analysed process X. In this way, we generate a process $X^{(m,\lambda)}$, for which each sample is an m-dimensional vector, and each vector is composed of consecutive elements of X separated by λ. The distance between consecutive m-dimensional vectors defining the embedded process, $X^{(m,\lambda)}$, is another important sampling distance λ_{th}. Then, from a process X, we reconstruct an embedded one $X^{(m,\lambda)}$, composed by m-dimensional vectors with a separation between consecutive vectors of λ_{th}. The classical Theiler prescription consists on choosing λ_{th} higher or equal to the correlation distance of the process, in order to eliminate all the dependencies between consecutive vectors of the embedded process. Therefore, only the static properties of the embedded process are taken into account. This correction, known as Theiler prescription [202], is usually used in order to avoid the effect of spurious correlations on the PDF of the studied signal [105].

We are interested in the study of the dynamics of the process at a given scale λ. With this purpose we modify the classical Theiler prescription to take into account the statistics and dynamics of the process at this scale, and we choose $\lambda_{th} = \lambda$. In this way, we force the distance between consecutive vectors to be exactly λ, the distance between consecutive samples inside the m-dimensional vectors, in order to eliminate the dependencies until the analysed scale λ. This is very important in order to take into account only desired interactions.

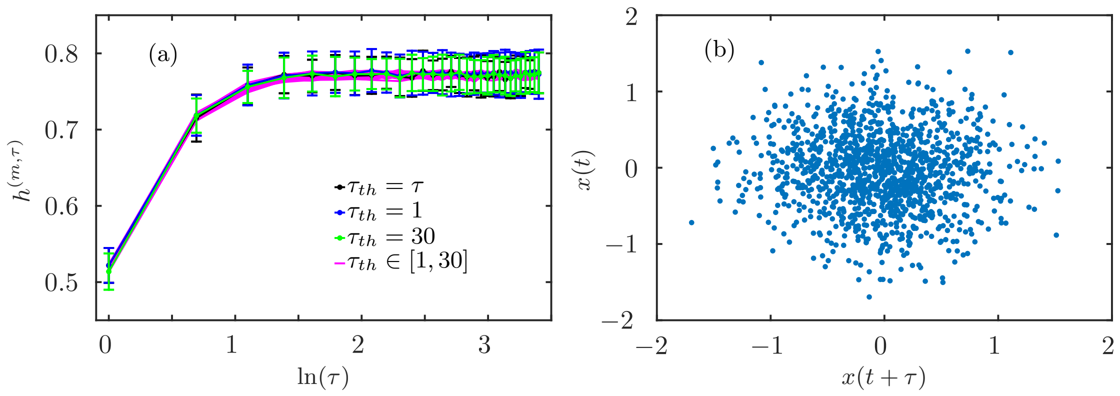

Figure 2.4: a) Entropy rate of a low-pass filtered Gaussian noise ($f_c = 0.1$) b) Phase space of a low-pass filtered Gaussian noise. The entropy rates have been computed with a constant number of $N = 2000$ samples and averaged over 50 realizations.

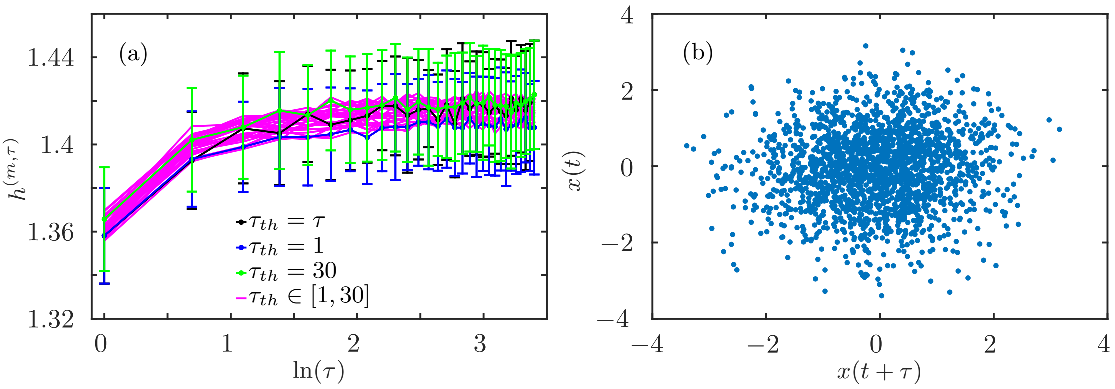

Figure 2.5: a) Entropy rate of fractional Gaussian noise ($\mathcal{H} = 0.7$) b) Phase space of fractional Gaussian noise. The entropy rates have been computed with a constant number of $N = 2000$ samples and averaged over 50 realizations.

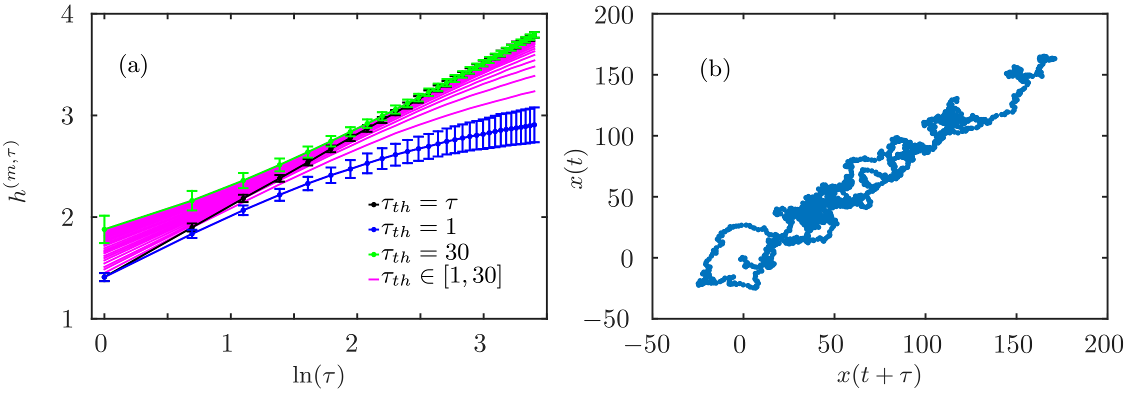

Figure 2.6: a) Entropy rate of fBm ($\mathcal{H} = 0.7$) b) Phase space of fractional Brownian motion. The entropy rates have been computed with a constant number of $N = 2000$ samples and averaged over 50 realizations.

We show, by analysing some temporal processes that, varying λ in the embedding procedure with an adapted Theiler prescription allows an efficient analysis of high order statistics across scales. We study the effect of performing an adapted Theiler prescription in four different temporal processes: a Gaussian filtered noise (the filter is defined in eq.(2.26)), a fractional Gaussian noise (section 3.1), a fractional Brownian motion (section 4.2) and Modane experimental turbulent velocity signal (section 6.1).

We can see in figures 2.4,2.5,2.6 and 2.8 that the adapted Theiler prescription is not equally important for every process. It depends on the importance of the spurious correlations remaining: the more correlated the process the more important the application of a Theiler correction to only take into account the desired statistics. In order to illustrate the importance of the spurious correlations on each analysed process, we plot in figures 2.4 b),2.5 b),2.6 b) and 2.8 b) their corresponding phase space. The more scattered the points in the phase space, the more uncorrelated the process.

We have analysed two noises with different dependence structure (figures 2.4 and 2.5),

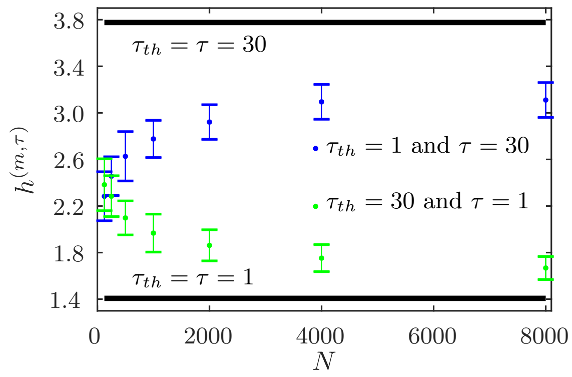

Figure 2.7: Entropy rate of fBm ($\mathcal{H} = 0.7$) in function of the effective size of the analyzed process. The entropy rates have been computed with a constant number of $N = 2000$ samples and averaged over 50 realizations.

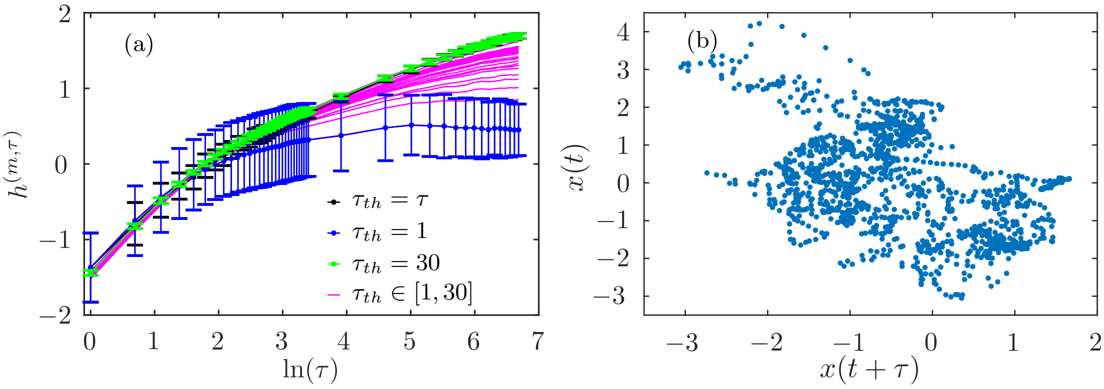

Figure 2.8: a) Entropy rate of a turbulent velocity signal b) Phase space of a turbulent velocity signal. The entropy rates have been computed with a constant number of $N = 1300$ samples and averaged over 20 realizations.

but for noises, where correlations are usually weak (figures 2.4 b),2.5 b)), Theiler prescription is not very important. Only for pathological cases where the noises are specially correlated Theiler correction starts to be necessary. We find the opposite situation when analysing motions (figure 2.6). Motions are defined as processes where each step is characterized by a corresponding noise, see section 4.2. Motions are more correlated than a noise (figure 2.6 b)), and Theiler prescription is therefore very important.

Figure 2.7 shows the effect of the size N of the analyzed process: increasing the size of the process it is possible to reduce the effect of spurious correlations. We can see in figure 2.7 that both limits $\tau_{th} = 1$, $\tau = 30$ and $\tau_{th} = 30$, $\tau = 1$ tend to their spected values, those at $\tau_{th} = \tau$, when N increases. Nevertheless the convergence towards the Theiler corrected measure when N increases is very slow.

We also present the results obtained when analysing an experimental signal of turbulent velocity 2.8. The importance of Theiler prescription in this case is not surprising as the turbulent velocity field is modeled in terms of correlations by a fractional Brownian motion, see section 5.2.1.

2.6 Kraskov-Stögbauer-Grassberger estimate for Conditional Mutual Information

Kraskov-Stögbauer-Grassberger estimates for mutual information, see section 2.3, can be generalized for conditional mutual information [65, 209, 74, 132]. As for the mutual information, this procedures allows to reduce the bias of the estimate by a bias compensation of the terms. The conditional mutual information of eq.(1.19) can be expressed as:

$$I(X,Y|Z) = H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z)$$

To extend the first Kraskov-Stögbauer-Grassberger estimate we calculate in the joint space $(X \cup Y \cup Z)$ the distance (using max norm) ϵ to the kth nearest neighbor. In the same way as for the mutual information, n_z is the number of neighbors lying strictly within the ball of diameter ϵ in the marginal space Z, and n_{xz} and n_{yz} are the number of neighbors strictly within a ball of diameter ϵ in the joint spaces $X \cup Z$ and $Y \cup Z$ respectively. We obtain:

$$\textbf{1)}\ \widehat{I}(X,Y|Z) = \psi(k) - \langle \psi(n_{xz}) - \psi(n_{yz}) + \psi(n_z) \rangle \tag{2.29}$$

The second estimate of Kraskov-Stögbauer-Grassberger for mutual information can be also extended to conditional mutual information, leading to:

$$\textbf{2)}\ \widehat{I}(X,Y|Z) = \psi(k) + \left\langle \psi(n_z) - \psi(n_{xz}) - \psi(n_{yz}) + \frac{1}{n_{xz}} + \frac{1}{n_{yz}} \right\rangle - \frac{2}{k} \tag{2.30}$$

As transfer entropy can be expressed as a conditional mutual information we can exploit these algorithms to compute it, and from here till the end of the manuscript every transfer entropy is computed following these algorithms.

CHAPTER 3

High order Statistics on non-Gaussian noises

The analysis of correlations and dependences covers an important fraction of the signal processing research. In order to characterize a process $X(t)$, evolving along a given dimension t, *e.g.*, the time, one needs to describe simultaneously: the one-point statistics of the process at a given time t and the dependences between n samples x_t at different times. The one-point statistics are given by the probability distribution, and amounts to the knowledge of all the moments. The dependences between n samples, or n-point dependences, are usually hard to assess and very often limited to 2 points.

To analyse one-point statistics and dependences, we assume, unless said otherwise, that the process is stationary. One-point statistics are then independent of the time t and n-point dependences depend on the $n-1$ time delays separating the n samples.

The correlation function in direct space, or the power spectrum in Fourier space, are commonly used to characterize dependences in stationary processes. These measures give insight into the underlying dynamics; they are essential tools in statistics and data analysis and have been widely used in many different fields. In neurosciences, the analysis of correlations between neurons [147, 44] or between stimulus and brain activity [16, 84, 142] is essential. In fluid mechanics, Kolmogorov's empirical theory of turbulence [108] characterizes the multiscale distribution of energy with the power spectrum of the velocity field [8]. The analysis of dependences also supports useful descriptions in terms of complex networks for, e.g., ecology [225], climatology [224] or traffic [151].

Correlation analysis and power spectrum probe two-point interactions and involve only the statistics of order two of the one point disctribution. This is satisfying for processes with Gaussian distribution: these are completely characterized by their two-point interactions together with the mean and variance of the distribution. On the contrary, the complete characterization of non-Gaussian processes requires higher order statistics.

A first way to explore higher order statistics is to generalize the two point correlation function and consider n different points. This increase of the dimensionality naturally involves the n first moments of the one-point distribution, and hence n-order statistics. Three-point correlation functions, defined as an extension of the classical two-point correlation function, have been used in cosmology, where they were related to cosmological parameters [195, 185]. Unfortunately the definition of three-point correlation functions is not unique, which may lead to ambiguities.

Another way to probe higher order statistics is to generalize the power spectrum: the bi-spectrum [86, 45] combines Fourier transforms at two different frequencies and therefore takes into account up to the third order statistics of the distribution. In the same spirit, the tri-spectrum [45, 36] probes the distribution up to the fourth order moment. Numerous applications of both bi-spectrum and tri-spectrum were proposed in cosmology [25, 139, 205], in medicine [148, 95, 106] and other fields [12, 48]. It is important to note that in order to probe higher order statistics with n-point correlations or multispectra, it is necessary to increase the dimension of the analysis.

An efficient way to probe high order statistics without increasing the dimensionality of the analysis is to use Shannon's information theory, see chapter 1. Shannon entropy characterizes the one-point distribution of the process and as such it naturally depends on all the moments of this probability density function, see section 1.3.1. Besides entropy, we are interested in the entropy rate, see section 1.4.1 and the auto mutual information, see section 1.4.2. Applications of the entropy rate have been reported in a wide range of domains, such as biology [92], dynamical systems [23, 129], fluid turbulence [80] or analysis of languages [58, 196]. Mutual information is also very popular, especially in computer sciences, e.g., in machine learning [204]. AMI has been used in very different domains [4, 5, 79, 73, 61].

Entropy rate and AMI probe dependences at a particular scale or delay, and until recently, very few papers explore their evolution with the scale [5, 80]. Finally, these quantities probe arbitrarily high order dependences when considering only two points, but they can be straigthforwardly generalized to consider interactions between more than two points.

The main goal of the chapter is to demonstrate the ability of information theory to describe high order statistics. To this end, we illustrate the use of entropy rate and auto-mutual information to characterize non Gaussian processes across scales in the same way as power spectrum and correlation function characterize Gaussian processes. As an example, we consider two synthetic log-normal processes with different dependence structure but with identical correlation function and identical marginal statistics. We show that we are able to measure differences in the high order statistics of the processes. In section 3.1, we present the synthetic processes that we study. In section 3.2, we report the convergence and the performance of the estimators of AMI and entropy rate. In section 3.3, we show their ability to measure high order dependences. We present their evolution across scales and emphasized that, contrary to classical correlation analysis, they allow a fine characterization of non Gaussian processes.

3.1 Data and procedure

3.1.1 Synthetic processes

In order to show the ability of entropy rate and auto mutual information in characterizing high order statistics of a signal, we synthesize two log-normal processes which have a different dependence structure, but identical one-point and two-point statistics. To show the generality of the method we use two frequently encountered covariance structures: exponential decay and long range dependence.

To synthesize a process Y, we need to impose the PDF and the correlation function. To obtain a log-normal PDF, we apply a transformation F on a Gaussian process X', which has the cumulative functionF_{G} and the correlation function $c_{x'}(\tau)$. This allows to map the Gaussian statistics of X' to the prescribed log-normal statistics of Y using $y = F_{\mathrm{LN}}^{-1} \circ F(x')$ where F_{LN}^{-1} is the inverse of the cumulative log-normal distribution. The correlation function $c_y(\tau)$ of the process Y can be related to $c_{x'}(\tau)$ using Hermite polynomial expansions. In some case, this relation can be inverted and $c_{x'}(\tau)$ can be obtained from the targeted correlation function $c_y(\tau)$ of the log-normal process.

The methodology proposed in [91] allows the synthesis of two log-normal processes by using two different transformation functions F. They lead to two log-normal processes with same marginal and correlation function but with different dependence structure.

Standard transformation We consider the bijective transformation

$$F(x') = F_{\mathrm{G}}(x') \tag{3.1}$$

Applied to a white Gaussian variable, this commonly used transformation leads naturally to the desired log-normal marginal distribution for Y. Due to the specific dependence structure of the log-normal process generated in this way, the correlation function of the Gaussian distribution $c_{x'}(\tau)$ is obtained by inversion of the targeted correlation function $c_y(\tau)$ [91].

$$c_{x'}(\tau) = \frac{1}{\sigma_{x'}^2} \ln((e^{\sigma_{x'}^2} - 1)c_y(\tau) + 1) \tag{3.2}$$

Even transformation Now, we consider the transformation

$$F(x') = 2(F_{\mathrm{G}}(|x'|) - 1/2) \tag{3.3}$$

which is even ($\forall x',\ F(-x') = F(x')$). In that case, the transformation that relates the correlation function $c_{x'}(\tau)$ of the underlying Gaussian process to the correlation of the log-normal process $c_y(\tau)$ is not known. However, it can be approximated using circulant embedding, as presented in [91].

Correlation structures

We impose for both log-normal processes, two different correlation structures with either short or long range dependences.

Exponential decay We impose the following covariance of the log-normal processes:

$$c_y(\tau) = \sigma_y^2 e^{-f_{\mathrm{c}}|\tau|}, \quad f_{\mathrm{c}} > 0.$$

where σ_y is the standard deviation of the process. The corresponding power spectrum is Lorentzian.

Power law decay A long range dependence process is often modelled by a fractional Gaussian noise (fGn) indexed by a coefficient $0.5 < \mathcal{H} < 1$, called Hurst exponent. We impose the covariance of the log-normal processes to be the one of the fGn, which is [164]:

$$c_y(\tau) = \frac{\sigma_y^2}{2}\left[(\tau-1)^{2\mathcal{H}} - 2\tau^{2\mathcal{H}} + (\tau+1)^{2\mathcal{H}}\right] , \tag{3.4}$$

The corresponding power spectrum behaves as a power law with exponent $-(2\mathcal{H}-1)$.

For each correlation structure, we synthesize two log-normal processes and a Gaussian one with the toolbox provided at www.hermir.org. We normalize these processes such that the standard deviation, σ_y, is equal to one.

In order to express analytically the auto mutual information defined in eq.(1.44) and the entropy rate in eq.(1.47), it's necessary to know the correlation of the process $X = \log Y$. The standard transformation amounts to choose $X' = \log Y = X$, so eq.(3.2) can be directly used to express Σ appearing in eq.(1.44) and eq.(1.47). However, for the even transformation, the complex dependence structure of Y doesn't allow to express explicitly $c_x(\tau)$. As a consequence, there is no analytical expression of the auto mutual information and the entropy rate, although the entropy, which only depend on the PDF, is identical to the one obtained with the standard transformation.

Throughout this chapter, we work only with temporal processes, and then we note the sampling distance λ as τ.

3.1.2 Estimation Setup

For two given correlation structures, exponential decay or power law decay, we consider two log-normal and a Gaussian process. For each process, we generate a set of realizations of size N. We choose $f_c = 0.1$ for the exponential decay correlation and a Hurst exponent $\mathcal{H} = 0.7$ for the power law decay.

First, we characterize the three estimators $\hat{H}$, $\hat{I}^{(m,p,\tau)}$ and $\hat{h}^{(m,\tau)}$ at a fixed scale τ, by measuring their bias and variance when varying both the signal size N and the number of nearest neighbors k. We also analyse their bias and variance when varying the correlation coefficient $c(\tau)$. We do so with $m = p = 1$ for the Gaussian process and the standard log-normal, for which we have the analytical expressions, see section 1.7. Then, we present a new log-normal process with the identical marginal and correlation function of the standard log-normal one, but with different dependence structure. For each correlation structure (exponential decay, power law decay) we present the behavior of our quantities across scales τ for all the processes. Finally we show how the AMI behaves when the embedding dimension m and p are increased. For all the estimations we compute the average of 100 independent realizations and use the standard deviation over realizations as error bars in the different graphs.

3.2 Characterization of the estimates

The study performed in this section allows to understand and characterize the effect that the non-physical parameters, such as the size of the analysed process N (section 3.2.1), or the number of neighbors used in the nearest neighbors search k (section 3.2.2), have on the measures. In section 3.2.3 we characterize the effect of the correlation strength on the estimations.

Figures 3.1, 3.2 and 3.3 present biases and standard deviations of the three different estimators $\hat{H}$ (upper row), $\hat{I}^{(1,1,1)}$ (middle row) and $\hat{h}^{(1,1)}$ (bottom row). Figure 3.1 shows the behaviour of the estimators with N, figure 3.2 with k, and 3.3 with $c(\tau)$. Troughout this chapter, the results for the standard log-normal process are depicted in blue, while the ones for the Gaussian are depicted in black.

3.2.1 Dependence on the data size N.

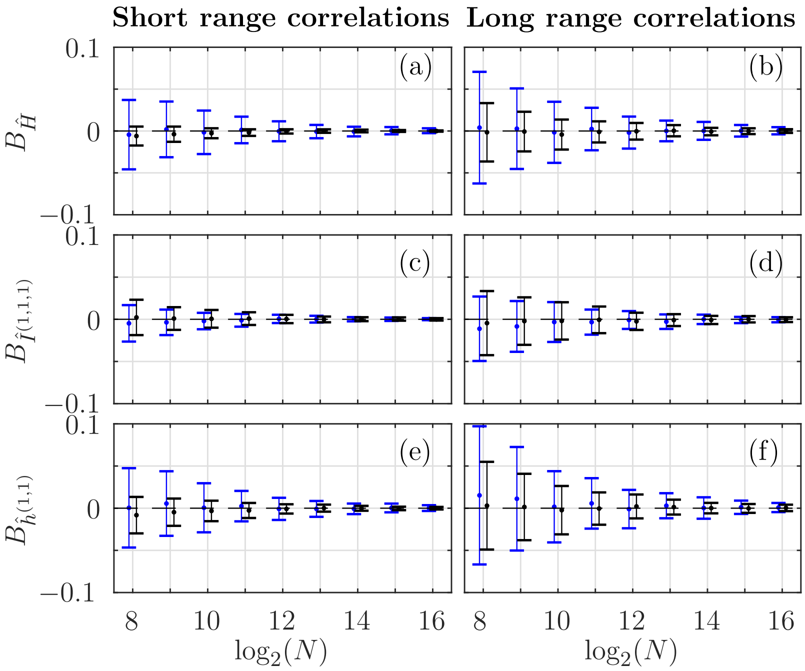

Figure 3.1: Bias of entropy, auto mutual information and entropy rate for $m = 1$, $p = 1$, $c(\tau) = 0.32$, $k = 5$ and varying N, for the standard log-normal (blue) and the Gaussian (black) processes. Left column: exponential decay $f_c = 0.1$. Right column: power law decay with $\mathcal{H} = 0.7$. The errorbars correspond to the standard deviation.

Figures 3.1 shows the behaviour of the estimates when N varies from $N = 2^8$ to $N = 2^{16}$, for a fixed $k = 5$. For the entropy estimator, figures 3.1(a) and (b) show that the bias $B_{\hat{H}} = \mathbb{E}[\hat{H} - H]$ goes to zero when the signal size N increases. The standard deviation of the estimate, represented as the error bar of the bias, also tends to zero when N increases. The auto mutual information and entropy rate estimators behave similarly, see figure 3.1(c),(d) and (e),(f), respectively. So, as a first conclusion, our three estimates are asymptotically unbiased with a vanishing variance, for any dependence structure (exponential or power law decay) and for any one-point distribution (Gaussian or log-normal). They are therefore robust [116, 69], and have (very) satisfying performance as soon as $N \geq 2^{10}$.

In the following, we use $N = 2^{16}$ to discard finite size effects.

3.2.2 Dependence on the number of neighbors k.

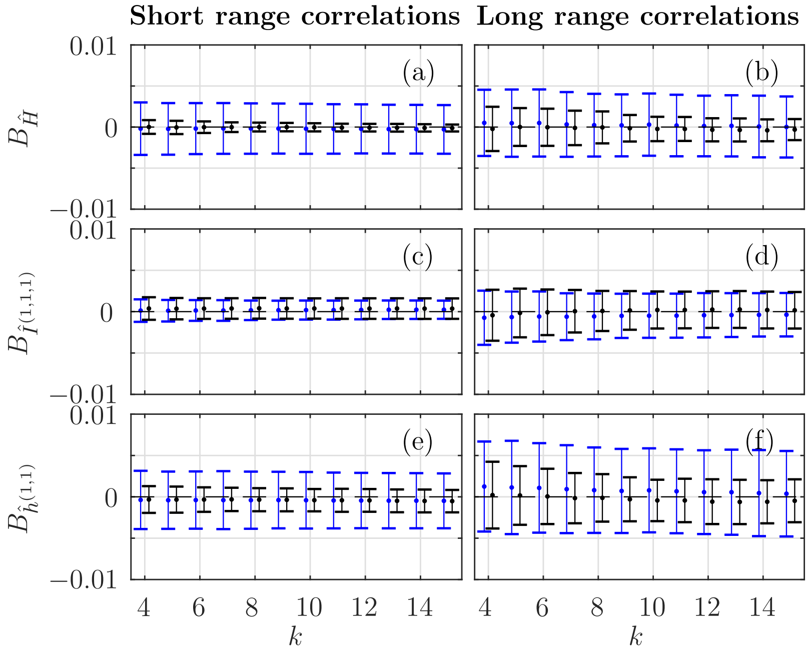

Figure 3.2: Bias of entropy, auto mutual information and entropy rate for $m = 1$, $p = 1$, $c(\tau) = 0.32$, $N = 2^{16}$ and varying k, for the standard log-normal (blue) and the Gaussian (black) processes. Left column: exponential decay $f_c = 0.1$. Right column: power law decay with $\mathcal{H} = 0.7$. The errorbars correspond to the standard deviation.

Figure 3.2 shows how our three estimators depend on k for a signal size $N = 2^{16}$. For any estimate, and any dependence structure, we barely see any evolution with k, although we have observed a small increase of bias when k increases. Because the choice of k does not impact significatively our estimates, we set $k = 5$ in the following to reduce the computation time.

3.2.3 Influence of the correlation strength

We now briefly study the behaviour of the AMI estimator when the correlation varies. With our choice of $(f_c, \mathcal{H})$, we have a correlation $c(\tau = 1) = 0.90$ for the exponential decay and $c(\tau = 1) = 0.32$ for the power law, when $\tau = 1$.

In figure 3.3, we observe that both the bias and the standard deviation of the estimator increase when the correlation increases. For $N = 2^{16}$ the bias doesn't increase significantly when $c(\tau)$ increases. Nevertheless, the standard deviation still depends on $c(\tau)$ for large sizes.

For both correlation structures (exponential decay and power law decay) the log-normal statistics lead to a larger standard deviation than the normal ones do, but a very similar bias. In addition, for a given correlation $c(\tau)$ the power law correlation structure shows a larger bias and standard deviation than the exponetial decay.

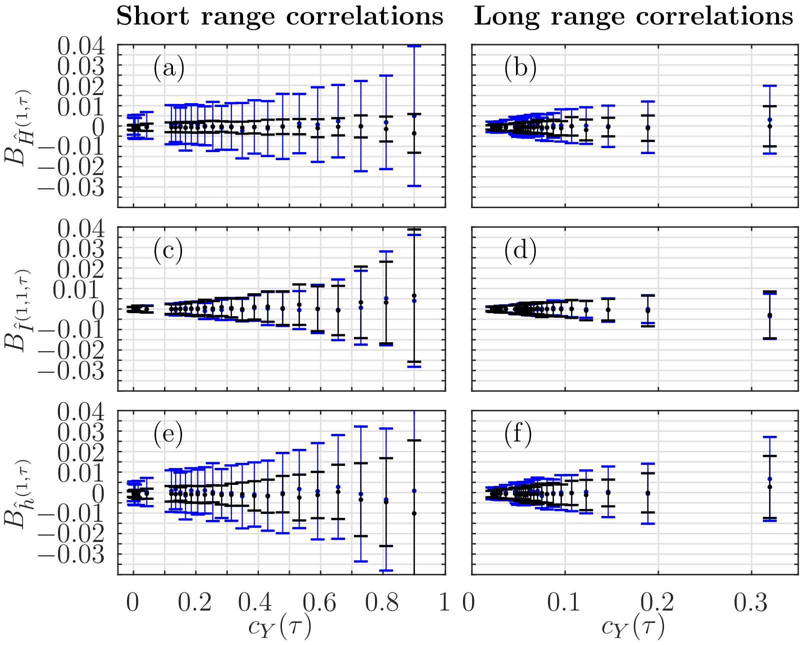

Figure 3.3: Bias of entropy (a,b), auto mutual information (c,d) and entropy rate (e,f) for $m = 1$, $p = 1$, $k = 5$, $N = 2^{12}$ and varying $c(\tau)$, for the standard log-normal (blue) and the Gaussian (black) processes. Left column: exponential decay $f_c = 0.1$. Right column: power law decay with $\mathcal{H} = 0.7$. The errorbars correspond to the standard deviation.

3.3 Dependence structure across scales

Once characterized our estimates, we explore their evolution on the scale τ, as well as on the embedding dimension. We show how these estimations can be used to discriminate processes that only differ on high-order dependences.

3.3.1 Dependence on the scale τ for fixed embedding dimensions.

Figures 3.4 and 3.5 compare the classical estimators (histogram, power spectrum and correlation function) with our estimators (entropy, entropy rate and AMI) for $m = p = 1$ when varying the scale τ. The results for the even log-normal process are reported in red.

Classical estimators cannot distinguish the two log-normal processes. Prescribing the PDF (figure (3.4,a)) univocally prescribe the entropy (figure (3.4,d)). It is therefore not surprising that the Gaussian and the log-normal processes have different entropies and that the entropy of the two log-normal processes are undistinguishable.

Exponential decay

For processes with the prescribed correlation function $c(\tau) = e^{-f_c\tau}$, the power spectrum decreases as $1/(1+(f/f_c)^2)$. The three synthetic processes have identical power spectrum — and hence identical correlation functions — so they cannot be distinguished with these estimators, see figures 3.4(b)(c).

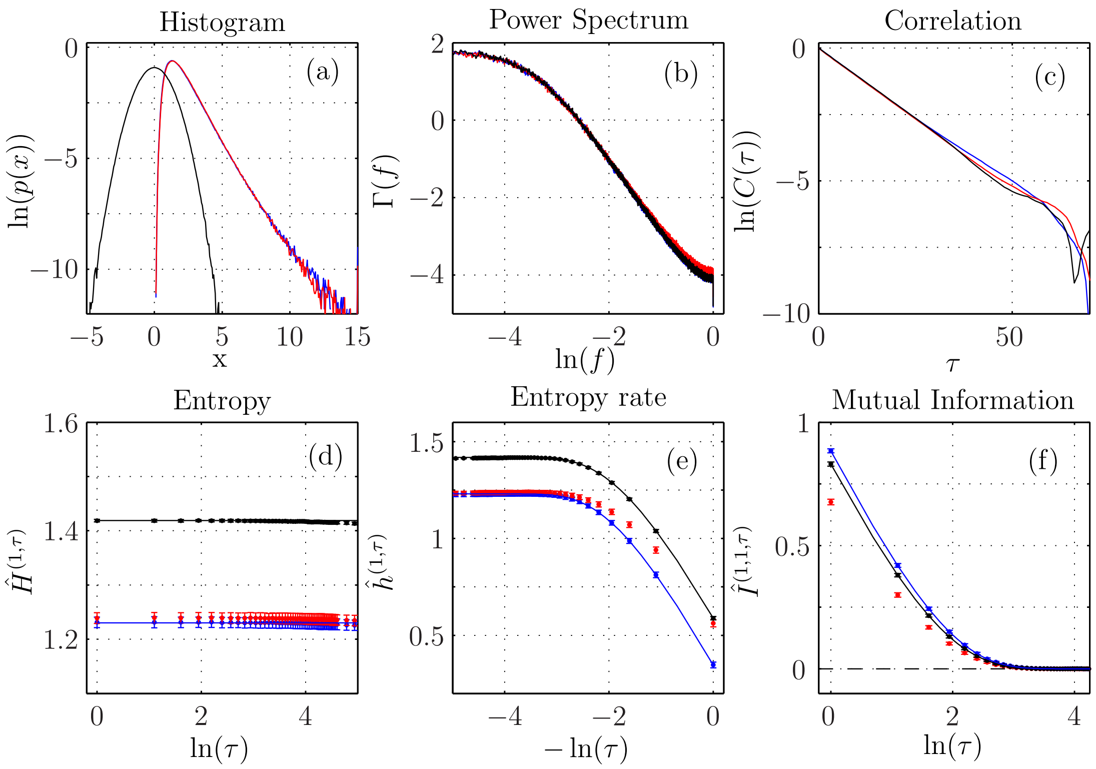

Figure 3.4: **Exponential decay** $f_c = 0.1$. Classical analysis: histogram (a), power spectrum (b) and correlation function (c). Information theory analysis for $m = 1$, $p = 1$: entropy (d), entropy rate (e) and auto mutual information (f). The color indicates the process: blue for standard log-normal, red for even log-normal and black for Gaussian. The continuous lines in (e,d,f) correspond to analytical expressions. We used $N = 2^{16}$ and $k = 5$.

On the contrary, AMI and entropy rate behave differently across scales for each of the three processes. Although AMI vanishes for larger τ, its value at $\tau = 1$ is different in each case. The observation is identical for the entropy rate: its value for smaller τ is also different for the three processes, although it converges to the entropy H for τ larger than the cut-off scale $1/f_c$. Analytical expressions (1.44 and 1.47) are presented as straight lines in 3.4(e) and 3.4(f) and they perfectly describe the scale-evolution of our estimations for the standard log-normal signal. The even (red) log-normal deviates significantly from the "standard" log-normal for smaller values of τ, where not only correlations are stronger, but also the complete dependence structure is different. We also report different behavior of the AMI of Gaussian and log-normal signals, although they all have the same correlation function (figure 3.4(c)). Again, this illustrates the sensitivity of AMI on higher order dependences. Studying the evolution of either AMI or entropy rate across scales reveals subtle differences in the dependence structures of the signals.

Power law decay

For scale invariant processes, there is no cut-off scale and both the power spectrum and the correlation function are power laws of the scale. As for the exponential correlation structure, classical measures are identical for both log-normal processes (figures 3.5(a,b,c)) but AMI and entropy rate evolve differently for each process ((figures 3.5(e,f)). Again, AMI always converges to zero for large τ but takes different values for $\tau = 1$. Again, entropy rate tends to the entropy for large τ while evolving differently from its value at $\tau = 1$, which depends noticeably on the process.

For stronger dependences (here the exponential decay, due to our choice of ($f_c = 0.1, \mathcal{H} = 0.7$), see section 3.2.3), AMI is larger and so are the differences between the processes. For scale invariant processes, AMI and entropy rate are again able to probe details of the dependence structure that originate from higher orders. In this case, the difference between both log-normal processes is smaller but still significant.

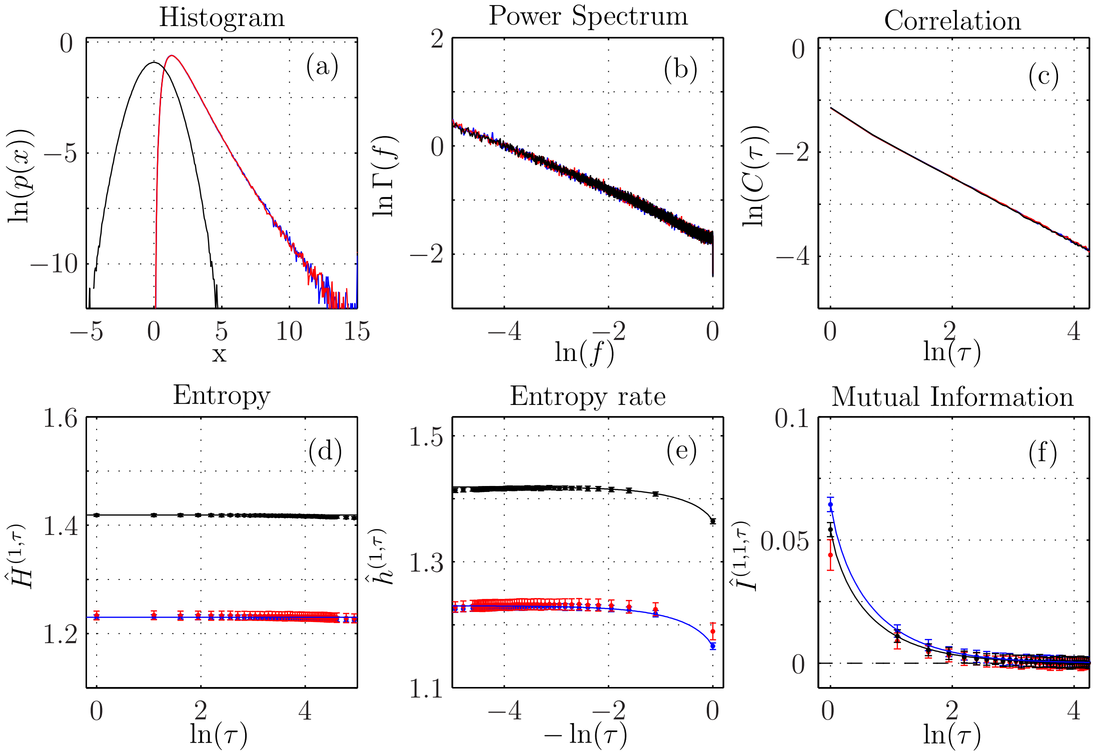

Figure 3.5: **Power law $\mathcal{H} = 0.7$.** Classical analysis: histogram (a), power spectrum (b) and correlation function (c). Information theory analysis for $m = 1$, $p = 1$: entropy (d), entropy rate (e) and auto mutual information (f). The color indicates the process: blue for standard log-normal, red for even log-normal and black for Gaussian. The continuous lines in (e,d,f) correspond to analytical expressions. We used $N = 2^{16}$ and $k = 5$.

Dependence structure beyond correlations

AMI measures statistical dependences between two consecutive points of a process, while correlation function measures is limited to only two-point correlations. Figures 3.4 and 3.5 show how information theory probes high order dependences, which allow to discriminate two very similar processes when performing an analysis across scales τ. We interpret AMI as a generalization of correlation function.

3.3.2 Dependence on embedding dimensions for fixed τ.

We now want to probe explicitly dependences between more than two points. To do so, we can increase the number of points that are involved in the definition of AMI, as this quantity provides a direct measure of dependences between its arguments, see eq.(1.17). A straightforward generalization is to increase any embedding dimension m or p appearing in eq.(1.26). This can be interpreted as probing how additional points increase the shared information, or in other words as measuring multiple-point dependences. We note that this shared information is a non decreasing function of m or p, *i.e.* it can be constant or increase. This increase of AMI hints at the evolution of multiple-point dependences when considering additional points, and we expect it to be larger for long-range than for short-range dependence processes.

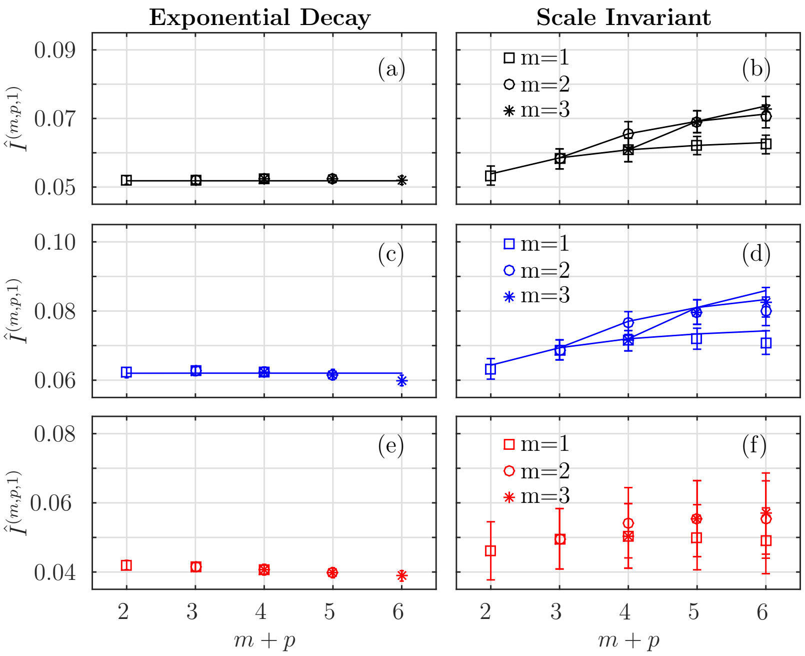

Figure 3.6: **Auto mutual information vs total embedding $m + p$.** Left column: exponential decay $f_c = 0.1$. Right column: power law $\mathcal{H} = 0.7$. The color indicates the process: blue for standard log-normal (c,d), red for even log-normal (e,f) and black for Gaussian (a,b). The continuous lines in (a-d) correspond to analytical expressions. We used $N = 2^{16}$, $k = 5$, $c(\tau) = 0.32$ ($\tau = 1$ for the scale invariant and $\tau = 11$ for the exponential decay).

For the exponential decay, the analytical curves (straight lines) in figure 3.6(a,c) show that AMI is independent of the total embedding dimension $m+p$. The estimations agree with the analytical curve, even if the existence of a bias unfortunately impairs an arbitrary increase of the embedding dimensions. This is known as the curse of dimensionality [21].

For the power law decay, figure 3.6(b,d,f) shows a small increase of analytical and estimated AMI when the total embedding dimension increases. Indeed, the estimations closely follow the analytical prediction and the increase is very similar for the three processes. In this case the bias is almost negligeable, nevertheless it slightly increases when increasing $m + p$.

A comparison between exponential and power law decay confirms that the increase of AMI with m and p indicates long range dependences. For long range dependence structures, any additional point contains new information. In that case, we can make another observation from figure 3.6: it is not equivalent to increase m or p. For a fixed $m+p$, AMI is larger when m and p are closer. This can be understood as a closer similarity between the two embedded vectors appearing as the arguments of the AMI in eq.(1.26).

All these observations indicate the more complex dependence structure of the long-range correlation processes.

3.4 Conclusion

2-point interactions: We proposed to explore the evolution of AMI and entropy rate on the scale parameter. AMI and entropy rate measure the strenght of the dependences between two points of a process. These quantities probe statistics of any order, and hence non-linear interactions, and as such appear as unambiguous generalizations of the correlation function and the power spectrum. As an illustration, we analyzed two log-normal processes with identical 1-point distribution and identical correlation function. We were able to discriminate these two processes, which appear as identical for classical analysis: so our measures enlight differences in high order dependences of the two processes. The behavior of our quantities along scales reveal the existence of stronger high-order dependences at smaller scale, which allows an easier separation of processes.

More than 2-point interactions: In addition, AMI generalizes easily to consider explicitly interactions between more than two points. We thus probe the complexity of the dependence structure above linear order, *i.e.*, the additional information measured by AMI when considering an extra point should indicate the existence and relevance of next-order interactions. As an example, comparing the effect of increasing the order of interactions for two dependence structures showed that while exponential decay — short range — dependences do not involve next order interactions, power law — long range — dependences do. This result could show that increasing the embedding allows to discriminate between short and long range dependences.

The same qualitative results were obtained when considering an exponential one-point distribution, which supports the validity of the methodology for any distribution and dependence structure. This generality should make the methodology very interesting to perform non-Gaussian processes characterization in several different applications.

CHAPTER 4

Information theory of non-stationary processes

In real life, a high number of processes such as: weather temperature, water ressources evolution, macroeconomic time series or Brownian motion, are non-stationnary. Consequently, the characterization and analysis of such processes become very important. However, because of the difficulty of studying non-stationnary processes, their characterization is usually done by studying their increments or time derivatives. During this chapter, we are interested in the direct analysis of non-stationnary processes.

Information theory was initially conceived to work with stationnary processes, where the measures have a clear interpretation in terms of information, see chapter 1. Nevertheless, two questions appear immediately:

- Can we apply information theory on non-stationnary processes?
- Do the interpretations on stationnary processes remain valid for non-stationnary ones?

For a non-stationnary time process $M(t)$ whose PDF and dependences evolve in time, its entropy $H(M(t))$ evolves with time.

We can measure the entropy of a process at a given time $H_t(M(t))$ by averaging over the realizations. In this framework, the interpretation of the entropy for stationnary processes remains valid [74]. In the same way, we can measure the dependences of the process at time t. Performing these analysis at fixed time for different times allows to characterize not only the distribution and dependences of the process at time t but also their evolution in time. Unfortunately, the analysis at a given time t requires having numerous realizations of the process to build its PDF without requiring ergodic hypothesis. This requirement is not easily resolve, specially when dealing with experimental signals.

In oder to avoid this requirement, we want to study a process $M(t)$ by analysing a experimental signal of length T. We can measure its entropy during the time window $[0, T]$, $H_T(M(t))$, but the interpretation of this entropy is not anymore the same. We interpret $H_T(M(t))$ as the complexity, or the total amount of information needed to characterize a temporal trajectory $M_t \; \forall \; t \in [0, T]$ of the process $M(t)$. In this way, we consider the temporal dimension as a new dimension of the process and then, the distribution of the process should take into account not only the distribution of the process in its phase space but also in the temporal dimension [211]. The main disadvantage of this viewpoint is the necessity of defining a window size T, restricting the evolution of the distribution in time.

Even if the natural environment of information theory is the analysis of stationnary processes, some scarce works have tried to perform an information theory analysis of non-stationnary processes [211, 165, 74]. In this chapter, the goal is to develop an information theory framework to analyse generic non-stationary processes. Once it is done we develop inside this framework the specific case of non-stationary self-similar motions with stationary increments, on which we are specially interested. Self similar power law decay motions are particularly interesting due to the high number of physical, biological or social systems (between others) that they can model. Once the theory is developed we verify it numerically and we use it to characterize some different motions.

4.1 Information Theory for Non-Stationary processes

4.1.1 General framework

Shannon entropy

We note $M(t)$ a continuous non-stationnary stochastic time process, and $p_{M_t}(x) = p(M(t) = x)$ its probability density function at time t. We can naturally build the time embedded process $M^{(m,\tau)}(t)$ [197].

The entropy of a non-stationnary time embedded process $M^{(m,\tau)}(t)$ at time t is defined as:

$$H(M^{(m,\tau)}(t)) = H_t(M^{(m,\tau)}) = -\int_{\mathbb{R}^m} p_{M_t^{(m,\tau)}}(\mathbf{x}) \log(p_{M_t^{(m,\tau)}}(\mathbf{x}))\mathrm{d}\mathbf{x} \tag{4.1}$$

Entropy rate

The entropy rate of order m at time t measures the variation of Shannon entropy between $M_t^{(m+1,\tau)}$ and $M_{t-\tau}^{(m,\tau)}$, two successive time-embedded versions of the non-stationnary stochastic process $M(t)$:

$$h^{(m,\tau)}(M(t)) = h_t^{(m,\tau)}(M) = H_t(M^{(m+1,\tau)}) - H_{t-\tau}(M^{(m,\tau)}). \tag{4.2}$$

Mutual Information

The mutual information measures the amount of information shared by two processes $L_t^{(m,\tau)}$ and $M_{t'}^{(p,\tau)}$ at time t and t', and is defined by :

$$I(M_{t'}^{(p,\tau)}, L_t^{(m,\tau)}) = \tag{4.3}$$

$$= \int_{\mathbb{R}^{m+p}} p_{M_{t'}^{(p,\tau)},L_t^{(m,\tau)}}(\mathbf{y},\mathbf{x}) \log\left(\frac{p_{M_{t'}^{(p,\tau)},L_t^{(m,\tau)}}(\mathbf{y},\mathbf{x})}{p_{M_{t'}^{(p,\tau)}}(\mathbf{y})p_{L_t^{(m,\tau)}}(\mathbf{x})}\right)\mathrm{d}\mathbf{y}\mathrm{d}\mathbf{x} \tag{4.4}$$

Auto Mutual Information

In order to analyse the nonlinear temporal dynamics of process M, we use auto mutual Information, defined in [77].

$$I^{(m,p,\tau)}(M(t)) = I_t^{(m,p,\tau)}(M) = I(M_t^{(p,\tau)}, M_{t-p\tau}^{(m,\tau)})\,. \tag{4.5}$$

$I_t^{(m,p,\tau)}(M)$ measures the shared information at time t between current p-points dynamics and past m-points dynamics.

Following eq.(1.25), the entropy rate can then be expressed as:

$$h_t^{(m,\tau)}(M) = H_t(M) - I_t^{(m,1,\tau)}(M)\,. \tag{4.6}$$

Relation between the entropy rate of a process and the entropy of its increments

For any process M, we have the following relation:

$$h^{(m,\tau)}(M(t)) = h^{(m-1,\tau)}(\delta_\tau M(t)) - I\left(M(t-\tau), \delta_\tau M(t) | \delta_\tau M^{(m-1,\tau)}(t)\right)\,. \tag{4.7}$$

In the special case $m = 1$, we have:

$$h^{(1,\tau)}(M(t)) = H(\delta_\tau M(t)) - I(M(t-\tau), \delta_\tau M(t))\,. \tag{4.8}$$

Proof. The demonstration does not involve any hypothesis on the process, which may not be scale invariant. The stationarity of the increments is not required.

Step 1: We say that:

$$\begin{aligned} H(M^{(m,\tau)}(t)) &\equiv H\left(M(t), M(t-\tau), M(t-2\tau), ..., M(t-(m-1)\tau)\right) \\ &= H\left(M(t), \delta_\tau M(t), \delta_\tau M(t-\tau), ..., \delta_\tau M(t-(m-2)\tau)\right)\,. \end{aligned} \tag{4.9}$$

This follows from expressing the movement M as a sum of its "recent" increments, and using chained conditioned probabilities.

Another way to derive eq.(4.9) is to transform the embedded vector $M^{(m,\tau)}(t)$, using a linear transformation:

$$M^{(m,\tau)} \mapsto M'^{(m,\tau)}(t) = Q_m.M^{(m,\tau)}(t),$$

where Q is the band matrix defined as:

$$Q_m \equiv \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 0 \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}\,. \tag{4.10}$$

Note that only the diagonal and the subdiagonal are populated. Note also that Q_m is not orthogonal, and not unitary, but its determinant is

$$\det(Q_m) = (-1)^{m-1}\,, \tag{4.11}$$

so $|\det(Q_m)| = 1$, which is enough to write $H(M'^{(m,\tau)}(t)) = H(M^{(m,\tau)}(t))$, which proves eq.(4.9).

Step 2: We combine the entropies according to eq.(4.2) and get:

$$h^{(m,\tau)}(M(t)) = H(\delta_\tau M^{(m,\tau)}(t)) - I\left(\left(M(t-\tau), \delta_\tau M^{(m-1,\tau)}(t-\tau)\right), \delta_\tau M^{(m,\tau)}(t)\right) \quad (4.12)$$

which can straightforwardly be rewritten as eq.(4.7).

Self Similar motions: Non-Stationary processes with Stationary increments

Along the above section we have setted a framework where applying information theory to non-stationnary processes. In this section we focus on self similar motions with stationnary increments and we deduce the behaviour of the above tools in function of the time t and the scale τ of analysis.

We suppose stationary increments

For stationary increments, the statistics of $\delta_\tau M(t)$ do not depend on time t and we can write $p_{\delta_\tau M(t)} = p_{\delta_\tau}$.

We suppose scale invariance

For scale-invariant processes there exists a value $\mathcal{H} \in \mathbb{R}$, $0 < \mathcal{H} < 1$ such that for all $a \in \mathbb{R}^+$, $x \in \mathbb{R}$ we have an equality in law between the statistics of $M(at)$ and $a^{\mathcal{H}} M(t)$.

$$p_{M_{at}}(x) = \frac{1}{a^{\mathcal{H}}} p_{M_t}(\frac{x}{a^{\mathcal{H}}}).$$

This is also true for the embedded signal.

$$p_{M_{at}^{(m,a\tau)}}(x) = \frac{1}{a^{m\mathcal{H}}} p_{M_t^{(m,\tau)}}\left(\frac{x}{a^{\mathcal{H}}}\right).$$

With the stationarity of the increments $p_{\delta_\tau M(t)} = p_{\delta_\tau}$ and setting $M(t=0) = 0$, $M(t)$ can be seen as an increment, $M(t) = M(t) - M(0) = \delta_t(t)$, which is itself scale invariant.

$$p_{\delta_{a\tau}}(x) = \frac{1}{a^{\mathcal{H}}} p_{\delta_\tau}\left(\frac{x}{a^{\mathcal{H}}}\right).$$

It is easy to show that a sum of scale invariant processes with parameter $\mathcal{H}$ is also scale invariant with the same parameter $\mathcal{H}$ [].

When there is scale invariance, the special time dependence can be recast to unit-time (t=1). This simply amounts to rescaling each dimension. This is easy to do with $M'^{(m,\tau)}(t)$. Unfortunately, all the results above do not give explicit formulas for the “constants” that appear.

Shannon entropy

The Shannon entropy defined in eq.(4.16), of the set of points $M^{(m,\tau)}(t)$ at time t is:

$$H(M^{(m,\tau)}(t)) = H_0(M^{(m,\tau/t)}(1)) + \mathcal{H} \ln t + (m-1)\mathcal{H} \ln \tau \quad (4.13)$$

Entropy rate

Once we know the Shannon entropy of a process M for any embedding m we can calculate the entropy rate using eq.(4.17) or eq.(4.19) arriving to:

$$h^{(m,\tau)}(M(t)) = h_0^{(m,\tau/t)}(M(1)) + \mathcal{H}\ln\tau \tag{4.14}$$

The "constants" H_0 or h_0 are given by the dependence structure of either the movement or its increments (*i.e.*, the noise) between times $(1, 1-\tau/t, ..., 1-(m-1)\tau/t)$. They have no reason to be independent on τ/t, but the dependence on τ/t is weak, hence the naming "constants".

4.1.2 T-length window framework

In order to work within this framework, the stationarity of the increments $\delta_\tau(M(t))$ of the analysed process is required. In this framework we study non-stationnary processes by fixing a window size T of analysis, and defining the PDF of apparition of a value $M^{(m,\tau)}$ during a time interval of length T.

$$\bar{p}_{M^{(m,\tau)},T}(\mathbf{x}) = \frac{1}{T}\int_{t_i}^{t_i+T} p_{M_t^{(m,\tau)}}(\mathbf{x})\mathrm{d}t \tag{4.15}$$

Shannon entropy

The entropy of this PDF is:

$$\bar{H}_T(M^{(m,\tau)}) = -\int_{\mathbb{R}^m} \bar{p}_{M^{(m,\tau)},T}(\mathbf{x}) \log(\bar{p}_{M^{(m,\tau)},T}(\mathbf{x}))\mathrm{d}\mathbf{x} \tag{4.16}$$

As such, it involves arbitrarily high order moments of the PDF $\bar{p}_{M^{(m,\tau)},T}(\mathbf{x})$, and therefore high order statistics. Nevertheless, as the entropy doesn't depend on the first moment of the PDF and the increments $\delta_\tau(M(t))$ are stationnary, $\bar{H}_T(M^{(m,\tau)})$ only depend on T and not on the initial time value t_i defining $\bar{p}_{M^{(m,\tau)},T}(\mathbf{x})$.

This entropy $\bar{H}_T(M^{(m,\tau)})$ represents the entropy of the set of values of the process $M^{(m,\tau)}(t)$ during a time interval of size T. As for stationnary processes, in the case of $m = 1$, there is no dependence of the entropy on τ, and then, from here till the end we will write $\bar{H}_T(M) = \bar{H}_T(M^{(1,\tau)})$.

Entropy rate

As for the Shannon entropy we can define the entropy rate during a time interval of size T:

$$\bar{h}_T^{(m,\tau)}(M) = \bar{H}_T(M^{(m+1,\tau)}) - \bar{H}_T(M^{(m,\tau)}). \tag{4.17}$$

Auto Mutual Information

We can define the auto mutual information during a time interval of size T as:

$$\bar{I}_T^{(m,p,\tau)}(M) = \bar{H}_T(M^{(m,\tau)}) + \bar{H}_T(M^{(p,\tau)}) - \bar{H}_T(M^{(m+p,\tau)}) \tag{4.18}$$

Following eq.(1.25), the entropy rate during a time interval of size T can then be expressed as:

$$\bar{h}_T^{(m,\tau)}(M) = \bar{H}_T(M) - \bar{I}_T^{(m,1,\tau)}(M)\,. \tag{4.19}$$

All the above time-averaged measures, $\bar{H}_T(M^{(m,\tau)})$, $\bar{I}_T^{(m,p,\tau)}(M)$ and $\bar{h}_T^{(m,\tau)}(M)$ depend on the window size T and the scale τ.

Analytical results for fractional Brownian motion

Fractional Brownian motion (hereafter fBm) is a continuous-time random process proposed by Mandelbrot and Van Ness [137] in 1968, which quickly became a major tool in various fields where concepts of self-similarity and long-range dependence are relevant. Although this signal is non stationary, its increments are Gaussian and stationary. The non-stationary covariance structure $\mathbb{E}\{B(t)B(t+\tau)\} = \sigma_0^2 c(t,\tau)$ is given by

$$\sigma_0^2 c(t,\tau) = \frac{\sigma_0^2}{2}\left[t^{2\mathcal{H}} + (t+\tau)^{2\mathcal{H}} - |t-(t+\tau)|^{2\mathcal{H}}\right] \tag{4.20}$$

The pre-factor σ_0 is a normalization constant.

The fBm has Gaussian statistics, we can therefore write its entropy following eq.(1.36). In addition, the fBm is non stationary, but if we consider a signal of finite temporal extension $T \gg \tau$ and Hölder exponent $\mathcal{H}$, if we note $H^0_{\mathrm{FBM}} \equiv \frac{1}{2}\ln\left(2\pi e\sigma_0^2\right)$, we find:

Entropy

$$\bar{H}_T^{(1,\tau)}(B) = H^0_{\mathrm{FBM}} + \mathcal{H}\ln T \tag{4.21}$$

Auto-mutual information

$$\bar{I}_T^{(1,1,\tau)}(B) = -\mathcal{H}\ln\left(\frac{\tau}{T}\right) + \mathcal{C}\left(\frac{\tau}{T}\right) \tag{4.22}$$

with

$$\mathcal{C}\left(\frac{\tau}{T}\right) = \frac{1}{2}\ln\left(\frac{\frac{\tau}{2T} + \frac{1}{2\mathcal{H}+1}}{\left(\frac{\tau}{T} + \frac{1}{2\mathcal{H}+1}\right)}\right)$$

which has been numerically shown to be very few dependent on the embedding.

Entropy rate

$$\bar{h}_T^{(m,\tau)}(B) \simeq H^0_{\mathrm{FBM}} + \mathcal{H}\ln\tau - \mathcal{C}\left(\frac{\tau}{T}\right) \tag{4.23}$$

which is independent of T up to corrections in τ/T, which are negligible if T is large enough compared to the range of τ used. The entropy rate of the fBm is therefore linear in $\ln(\tau)$, with a constant slope $\mathcal{H}$, independent of the temporal extension. This analitycal result has been numerically probed to be very few dependent of the embedding. We can observe how the main dependance on $\ln(T)$, appearing in the expressions of $\bar{H}_T(B)$ and $\bar{I}_T^{(m,p,\tau)}(B)$ dissapears in the entropy rate $\bar{h}_T^{(m,\tau)}(B)$, just because of a compensation of the dependances in the entropy and the auto mutual information.

Relation between the entropy rate of fBm and the entropy of its increments

We can write the entropy of the increments of a fBm, which are stationnary and Gaussian, as:

$$H(\delta_\tau B) = H^0_{\mathrm{FBM}} + \mathcal{H}\ln(\tau) \tag{4.24}$$

Looking at eq.(4.23) and eq.(4.24), we can straightforwardly write:

$$\bar{h}_T^{(m,\tau)}(B) \simeq H(\delta_\tau B) - \mathcal{C}\left(\frac{\tau}{T_{\max}}\right) \tag{4.25}$$

To obtain these relations for a fBm, we have calculated the different coefficients of the covariance matrix of eq.(1.36). Then, we can calculate the determinant of this matrix and the mean over time of the determinant between $t_{min} = 0$ and $t_{max} = T$. With the determinant and eq.(1.36), we can obtain the different entropies involved in the different equations of $\bar{H}_T$, $\bar{I}_T$ and $\bar{h}_T$ respectively. It's important to suppose that $T \gg \tau$.

4.2 Data and procedure

In order to build different self-similar non-stationary processes $M(t)$ with stationary increments $x(t)$, we use eq.(4.26) with the three different processes defined in section 3.1 as increments.

$$M(t) = \sum_{i=1}^{t} x(i) \tag{4.26}$$

Using the fractional Gaussian noise of section 3.1, we generate a fractional Brownian motion (fBm), which is the only Gaussian self-similar process with stationary increments [137]. Using the two log-normal noises of section 3.1, we generate two different self-similar motions, that we call "standard" and "even" log-normal motions depending on the respective transformation used to generate them (section 3.1).

We characterize each motion by computing its entropy, auto mutual information and entropy rate, using the Kozachenko-Leonenko and Kraskov-Stögbauer-Grassberger algorithms for entropy and mutual information (see chapter 2).

4.3 Numerical analysis: T-length window framework

4.3.1 Characterization of the estimates

Standard deviation characterization

We present the variation of the standard deviation (std) of $\bar{h}_T^{(m,\tau)}$ when N is varied, and in particular we compare it to the std of the two separate terms $\bar{H}_T$ and $\bar{I}_T^{(m,\tau)}$ in eq.(4.19). We observe in figure 4.1 that the std of each contribution separately is much higher than the std of $\bar{h}_T^{(m,\tau)}$: There is a partial cancellation of the standard deviations in the estimation of the entropy and the mutual information, respectively, such that the standard deviation of $\bar{h}_T^{(m,\tau)}$ is much smaller.

Bias characterization

We have tried to quantify how the bias depends on both N and k and observed that the convergence seems to be in $k/N^{\frac{1}{m+1}}$ — in agreement with the results of [116, 69] for the mutual information estimator that we use — as can be seen in figure 4.2 where we have gathered all possible values of the couples $(\log_2(N), k) \in [9, ..., 17] \times [4, ..., 18]$.

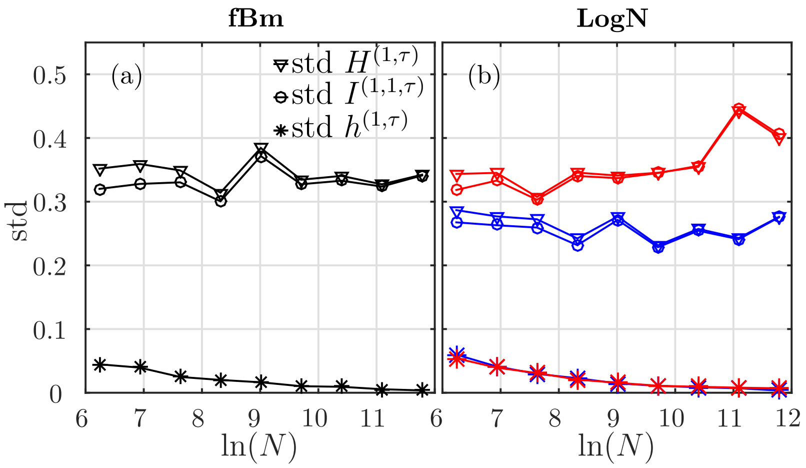

Figure 4.1: Standard deviations of $\bar{H}_T^{(1,\tau)}$ (triangles), $\bar{I}_T^{(1,1,\tau)}$ (circles) and $\bar{h}_T^{(1,\tau)}$ (stars), as functions of N, in a) for the fBm and in b) for the two log-normal processes. In blue the standard log-normal process; in red the even log-normal process.

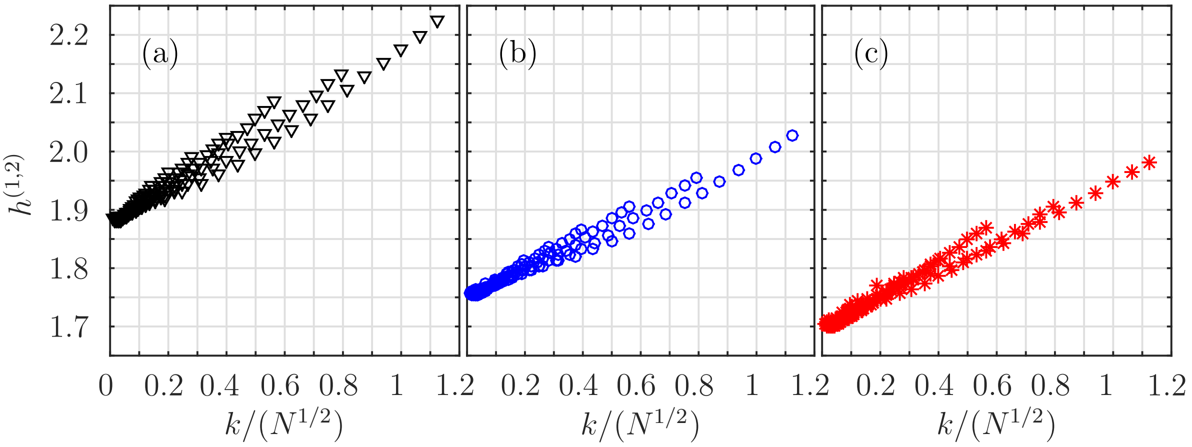

Figure 4.2: Dependence of $\bar{h}_T^{(1,\tau=2)}$ on $k/N^{\frac{1}{2}}$ for the fBm (a) and for the standard (b) and even (c) log-normal processes.

Conclusions

The standard deviation of $\bar{h}_T^{(m,\tau)}$ can be made arbitrarily small, especially compared to the ones of the entropy or the auto mutual information. Both the bias and the standard deviation of our estimator increase when k is increased or N is reduced. We choose $k = 5$ in the remainder of this chapter, and use datasets of $N = 2^{16}$ points.

4.3.2 Dependence of $\bar{h}_T^{(m,\tau)}$ and $\bar{H}_T^{(m,\tau)}$ on T and the scale τ

In this section, we want to numerically verify if the analytical behaviors of entropy and entropy rate obtained in the general framework for self similar motions remain valids in the T-length window framework. To do that, we compute numerically the entropy and entropy rate of a fBm for different window sizes T, and independently, for different scales

τ.

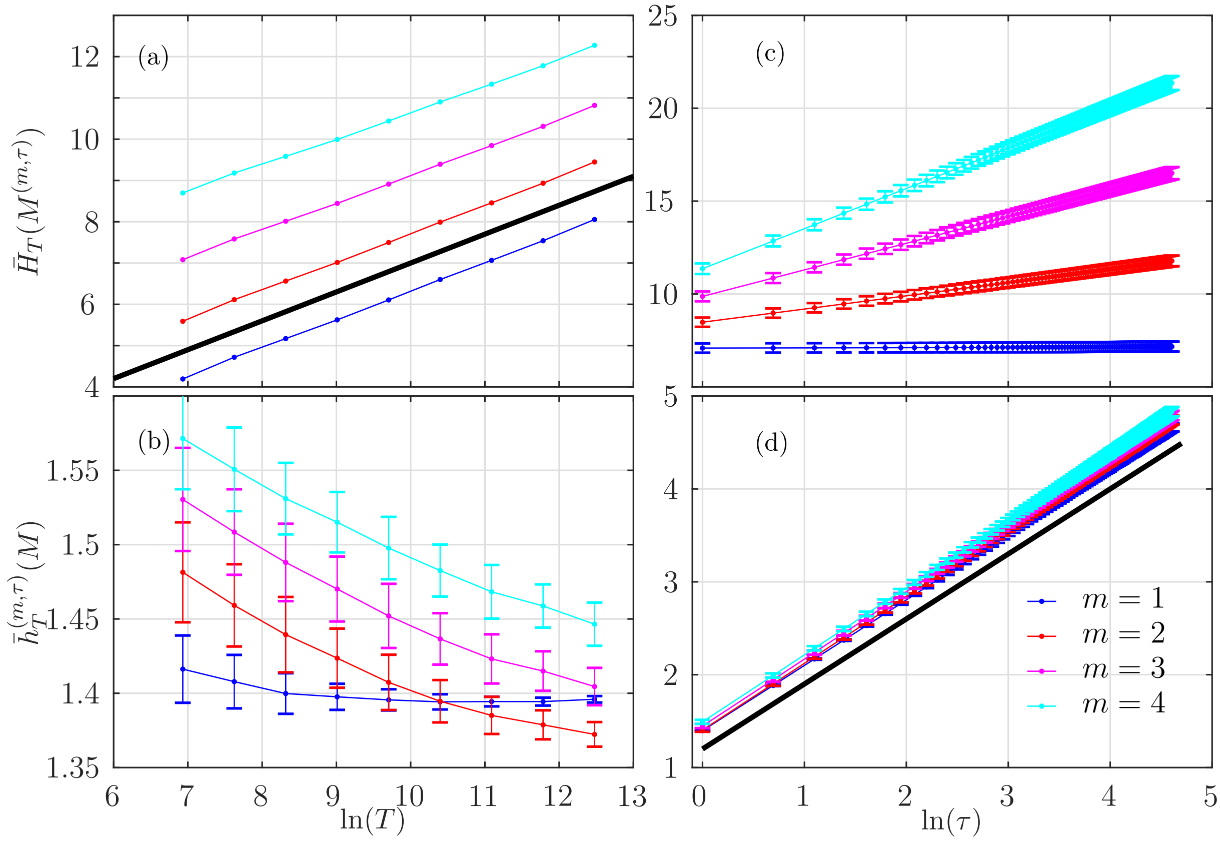

Figure 4.3: a) Entropy $\bar{H}_T(M^{(m,\tau)})$ and b) entropy rate $\bar{h}_T^{(m,\tau)}(M)$ in function of the logarithm of the window size $\ln(T)$ of the analysis for a fixed scale τ. c) Entropy $\bar{H}_T(M^{(m,\tau)})$ and d) entropy rate $\bar{h}_T^{(m,\tau)}(M)$ in function of the logarithm of the scale of analysis $\ln(\tau)$ for a fixed $T = 2^{16}$. All the estimators were applied on a fBm with $\mathcal{H} = 0.7$. In a) and d) the black line represents a straight line of slope $\mathcal{H} = 0.7$. In c) the different measures follow a linear behaviour of slope $(m-1)\mathcal{H}$.

The behavior of $\bar{H}_T^{(m,\tau)}$ in figure 4.3 (a) agrees with the behavior of $H_t^{(m,\tau)}$ in eq.(4.13). The Shannon entropy $\bar{H}_T^{(m,\tau)}$ at a given scale, in function of T, shows a slope equal to $\mathcal{H}$ independently of the embedding. On the other hand, in figure 4.3 (b), the entropy rate $\bar{h}_T^{(m,\tau)}$ shows an almost constant behaviour for any embedding, in agreement with eq.(4.14) for $h_t^{(m,\tau)}$. The small dependence of the entropy rate on T can be due to bias effects, whose importance increases with the embedding, but it can also be due to a different behavior of the entropy rate in the different frameworks when the embedding increases.

Figures 4.3 c) and d) show a behaviour of $\bar{H}_T^{(m,\tau)}$ and $\bar{h}_T^{(m,\tau)}$ in function of τ according to eq.(4.13) and eq.(4.14) for $H_t^{(m,\tau)}$ and $h_t^{(m,\tau)}$. Shannon entropy $\bar{H}_T^{(m,\tau)}$ behaves as $(m-1)\mathcal{H}\ln(\tau)$ while entropy rate behaves as $\mathcal{H}\ln(\tau)$ for any embedding m. Figure 4.4 shows the differences between the computed entropy rates $\bar{h}_T^{(m,\tau)}$ in function of τ and their main supposed behaviours $\mathcal{H}\ln(\tau)$ for different embedding. When the embedding dimension increases, the difference between the behaviour of $\bar{h}_T^{(m,\tau)}$ and $\mathcal{H}\ln(\tau)$ increases.

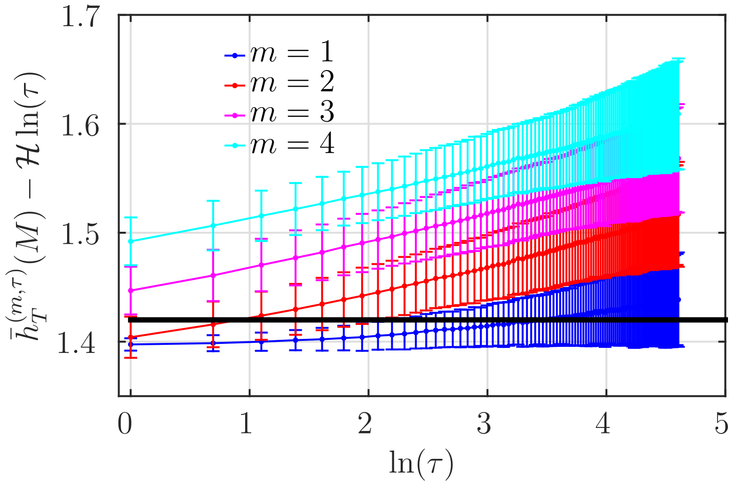

Figure 4.4: Entropy rate $\bar{h}_T^{(m,\tau)}(M) - \mathcal{H}\ln(\tau)$ for a fBm with $\mathcal{H} = 0.7$. The embedding dimension m varies from 1 to 4. The time window size is fixed to $T = 2^{16}$.

This result is in agreement with figure 4.3 b), and probably due to a different behavior of the entropy rate in the different frameworks when the embedding increases.

We find the same behaviours of $\bar{h}_T^{(m,\tau)}$ and $\bar{H}_T^{(m,\tau)}$ in function of T and τ for the different self-similar motion that we studied.

4.3.3 Relation between the entropy rate of the motion and the entropy of its increments

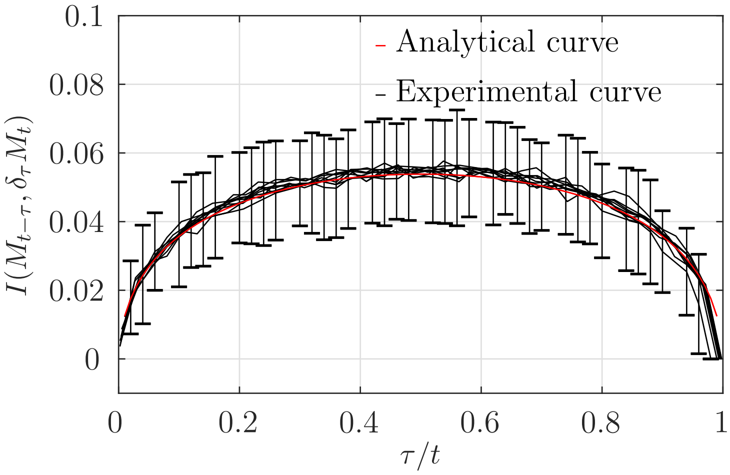

Figure 4.5: Mutual information between the motion and its increments $I(M_{t-\tau}, \delta_\tau M_t)$ in function of the rate τ/t. The motion is a fBm with $\mathcal{H} = 0.7$.

Returning to the general framework, figure 4.5 shows the behaviour of $I(M_{t-\tau}, \delta_\tau M_t)$ in function of τ/t which varies from 0 in both limits to a maximum value of 0.08 when $\tau/t = 0.5$. In figure 4.5, the red line correspond to the analytical prediction, while the black lines correspond to multiple curves for different t and different τ. Consequently, the only important parameter defining the magnitude of $I(M_{t-\tau}, \delta_\tau M_t)$ is τ/t and the importance of this term in eq.(4.8) is very few for any rate τ/t.

Figure 4.6 a) shows that eq.(4.8) is valid in the T-window framework for scale invariant motions with stationnary increments. The black line indicates the entropy of the different increments of size τ of a fBm M. The blue line is the entropy rate of the fBm at scale τ. The red line is the entropy rate of the fBm computed as the right hand side of eq.(4.8). Finally, the magenta line is the mutual information between the increments of the motion and the motion itself. This mutual information term is negligeable for $\tau \ll T$. In addition, the behaviour of $\bar{I}_T(M_{t-\tau}, \delta_\tau M_t)$ in τ/T is expected to be similar to that of $I(M_{t-\tau}, \delta_\tau M_t)$ (figure 4.6), and numerical computations show that both $I(M_{t-\tau}, \delta_\tau M_t)$ and $\bar{I}_T(M_{t-\tau}, \delta_\tau M_t)$ are of the same maginitude for $\tau/t = \tau/T$.

Figure 4.6 b) shows how the entropy rate of the fBm match almost perfectly (in the errorbars) with the entropy of the increments of the fBm.

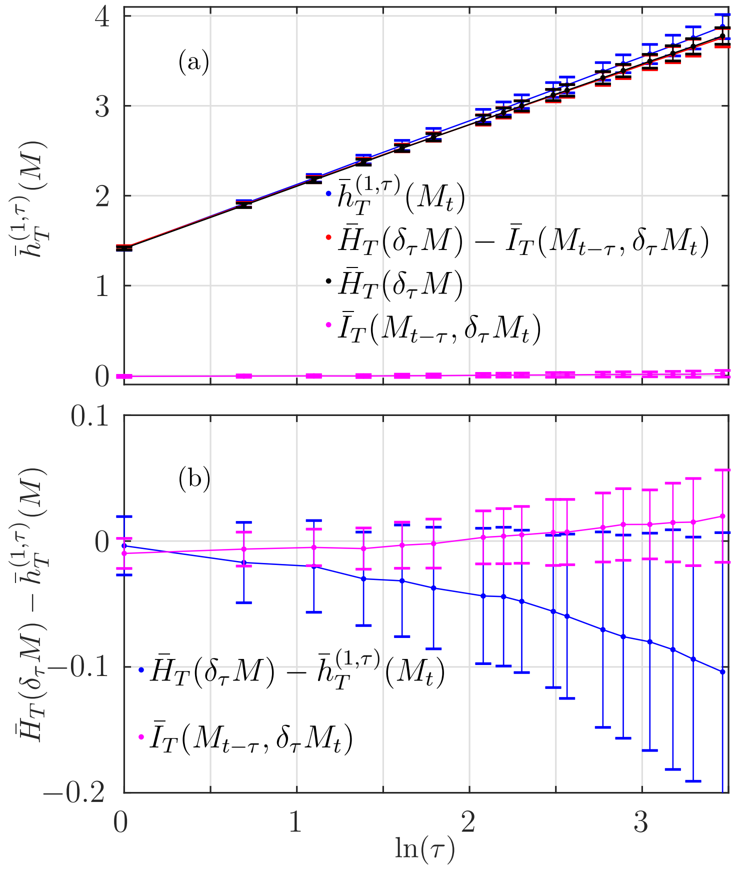

Figure 4.6: a) Entropy rate of the motion $\bar{h}_T^{(1,\tau)}(M_t)$, entropy of the increments $\bar{H}_T(\delta_\tau M)$, mutual information between the motion and the increments $\bar{I}_T(M_{t-\tau}, \delta_\tau M_t)$, and the difference between the two $\bar{H}_T(\delta_\tau M) - \bar{I}_T(M_{t-\tau}, \delta_\tau M_t)$. b) Entropy rate of the motion $\bar{h}_T^{(1,\tau)}(M_t)$ minus entropy of the increments $\bar{H}_T(\delta_\tau M)$, and mutual information between the motion and the increments. The motion is a fBm with $\mathcal{H} = 0.7$. The time window size is $T = 1024$.

4.4 Application: Discrimination between processes

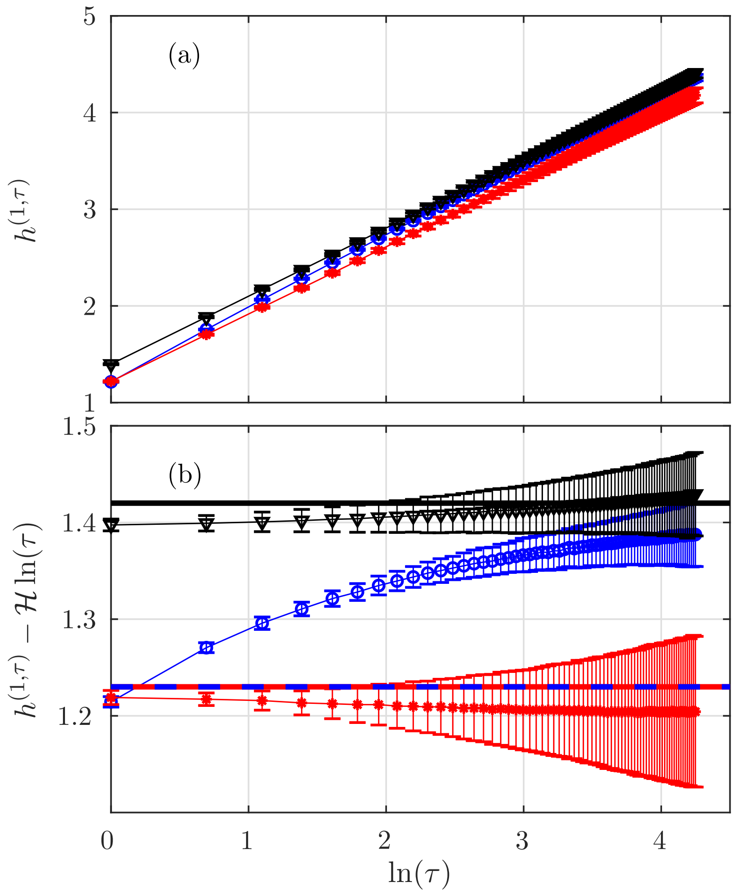

Figure 4.7: $\bar{h}_T^{(1,\tau)}$ value for an standard (blue) and even (red) log normal signals and a fBm (black) in function of the logarithm of the τ. Number of neighbors $k = 5$ and data size $N = 2^{16}$. The standard deviation is obtained with 100 realisations. The straight lines in (b) indicate the theoretical values of the entropy of the processes.

We now study the dependence of $\bar{h}_T^{(m,\tau)}$ on the time lag τ for the three motions presented in section 4.2. Varying τ is analogous to varying the time scale at which the signal is sampled [80], and then, we can analyse the self-similarity properties of the processes and look for fine differences between them.

Results for fixed size, $N = 2^{16}$, $k = 5$ neighbors and embedding dimension $m = 1$, are reported in figure 4.7 (a). The behaviors accross scales of the three motions are linear in $\ln(\tau)$. To get a more precise insight, we substract, for the three motions, the dependence $\mathcal{H}\ln(\tau)$ obtained numerically for self-similar processes in the precedent section. Results are presented in figure 4.7 (b).

We obtain for the fBm a constant value very close to the theoretical value $\frac{1}{2}\log(2\pi e) \simeq 1.42$. The red, even log-normal motion does not depend on τ anymore. So, just like in the fBm case, the slope is exactly $\mathcal{H}$, as expected for a self-similar process. On the contrary, the blue, standard log-normal motion still depends slightly on τ, sharing the same value with the other log-normal process for $\tau = 1$, and the same value with the fBm for large τ.

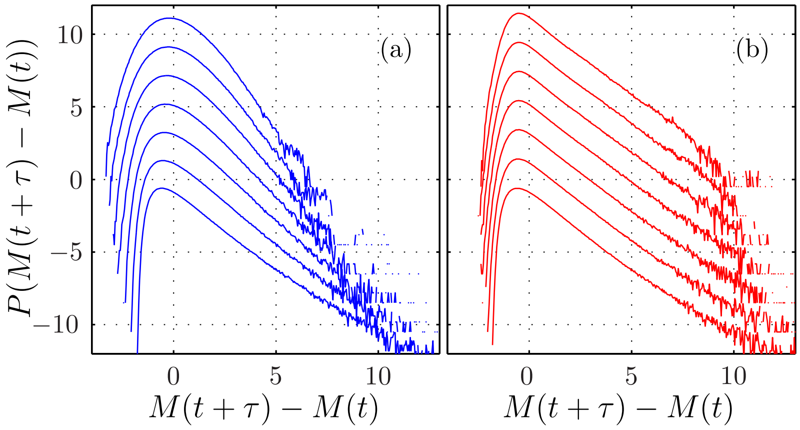

Figure 4.8: PDF of the normalized increments with $\tau = 2^j$, from $j = 0$ (bottom) up to $j = 6$ (up) of (a) Hermitian and (b) even Hermitian log-normal motions. Curves have been arbitrarily shifted on the Y-axis for clarity.

We can state that the fBm and the even log-normal (red) motions are self-similar with Holder exponent $\mathcal{H} = 0.7$. Their values at $\tau = 1$ are directly related to the probability density function of the noises defining the motions, see eq.(4.8) for the *general framework*. We can state also that the standard log normal process (blue) is not self-similar with an evolution of its pdf across scales.

The different behaviors of $\bar{h}_T^{(m,\tau)}$ observed for the two log-normal motions should be therefore related to different evolutions of the pdfs of the increments when τ is increased. To check this, we plot in figure 4.8 the normalized pdfs (std=1) of the increments $\delta_\tau M(t) = M(t) - M(t-\tau)$, for various values of τ.

While the pdf of the increments of the, even log-normal process (red) does not change when τ is increased, this is not the case for the increments of the standard log-normal motion (blue): the shape of the pdf changes from log-normal for small τ to Gaussian for large τ. Note that for $\tau = 1$, the pdf of the increments of the motion is nothing but the pdf of the noise, which is the same for both log-normal processes. Indeed, the entropy rate has very similar values for the two log-normal motions when $\tau = 1$. While for the red, even log-normal process, $\bar{h}_T^{(m,\tau)}$ remains constant across scales τ, this is not the case for the blue, standard process, for which the entropy rate converges to the value obtained for the Gaussian process, see figure 4.7 (b).

4.5 Discussion and Conclusions

In the first part of the chapter, we develop analytically the expression of entropy rate $h_t^{(m,\tau)}(M)$ at a given time t and scale τ for a *generic non-stationary* process $M(t)$, and we obtain the analytical relationship between this entropy rate and the entropy of its increments $H(\delta_\tau M)$, which are very strongly connected (eq.(4.8)). This connection allows us to argue that the self-similarity properties of any motion can be studied by measuring its entropy rate across scales. In the case of monofractal motions, we show analytically that the entropy rate across scales gives a measure of the Hurst exponent. However, we don't show the specific analytical expressions for multifractal processes.

In the second part of the chapter, we propose a new framework where analysing non-stationnary processes using information theory. In this framework, we analyse time windows of size T of the processes. In order to do that, we define the probability density function of a process in a time window of size T, as the mean of the different PDF's of the process at the different times inside the window. Then, the information theory quantities are computed using this averaged PDF. The interpretation of these information theory quantities is not anymore the same as in the general framework. We interpret $\bar{H}_T(M(t))$ as the complexity, or the total amount of information needed to characterize a temporal trajectory $M_t \; \forall \; t \in [0, T]$ of a process $M(t)$.

In this framework, we develop the analytical expressions of entropy, auto mutual information and entropy rate of a fractional Brownian motion, for which the covariance is known. We also show numerically the behaviours of entropy, and entropy rate in function of T and τ for self similar processes and conclude that these behaviours are in agreement with the ones predicted for $H_t^{(m,\tau)}(M)$ and $h_t^{(m,\tau)}(M)$. We show numerically the validity in this framework of the analytical relationship between the entropy rate of a process and the entropy of its increments, obtained in the generic framework. The validity, in this new framework, of this analytical relationship is shown for a fBm. However, it is valid for any process, either monofractal or multifractal. This methodology, based on the measure of the entropy rate computed in a window size of time T, is strongly related to a new methodology that we develop in chapter 7. This new methodology works on the increments of the analysed processes, and both are able to describe the self-similarity properties of any process.

Finally, we use $\bar{h}_T^{(m,\tau)}(M)$ to analyse three different *a priori* self similar non-stationnary processes, and obtain fine differences in their self similarity properties (figure 4.7). We observe that while the fBm and the even log-normal motions are self-similar with $\bar{h}_T^{(m,\tau)}(M) = H^0 + \mathcal{H} \ln(\tau)$, the standard log-normal motion presents a non-linear entropy rate across scales. This application allows to corroborate the close relationship between the entropy rate across scales of a process and the entropy, and consequently the PDF, of its increments. In this application, the three analysed motions have been built with the noises of chapter 3 as smaller increment. Comparing the results of chapter 3 with the results obtained in section 4.4, we can conclude that analysing the motion is not equivalent to analysing the generator noise. More precisely, the entropy rate of the motions characterize their self-similarity properties and the entropy rate of the noise doesn't.

Part II

Turbulence

CHAPTER 5

Fully developed turbulence

Two different descriptions of a fluid flow are usually adopted to study fluid mechanics: the Eulerian framework and the Lagrangian one. In both, one supposes the fluid to be composed of particles of size larger than molecular scales but smaller than dissipative ones (defined below).

- In the Eulerian framework, the observables of the fluid particles: velocity, pressure, temperature, *etc* depend on space **r** and time t.
- In the Lagrangian framework, one follows the particles and studies the observables of each particle over time.

During my PhD, I focused on Eulerian turbulence and so, in the following, we develop the theory using the Eulerian framework. In this framework a flow of a fluid of viscosity ν is characterised by:

- Conservation of momentum or Navier-Stokes equations:

$$\rho \left[\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} \right] = -\nabla p + \nabla \cdot \boldsymbol{\tau} \tag{5.1}$$

  where p is the pressure, ρ the density of the fluid, $\mathbf{v}$ the velocity and $\boldsymbol{\tau}$ is the viscous stress tensor.

- Conservation of mass or continuity equation:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0 \tag{5.2}$$

- Conservation of energy:

$$\rho c_p \left[\frac{\partial T}{\partial t} + (\mathbf{v} \cdot \nabla)T \right] = k\nabla^2 T + \phi \tag{5.3}$$

where c_p is the specific heat of the fluid, ϕ is the dissipation function representing the work done against viscous forces, and T is the temperature.

In the following we develop the theory only for incompressible and isothermal flows. For an incompressible flow ($\rho \equiv$ cte) with constant temperature ($T \equiv$ cte):

- Conservation of momentum or Navier-Stokes equations becomes:

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}.\nabla)\mathbf{v} = -\frac{\nabla p}{\rho} + \nu \nabla^2 \mathbf{v} \tag{5.4}$$

- Conservation of mass or continuity equation becomes:

$$\nabla.\mathbf{v} = 0 \tag{5.5}$$

- Conservation of energy becomes:

$$\phi = 0 \tag{5.6}$$

Eq.(5.4) describes the dynamics of the velocity $\mathbf{v}(\mathbf{r}, t)$ of an incompressible fluid of viscosity ν and density ρ. It shows the competition between different forces acting on the fluid:

- $\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}.\nabla)\mathbf{v}$ are the inertial terms. The first one indicates the variation in time of the velocity, while the second one, which is non-linear, describes advection.
- $-\frac{\nabla p}{\rho}$ is the internal force.
- $\nu \nabla^2 \mathbf{v}$ is the diffusion term due to the viscosity of the fluid.

The competition between the non-linear advective and the linear diffusion terms defines the Reynolds number:

$$\mathbb{R} = \frac{(\mathbf{v}.\nabla)\mathbf{v}}{\nu \nabla^2 \mathbf{v}} \tag{5.7}$$

The Reynolds number, initially introduced by Stokes and later generalized by Reynolds, characterises the behaviour of the flow. A low Reynolds number corresponds to a laminar flow and a high Reynolds number indicates a turbulent flow. Fully developped turbulence is defined in the limit of infinite Reynolds number. With dimensional analysis one can approximate the Reynolds number as [66]:

$$\mathbb{R} \approx \frac{\sigma L}{\nu} \tag{5.8}$$

where the characteristic velocity σ is the r.m.s velocity fluctuation of the fluid at the characteristic scale L of the flow, defined as $\sigma = \sqrt{\langle (\delta_L v(x))^2 \rangle}$.

In 1921 Richardson presented an account of turbulence, which still prevails. In Richardson's description, turbulence is presented as a hierarchy of whirls of different sizes, with an energy cascade from larger eddies down to smaller ones [167], see figure 5.1.

This description defines inherently two important domains: the integral (large scales) domain where energy is injected and the dissipative (small scales) domain where energy is mostly dissipated. So, from Richardson's picture, we note a very important characteristic of turbulence: it is multi-scale, see figure 5.2.

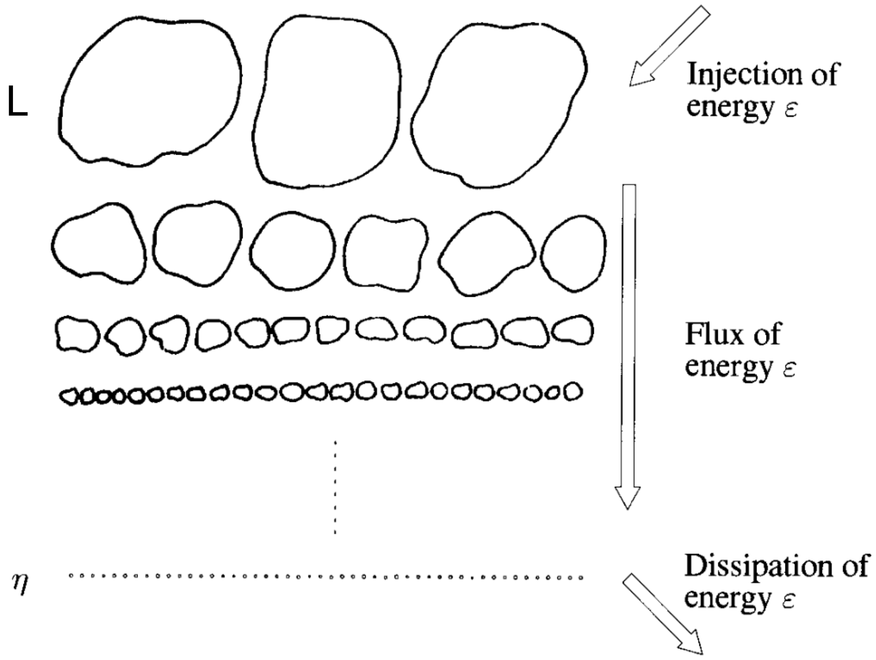
L
Injection of energy ε
Flux of energy ε
η
Dissipation of energy ε

Figure 5.1: Richardson cascade description of turbulence [66].

Inspired by Richardson's vision of turbulence, A.N. Kolmogorov built a revolutionary theory of turbulence, Kolmogorov 1941 theory, abbreviated as K41.

5.1 Kolmogorov's theory

Influenced by Richardson's description of turbulence, Kolmogorov differentiated three different domains of scales in turbulence:

- **The integral region:** large scales, equal to or larger than the integral scale L. The integral domain is defined as the domain where energy is injected in the system.
- **The inertial range:** scales smaller than L but larger than the Kolmogorov dissipative scale η_K. In the inertial domain, energy cascades from large to small scales.
- **The dissipative domain:** scales smaller than η_K. In the dissipative domain, energy is dissipated.

The inertial region is characterized by a clear predominance, in the Navier-Stokes equation, of the advective term over the diffusive one, while in the dissipative domain both terms are equally important.

In order to develop his theory, Kolmogorov made two hypotheses [108]:

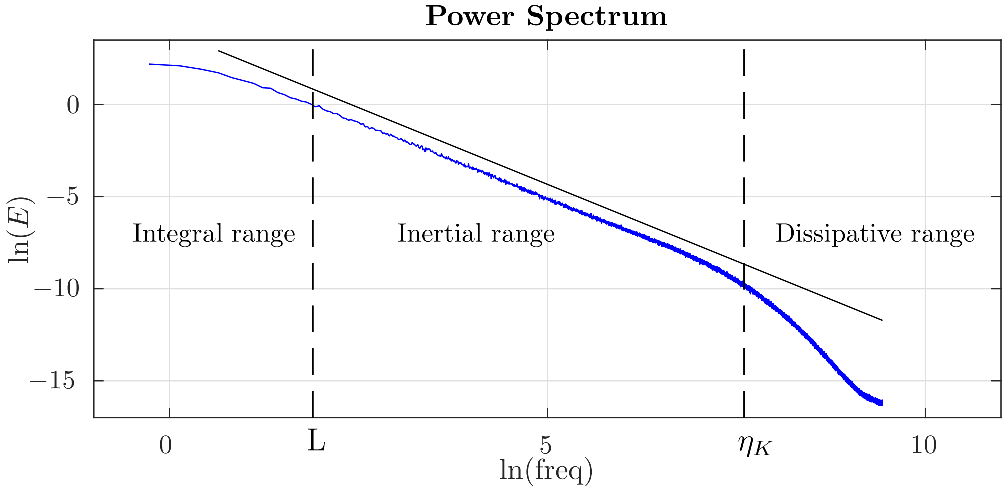

Figure 5.2: P.S.D. of turbulence.

- **The first similarity hypothesis:** "At very high, but not infinite, Reynolds number, all the small scale statistical properties are uniquely and universally determined by the scale l, the mean energy dissipation rate $\langle \epsilon \rangle$ and the viscosity ν."
- **The second similarity hypothesis:** "In the limit of infinite Reynolds number, all the small-scale statistical properties are uniquelly and universally determined by the scale l and the mean energy dissipation rate $\langle \epsilon \rangle$."

which we can summarize as: for Reynolds number going to infinity, the small scales (scales smaller than the integral scale L) of turbulence are statistically isotropic and independent of the large scales [66, 37].

The mean energy dissipation rate is defined as [108]:

$$\langle \epsilon \rangle = \frac{1}{2}\nu \sum_{i,j} \left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right)^2 \tag{5.9}$$

Using only his first similarity hypothesis, Kolmogorov shows in [108] that the dissipation scale η_K, below which diffusion is relevant, can be expressed as:

$$\eta_K \approx \frac{\nu^{3/4}}{\langle \epsilon \rangle^{1/4}} \tag{5.10}$$

Using the two universality assumptions and phenomenological analysis [66], Kolmogorov statistically characterized the behaviour of the velocity increments in the inertial region [109]:

$$\delta_l v(x) = v_x(x + l) - v_x(x) \approx \epsilon^{1/3} l^{1/3} \tag{5.11}$$

Eq.(5.11) leads to specific scaling behaviours of the moments of the velocity increments, called structure functions [109]:

$$S_p(l) \equiv \langle (\delta_l v)^p \rangle = C_p(l \langle \epsilon \rangle)^{p/3} \tag{5.12}$$

where p is the order of the structure function and C_p are the universality constants that only depend on the order of the structure function.

From the scaling behaviour of the second order structure function $S_2(l)$ one obtains the spectral energy distribution of Eulerian fully developed turbulence (see figure 5.2):

$$E(k) = \left\langle \|v_x(k)\|^2 \right\rangle \propto k^{-5/3} \tag{5.13}$$

where $l = 1/k$.

The relations above (eq.(5.11), eq.(5.12) and eq.(5.13)) are valid for any scale l in the inertial region, where there is neither direct energy injection nor direct energy dissipation, but a flux of energy Π from integral scale L to dissipative scale η_K. Consequently with the eq.(5.11), the flux of energy Π should be independent of the scale and equal to the mean energy dissipation rate:

$$\Pi \approx \frac{\langle (\delta_l v(x))^3 \rangle}{l} \approx \langle \epsilon \rangle \tag{5.14}$$

In particular, very close to the integral scale L we arrive to

$$\langle \epsilon \rangle \approx \frac{\langle (\delta_L v(x))^3 \rangle}{L}\,, \tag{5.15}$$

showing the finiteness of the mean energy dissipation rate.

Four-fifths law

The *a priori* phenomenological definition of the mean energy dissipation rate of eq.(5.15) was directly derived by Kolmogorov from the Navier-Stokes equations. Kolmogorov obtained an *exact* relation for the third order structure function $S_3(l)$ [107], which should be respected for any model of turbulence:

In the limit of infinite Reynolds number, the third order structure function of homogeneous isotropic turbulence, evaluated for increments l small compared to the integral scale, is given in terms of the mean energy dissipation rate by

$$S_3(l) = -\frac{4}{5} \langle \epsilon \rangle \, l \tag{5.16}$$

The four-fifths law shows the existence of an energy cascade in fully developed turbulence from large scales to smaller ones.

From the estimation of the Kolmogorov scale in eq.(5.10) and the approximation of the mean energy dissipation rate of eq.(5.15) it is possible to show that for large enough Reynolds number the inertial range is sufficiently large [125, 112].

$$\frac{\eta_K}{L} \approx \mathbb{R}^{-3/4} \tag{5.17}$$

This relation shows that for Reynolds number going to infinity the distance between L and η_K becomes infinite (infinite inertial range) and one is in the framework of fully developped turbulence.

In order to avoid anisotropic effects of integral scales, L, in the definition of the Reynolds number, it is useful to define the Reynolds number at a smaller scale, called the Taylor scale λ. This Taylor scale is defined through the relation:

$$\frac{1}{\lambda^2} = \frac{\langle (v - \langle v \rangle)^2 \rangle}{\langle (\partial_x v(x))^2 \rangle} \tag{5.18}$$

The associated Reynolds number is defined as:

$$\mathbb{R}_\lambda \approx \frac{\sigma_\lambda \lambda}{\nu} \tag{5.19}$$

where σ_λ is the r.m.s velocity fluctuation of the fluid at the scale λ.

Using a phenomenological approach and eq.(5.15) (result obatined directly from the finiteness of the energy dissipation) one can shows [66]:

$$\mathbb{R}_\lambda \approx \sqrt{\mathbb{R}} \tag{5.20}$$

This definition of the Reynolds number is specially advantageous for experimentalists, who can measure the Reynolds number taking advantage of the isotropy of the flow at the Taylor scale, which is located in the inertial range.

Kolmogorov-Oboukhov 1962 correction

Kolmogorov's 1941 theory assumes that "*Since* $\langle \epsilon \rangle$ *is almost constant in regions which are small in comparison with the external scale* L*, when* $l \ll L$ *it may be supposed that* $\langle \epsilon_l \rangle = \langle \epsilon \rangle$". This assumption was quickly constested by Landau, as it does not take into account that with the increase of the ratio L/l the variation σ_ϵ^2 of the dissipation of energy ϵ defined in the K41 theory (see eq.(5.9)) would increase without limit [125, 112, 66].

Experimental support for the criticism of K41 theory came later with observations of turbulent intensity variations when analysing spectral fluctuations [88, 150]. These observations indicated the inhomogeneity of the dissipation rate. In the seventies, experiments analysing the deformation of the PDFs of velocity increments across scales [199] (see figure 5.4) and later the scalings of the structure functions of order higher than 3 [8] corroborate the innacuracy of K41 theory.

Kolmogorov and Oboukhov corrected the previous K41 theory by supposing an asymptotic behaviour for the dispersion of the logarithm of ϵ, $\sigma_{\log(\epsilon)}$. The main correction made by Oboukhov and Kolmogorov to the K41 theory was the definition of a local energy dissipation ϵ_l instead of a global one ϵ [112, 150]. They defined ϵ_l distributed with log-normal statistics and they found:

$$\delta_l v(x) = v_x(x + l) - v_x(x) \approx (\epsilon_l l)^{1/3} \tag{5.21}$$

$$S_p(l) = \langle (\delta_l v)^p \rangle = C_p(l \langle \epsilon_l \rangle)^{p/3} \tag{5.22}$$

The assumption of a non-constant dissipation rate leads to different scalings of the increments, and hence of the structure functions, across scales. This phenomenon is called intermittency.

5.2 Multifractal approach of Kolmogorov's theory

Frisch and Parisi [67], recovered all of the Kolmogorov theory of turbulence using the multifractal approach [66, 37, 191]. This is done by reformulating the scaling behaviour of increments and structure functions across scales:

$$\delta_l v(x) \approx l^{\mathfrak{h}} \tag{5.23}$$

$$S_p(l) \approx l^{\zeta(p)} \tag{5.24}$$

where $\mathfrak{h}$, called the Hölder exponent, indicates the order of the singularities characterising the turbulent velocity and $\zeta(p)$ defines the scaling exponents of turbulent velocity. Any model is characterized by the set of their scaling exponents, $\zeta(p)$, as they appear in eq.(5.24). The four-fifths law imposes $\zeta(3) = 1$ and this should be respected by any model or process representing turbulence, see figure 5.3 a).

In multifractal theory $\mathfrak{h}$ is a random variable with probabilty density:

$$P_l(\mathfrak{h}) \approx l^{3-D(\mathfrak{h})} \tag{5.25}$$

with $D(\mathfrak{h})$ the singularity spectrum (see figure 5.3 b)), which is related to the probability of finding the Hölder exponent $\mathfrak{h}$ [67, 208]. The number 3 in eq.(5.25) indicates the dimension of the flow.

The scaling exponents and the singularity spectrum are related by a Legendre transform [67, 208]:

$$\zeta(p) = \min_{\mathfrak{h}} \left[p\mathfrak{h} + 3 - D(\mathfrak{h})\right] \tag{5.26}$$

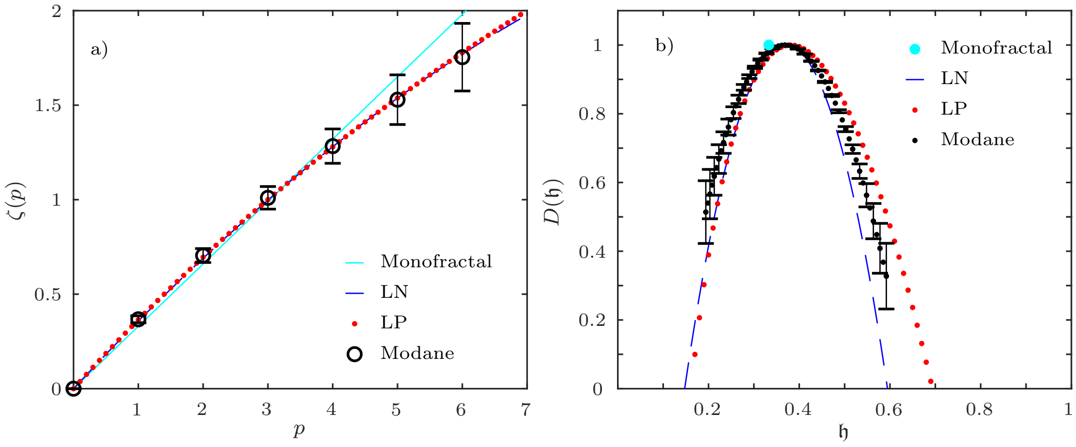

Figure 5.3: a) Scaling exponents $\zeta(p)$ versus order p and b) Singularity spectrum $D(\mathfrak{h})$ versus Hölder exponent $\mathfrak{h}$, both for three different models of turbulence in the inertial domain, together with an experimental Eulerian velocity measure (Modane, black symbols). Models are: monofractal fractional Brownian motion (cyan, continuous straight line), multifractal log-normal (blue, dashed line) and multifractal log-Poisson (red, dotted line). The scaling exponents $\zeta(p)$ are computed from the partition functions of the increments. In order to compute $D(\mathfrak{h})$ we use the wavelet leaders methodology [215].

Another way to describe the scaling exponents is to use the following expansion [52]:

$$\zeta(p) = c_1 p - c_2 \frac{p^2}{2!} + c_3 \frac{p^3}{3!} - \cdots \tag{5.27}$$

where the coefficients c_p of the expansion are called log-cumulants.

A linear behavior of the scaling exponents with p characterizes a monofractal process. In the case of a monofractal process $c_1 = \mathfrak{h}$ is called the Hurst exponent, and noted $\mathcal{H}$. In contrast, the existence of non-zero log-cumulants c_p of order $p \geq 2$, and hence a non linear behavior of the scaling exponents, reveals multifractality, see figure 5.3 a).

5.2.1 Monofractal model

Kolmogorov's 1941 theory states $\mathfrak{h} \equiv \mathcal{H} = 1/3$ which places turbulence in the monofractal regime. This assumption can be formulated as:

- $D(\mathfrak{h}) = \delta(\mathfrak{h} - 1/3)$ is a single-valued function. For a monofractal process there is only one possible value for the Hölder exponent $\mathfrak{h}$, the Hurst exponent $\mathcal{H}$, see figure 5.3 b).
- $\zeta(p) = p\mathcal{H} = \frac{p}{3}$. The scaling exponent are linear in p of slope 1/3 (see figure 5.3) a), so only the first log-cumulant c_1 is non zero: $c_1 = \mathcal{H} = 1/3$.
- $S_p(l) \approx l^{p/3}$ normal power laws.
- Self similarity: statistical properties of the increments remain unchanged across scales. The scale invariance implies the following relation between the probability distributions $p_{\delta_l X}$ and $p_{\delta_{l_0} X}$ of the increments of scales l and l_0:

  $$p_{\delta_l X}(\delta_l X) = \left(\frac{l_0}{l}\right)^{\mathcal{H}} p_{\delta_{l_0} X}\left(\left(\frac{l_0}{l}\right)^{\mathcal{H}} \delta_l X\right) , \tag{5.28}$$

  This relation is valid for all pairs of scales (l, l_0).

Figure 5.3 a) shows that, when p increases, the scaling exponents of experimental turbulent velocity deviate from the linear behaviour predicted by the monofractal model. Therefore, following K41 to model turbulence is not very satisfying for larger p.

5.2.2 Multifractal models

Intermittency is manifested in different ways, which are related. First, one can understand intermittency as the multivaluation of $\mathfrak{h}$ which leads to the non-linearity of the scaling exponents $\zeta(p)$ [13, 8]. This non-linearity directly implies the abnormal power law behaviour of the structure functions of the velocity increments [8].

A related manifestation of intermittency is the evolution of the PDF of the velocity increments when the size of the increment varies, see figure 5.4. The PDF of the velocity increments, close to Gaussian for scales higher or equal to the integral scale L, deforms across the inertial domain: it develops larger tails and a dissymetry [199, 13, 149]. The deformation of the PDF is faster when entering the dissipative range [39].

Intermittency corrections give a multifractal status to turbulence and then:

- $\mathfrak{h}$ takes different values. Consequently, the singularity spectrum $D(\mathfrak{h})$ is multivalued, see figure 5.3 b). In the case of KO62 theory the singularity spectrum is parabolic.
- $\zeta(p) \neq \frac{p}{3}$ non linear.
- $S_p(l) \approx l^{\zeta(p)}$ modified power laws.
- No self similarity: statistical properties of the increments change across scales. Their pdfs deform across scales.

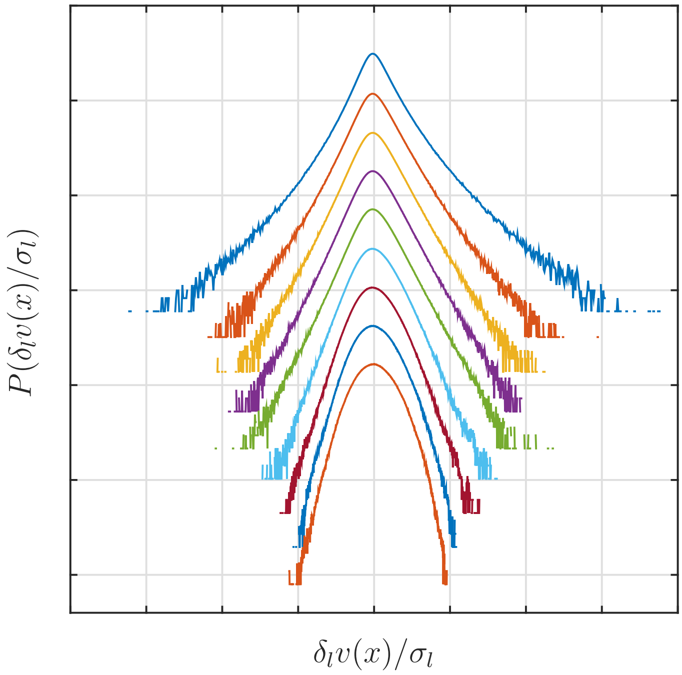

Figure 5.4: Deformation of the PDF of velocity increments across scales, from Gaussian at large scales to strongly non Gaussian at small scales.

Different multifractal models of turbulence exist. Each one of them takes into account intermittency in a slightly different way. Between the most famous models are the log-normal model and the log-Poisson one.

Intermittent log-normal model

This model for turbulence was introduced by Kolmogorov and Oboukhov in 1962 [112, 150] and extensively studied by Chevillard et al. [38]. It was the first intermittent model of turbulence, with the following scaling exponents:

$$\zeta(p) = c_1 p - c_2 \frac{p^2}{2}$$

The non-linear dependence of the scaling exponents in p indicates the multifractal nature of the model, which is quantified by c_2. All log-cumulants c_p of order $p \geq 2$ are zero. Its singularity spectrum is:

$$\mathcal{D}(\mathfrak{h}) = 1 - \frac{(\mathfrak{h} - c_1)^2}{2c_2} \tag{5.29}$$

This multifractal process offers a satisfying representation of the scaling exponents of turbulence if $c_2 = 0.025$ and $c_1 = 1/3 + 3c_2/2 = 0.371$. The c_2 value for a correct log-normal modeling of turbulence was obtained experimentally. Once c_2 is fixed, the four-fifths law imposes the c_1 value for the model.

Intermittent log-Poisson model

This model was introduced by She and Leveque [178]. This heuristic model leads to scaling exponents of the form:

$$\zeta(p) = -\gamma p - \lambda(\beta^p - 1)$$

The four-fifths law, together with the assumption of a finite support of dissipation when the viscosity tends to zero [178], imposes $\lambda = 2$, $\beta = \left(\frac{2}{3}\right)^{(1/3)}$ and $\gamma = -1/9$. For these values the model describes the scaling exponents $\zeta(p)$ as well as the log-normal model does.

It has later been interpreted as a log-Poisson model with a singularity spectrum

$$\mathcal{D}(\mathfrak{h}) = 1 - \lambda + \frac{\mathfrak{h} + \gamma}{\ln(\beta)} \left(\ln \left(\frac{\mathfrak{h} + \gamma}{-\lambda \ln(\beta)} \right) - 1 \right) \tag{5.30}$$

The corresponding log-cumulants are:

$$c_1 = -\gamma - \lambda \ln(\beta) \tag{5.31}$$

$$c_m = \lambda \ln(\beta)^m, \quad m \geq 2 \tag{5.32}$$

5.2.3 Conclusion

This multifractal formalism of turbulence developed by Frisch and Parisi can describe both Kolmogorov 1941 and Kolmogorov-Oboukhov 1962 theories. This formalism is also able to characterize modern physical theories such as the β-model [68, 22, 66] or She and Leveque model of turbulence [178].

A possible objection to the formalism is that it moves away from physical phenomena. Nevertheless, it provides an ideal framework to statistically characterize processes and signals. In part II, we characterize turbulence by analysing experimental and synthetic processes together with theoretical models, using information theory. We rely on multifractal formalism to describe turbulence and to give an interpretation to our measures.

CHAPTER 6

An information theory viewpoint of K41 theory

We propose in this chapter an information theory perspective on turbulence, and especially on K41 theory (section 5.1). We first describe the distribution of information (in Shannon's sense) across scales in turbulence and then show and explain its close connexion with the energy scaling. We also explore the existence of an energy cascade in turbulence *via* information theory.

In order to study turbulence, and recover the K41 theory in the information theory framework, we analyse experimental velocity signals obtained from different experimental setups and a fractional Brownian motion, which is the monofractal synthetic process proposed by Kolmogorov to model turbulence (section 5.2).

In a first stage, we report how the entropy rate is able to describe the distribution of information amongst scales, and the relation between the entropy rate of the analysed process and the entropy of its increments (chapter 4). In a second stage, we design a conditioning procedure of the signal and we show that this allows for a very fine exploration of the asymmetry of the increments distribution and hence the inference of the existence of a finite skewness and an energy cascade.

6.1 Turbulent experimental systems and signals

We analyze two different sets of experimental turbulent data, in order to show the ability of our measures to grasp inherent properties of turbulence.

The first system consists of a temporal measurement of the longitudinal velocity (V) at one location in a grid turbulence setup in the wind tunnel of ONERA at Modane [104]. The Taylor-scale based Reynolds number $\mathbb{R}_\lambda$ is about 2500, with a turbulence intensity of about 8%. The inertial region spans approximately three decades. The sampling frequency is $f_s = 25$ kHz and the mean velocity of the wind in the tunnel is $\langle v \rangle = 20.5$ m/s. The probability density function of the data is almost Gaussian although there is some visible asymmetry: the skewness is about 0.175 ± 0.001.

The second system is a set of temporal velocity measurements at different Reynolds numbers in a jet turbulence experiment with Helium [34]. The Taylor-scale based Reynolds number $\mathbb{R}_\lambda$ is respectively 89, 208, 463, 703, 929, with a turbulence intensity about 23%.

Using the Taylor hypothesis [66] and the mean velocity $\langle v \rangle$ of the flow, we interpret these time series as the spatial evolution of the longitudinal velocity. The time scale τ and the spatial scale l are related by $l = \langle v \rangle \tau$. We note the integral time scale T and the integral spatial scale L, and we have $L = \langle v \rangle T$. During this chapter, we represent our results in function of $\ln \frac{1}{\tau}$. We define the temporal variable τ being multiple of dt where $dt = 1/f_s$ with f_s the sampling frequency in the case of experimental signals. This representation together with Batchelor model presented in section 6.2.1 allows to estimate the integral and Kolmogorov scales of the studied flow. In the case of the generated fBm we analysed in this section, we can suppose $f_s = 1$. In chapter 7 we present all our results as functions of the ratio $\tau/T = l/L$ between the scale of the increment and the integral scale.

6.2 Scaling of information in turbulence

In order to study the dynamics of the signals, we compute the entropy rate $h^{(m,\lambda)}$ across scales (see sections 1.5 and 2.5).

The size of the analysed processes $N = 2^{17}$ is kept constant for all the scales τ. The entropy rate is computed, and then averaged, over 195 independent realizations in the case of the Modane turbulent velocity data and the fractional Brownian motion and over 10 realizations in the case of the jet turbulence.

6.2.1 Grid turbulence

Our results are presented in figure 6.1 as a function of $\log(1/\tau)$. Three regions are observed: above 36ms are the integral scales, below ~ 0.18ms is the dissipative range, and in-between them is the inertial range. At larger (integral) scales, the auto-correlation function vanishes and the entropy rate is equal to the entropy H of the signal ($m = 1$) which depends only on one-point statistics. We can interpret this first result as indicating that the integral scales, where energy is injected and turbulence is generated, are the most disorganized. In the inertial range, the entropy rate decreases almost linearly with a slope close to -1/3, represented by a straight line in figure 6.1. As the time — or space — scale is decreased, the flow appears more and more organized in the sense that the amount of "new" information brought to one point by another point at a distance τ decreases with τ. In the dissipative range, the entropy rate decreases faster and faster, as the dissipation become stronger and stronger. We interpret this as a consequence of the flow being more and more organized.

The entropy rate displays differently from the power spectrum the separation between the different domains, as can be seen in figure 6.1. Remaining in the real space we avoid some perturbing efects of the Fourier transform over the spectrum [134]. The integral and Kolmogorov scales shown in this figure have been obtained using the Batchelor model [20] for fully developed turbulence, as detailed below.

We checked that our estimation of the entropy rate does not depend on the number of neighbors used in the k-nn search algorithm ($k = 5$ and 10) nor on the sample size N

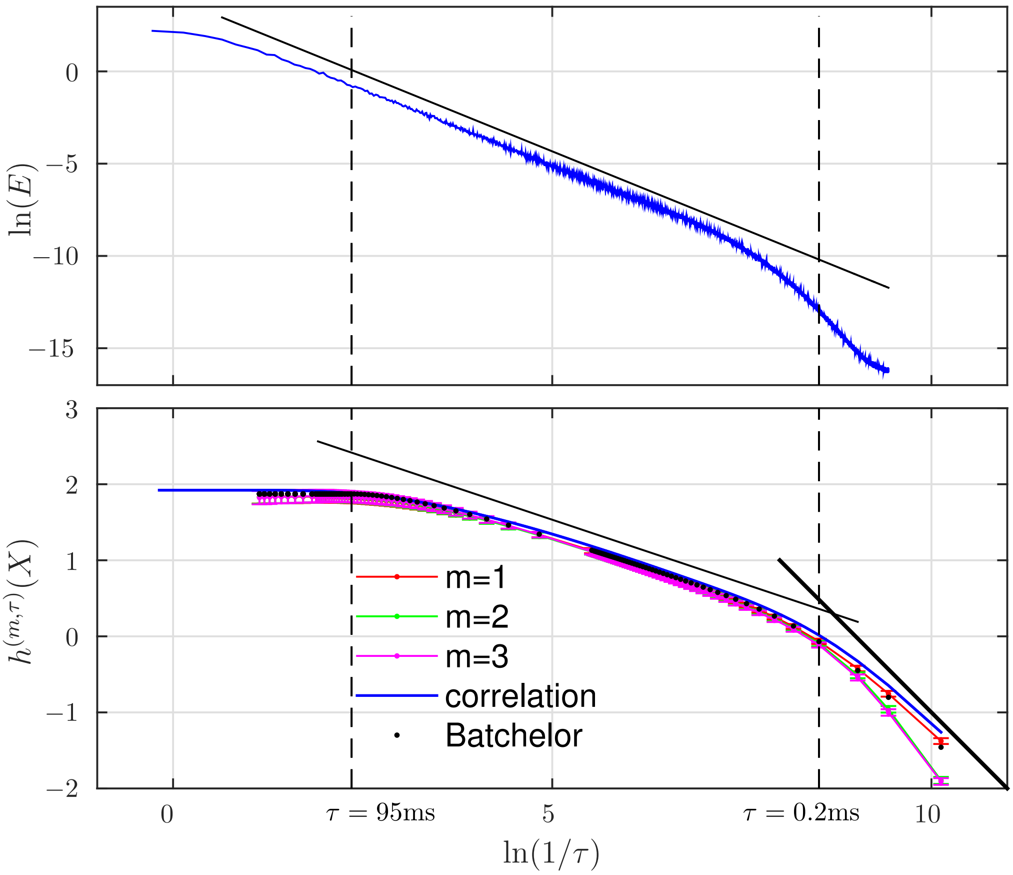

Figure 6.1: Top: Power Spectral Density of the experimental signal. The straight line corresponds to the K41 scaling. Bottom: entropy rate $h^{(m,\tau)}$, as a function of the scale ($f = 1/\tau$) for different embedding dimensions $m \in \{1, 2, 3\}$. The blue curve corresponds to eq.(1.40) using the autocorrelation function. The thin straight line is a line of slope -1/3, the thick straight line has a slope -1.

($N = 2^{16}, 2^{17}$ and 2^{19}). The standard deviation of the entropy rate estimation is of the order of 0.03 for small τ and increase to 0.08 for large τ.

We varied the embedding dimension m and observed a dependance of the entropy rate in the dissipative range only, for scales τ smaller than 0.18ms. Going deeper and deeper in this range, the signal is more and more continuous so the knowledge of an increasing number m of points in the past (separated by the scale τ) decreases significantly the ”new” information brought by a $(m+1)$th point in the future (figure 6.1). But for $m > 2$, there is no measurable evolution anymore, even in the dissipative range.

Batchelor Model: Entropy rate and autocorrelation

Here we present the Batchelor model for fully developed turbulence [20]. Batchelor provided a model for the inertial and dissipative regions of $S_2(l)$, the second-order structure function:

$$S_2(l) = \frac{(l/L)^{2/3}}{\left(1 + (\eta_k/l)^2\right)^{2/3}} \,. \tag{6.1}$$

L is the integral scale and η_k the dissipation scale, l is the current scale.

Eq.(6.1) imposes on the second-order structure function to have a slope 2/3 in the inertial range and 2 in the dissipative range, with a smooth transition between these two regions. Noting that $S_2(l) = 1 - c(l)$ and assuming Gaussian distribution for the turbulent velocity, we can use eq.(1.40) and obtain:

$$h^{(1,l)}(X) = H(X) + \frac{1}{2}\log(S_2^2(l) - 2S_2(l)) \tag{6.2}$$

Substituting the expression of the second order structure function of eq.(6.1) into eq.(6.2) one derives for the entropy rate $h^{(1,l)}$ two linear behaviors in $\log(1/l)$: one with a slope $-1/3$ in the inertial region and another one with a slope of -1 in the dissipative range. This analytical development shows the narrow connection between the traditional Kolmogorov 1941 scaling law (section 5.1) and our scaling law for the entropy rate.

Fitting S_2, the structure function, or fitting the entropy rate (with the Batchelor model) can be used to estimate the scales L and η_k. Thus, for a serie of measures where L is kept constant and the Reynolds number ($\mathbb{R}e$) varied, Batchelor model allows to verify, the well known relation between the Kolmogorov scale (η_k) and the integral one (L) showed in eq.(5.17).

This can be understood as another test (and application) for our entropy rate estimator, in order to check if the physics of turbulence is well grasped by the entropy rate. Theoretically, being directly related with the second-order structure function by eq.(6.2), the entropy rate can also be used to corroborate eq.(5.17).

Fractional Brownian motion

As the turbulent velocity distribution is almost Gaussian, in the K41 theory Kolmogorov proposed fractional Brownian motion (see section 4.2) with Hurst exponent $\mathcal{H} = 1/3$ to model the dependency structures of turbulent velocity in the inertial range, see section 5.2.1. Fractional Brownian motion with Hurst exponent $\mathcal{H} = 1/3$ presents a power spectral density with a power law of exponent $-5/3$, like that of the turbulent velocity, see figure 5.2.

The entropy rate of the fBm is therefore linear in $\ln(\tau)$, with a constant slope $\mathcal{H}$, independent of the temporal extension, as we can see in figure 4.3. Figure 6.5 shows the entropy rate $h^{(1,\tau)}$ of a synthesized fBM with $\mathcal{H} = 1/3$. The measured slope of the entropy rate in function of $\ln(\tau)$ is 0.32 ± 0.01, in agreement with the theoretical value $\mathcal{H} = 1/3$. We have also computed $h^{(m,\tau)}$ for $2 \leq m \leq 4$ and observed a small deviation from the linear behavior for large τ, while the slope slightly increases to reach 0.34 ± 0.01 for $m = 4$. We attribute this to finite size effects, especially in the corrections to eq.(4.23).

The fBm, being monofractal (see section 5.2.1), has no characteristic scales (neither integral nor Kolmogorov scale), so the slope of its entropy rate is unperturbed. Therefore, analysing the behaviour of such a process with Hurst exponent $\mathcal{H} = 1/3$ allowed us to explain the scaling behaviour of the entropy rate of our turbulence data in the inertial range.

Entropy rate of the velocity and the entropy of its increments

The turbulent velocity is stationary, and eq.(4.8) can be used to show the analytical relationship between the entropy rate of the velocity and the entropy of its increments. Figure 6.2 shows numerically the correctness of eq.(4.8). Figure 6.2 b) shows the border effects in both limits: when τ is small, under the dissipative scale, the filtering effects of the experimental set-up affects differently the velocity signal and its increments; for large scales, the effects of the available statistics are also different. Nevertheless, figures 6.2 a) and b) show the close relationship between the entropy rate of the velocity and the entropy of its increments.

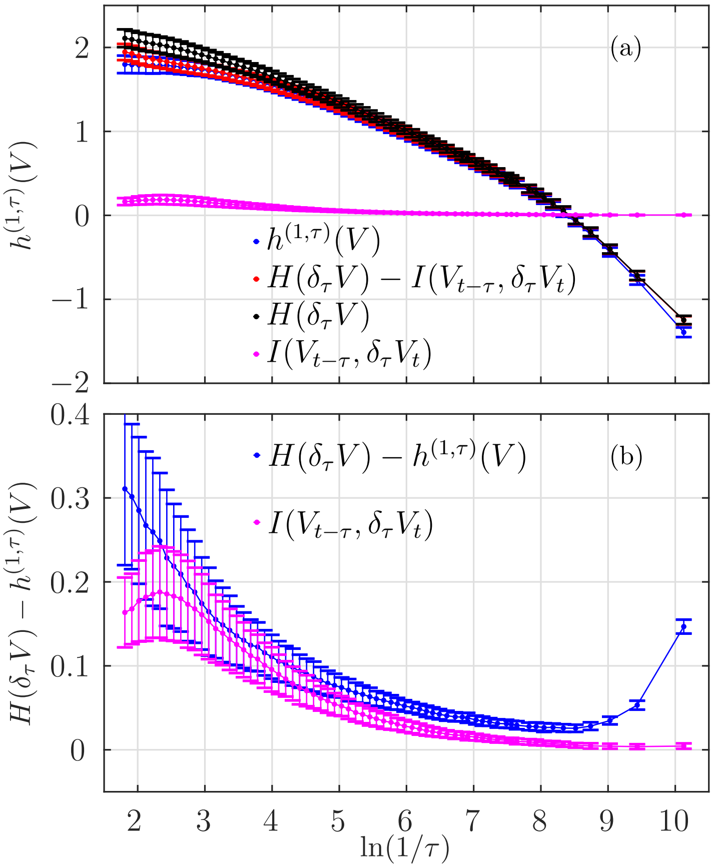

Figure 6.2: a) Entropy rate of the velocity $h^{(1,\tau)}(V)$, entropy of its increments $H(\delta_\tau V)$, mutual information between the velocity and its increments $I(V_{t-\tau}, \delta_\tau V_t)$, and the difference between the two last $H(\delta_\tau V) - I(V_{t-\tau}, \delta_\tau V_t)$. b) Entropy rate of the velocity $h^{(1,\tau)}(V)$ minus entropy of its increments $H(\delta_\tau V)$, and mutual information between the velocity and its increments. The velocity signal is the one from the wind tunnel at Modane . The time window size is $T = 2^{16}$.

The case of fractional Brownian motion, being non-stationary, is completely studied in chapter 4, sections 4.1.2 and 4.3.

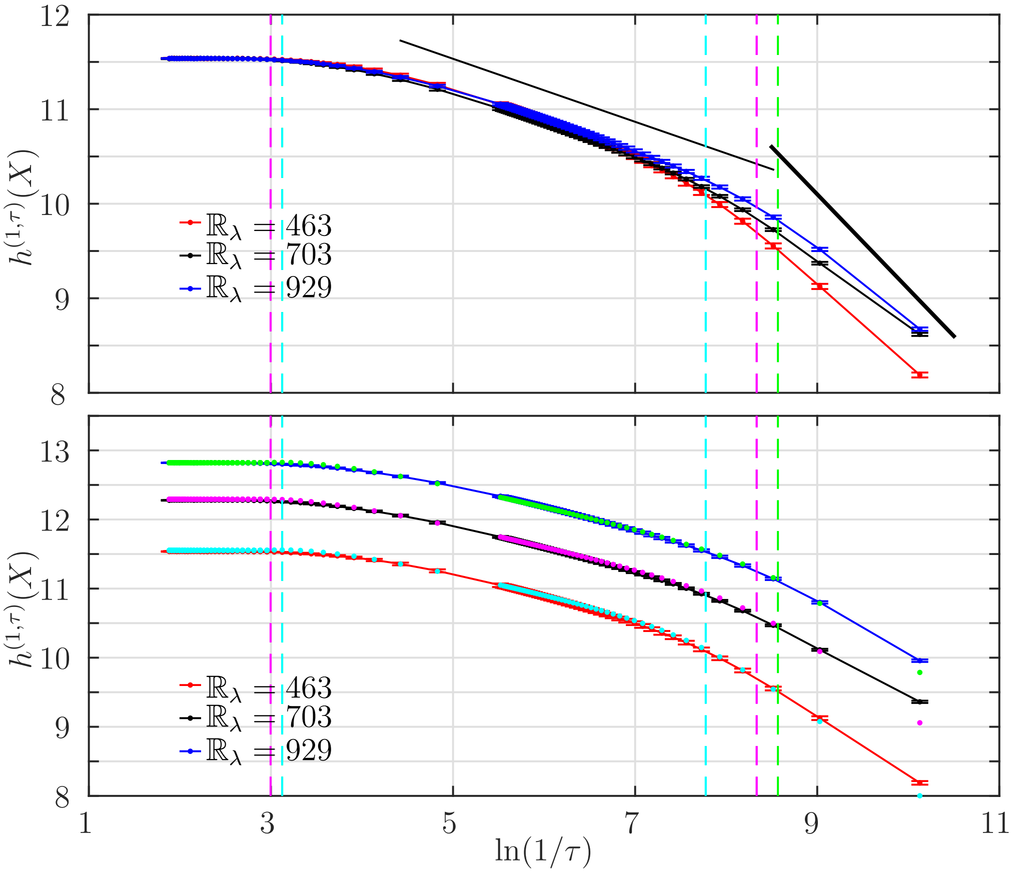

Figure 6.3: Top: entropy rates $h^{(1,\tau)}$ for different Reynolds numbers, as a function of the scale ($f = 1/\tau$). The entropy has been normalised to put all the curves over the lowest Reynolds number measure. The thin straight line is a line of slope -1/3, the thick straight line has a slope -1. Bottom: entropy rates $h^{(1,\tau)}$, as a function of the scale ($f = 1/\tau$) for different Reynolds numbers, and the Batchelor fits for each of them (points).

6.2.2 Jet turbulence in Helium

As in the grid turbulent system, in the jet turbulence on Helium we find three different regions, figure 6.3: integral, inertial and dissipative, with three different behaviors. As predicted by the theory developed by Batchelor we find a slope $-1/3$ in the inertial range, a slope tending to -1 in the dissipative region and a convergence to the entropy value of the velocity signal in the integral domain. The interpretation is the same than in the first system. We can see in figure 6.3 how the Batchelor Model is able to fit the behavior of the entropy rate in the inertial and dissipative domains, and the transition between them. Another important result is the maximum value of the entropy rate increasing with the Reynolds number. That can be understood as the total information contained in the system increasing with the Reynolds number, and then, increasing with the development of turbulence. The higher the Reynolds the more developed the turbulence, the larger the value of information contained in the system. It's important to remark that the maximum value of the entropy rate is the value of the entropy of the signal and mainly depends on the standard deviation of the signal.

In this system we have access to different Reynolds number flows, and thus we can analyse the variation of the inertial region length of turbulence as measured by the entropy rate in function of $\mathbb{R}_\lambda$. We can observe in figure 6.3 how qualitatively the inertial region length of turbulence measured by the entropy rate decreases when the Reynolds number decreases, as expected. Using Batchelor model we can fit the different Reynolds number curves and obtain the integral and Kolmogorov scales predicted by the model for each curve. The integral scale L is the same for the different analysed signals as they come from the same experimental set-up, in which $\mathbb{R}_\lambda$ was changed without modifying L.

When the embedding dimension increases we find similar results to the obtained in the first system. The estimation doesn't vary, except in the dissipation domain.

6.3 Conditioned Entropy rate

In a general way we can define the conditioned entropy rate $h^{(m,\lambda)}(X^{\text{cond}})$ as the entropy rate applied over a subsampled signal X^{cond} obtained by conditioning the mother signal X.

In chapters 3 and 4 we showed that the entropy rate of a random variable X is able to show the distribution of information along the scales. Even if the entropy rate, as defined in eq.(1.25), is related to all order statistics, it cannot probe finely the existence of a non zero skewness because of the effect of a very large 2-point correlation hiding the higher order statistics of the random variable distribution.

To probe more accurately the symmetry of the two points PDF of a signal X, we propose a conditioning procedure over its increments. Given a random variable X composed by a set of points $x(i)$ we can define the increment signal $\delta_\lambda X$ composed by the points $\delta_\lambda(i) = x(i) - x(i-\lambda)$.

We define the signal X_+, resp. X_-, as the subset of points $x(i)$ from X such that $\delta_\lambda(i) = x(i) - x(i-\lambda) > 0$, resp. $\delta_\lambda(i) < 0$. We then define the conditioned PDF p_+, resp. p_-, of the signal X_+, resp. X_-. Although signals X_+ and X_- are one-dimensional and take their values in the same vector space S as X does, they contain some information provided by the increments, namely the sign of the local increment $\delta_\lambda(i)$ associated with $x(i)$. And then the PDF's of the signals X_+ and X_- will be, generically, different from the PDF of X.

It is important to note here that if the statistics of the increments $\delta_\lambda(i)$ of the signal X are skewed, then the joint PDF of $(x(i), x(i-\lambda))$ is not symmetrical with respect to the origin. The reciprocal may not hold.

We then define the increment conditioned entropy rate $h^{+,(m,\lambda)}$, resp. $h^{-,(m,\lambda)}$, of the signal X as the entropy rate $h^{(m,\lambda)}$ of the conditioned signal X_+, resp. X_-. In practice, we compute the same quantity as before, defined by eq.(1.25), but using only a subset of all data points; this subset is obtained by retaining points $x(i)$ with a given sign of $\delta_\lambda(i)$. The conditioning is performed at the single date i, whatever the embedding dimension m is, so not looking at the (sign of) increments $\delta_\lambda(i-k\lambda), 1 \leq k \leq m-1$.

The entropy rate considers an embedded signal of dimension $m+1$, so even for the smallest $m=1$, the entropy rate probes 2-points dependences (see chapter 3), between $x(i)$ and $x(i-\lambda)$. In that case, the additional conditioning on the sign of the increment allows the conditioned entropy rate to probe the asymmetry of the joint PDF $p(x(i), x(i-\lambda))$, which is related to the skewness of the increments.

If the joint PDF is symmetrical with respect to the origin, it is easy to check that $p_+(x) = p_-(-x)$ and therefore

$$h^{+,(m,\lambda)}(X) = h^{-,(m,\lambda)}(X).$$

On the contrary, if the statistics of the increments are skewed, then the joint pdf does not have the central symmetry and we may have $h^{+,(m,\lambda)}(X) \neq h^{-,(m,\lambda)}(X)$.

To check the robustness of our results, we perform two different tests using skewed signals. First, because of the skewness of the increments, the fraction of points of X in subsets X^+ and X^- can be quite different. As this may cause a difference in conditioned entropy rates, we recomputed $h^{-,(m,\lambda)}$ when imposing that X^- has the same number of points as X^+. To do so, we simply discarded the extra points from X^- (randomly chosen). This procedure does not change the correlations of X^-, nor the statistics and hence the pdf. Again, we will find the same significant difference between $h^{+,(m,\lambda)}$ and $h^{-,(m,\lambda)}$. Second, we replaced the conditioning on the sign of the increments by a random sub-sampling of X, in order to obtain a subsampled signal X^{rand}. In that case we will obtain $h^{(m,\lambda)}(X^{\text{rand}}) = h^{(m,\lambda)}(X)$, as expected, although the number of points in X and X^{rand} differ by a factor 2.

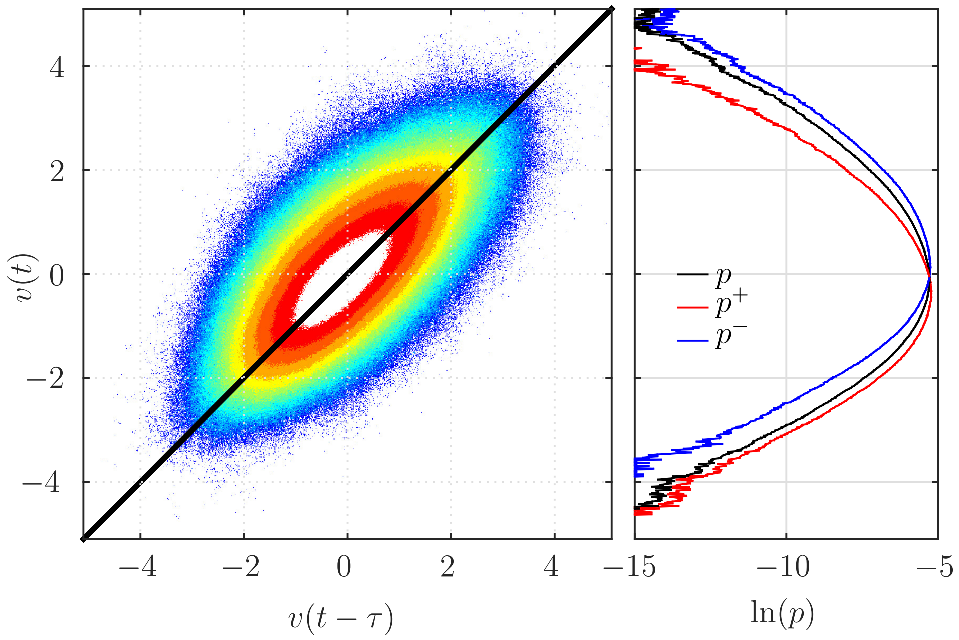

Figure 6.4: Left: bi-variate PDF of $(v(t), v(t-\tau))$, the velocity field. Right: $p(X)$ (in red), $p_+(X)$ (blue) and $p_-(X)$ (black).

6.4 Skewness of turbulent velocity increments

We compute the conditioned entropy rate, $h^{+,(m,\tau)}$, resp. $h^{-,(m,\tau)}$ (section 6.3) with the embedding dimension always kept as one ($m = 1$). The size of the analysed signals is $N = 2^{19}$, constant. The conditioned entropy rate is averaged over over 195 independent

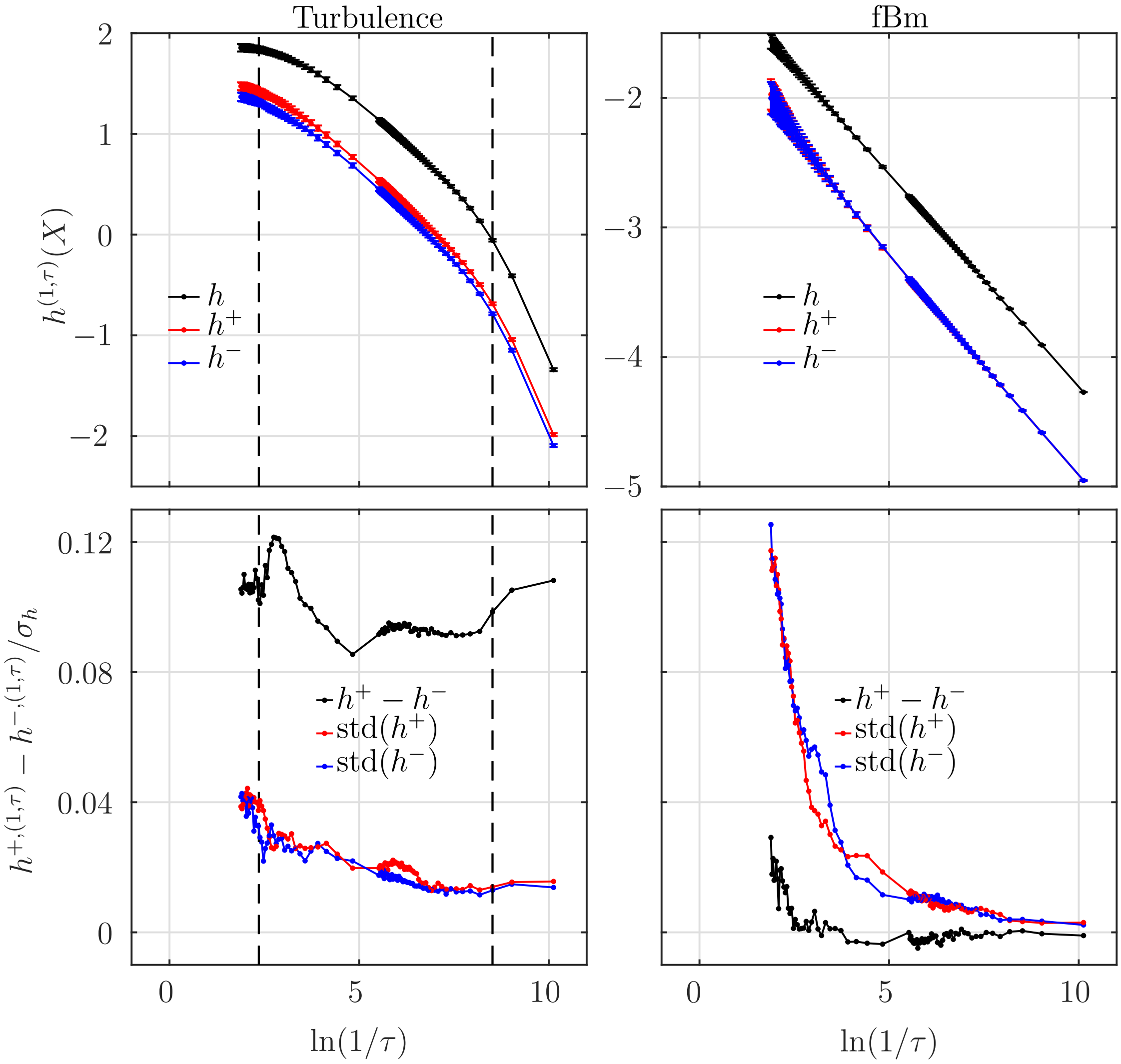

Figure 6.5: Top: conditioned entropy rates $h^{+,(1,\tau)}$ (red) and $h^{-,(1,\tau)}$ (blue) as function of scale for turbulence data for embedding $m = 1$. The entropy rate $h^{(1,\tau)}$ (black) is reported for comparison. Bottom: difference $(h^{+,(1,\tau)} - h^{-,(1,\tau)})$ (black) and standard deviation of $h^{+,(1,\tau)}$ (blue) and $h^{-,(1,\tau)}$ (red).

realizations in the case of the Modane turbulent velocity data and the fractional Brownian motion and over 10 realizations in the case of the jet turbulence.

Grid turbulence:

Above we showed that the entropy rate of a turbulent velocity signal can be well described using only the autocorrelation function (blue line in figure 6.1), because the signal statistics are very close to Gaussian. Nevertheless, the statistics of the increments $\delta_\tau(t) \equiv x(t) - x(t-\tau)$ of a turbulent velocity signal can be far from Gaussian, especially for smaller scales τ. In particular, the PDF of increments is skewed (see eq.(5.16)). A non-vanishing skewness of the increments or a non-vanishing order-3 structure function,

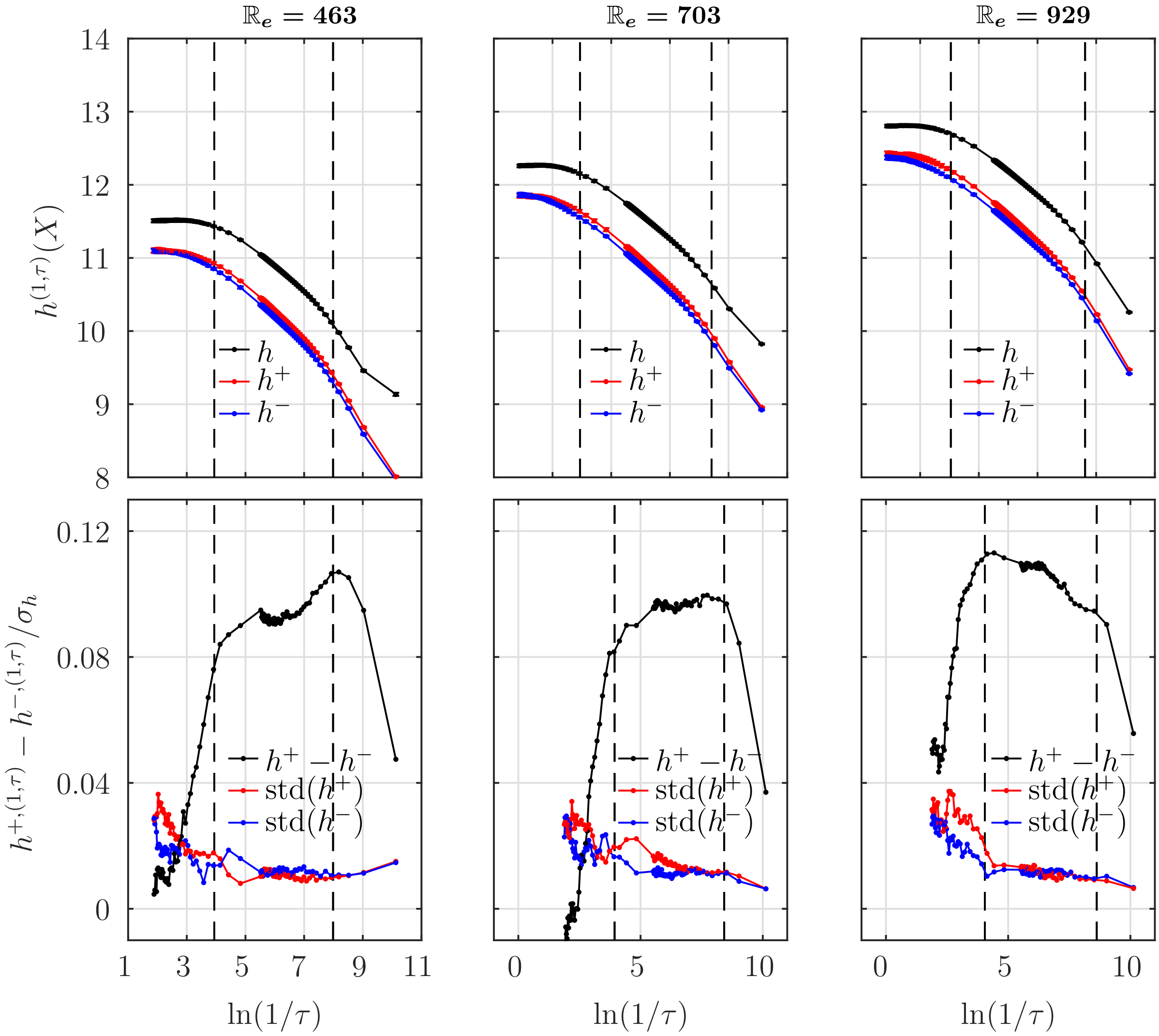

Figure 6.6: Top: conditioned entropy rates $h^{+,(1,\tau)}$ (red) and $h^{-,(1,\tau)}$ (blue) as function of scale for turbulence data (Increasing Reynolds form left to right) for embedding $m = 1$. The entropy rate $h^{(1,\tau)}$ (black) is reported for comparison. Bottom: difference $(h^{+,(1,\tau)} - h^{-,(1,\tau)})$ (black) and standard deviation of $h^{+,(1,\tau)}$ (blue) and $h^{-,(1,\tau)}$ (red).

$S_3(l) \equiv \mathbb{E}\{(x(t) - x(t-\tau))^3\}$ results in an energy cascade [108, 109, 107] (see section 5.1). The PDFs of the subsampled conditioned signals and the joint PDF of the signal are shown in figure 6.4. In order to probe this energy cascade, measurements of conditioned entropy rates are reported in figure 6.5.

For the turbulent velocity signal, the two conditioned entropy rate give significant different results, with an almost constant difference around 0.1 ± 0.01 (compared to the standard deviation of $h^{(m,\tau)}(X^{\pm})$ which is around 0.03 ± 0.01), see figure 6.5.

Jet turbulence:

As in the precedent system, the difference between the two conditioned entropy rates, $h^{+,(1,\tau)}$ and $h^{-,(1,\tau)}$, is significant for all the different reynolds numbers, see figure 6.6. For all the different Reynolds numbers, the behaviour of the entropy rate and the conditioned entropy rates is the same, being the constant value at large scales, and then the entropy of the velocity signal, the only variation.

Fractional Brownian motion

Figure 6.5 shows how for the fBm, the conditioned entropy rates are indistinguishable and follow the same linear behavior as the entropy rate $h^{(m,\tau)}$, as expected for a Gaussian (no-skewed) distribution.

6.5 Conclusions

We measured the information content of several turbulent velocity signal by computing the entropy rate as a function of the scale at which the signal is considered. We found that the distribution of information is reminiscent of the energy scaling, and we related it to the second order structure function. The entropy rate is able to separate properly the different domains as the second order structure function does. We argued that the entropy rate is more sensitive to dependences than the Power Spectrum (chapter 3), in particular because it can take into account higher order correlations, especially for large embeddings. For this reason, Information Theory perspective may give some new insight on turbulence.

In the entropy rate measure of the longitudinal velocity, the distinction between the contribution of the second order and the higher order moments (concretely the skewness) is unclear. As a consequence, we are not able to probe either the weak skewness of the turbulence signal or the larger skewness of its increments *via* entropy rate. We then designed a conditioning procedure of the data, based on the sign of the increments. Applying the entropy rate to this conditioned data, we were able to illustrate an effect of the skewness of the velocity increments.

The procedures described here are of general interest for the study of complex systems, especially those having multiscale dynamics, as can be found in, *e.g.*, economy, ecology, neuroscience, and of course fluid dynamics. Given the plethora of laws governing such different systems, the model-free and nonlinear nature of information theory makes it a very interesting approach. In the particular case of turbulence, we showed not only that the entropy rate allows one to measure in all generality the information distribution amongst scales, in perfect agreement with known models, but also that a well-chosen conditioning of the data allows one to prove the existence of the energy cascade. Using information theory only, we recovered all classical characteristics of second order moment of turbulence, as well as the existence of an energy cascade via the third order moment.

CHAPTER 7

Rediscovering intermittency

As we explain in chapter 5, sections 5.1 and 5.2, turbulence displays intermittency which manifests in a non linear behaviour of the scaling exponents $\zeta(p)$ and also in the deformation of the PDF across scales. This deformation of the PDF has been conventionally quantified by the evolution of its flatness [66], measured as the normalized kurtosis of the distribution: $\langle(\delta_l v)^4\rangle / \langle(\delta_l v)^2\rangle^2 = S_4(l)/S_2(l)^2$. At larger scales, about or above the integral scale L, the PDF of the velocity increments is almost Gaussian and has a flatness very close to 3. For smaller and smaller scales, the PDF is less and less Gaussian as the PDF of the normalized increments becomes wider and wider; therefore the flatness increases. Fine evolutions of the PDF, and hence intermittency, have been studied with the flatness, such as a rapid increase of intermittency when the scale is reduced down or below the Kolmogorov dissipative scale [39]. S_2 evolves according to the 2/3 law predicted by K41 theory (see section 5.1), so, the kurtosis only involves one higher-order structure function, namely S_4, and as such it does not describe the deviation of all the scaling exponents $\zeta(p)$ from their linear behavior in p. This is why, we propose in this chapter a measure of intermittency that involves all structure functions.

We propose to measure intermittency, interpreted as the deformation of a PDF which is Gaussian at large scales. To do so, we consider the Kullback-Leibler (KL) divergence [120] (see section 1.3.2) between the PDF and the Gaussian PDF defined with the same standard deviation. By comparing the PDF — defined by all its moments — and the Gaussian approximation of this PDF — defined by the second order moment only — we measure not only the growth of the pth order moment with respect to the variance, but also the evolution of all the moments with respect to the variance, *i.e.*, we exhaustively characterize the deformation of the PDF. Measuring the intermittency with a KL divergence provides a generalization of measures such as flatness ($p = 4$), hyperflatness ($p = 6$), *etc.*

Although we propose to study turbulence as an application of our framework, our methodology is very general, and it does not require any *a priori* knowledge of the signal, neither any underlying model of the system that produced the signal. As such, it can prove a very powerful tool to analyze complex systems exhibiting power law behaviors or multiscale dependencies. We apply it to characterize intermittency in turbulence.

This chapter is organized as follows. In section 7.1, we define our information theoretical measure of intermittency that involves Shanon entropy and a well chosen Kullback-Leibler divergence. In section 7.2, we compute this quantity for experimental measurements of the Eulerian velocity field in several setups and several Reynolds numbers. We then turn in section 7.3 to some phenomenological modelings in order to better understand and describe our observations.

7.1 Definitions

7.1.1 KL divergence from Gaussianity

Shannon entropy, $H(X)$, of a process X of PDF $p_X(x)$ is the total information that defines the process, see section 1.3.1. Eq.(1.2) shows that entropy depends on all the moments of the PDF $p_X(x)$ except the first order one.

We know that a Gaussian process, X_G, is uniquely defined by the prescription of its mean, variance, and two-point correlation function. Therefore, its Shannon entropy only depends on its variance $\sigma_{X_{\mathrm{G}}}$, as eq.(1.38) shows. For a generic process X which is *a priori* non-Gaussian and has the variance σ_X^2, we define the “entropy under Gaussian hypothesis” $H_{\mathrm{G}}(X)$ as the entropy that one would get assuming the process is Gaussian and using eq.(1.38):

$$H_{\mathrm{G}}(X) = \frac{1}{2}\ln(2\pi e \sigma_X^2)\,, \tag{7.1}$$

where σ_{X} is the standard deviation (std) of the generic process X. So, the ”entropy under Gaussian hypothesis” of X is a measure of the entropy of a Gaussian PDF with same std as the real PDF of X. If X is Gaussian, obviously $H_{\mathrm{G}}(X_{\mathrm{G}}) = H(X_{\mathrm{G}})$.

For any process X with probability density function $p_X(x)$, we can measure the difference between the ”real” PDF $p_X(x)$ of X and the Gaussian approximation $p_{\mathrm{G}}(x)$ using the Kullback-Leibler divergence defined in section 1.3.2:

$$\mathcal{K}_{p_X||p_{\mathrm{G}}}(X) = \int_{\mathbb{R}} p_X(x) \ln\left(\frac{p_X(x)}{p_{\mathrm{G}}(x)}\right) \mathrm{d}x\,. \tag{7.2}$$

Using the definitions of $H(X)$ and $H_G(X)$, we have:

$$\mathcal{K}_{p_X||p_{\mathrm{G}}}(X) = H_{\mathrm{G}}(X) - H(X) \geq 0\,. \tag{7.3}$$

$\mathcal{K}_{p_X||p_{\mathrm{G}}}(X)$ is a measure of the distance from Gaussianity of the process X, *i.e.*, the distance between the PDF $p_X(x)$ of X, and a Gaussian PDF $p_{\mathrm{G}}(x)$ which has the same std. The maximum entropy principle [100, 101] states that for a given standard deviation, the Gaussian PDF maximizes the entropy, see also [49]. So this distance is also a comparison between the total information needed to define the process and the total information defining the most ambiguous process with same std. The maximization of the entropy for the Gaussian case ensures that the difference $H_{\mathrm{G}}(X) - H(X)$ is always positive, as expected for a KL divergence (see section 1.3.2) and vanishes only when X has a Gaussian distribution.

7.1.2 Distance from Gaussianity across scales

We analyze the process X at scale τ by studying its increments of size τ:

$$\delta_\tau X(t) = X(t+\tau) - X(t) \tag{7.4}$$

We note $D_\tau(X)$ the KL divergence $\mathcal{K}_{p_X||p_{\mathrm{G}}}(\delta_\tau X)$ which measures the distance from Gaussianity of the increments at scale τ of a process X:

$$D_\tau(X) = \mathcal{K}_{p_X||p_{\mathrm{G}}}(\delta_\tau X) = H_{\mathrm{G}}(\delta_\tau X) - H(\delta_\tau X) \tag{7.5}$$

This quantity measures the deformation of the PDF of the increment as a function of the size τ of the increment: it quantifies the evolution of the shape of the PDF, which depends on all the moments of the process, except its mean.

Indeed, at each scale τ, the increment $\delta_\tau X$ has a different standard deviation. The larger the scale τ the higher the standard deviation. So, changing τ changes quantitatively the entropies $H(\delta_\tau X)$ and $H_G(\delta_\tau X)$, which both depend strongly on the standard deviation. Subtracting the two entropies eliminates most of this quantitative variation because the std is by construction the same in both expressions $H_G(\delta_\tau X)$ and $H(\delta_\tau X)$). $D_\tau(X)$ thus only measures subtle and delicate evolutions of the shape of the PDF than the trivial rescaling induces by the std.

In the specific case of turbulence, the PDFs of the increments of size equal or larger than the integral scale L are almost Gaussian. As a consequence, we expect that the distance from Gaussianity $D_\tau(X)$ tends to zero when τ approaches the integral scale. Conversely, it is expected to increase in the inertial range down to the dissipative scale where it should increase faster [39]. Our distance from Gaussianity should therefore be able to probe intermittency of turbulence by measuring the deformation of the PDF of velocity increments.

7.1.3 Methodology

In order to compute accurately H from experimental data, we use the estimator described in section 2.2. We chose the usual value $k = 5$ which is large enough to estimate the Shannon entropy correctly within a reasonable computational time.

Following Theiler [202], when computing the entropy of $\delta_\tau X(t)$, we only retain data points separated in time by a delay time τ_{max}, defined as the size of the largest increment that we compute. This prescription has two benefits. Firstly, two successive points of the subsampled dataset are uncorrelated, because the increments of size τ are typically correlated over a time $\tau \leq \tau_{\mathrm{max}}$. Secondly, the number N of points used in the computation of the entropy of $\delta_\tau X$ is independent of τ, so the bias due to finite size effects is constant when τ is varied.

To compute the entropy under Gaussian hypothesis H_{G}, we compute the standard deviation of X and then use eq.(1.38).

In the remainder of this chapter, all quantities are computed using $N = 512$ points ($\mathrm{Bias}_{\hat{H}} \lesssim 0.01$), $\tau_{\mathrm{max}} = 4096$, so signals with a total of $N\tau_{\mathrm{max}} = 2^{21}$ points. We also average our results over independent realizations, in order to compute the standard deviation of the quantities and provide error bars to the estimations. We use 12 realizations for experimental signals and 8 realizations for synthetic processes.

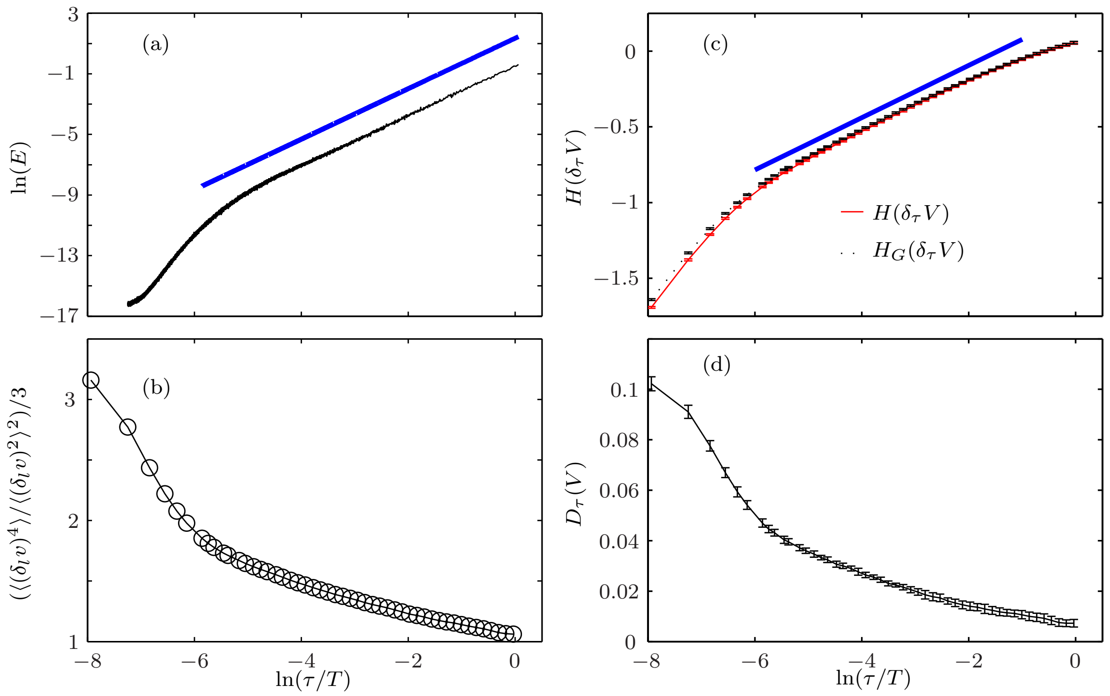

Figure 7.1: (a) Power spectrum (b) Flatness (c) Entropies (d) KL distance from Gaussianity, for the Modane experimental data, as functions of $\ln(\tau/T) = \ln(l/L)$, the logarithm of scale normalized by the integral scale. In (a) and (c), the straight lines indicate the theoretical scaling in the inertial region predicted by Kolmogorov K41 theory.

7.2 Application to turbulent velocity signals

In figure 7.1, we present the analysis of the Modane data (section 6.1). In the left column, we report the classical viewpoint and compare it to the information theory viewpoint in the right column.

We first plot the power spectrum of the velocity signal V as a function of the inverse frequency $1/f = \tau$ in figure 7.1(a) : it shows the distribution of energy across scales following the well known 5/3 Kolmogorov law. In order to measure the deformation of the PDF of the velocity increment when the scale τ is varied, we follow Frisch [66], and compute the flatness as the kurtosis $\langle(\delta_\tau v)^4\rangle/\langle(\delta_\tau v)^2\rangle^2$ of the velocity increments normalized by 3, the kurtosis of a Gaussian PDF. Results are reported in figure 7.1(b). For $\tau \gtrsim T$, *i.e.*, $l \gtrsim L$, the flatness is 1, as expected for a Gaussian PDF. Reducing τ, the flatness increases. When τ is smaller than the dissipative scale [39], the increase of the flatness is sharper. Three different regions can be distinguished in both figures: integral, inertial and dissipative.

The right column of figure 7.1 is devoted to the information theory viewpoint. We first plot the entropy of the increments in figure 7.1(c) and compare it with the PSD in Fig. a. We then plot $D_\tau(V)$ in figure 7.1(d) and compare its behavior in τ with the flatness.

In figure 7.1(c) we see that the entropy of the increments $(H(\delta_\tau V))$ increases with τ. The larger the scale, the higher the total Shannon information needed to completely characterize the increments. We can distinguish three different ranges with different dependence of the entropy on the scale. For the large scales, larger than the integral scale, the entropy reaches its highest value and is then constant. So, the most disorganized or complex scales — the ones requiring more information to be completely characterized — are the scales in the integral domain. Within this region, the characterization of the scale does not require more entropy when the size of the increment increases . A linear behavior of the entropy in $\ln(\tau/T)$ is found in the inertial region, $\tau \in [10, 400]$ samples *i.e.* $\tau \in [0.0004, 0.016]$ seconds, $l \in [0.0082, 0.328]$ meters. The complexity of the scales, as measured by $H(\delta_\tau V)$, decreases linearly in $\ln(\tau/T)$ between the integral and the Kolmogorov scales. For the smallest scales, below the Kolmogorov scale — which we can measure at $\ln(\tau/T) \approx -5$ — we observe a steeper decrease of the disorganization when the scale decreases. So, using the entropy of the increments, we recover the same qualitative results as using the entropy rate of the velocity measures (chapter Chap7).

Both entropies $H(\delta_\tau V)$ and $H_G(\delta_\tau V)$ in figure 7.1,c) are indistinguishable in the integral domain. They start to separate when they enter into the inertial region. In figure 7.1,d, we plot the difference between these two entropies, which, according to eq.(7.5) is the distance from Gaussianty $D_\tau(V)$. Starting from 0 at scales larger than the integral one, it increases when the scale decreases. The vanishing of $D_\tau(V)$ for scales larger than the integral scale implies that the PDF of the velocity increments is almost Gaussian, as expected. Below this integral scale the PDF starts to deform, and becomes less and less Gaussian when the scale decreases. The evolution of $D_\tau(V)$ is almost linear between the integral and the Kolmogorov scales. Finally, in the dissipative range, we observe an abrupt deformation of the PDF, in perfect agreement with the rapid increase of the flatness in figure 7.1,b). [39]

In the four sub-plots of figure 7.1, the three different domains of turbulence are distinguishable: integral, inertial and dissipative. Figure 7.1c) allows us to interpret these three domains in terms of organization and complexity of velocity increments. Figure 7.1,d) shows that the KL divergence allows us to quantify the evolution of intermittency amongst scales τ. We not only recover the three different ranges with our measures based on information theory, but the qualitative behavior of intermittency in each domain is also in perfect agreement with previous studies. Moreover, our measure of intermittency doesn't depend on a specific ratio between selected moments of the PDFs like the kurtosis. $D_\tau(V)$ takes into account all the moments defining the PDFs: this makes our KL distance from Gaussianity across scales a good candidate for a quantitative measure of intermittency.

In figure 7.2, we have compiled information theory results for all the experimental signals presented in section 6.1, in order to study the influence of the Reynolds number. The entropy as a function of the scale is reported in figure 7.2(a); we observe how the size of the inertial range varies with the Reynolds number, with the Kolmogorov scale increasing when the Reynolds number decreases. This classical behavior of the Kolmogorov scale is also recovered with the KL divergence, represented in figure 7.2(b). The steeper slope — that indicates the dissipative domain — appears at higher scales when the Reynolds number is lower; we recover the dependence of the Kolmogorov scales with the Reynolds number.

The behaviors of both the entropy and the distance from Gaussianity are qualitatively the same for different experimental setups and for any Reynolds number. The dependence of the entropy $H(\delta_\tau V)$ of the increments is, at first order, in agreement with the K41 theory: we recover the scaling law in the inertial domain [80]. The KL divergence $D_\tau(V)$ then enlightens the deformation of the PDF across scales, which is qualitatively compatible with the K062 theory and hence the intermittency in turbulence.

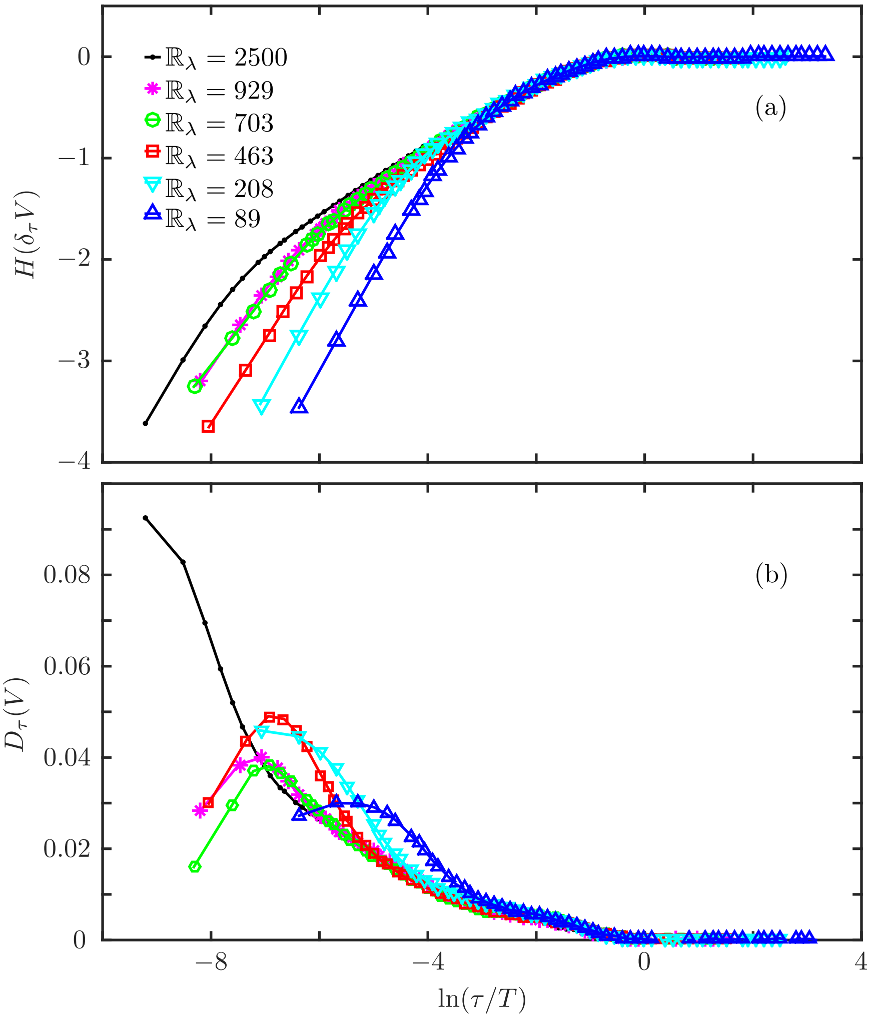

Figure 7.2: a) Entropy $H(\delta_\tau V)$ of the Eulerian velocity increments as a function of $\ln(\tau/T) = \ln(l/L)$. b) KL divergence $D_\tau(V) = H_G(\delta_\tau V) - H(\delta_\tau V)$. Different experimental signals with various Reynolds numbers have been used.

7.3 Modeling

In order to get some insight on the quantitative results obtained with our Kullback-Leibler divergence D_τ, we now turn to some theoretical descriptions of the inertial domain of fully developed turbulence.

First, we study different processes generated to imitate the behavior of turbulence in the inertial region according to a corresponding model. Amongst the simplest, popular and most important is fractional Brownian motion (fBm) [108, 137] which, as a monofractal process, doesn't display intermittency. We also explore multifractal processes, which exhibit intermittency: Multifractal Random Walk (MRW) [15, 126], Random Wavelet Cascade (RWC) with log-normal [126, 9] or log-Poisson distribution of multipliers [126, 9, 178]. We then examine the propagator formalism [32], a phenomenological model that provides an analytical expression of the PDF of the velocity increments [38].

7.3.1 Synthetic processes

We now briefly present the different processes that we numerically generated, according to the prescriptions of section 5.2.

Monofractal process We use the spectral procedure presented by Helgason to synthesize a fractional Brownian motion (fBm) [90] with Hurst exponent $\mathcal{H} = 1/3$, as introduced by Kolmogorov [108].

Log-normal multifractal processes We use two different synthetic processes with log-normal statistics: a Random Wavelet Cascade (RWC) [126, 9] and a Multifractal Random Walk (MRW) [15]. Multifractality requires the existence of an integral scale T, from or towards which the PDF evolves. For both processes, we impose the integral scale T to be equal to the size of the signal.

Log-Poisson multifractal process We use a RWC with log-Poisson statistics [9]. Again, our synthesis fixes the integral scale T to the size of the generated signal.

7.3.2 Results

Classical multifractal analysis offers a way to estimate the log-cumulants c_1 and c_2, but fails to estimate c_3 and higher order log-cumulants. It can therefore be interpreted as projecting the different models onto their log-normal approximation, with varying (c_1, c_2). For example, the multifractal analysis of a realistic log-Poisson model of turbulence leads the couple of values given in table 7.1, and no additional higher order log-cumulant. As a consequence, such an analysis is not able to discriminate which process — log-normal or log-Poisson —better represents turbulence. For this reason, we compute the KL divergence D_τ which takes into account all moments of the PDF of increments, and hence higher order log-cumulants [208], in order to obtain a finer analysis of the inertial domain of turbulence.

In figure 7.3(a) we plot for the four synthetic signals the entropy $H(\delta_\tau X)$ as a function of $\ln(\tau/T)$, the logarithm of the scale. We also plot the entropy under Gaussian hypothesis, $H_\mathrm{G}(\delta_\tau X)$, but it is indistinguishable from $H(\delta_\tau X)$.

From eq.(5.28), it is possible to show that for a monofractal, and then non-intermittent (K41 theory), process:

$$H(\delta_\tau X) = H(\delta_{\tau_0} X) + \mathcal{H} \ln(\tau/\tau_0)\,. \tag{7.6}$$

For any process, the entropy under Gaussian hypothesis H_G is computed using eq.(7.1). It involves the second order moment $S_2(\tau)$ only, which we express using eq. (5.24) as

$$S_2(\tau) = \sigma_\tau^2 = \sigma_T^2 \left(\frac{\tau}{T}\right)^{\zeta(2)}\,.$$

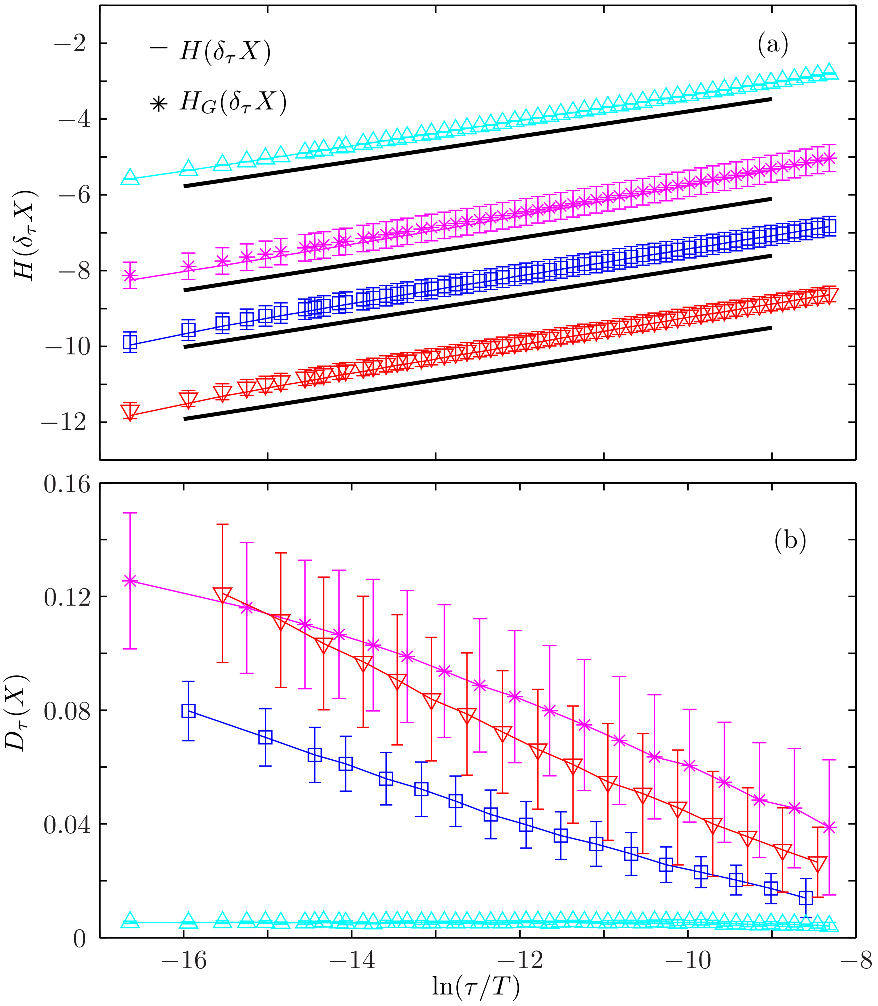

Figure 7.3: a) Entropy H (continuous curve) and entropy under Gaussian hypothesis H_G (symbols) for mono- and multifractal processes. Black lines represent the theoretical slope (1/3) for information expected for turbulence within K41 theory. b) Distance $D_\tau(X) = H_\mathrm{G}(\delta_\tau X) - H(\delta_\tau X)$ from Gaussianity. Four different models are used: fBm (cyan △), MRW (magenta ∗), log-normal RWC (blue □) and log-Poisson RWC (▽ red).

	fBm	MRW	log-N	log-P
c_1	1/3	0.371	0.371	0.381
c_2	0	0.025	0.025	0.036
$\zeta(2)/2$	1/3	0.345	0.345	0.345
$\hat{c}_1$	0.333	0.42	0.372	0.382
$\hat{c}_2$	$1e^{-4}$	0.038	0.026	0.035
$\hat{\zeta}(2)/2$	0.332	0.363	0.353	0.356
$\Delta_{\ln(\tau)} H_G(\delta_\tau X)$	0.33±0.01	0.37±0.01	0.35±0.01	0.35±0.01

Table 7.1: The first three lines indicate the values of parameters (c_1 and c_2 and hence $\zeta(2)$) used in the generation. Estimates $\hat{c}_1$, $\hat{c}_2$ and $\hat{\zeta}(2)$ are obtained by classical multifractal analysis. Last line reports the slopes $\Delta_{\ln(\tau)} H_G(\delta_\tau X)$ of the entropy $H_G(\delta_\tau X)$ as a function of $\ln(\tau/T)$, for the four different models, which according to eq.(7.7) provides another estimate of $\zeta(2)/2$.

We then obtain the dependence of H_{G} on the scale τ:

$$H_{\mathrm{G}}(\delta_\tau X) = H_{\mathrm{G}}(\delta_T X) + \frac{\zeta(2)}{2} \ln(\tau/T) \; . \tag{7.7}$$

In figure 7.3(a), we observe that the slope of the curves, which should be $\frac{\zeta(2)}{2}$ is very similar for all processes: we report in table 7.1 the different values we measured, and compare them to the prescribed value (1/3 for fBM and 0.345 for all three multifractal processes). The distribution of information along the scales for the four different models is in agreement with the prescribed Kolmogorov K41 scaling [80].

Up to this point, looking at the entropies, the four models cannot be distinguished in the inertial domain. In figure 7.3(b) we plot the Kullback-Leibler divergence $D_\tau(X)$ as a function of $\ln(\tau/T) = \ln(l/L)$, for scales ranging from $\tau/T = 1/2^{24}$ to $\tau/T = 4096/2^{24}$ where the integral scale is $T = 2^{24}$.

For a monofractal process, the entropy is given by eq.(7.6), and the entropy under Gaussian hypothesis is given by eq.(7.7) with $\mathcal{H} = \zeta(2)/2$, so $D_\tau(X) = H_{\mathrm{G}}(\delta_\tau X) - H(\delta_\tau X)$ is constant and does not depend on the scale τ. If the monofractal process has Gaussian statistics — which defines the fBm — $D_\tau(X) = 0$ by construction. Looking at figure 7.3(b), D_τ for the fBm is not exactly zero; this is due to the bias in the estimation of $H(\delta_\tau X)$ and $H_G(\delta_\tau X)$. This bias is constant across scales, because our procedure was built to use a constant number of points in the range of τ we use.

For the three multifractal processes, $D_\tau(X)$ decreases monotonically when τ increases, and tends to zero when the scale tends to the integral scale. So in the three multifractal models, the PDF of the increments deforms into a Gaussian PDF when approaching the integral scale. Moreover, in figure 7.3(b), we observe that the three processes, which indeed have different statistics, do not converge to zero in the same way. The distance from Gaussianity D_τ, by involving all the moments of the probability distributions, is able to reveal fine differences between processes.

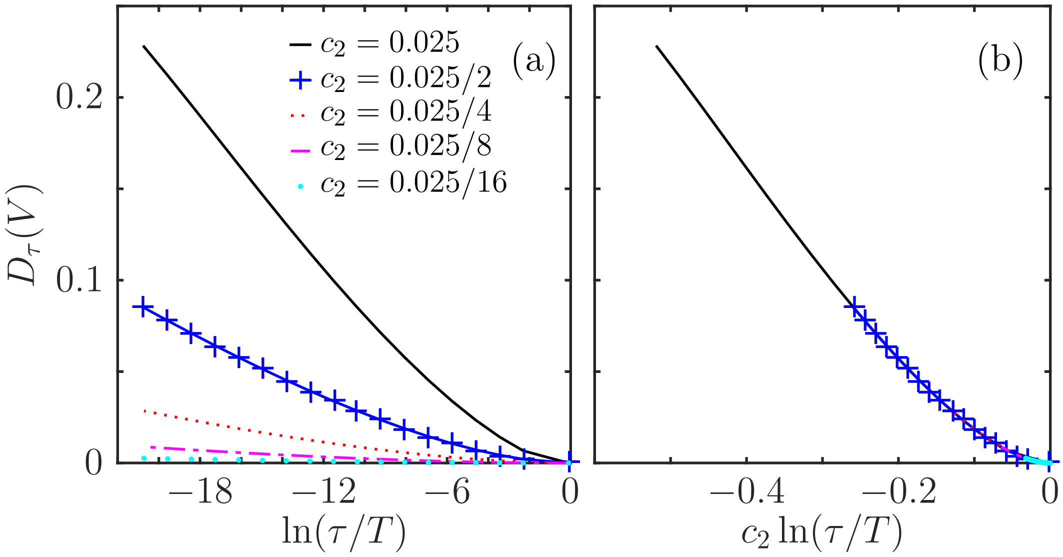

Figure 7.4: Kullback-Leibler divergence D_τ for the log-normal propagator model, for varying values of the log-cumulant c_2, as a function of $\ln(\tau/T) = \ln(l/L)$ (a) or as a function of $c_2 \ln(\tau/T)$ (b).

The synthetic processes used above are good representations of the inertial range only. They do not properly take into account either the dissipative nor the integral scales. Nevertheless, the synthesis imposes an effective integral scale that corresponds to the size of the generated signal. In order to study more precisely the deformation of the PDFs at large scale, we now turn to descriptions that explicitly involve the integral scale.

7.3.3 Phenomenological model : the propagator formalism

First introduced by Castaing [32], the propagator formalism describes the statistics of the Eulerian velocity increment $\delta_l v$ as identical, in the probabilistic sense, to the statistics of the product of two random variables: the large scale fluctuations $\sigma_L \delta$ and the propagator $(l/L)^{\mathfrak{h}}$. The large scale fluctuations are supposed Gaussian, with standard deviation σ_L, and δ is therefore a Gaussian variable with unit variance. The propagator deforms the large-scale statistics when the scale l is reduced below the integral scale L. In the simple situation where no dissipative scale is taken into account, and where the propagator is supposed independent of large scale statistics, one can write formally the PDF of the Eulerian velocity increments $\delta_l v = \sigma_L (l/L)^{\mathfrak{h}} \delta$ as [38]:

$$p_{\delta_l v}(\delta_l v) = \int_{-\infty}^{\infty} \frac{1}{\sigma_L} \left(\frac{l}{L}\right)^{-\mathfrak{h}} \mathcal{P}_\delta \left[\frac{\delta_l v}{\sigma_L} \left(\frac{l}{L}\right)^{-\mathfrak{h}} \right] \mathcal{P}_{\mathfrak{h}}[\mathfrak{h}] d\mathfrak{h} \tag{7.8}$$

where $\mathfrak{h}$ is the Hölder exponent. We have noted $\mathcal{P}_\delta(\delta)$ and $\mathcal{P}_{\mathfrak{h}}(\mathfrak{h})$ the probabilities of the independent random variables δ and $\mathfrak{h}$. The PDF $\mathcal{P}_{\mathfrak{h}}(\mathfrak{h})$ depends only on the singularity spectrum $\mathcal{D}(\mathfrak{h})$. See [38] for a detailed explanation.

We integrate numerically eq.(7.8) to get the PDF of the increments $\delta_l v$, and then compute the KL divergence D_τ by introducing this PDF in eq.(7.5) for several singularity spectra, either log-normal or log-Poisson.

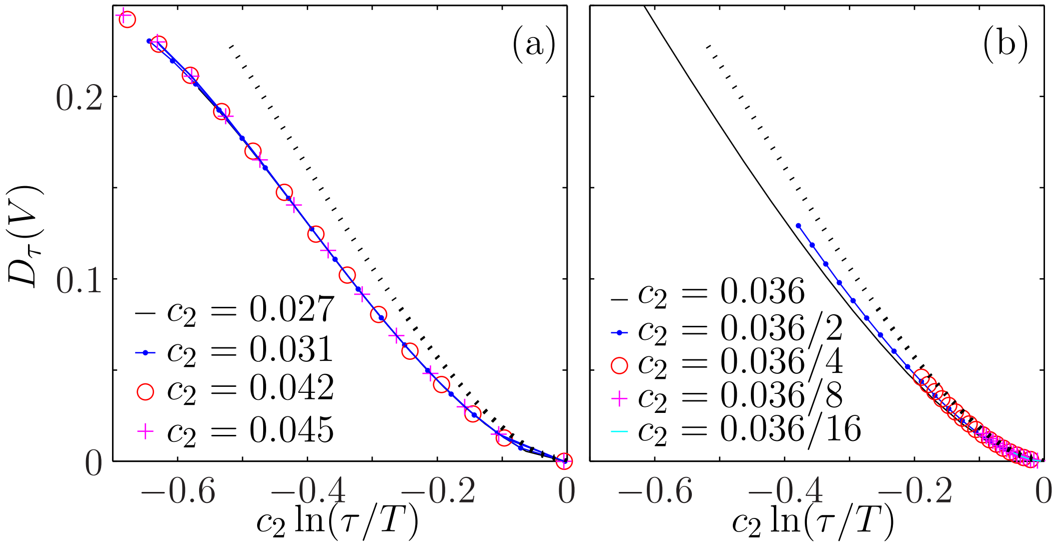

Figure 7.5: Kullback-Leibler divergence D_τ as a function of $c_2 \ln(\tau/T)$ for the log-Poisson propagator model, for varying values of λ (a) and β (b). The dotted line represents the single curve obtained for the log-normal propagator (figure 7.4(b)).

log-normal model . We varied the value of the log-cumulant c_1 and didn't observe any dependence of D_τ on c_1. On the contrary, varying c_2 strongly changes the convergence. Results are presented in figure 7.4(a). We observe and report in figure 7.4(b) that curves for different values of c_2 can be collapsed into a single curve when plotted as a function of $c_2 \ln(\tau/T) = c_2 \ln(l/L)$.

To understand this scaling behavior, we performed a saddle-node expansion of expression (7.8) in the log-normal case, and obtained the following simplified expression for the PDF of the normalized increments $y = \delta_l v/\sigma_l$ at scale l:

$$p_y(y) = \frac{e^{\frac{3}{2}c_2 x}}{\sqrt{2\pi}} \frac{e^{-\frac{2W+W^2}{8c_2 x}}}{\sqrt{1+W}} \tag{7.9}$$

where we have noted $x \equiv -\ln(l/L)$ the logarithmic scale, and W the value of the Lambert W-function of argument $2c_2 x y^2 e^{4c_2 x}$. Eq.(7.9) is a non-Gaussian PDF which converges to the Gaussian PDF of variance σ_L^2 when $x \to 0$. From eq.(7.9), the PDF of the increments only depends on $\ln(l/L) = \ln(\tau/T)$ and c_2 via the product $c_2 \ln(l/L)$. As a consequence, the entropy of the increments depends on the scale l as $c_2 \ln l/L$ only. This implies that the KL divergence D_τ for the log-normal process has the scaling observed in figure 7.4(b).

log-Poisson model We varied independently γ, λ and β. We didn't observe any change of D_τ when γ was varied. This can be understood as γ only changes the value of c_1 (see eq.(5.31)), which does not impact D_τ. Varying λ changes the convergence, as this amounts to change c_2 (see eq.(5.32)), but we observe again that D_τ depends only on $c_2 \ln(\tau/T)$, see figure 7.5(a). This can be understood by noting that all log-cumulants are linear in λ; thus varying λ amounts to a change of c_2 while keeping higher order cumulants within the same ratio. On the contrary, varying β has more impact on the convergence, and the re-scaling in $c_2 \ln(\tau/T)$ is then not perfect, albeit still relevant, see figure 7.5(b). This can be understood by noting that changing β does not only change c_2, but also the ratio of all higher order cumulants.

Comparison between models The rescaling in $c_2 \ln(\tau/T)$ — which absorbs most, if not all, the dependence of D_τ on c_2 — allows a direct comparison of models. As can be seen in figure 7.5, all curves obtained with the log-Poisson propagator model clearly differs from the ones obtained with the log-normal propagator, especially for smaller scales. This probably results from the presence of higher order log-cumulants c_p, $p > 2$ in the log-Poisson propagator. As a consequence, whatever the choices of the propagators parameters, the KL divergence behaves very distinctly in the log-normal and log-Poisson models.

Comparison with turbulence data Amongst open questions regarding statistical descriptions of Eulerian turbulence is the choice of a log-normal or log-Poisson modeling of its multifractal nature. We of course want to address this issue, and we explore both the flatness and the KL divergence to compare the two propagator models with experimental data in figure 7.6. In both models, parameters are set to the values acknowledged for turbulence (see section 5.2). For experimental data, we remove the bias from the KL divergence estimation by subtracting the small constant value that we measured for a fBm signal, see figure 7.3(b). As the c_2 value for turbulence is *a priori* unknown, we do not rescale the x-axis with c_2. Let us remark though that the c_2 value we have used in the log-normal propagator ($c_2 = 0.025$) is exactly the one measured in experimental data, using multifractal analysis [38].

In the inertial range, although the flatness behaves differently for the two propagator models, the difference is small and remains within the error bars of the estimation performed on experimental data. On the contrary, when looking at the KL divergence, our results show a much better agreement of the log-normal model with the experimental data. Although this may be due to the very appropriate choice of c_2 in the log-normal model, the log-Poisson model does not allow such a choice and fixes all the log-cumulants [178]. As a consequence, we can state that the deformation of the experimental velocity increments PDF in the inertial range is better modelled by a multiplicative cascade with log-normal multipliers.

Similar results are obtained by using wavelet multifractal analysis, which allows to probe both sides of the singularity spectrum, see figure 5.3 b). Measures of $D(\mathfrak{h})$ for experimental turbulent velocity together with log-normal and log-Poisson models suggests that log-normal can model turbulence better than log-Poisson does [216, 10]. Nevertheless, the results obtained by performing this analysis depend a lot on the chosen parameters in the estimation, and then the interpretation is difficult.

In the dissipative range, *i.e.*, for smaller scales $\ln(\tau/T) \lesssim -5$, we observe a rapid increase of D_τ for the experimental data, which the two models (continuous lines in figure 7.6) fail to reproduce. This is expected, as both propagator models were proposed to describe the inertial range only, and as such do not incorporate any modeling of the dissipative scales. To probe the dissipative range, we use the extension described in [38], and indicate the results for both propagators with dashed lines in figure 7.6. The rapid increase of intermittency in the dissipation range is captured by the flatness and the KL divergence. Although great care should be taken when commenting on the lowest

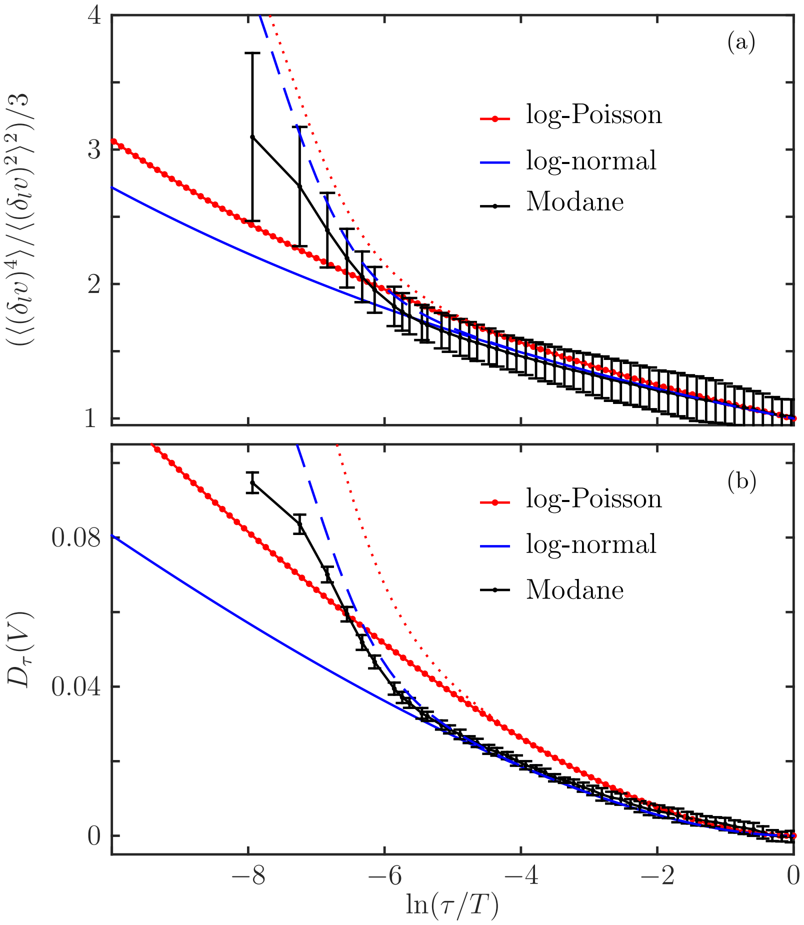

Figure 7.6: Comparison of the two propagator models of turbulence (log-normal (continuous blue line) and log-Poisson (red dotted line)) with Modane experimental data (in black with errorbars), as a function of $\ln(\tau/T)$: (a) normalized flatness, (b) KL divergence. The dashed (blue) and dotted (red) lines represent the modified models that take into account the dissipative range [38].

accessible scales in the experimental data (mainly due to the acquisition process), we observe once again that the KL divergence indicates — much more clearly than the flatness does — that the log-normal propagator model is closer to the experimental observations.

7.4 Discussion and Conclusions

We have measured the Shannon entropy of the Eulerian turbulent velocity increments, and studied its dependence on the scale. We have recovered the same qualitative results as with the entropy rate of the velocity (chapter 6), in perfect agreement with K41 theory. A closer look at the entropy of the increments, and especially a comparison with its Gaussian approximation, which only takes into account the variance of the signal – exactly as the PSD does — allows a much finer description and in particular, a measure of intermittency, as introduced in KO62.

We have proposed a quantitative measure of intermittency. Although some quantities were already used as an intermittency coefficient, most, if not all, were ratios of structure functions [66], and as such, they were depending on the chosen ratio: flatness, hyper-flatness [8] or higher order ratios. We interpret intermittency as the distance from Gaussianity, and measure it as D_τ, the Kullback-Leibler divergence between the complete PDF $p(\delta_\tau V)$ and its Gaussian approximation $p_{\mathrm{G}}(\delta_\tau V)$; the first involves all the statistical moments while the second one only depends on the variance. Our measure of intermittency, by comparing complete PDFs, takes into account all the moments of the distributions, which leaves no room for ambiguity on the choice of the moments.

We have checked the robustness of our approach by analyzing several experimental datasets, from two different experimental setups, and with varying Reynolds numbers.

The quantity D_τ is not only able to measure intermittency in turbulence, but also to discriminate very easily monofractal from multifractal processes. Furthermore, the evolution of D_τ with the scale depends on the process: this provides a much more precise characterization of the process than the bare set of log-cumulant values (c_1, c_2) given by a regular multifractal analysis. This may be exploited to discriminate log-Poisson from log-normal models of intermittency in turbulence.

We have investigated the dependence of D_τ on the log-cumulants. D_τ does not depend on c_1 and we have captured its dependence on c_2, and especially how it affects the convergence to 0 at large scales. Because D_τ appears to mainly depend on $c_2 \ln(\tau/T)$, we can state that the speed of the deformation of the PDF, starting from a Gaussian at large scale L, depends on c_2. For a given scale l/L, or equivalently τ/T, the deformation of the PDF, and hence the intermittency, is an increasing function of c_2. Conversely, for a fixed value of c_2, the influence — or reminiscence— of the integral scale persists down to scales l/L smaller and smaller when c_2 is reduced. Because the typical c_2 of turbulence is small, the influence of the integral scale persists in the inertial domain, down to the dissipative domain, unless the Reynolds number tends to arbitrarily large values. We have shown that D_τ depends on higher order log-cumulants c_p, for $p > 2$, by looking at the special case of log-Poisson statistics (figure 7.5). The dependence seems weak, but is nevertheless present, and could be exploited.

Our measure of intermittency substantially differs from the existing measure involving the flatness. KL divergence gives a sharper contrast between propagator models; this is

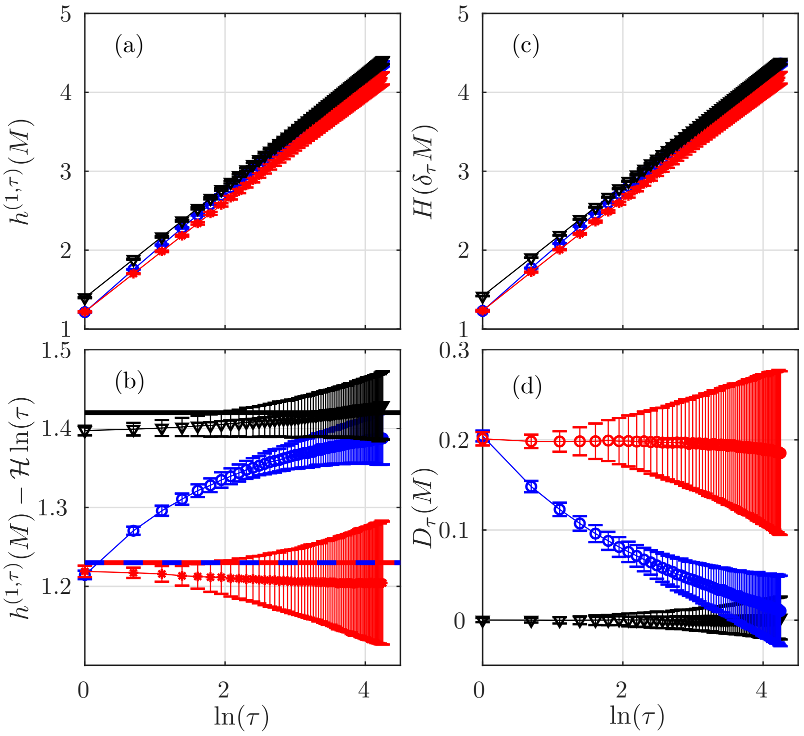

Figure 7.7: a) $\bar{h}_T^{(1,\tau)}(M)$ b) $H(\delta_\tau M)$ c) $\bar{h}_T^{(1,\tau)}(M) - \mathcal{H}\ln(\tau)$ and d) $D_\tau(M)$ for an standard (blue) and even (red) log normal signals and a fBm (black) in function of the logarithm of τ. Number of neighbors $k = 5$ and data size $N = 2^{16}$. The standard deviation is obtained with 100 realizations. The straight lines in (b) indicate the theoretical values of the entropy of the processes.

not surprising as KL divergence offers a complete perspective on the PDF and all its moments. In addition, the errorbars of the KL divergence of experimental data are sufficiently small to allow a quantitative comparison of experiments with model predictions. This comparison was not instructive using previously existing tools. Our results show that KL divergence offers quantitative arguments in favor of a log-normal description of the propagator for modelling the distribution of Eulerian velocity increments across scales.

Although we have put a strong emphasis on turbulence, we want to point out that our approach is extremely general and should find successful applications in many other fields. It should prove particularly interesting for non-Gaussian processes, the most common in Nature and Society. Figure 7.7 shows the performance of this approach (right column), together with the entropy rate across scales $\bar{h}_T^{(m,\tau)}$ presented in chapter 4 (left column), when analysing the non-stationnary processes presented in chapter 4. Both methodologies describe the processes in a very similar way. However, while the methodology presented in chapter 4 requires the previous knowledge of $\mathcal{H}$ to obtain figure 7.7 b), the approach presented in this chapter doesn't require any *a priori* knowledge of the process.

We can conclude that, the self-similarity properties of any process can be analyzed combining $H(\delta_\tau)$ and D_τ. The local intermittency measure that D_τ provides can be used to characterize the process at any scale.

CHAPTER 8

Wiener Causality across scales of turbulence

The complete characterization and understanding of the energy cascade in turbulence is far from being reached. Is this energy cascade due to the splitting of large eddies into smaller ones? How much time does the energy take to go from one scale to another? Does this time depend on the range of scales? Within the current research frameworks the discussion is open, and the conclusions are not always clear. So, we propose in this chapter to look for conclusive results using information theory, and more precisely to use transfer entropy (section 1.6.2), to analyse turbulent velocity measures and synthetic processes modelling turbulence. We then propose to study interactions between scales in turbulence from this new viewpoint.

Transfer entropy, defined in eq.(1.31), is a measure of Wiener causality between time series, and has been frequently interpreted as a measure of information flows, (section 1.6). Then, the question naturally arises, do information flows between scales of turbulence exist? or, at least, are there Wiener causality relationships between different scales of turbulence? And if we can measure them, do they bring information about the energy cascade?

In order to probe the existence of Wiener causality relationships between scales of turbulence, we study different time series: on one hand, the experimental grid turbulent velocity measures presented in section 6.1, on the other hand fractional Brownian motions, (section 4.2), with different Hurst exponents. The third order structure function of turbulent velocity data is non-zero and thus it shows the existence of an energy cascade. This is not the case for fBm, where no energy cascade takes place.

8.1 Definitions of the scale

In order to measure causality relationships between scales of a time series, first, we have to define these scales. This step is very delicate, as the definition of the scales can influence the transfer entropy measures between them, and so, distort the results. We define the scales of a time-series $X(t)$ by applying a transformation f_τ on it. Then, we define

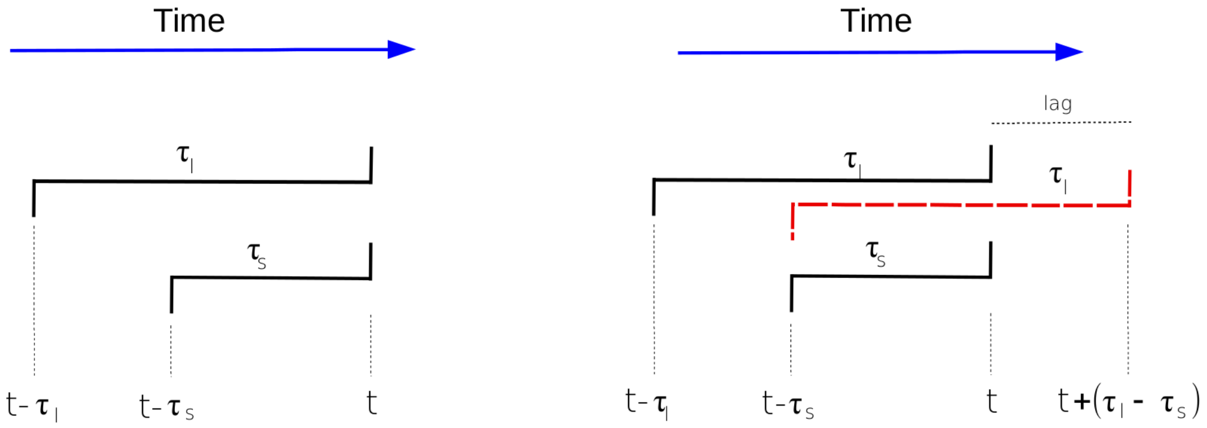

Figure 8.1: Left:Increments initial emplacement. Right: Increments coincidence. The red increment is the lagged one.

the diferent scales of $X(t)$ as $f_\tau(X(t))$ where τ indicates the size of the scale. Therefore, the scales are also time-series. However, the transformation f_τ can itself introduce Wiener causality relationships between its different versions f_τ, $f_{\tau'}$. Relationships that are unconnected to the physical system.

We can generate the different scales of a time-series by using different transformations: increments, wavelets, filtering, embedding ... Each of these transformations lead to different time series defining the scales of the initial time series $X(t)$. As a consequence, even if the transfer entropy interpretation remain constant, concerning Wiener causality between time series, the question is: how much part of the measure is due to the way of generating the scales? Can we really talk about Wiener causality between scales or just between these transformed time series?

To answer these questions, we perform a complete study of transfer entropy between scales of a time series by using three different scale definitions: increments, filtering and embedding. Having seen the wide range of possibilities in the definition of scales, we have decided to use increments, completely localized in the direct space, filtered signals, completely localized in the Fourier space, and embedded signals, in which the scale is defined by a couple of points of the initial time series $X(t)$ separated by τ.

8.1.1 Increments

From an initial time-series, we generate different scales by generating increments of different sizes. The length of the increments indicates the size of the scale. The corresponding transformation $f_\tau(X(t))$ is defined as:

$$f_\tau(X(t)) = \delta_\tau X(t) = X(t) - X(t-\tau) \tag{8.1}$$

Then, for two different scales we have two increments $\delta_{\tau_1}X(t) = X(t) - X(t-\tau_1)$ and $\delta_{\tau_2}X(t) = X(t) - X(t-\tau_2)$ defined in order to fix the same present t, for both scales, see figure 8.1 (left).

In the example of figure 8.1 (right), transfer entropy definition of eq.(1.31) is a calculation of the information of the red increment ($\delta_{\tau_l}X(t+lag)$) that is contained in the small increment ($\delta_{\tau_s}X(t)$) but not in the large one ($\delta_{\tau_l}X(t)$).

We define lag as the time delay which separates a future sample at $t+lag$ from the present one at t. We interpret it as a time of interaction. With the above definition of transfer entropy, the delay lag is the same between the caused increment ($\delta_{\tau_l}X(t+lag)$) and both, the condition ($\delta_{\tau_l}X(t)$) and the causal increment ($\delta_{\tau_s}X(t)$). We can see in figure 8.1(right) the effect of coincidence of increments when $lag=\tau_l-\tau_s$ and $lag=\tau_l$, that will define the emplacement of the peaks of transfer entropy between scales.

In the definition of transfer entropy of eq.(1.33), lag only indicates the distance between the caused ($\delta_{\tau_l}X(t+lag)$) and the causal increment ($\delta_{\tau_s}X(t)$). With this definition, the distance between the caused ($\delta_{\tau_l}X(t+lag)$) and the conditioned increment is always the smallest possible. The effect of coincidence of increments, when $lag=\tau_l-\tau_s$ and $lag=\tau_l$, still defines the emplacement of the peaks of transfer entropy between scales.

This definition of scale allow to precisely localize the position and size of the scale in the direct space, nevertheless the localization of the scale in the Fourier space is unclear.

8.1.2 Filtering

Another way to define a scale is to filter the initial time series. In order to do that, three different kind of causal filter transformations have been used: Low pass (LP), defined in eq.(8.2), High Pass (HP), defined in eq.(8.3) and Band Pass (BP) filter, defined as the addition of a High and a Low Pass. The idea behind this generation of scales is to completely localize the scales in frequency, but not at all in space, so, the contrary to the increments case.

$$F(f)_{LP}=\frac{1}{1+i\frac{f}{fc}} \tag{8.2}$$

$$F(f)_{HP}=\frac{i\frac{f}{\mathrm{fc}}}{1+i\frac{f}{fc}} \tag{8.3}$$

where fc is the cutoff frequency, indicating the frequencies contained in the filtered signal. In the case of the LP filter the frequencies contained in the signal are the frequencies smaller than fc, in the case of HP filter the frequencies higher than fc, and in the case of a BP filter the frequencies between fc_{HP} and fc_{LP}.

For all the above filters, LP, HP and BP, we consider that the higher the cutoff frequency the smaller the scale generated.

8.1.3 Embedding

Increasing the dimension of the initial time series also allows to define a scale. We use the embedding procedure proposed by Takens:

$$f_\tau(X(t))=X^{(2,\tau)}(t)=(X(t),X(t-\tau)) \tag{8.4}$$

Thus, we define a scale τ as a 2-dimensional vector composed by two samples of the initial time series separated by τ. As in the case of the increments, the effect of coincidence of samples, will define the emplacement of the peaks of transfer entropy between scales (figure 8.1 (right)).

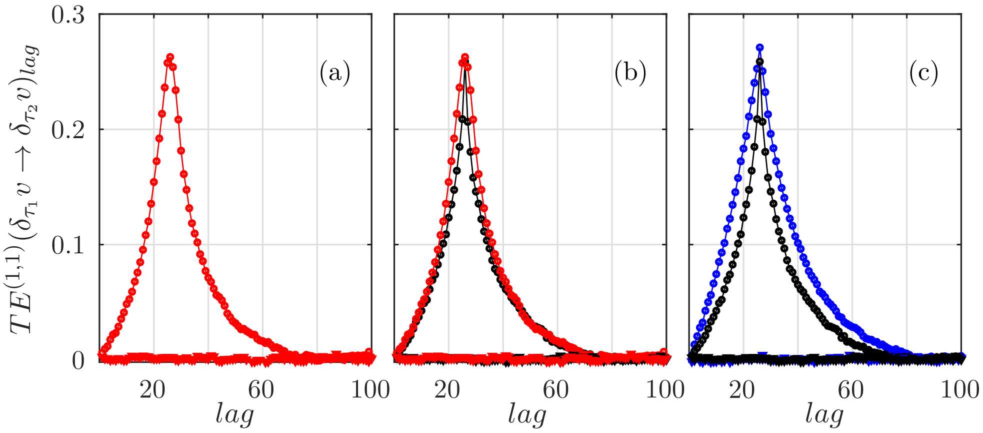

Figure 8.2: Transfer entropy between scales δ_{τ_l} and δ_{τ_s}, with $\tau_l = 92$ and $\tau_s = 66$m for turbulence (red), fBm with $\mathcal{H} = 0.3$ (black), and fBm with $\mathcal{H} = 0.5$ (blue). Circles indicate the TE from the small to the large scale, and triangles indicate the TE from the large to the small scale. a) Turbulence b) Comparison between turbulence and fBm with $\mathcal{H} = 0.3$ c) Comparison between fBm with $\mathcal{H} = 0.3$ and fBm with $\mathcal{H} = 0.5$

8.2 Results for different scale functions

In this section, we characterize the information flow between scales for three different systems: Modane velocity data, a fBm with $\mathcal{H} = 0.3$ and a fBm with $\mathcal{H} = 0.5$ and for the three different ways of generating the scales presented before.In order to analyse the dynamics of information in a given system, we choose an observable of the system, in this section velocity, and we build the scales by applying the corresponding function f_τ to the observable. Then, we calculate the TE between two scales, from the larger to the smaller and vice versa.

8.2.1 Increments

We build the scales of $V(t)$ by generating the increments of the observable. We calculate the TE, defined in eq.(1.31), where $X = \delta_{\tau_1}V(t)$ and $Y = \delta_{\tau_2}V(t)$.

Figure 8.2 shows that TE is different from zero only from small scales to large ones, and is zero from large to small scales. This can be seen for turbulent velocity in figure 8.2 a), but it's general for any system, see figures 8.2 b) ,8.2 c). For all the studied systems: turbulence and fBm with different Hurst exponents the small scales Wiener cause the large ones.

Figures 8.2 a) b) and c) also show the behavior of the time of interaction between scales (lag). We find that the maximum intensity of the interaction between two scales appears at time delay $lag = |\tau_l - \tau_s|$. The time of interaction between two scales depends directly on the difference of the length of the scales; the larger the diference of length the longer the time of interaction. This result is directly related to the definition of the scales as increments.

Another general result is, the closer the scales, the strong the interaction, *i.e.* the interaction between two scales is stronger when the sizes of the scales interacting are

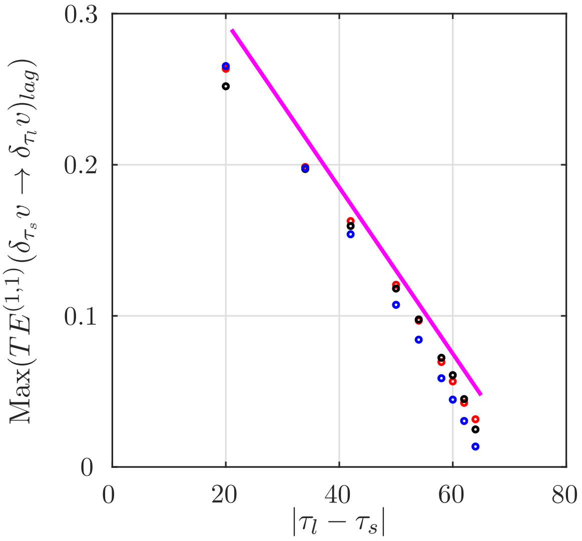

Figure 8.3: Maximum amplitude of the interaction between two scales, τ_{large} and τ_{small}, in function of the time of interaction ($\approx \tau_{large} - \tau_{small}$). τ_{large} is fixed to 66 and τ_{small} varies. Black is for fBm $\mathcal{H} = 1/3$ blue is for fBm $\mathcal{H} = 1/2$ and red is for 3D-Turbulence. The magenta line is guideline of slope 0.0055.

closer, see figure 8.3.

We now consider separately each system. It's well known that a fBm with Hurst exponent $\mathcal{H} = 1/3$ models the distribution of energy of turbulence in the inertial range. We compare the results obtained for two interacting scales of turbulent velocity, both of them in the inertial region, and the same interacting scales of a fBm with Hurst exponent $\mathcal{H} = 1/3$. We find the same time of interaction for both systems, as we expected from the coincidence of increments at $lag = \tau_l - \tau_s$. In addition, the maximum of the intensity of the interaction is the same for both systems. However, the way of reaching the maximum value of the interaction is different, being smoother for the turbulent system, see figure 8.2 b). This smoothness can be due to the dissipation existing in turbulence but not in the fBm, or it can be due to intermittency or even to the energy cascade. Another observation from figure 8.2 is the non-symmetry of the interaction between the two sides of the maximum of the interaction, *i.e.* the way of reaching the maximum is different from the left and from the right, but it seems to be symmetric when we are very close to the maximum for both systems turbulence and fBm.

If we compare two fBm with $\mathcal{H} = 1/3$ and $\mathcal{H} = 1/2$ we see that there are differences in the measures, see figure 8.2 c). The Wiener causality relationships between scales depend on the self-similarity exponent of the system.

8.2.2 Scales obtained with filtering

In this section we generate scales by filtering signals. We filter the same turbulent velocity signal analysed above, and the same fBm with $\mathcal{H} = 0.3$.

We generate the scales using three different filters: low pass, band pass and high pass. Using LP and BP filtering we obtain that the small scales cause the large ones for both

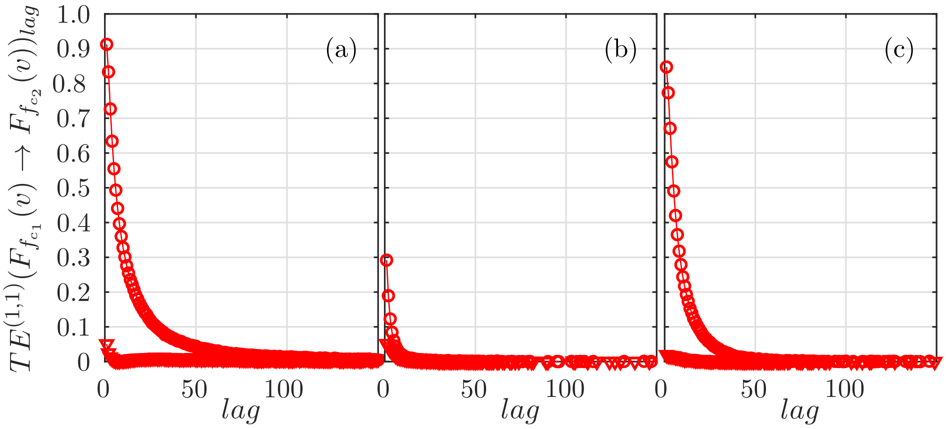

Figure 8.4: Transfer entropy between scales of turbulent velocity. Red circles indicate the TE from the small scale ($\frac{1}{fc} = 10$) to the large one ($\frac{1}{fc} = 100$). Red triangles indicate the TE from the large scale to the small one. The scales are generated using three different filters: a) Low pass filter, b) High Pass filter, c) Band Pass filter.

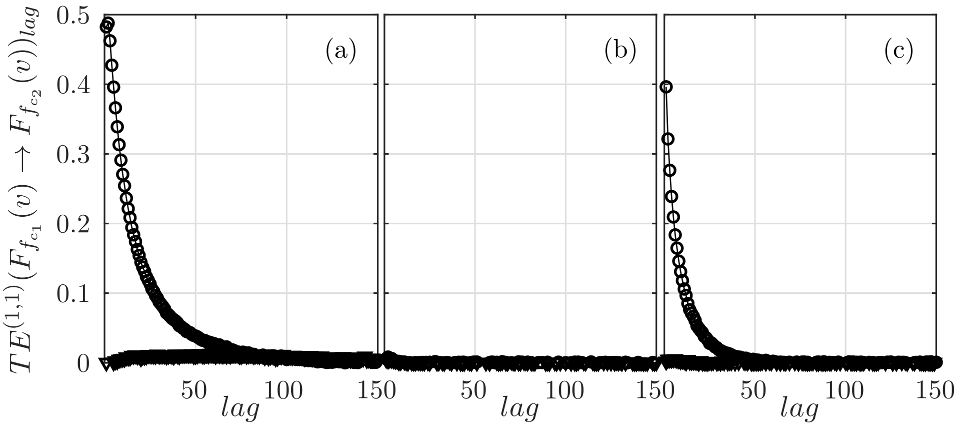

Figure 8.5: Transfer entropy between scales of fBm with $\mathcal{H} = 0.3$. Black circles indicate the TE from the small scale ($\frac{1}{fc} = 10$) to the large one ($\frac{1}{fc} = 100$). Black triangles indicate the TE from the large scale to the small one. The scales are generated using three different filters: a) Low pass filter, b) High Pass filter, c) Band Pass filter.

systems: turbulence (figures 8.4 a) and c)) and fBm (figures 8.5 a) and c)). In the case of the HP filtering we obtain no clear causality relationships between scales, see figure 8.4 b) and figure 8.5 b).

In this case, the maximum of the TE between scales is always at $lag = 0$ and the amplitude of the maximum decreases when the scales are closer, see figure 8.6. When scales are obtained by filtering, the cut-off frequency associated with the large scale defines the decay length of the TE measure, see figure 8.4 and figure 8.5.

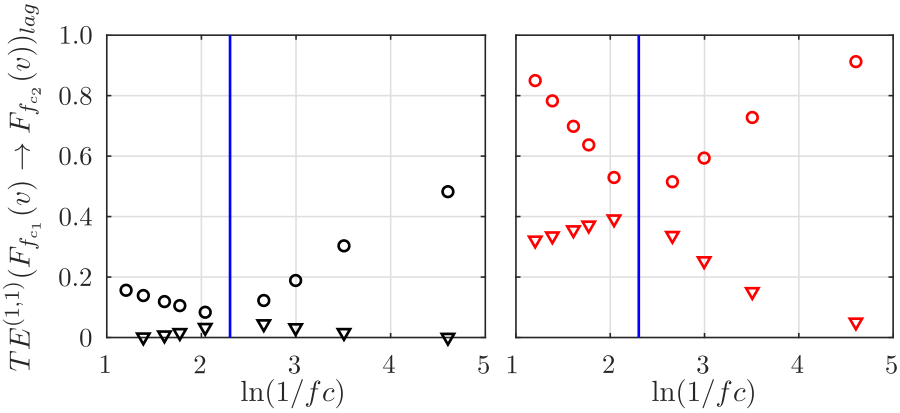

Figure 8.6: Maximum amplitude of TE between the scales generated at $1/fc_{\text{ref}} = 10$ (indicated by the blue line) and $1/fc$ variable (x axis). FBm $\mathcal{H} = 0.3$ in black and turbulent velocity in red. For both plots, triangles indicates TE from large to small scales, and circles TE from small to large scales.

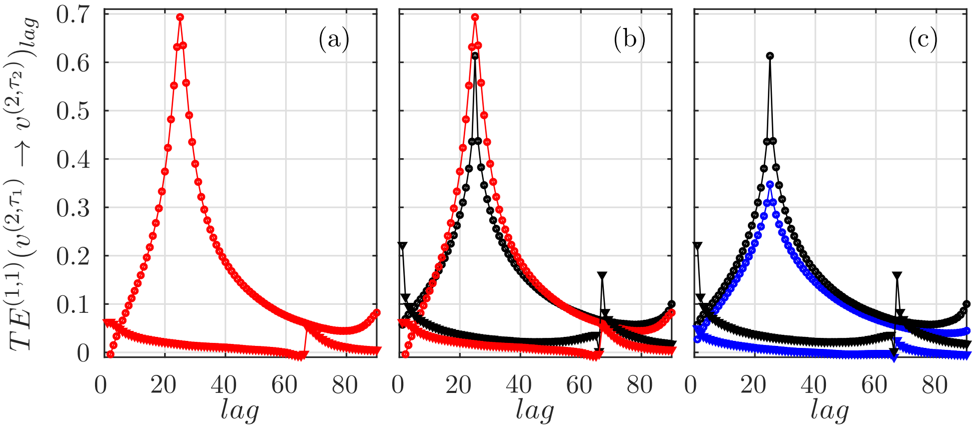

Figure 8.7: Transfer entropy between scales $v^{(2,\tau_l)}$ and $v^{(2,\tau_s)}$, with $\tau_l = 92$ and $\tau_s = 66$m for turbulence (red), fBm with $\mathcal{H} = 0.3$ (black), and fBm with $\mathcal{H} = 0.5$ (blue). Circles indicate the TE from the small to the large scale, and triangles indicate the TE from the large to the small scale. a) Turbulence b) Comparison between turbulence and fBm with $\mathcal{H} = 0.3$ c) Comparison between fBm with $\mathcal{H} = 0.3$ and fBm with $\mathcal{H} = 0.5$

8.2.3 Scales obtained by embedding

The third way of generating the scales is by embedding the data. As in the case of the increments, the maximum amplitudes of TE measures are localized for *lag* values where a sample of the caused vector coincide with a sample of the causal vector. We find qualitatively similar results, with an information flow from small to large scales, (figure 8.7). Even if this time a new peak appears in the TE from large to small scales, this new peak is much smaller than the peak from small to large (figure 8.7). The qualitative results are the same for turbulence and both fBm, but we can observe some differences in the amplitudes and the shapes of the peaks.

8.3 Results for different observables of turbulence

In this section, we study the causality relationships between scales of different observables of turbulence. We use increments as scales and we analyse: the turbulent velocity increments $\delta_\tau v(t)$ analysed above, the increments of the square of the velocity $\delta_\tau v^2(t)$, and the square of the velocity increments $(\delta_\tau v(t))^2$.

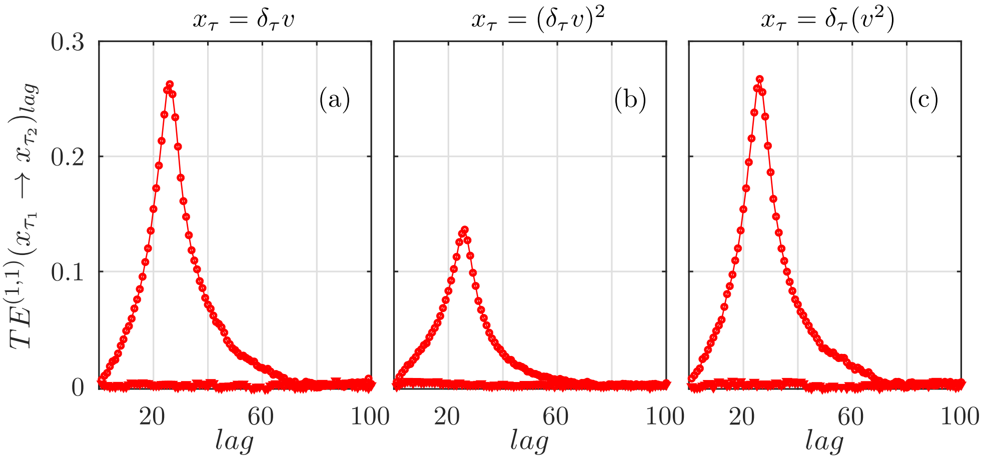

Figure 8.8: Transfer entropy between scales for different observables of turbulence. Circles indicate TE from small to large scales. Triangles indicate TE from large to small scales. a) TE for turbulent velocity increments $\delta_\tau(v(t))$. b) TE for the square of turbulent velocity increments $(\delta_\tau v(t))^2$. c) TE for the increments of the square of the turbulent velocity $\delta_\tau(v(t))^2$.

Figure 8.8, shows the TE between scales for these three observables of turbulence. The qualitative behavior of the causality is common for the three analysed observables: we find that small scales cause large scales, causality from large scales to small scales is null, the maximum amplitude of the causality appears at $lag = \tau_l - \tau_s$, the causality is more important when the scales are closer. Figure 8.8, shows that the TE between the increments of the velocity $\delta_\tau v(t)$ and the TE between the increments of the square of the velocity $\delta_\tau v^2(t)$ are almost identical. However, the TE between the square of the velocity increments $(\delta_\tau v(t))^2$ presents lower values. We find the same results for fBm with both Hurst exponents.

8.4 Conclusions and perspectives

In this chapter, we study several signals: velocity and square of the velocity, obtained from various complex systems: turbulence and fBm with two different Hurst exponents $\mathcal{H}$. We generate the scales of the studied processes using three different methods: computing the increments of the process, filtering the process, or embedding the process. Next, we compute the transfer entropy between scales of the time process, and we find for any system and any scale function the same qualitative result: transfer entropy is larger from small to large scales than from large to small scales. We interpret transfer entropy as a measure of the Wiener causality between two time series, and eventually as a measure of information flows. With this interpretation of TE, our results indicate that the small scales of a complex system cause the large scales.

We show the existence of an information flow from small to large scales, independently of the system, the definition of the scales and the observable. Nevertheless, generating the scales of a process by applying a transformation to the initial process influences the causality measure and we cannot distinguish if this causation comes from the physics of the initial process or from the transformation used to generate the scales. Consequently, our search of a flow of information between scales is not meaningful for the moment. Nevertheless, the measures are not meaningless, and the shapes of the obtained peaks can be used to study the dependences of the process.

In order to confirm the causality from small to large scales, we have tested different causality measures: transfer entropy of eq.(1.32), transfer entropy of eq.(1.33) and directed information. The qualitative results obtained with the different measures are the same: small scales Wiener cause large scales. We propose as a next step, to use experimental time series allowing to generate the scales without the above transformations, or even experimental time series with scale meaning. These experimental processes would allow to isolate the origin of the Wiener causality measures to the physics of the studied processes.

Part III

Biomedical Signal Processing

CHAPTER 9

Information Theory to Probe Intrapartum Fetal Heart Rate Dynamics

Intrapartum fetal heart rate monitoring: Because it is likely to provide obstetricians with significant information related to the health status of the fetus during delivery, intrapartum fetal heart rate (FHR) monitoring is a routine procedure in hospitals. Notably, it is expected to permit detection of fetal acidosis, which may induce severe consequences for both the baby and the mother and thus requires a timely and relevant decision for rapid intervention and operative delivery [35]. In daily clinical practice, FHR is mostly inspected visually, through training by clinical guidelines formalized by the International Federation of Gynecology and Obstetrics (FIGO) [14, 194]. However, it has been well documented that such visual inspection is prone to severe inter-individual variability and even shows a substantial intra-individual variability [94]. This reflects both that FHR temporal dynamics are complex and hard to assess and that FIGO criteria lead to a demanding evaluation, as they mix several aspects of FHR dynamics (baseline drift, decelerations, accelerations, long- and short-term variabilities). Difficulties in performing objective assessment of these criteria has led to a substantial number of unnecessary Caesarean sections [6]. This has triggered a large amount of research world wide aiming both to compute in a reproducible and objective way the FIGO criteria [14] and to devise new signal processing-inspired features to characterize FHR temporal dynamics (cf. [187, 89] for reviews).

Related works: After the seminal contribution in the analysis of heart rate variability (HRV) in adults [3], spectrum estimation has been amongst the first signal processing tools that has been considered for computerized analysis of FHR, either constructed on models driven by characteristic time scales [76, 207, 181] or the scale-free paradigm [63, 55, 57]. Further, aiming to explore temporal dynamics beyond the mere temporal correlations, several variations of nonlinear analysis have been envisaged both for antepartum and intrapartum FHR [59], based, e.g., on multifractal analysis [55], scattering transforms [42], phase-driven synchronous pattern averages [136] or complexity and entropy measures [160, 159, 124]. Interested readers are referred to e.g., [187, 89] for overviews. There have also been several attempts to combine features different in nature by doing multivariate classification using supervised machine learning strategies (cf. e.g., [71, 187, 50, 214, 188]).

Measures from complexity theory or information theory remain, however, amongst the most used tools to construct HRV characterization. They are defined independently from (deterministic) dynamical systems or from (random) stochastic process frameworks. The former led to standard references, both for adult and for antepartum and intrapartum fetal heart rate analysis: approximate entropy (ApEn) [160, 51] and sample entropy (SampEn) [168], which can be regarded as practical approximations to Kolmogorov–Sinai or Eckmann–Ruelle complexity measures. The stochastic process framework leads to the definitions of Shannon and Rényi entropies and entropy rates. Both worlds are connected by several relations, cf., e.g., [87, 122] for reviews. Implementations of ApEn and SampEn rely on the correlation integral-based algorithm (CI) [160, 83], while that of Shannon entropy rates may instead benefit from the k-nearest neighbor (k-NN) algorithm [114], which brings robustness and improved performance to entropy estimation [161, 190, 223].

Labor stages: Automated FHR analysis is complicated by the existence of two distinct stages during labor. The dilatation stage (Stage 1) consists of progressive cervical dilatation and regular contractions. The active pushing stage (Stage 2) is characterized by a fully-dilated cervix and expulsive contractions. The most common approaches in FHR analysis consist of not distinguishing stages and performing a global analysis [46, 214] or of focusing on Stage 1 only, as it is better documented and usually shows data with better quality, cf., e.g., [70, 188]. Whether or not temporal dynamics associated with each stage are different has not been intensively explored yet (see a contrario [186, 131]). However, recently, some contributions have started to conduct systematic comparisons [189, 79].

Goals, contributions and outline: The present contribution remains in the category of works aiming to design new efficient features for FHR, here based on advanced information theoretic concepts. These new tools are applied to a high quality, large (1404 subjects) and documented FHR database collected across years in an academic hospital in France and described in section 9.1. The database is split into two datasets associated each with one stage of labor, which enables us first to assess and compare acidosis detection performance achieved by the proposed features independently at each stage and second to address differences in FHR temporal dynamics between the two stages. Reexamining formally the definitions of entropy rates in information theory, section 1.4 first establishes that they can be split into two components: Shannon entropy, which quantifies static data properties, and auto-mutual information (AMI), which characterizes temporal dynamics in data, combining both linear and nonlinear (or higher order) statistics. ApEn and SampEn, defined from complexity theory, are then explicitly related to entropy rates and, hence, to AMI (cf. section 9.2.2). Estimation procedures for Shannon entropy, entropy rate and AMI, based on k-nearest neighbor (k-NN) algorithms [114], are compared to those of ApEn and SampEn, constructed on correlation integral algorithms [160, 47, 83]. Acidosis detection performances are reported in section 9.3.1. Results are discussed in terms of quality versus analysis window size, k-NN or correlation integral-based procedures and differences between Stages 1 and 2. Further, a longitudinal study consisting of sliding analysis in overlapping windows across the delivery process permits showing that processes characterizing Stages 1 and 2 are different (section 9.3.4).

9.1 Datasets: Intrapartum Fetal Heart Rate Times Series and Labor Stages

Data collection: Intrapartum fetal heart rate data were collected at the academic Femme-Mère-Enfant hospital, in Lyon, France, during daily routine monitoring across the years 2000 to 2010. They were recorded using STAN S21 or S31 devices with internal scalp electrodes at 12-bit resolution, 500-Hz sampling rate (STAN, Neoventa Medical, Sweden). Clinical information was provided by the obstetrician in charge, reporting delivery conditions, as well as the health status of the baby, notably the umbilical artery pH after delivery and the decision for intervention due to suspected acidosis [56].

Datasets: For the present study, subjects were included using criteria detailed in [56, 188], leading to a total of 1404 tracings, lasting from 30 min to several hours. These criteria essentially aim to reject subjects with too low quality recording (e.g., too many missing data, too large gaps, too short recordings, etc.). As a result, for subjects in the database, the average fraction of data missing in the last 20 min before delivery is less than 5%. The first goal of the present work is to assess the relevance of new information theoretic measures; their robustness to poor quality data is postponed for future work.

The measurement of pH, performed by blood test immediately after delivery, is systematically documented and used as the ground-truth: When pH ≤ 7.05, the newborn is considered has having suffered from acidosis and is referred to as acidotic (A, pH ≤ 7.05). Conversely, when pH > 7.05, the newborn is considered not having suffered from acidosis during delivery and is termed normal (N, pH > 7.05). In order to have a meaningful pH indication, we retain only subjects for which the time between end of recording and birth is lower than or equal to 10 min.

Following the discussion above on labor stages, subjects are split into two different datasets. Dataset I consists of subjects for which delivery took place after a short Stage 2 (less than 15 min) or during Stage 1 (Stage 2 was absent). It contains 913 normal and 26 acidotic subjects. Dataset II gathers FHR for delivery that took place after more than 15 min of Stage 2. It contains 450 normal and 15 acidotic subjects.

Beats-per-minute time series and preprocessing: For each subject, the collection device provides us with a digitalized list of RR-interarrivals Δ_k in ms. In reference to common practice in FHR analysis and for the ease of comparisons amongst subjects, RR-interarrivals are converted into regularly-sampled beats-per-minute times series, by linear interpolation of the samples $\{\ldots, 36{,}000/\Delta_k, \ldots\}$. The sampling frequency has been set to $f_s = 10$ Hz as FHR do not contain any relevant information above 3 Hz.

9.2 Methods

Outline: We describe in this section the five features that we use to analyze heart rate signals. We propose to apply information theory, as defined by Shannon, to the analysis of cardiac signals. We do so by computing three features from information theory, the Shannon entropy, see section 1.3.1, the Shannon entropy rate, see section 1.4.1 and the auto-mutual information, see section 1.4.2. The first section reports the definitions of two features rooted in complexity theory: approximate entropy (ApEn) and sample entropy (SampEn), which are classically used in cardiac signal analysis. Although we use them in practice only as benchmarks, we devote the second section to their relation with the new features we propose.

Information theory and complexity theory only differ in the nature of the objects under study. Information theory, on the one hand, aims to analyze random processes and defines functionals of probability densities. Complexity theory, on the other hand, aims to analyze signals produced by dynamical systems and assumes the existence of ergodic probability measures to describe the density of trajectories in phase space, so that they can be manipulated as probability densities. In this spirit, we consider throughout this chapter the signals to analyze as random processes, although they indeed originate from a dynamical system.

Assumptions: For the sake of simplicity in the description of the features and for practical use, we assume that signals are monovariate (unidimensional) and centered (zero mean) because the five features we use are independent of the first moment of the probability density function. We also assume that signals are stationary. Although this may seem at first a very strong assumption, it is very reasonable when examining time windows smaller than the natural time scale of the evolution between Stages 1 and 2, as we discuss in section 9.3.4, and larger than events such as contractions. Finally, we also assume that the signals contain N points, sampled at a constant frequency. All estimates depend on N via finite size effects. In the following, we do not mention this dependence explicitly in the notations and only compare features computed over the same window size.

Time-embedding: Because we are interested in the dynamics of the signal, we use the delay-embedding procedure introduced by Takens [197] in the context of dynamical systems, see section 1.2.

9.2.1 Features from Complexity Theory

In the 1960s, Kolmogorov and Sinai adapted Shannon's information theory to the study of dynamical systems. The divergence of trajectories starting from different, but undistinguishable initial conditions can be pictured as creating uncertainty, so creating information. Kolmogorov complexity (KC), also known as the Kolmogorov–Sinai entropy and denoted $h_{\mathrm{KS}}(\rho)$ in the following, measures the mean rate of creation of information by a dynamical system with ergodic probability measure ρ. KC is constructed exactly as the Shannon entropy rate from information theory, using eq.(1.21) and the same functional form as in eq.(1.7), but using the density ρ of trajectories in phase space instead of the probability density p. In the early 1980s, the Eckmann–Ruelle entropy $K_2(\rho)$ [83, 60] was introduced following the same steps, but using the functional form of the Rényi order-two entropy [166]. The interest of K_2 relies in its easier and hence faster computation from experimental time series, which was at the time a challenging issue.

Kolmogorov–Sinai and Eckmann–Ruelle entropies: The ergodic theory of chaos provides a powerful framework to estimate the density of trajectories in the phase space of a chaotic dynamical system [60]. For an experimental or numerical signal, it amounts to assimilating the phase space average to the time average. Given a distance $d(.,.)$, usually defined with the $\mathcal{L}^2$ or the $\mathcal{L}^\infty$ norm, in the m-dimensional embedded space, the local quantity:

$$C_i^m(\epsilon) = \frac{\text{number of } j \text{ such that } d\left(x_i^{(m,\tau)}, x_j^{(m,\tau)}\right) \leq \epsilon}{N-m+1}, \tag{9.1}$$

provides, up to a factor ϵ, an estimate of the local density ρ in the m-dimensional phase space around the point $x_i^{(m,\tau)}$. The following averages:

$$\Phi^m(\epsilon) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \ln C_i^m(\epsilon), \tag{9.2}$$

$$C^m(\epsilon) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} C_i^m(\epsilon), \tag{9.3}$$

are then used to provide the following equivalent definitions of the complexity measures [60]:

$$h_{\mathrm{KS}}(\rho) = \lim_{\epsilon \to 0} \lim_{m \to \infty} \lim_{N \to \infty} \left(\Phi^m(\epsilon) - \Phi^{m+1}(\epsilon)\right), \tag{9.4}$$

$$K_2(\rho) = \lim_{\epsilon \to 0} \lim_{m \to \infty} \lim_{N \to \infty} \ln\left(\frac{C^m(\epsilon)}{C^{m+1}(\epsilon)}\right). \tag{9.5}$$

Approximate Entropy

Approximate entropy (ApEn) was introduced by Pincus in 1991 for the analysis of babies' heart rate [158]. It is obtained by relaxing the definition (9.4) of h_{KS} and working with a fixed embedding dimension m and a fixed box size ϵ, often expressed in units of the standard deviation σ of the signal as $\epsilon = r\sigma$. ApEn is defined as:

$$\mathrm{ApEn}(m, \epsilon) = \Phi^m(\epsilon) - \Phi^{m+1}(\epsilon). \tag{9.6}$$

On the practical side, and in order to have a well-defined $\Phi^m(\epsilon)$ in eq.(9.2), the counting of neighbors in the definition (9.1) allows self-matches $j = i$. This ensures that $C_i^m(\epsilon) > 0$, which is required by eq.(9.2). ApEn depends on the number of points N in the time series. Assuming N is large enough, we have:

$$\lim_{\epsilon \to 0} \lim_{m \to \infty} \mathrm{ApEn}(m, \epsilon) = h_{\mathrm{KS}}. \tag{9.7}$$

We interpret ApEn as an estimate of the m-order Kolmogorov–Sinai entropy h_{KS} at finite resolution ϵ. The larger N, the better the estimate. More interesting is that the non-vanishing value of ϵ in its definition makes ApEn insensitive to details at scales lower than

ϵ. On the one hand, this is very interesting when considering an experimental (therefore noisy) signal: choosing ϵ larger than the rms of the noise (if known) filters the noise, and ApEn is then expected to measure only the complexity of the underlying dynamics. This was the main motivation of Pincus and explains the success of ApEn. On the other hand, not taking the limits $\epsilon \to 0$ and $m \to \infty$ makes ApEn an ill-defined quantity that has no reason to behave like h_{KS}. In addition, only very few analytical results have been reported on the bias and the variance of ApEn.

Although m should in theory be larger than the dimension of the underlying dynamical system, ApEN is defined and used for any possible value of m, and most applications reported in the literature use small m (1 or 2) without any analytical support, but with great success [158, 119].

Sample Entropy

A decade after Pincus's seminal paper, Richman and Moorman pointed out that ApEn contains in its very definition a source of bias and was lacking in some cases "relative consistency". They defined sample entropy (SampEn) on the same grounds as ApEn:

$$\mathrm{SampEn}(m, \epsilon) = \ln \left(\frac{C^m(\epsilon)}{C^{m+1}(\epsilon)} \right) . \tag{9.8}$$

So that:

$$\lim_{\epsilon \to 0} \lim_{m \to \infty} \mathrm{SampEn} = K_2 . \tag{9.9}$$

On the practical side, the counting of neighbors in eq.(9.1) does not allow self-matches. $C_i^m(\epsilon)$ may vanish, but when averaging over all points in eq.(9.3), the correlation integral $C^m(\epsilon) > 0$. In practice, SampEn is considered to improve on ApEn as it shows lower sensitivity to parameter tuning and data sample size than ApEn [169, 123].

We interpret SampEn as an estimate of the m-order Eckmann–Ruelle entropy K_2 at finite resolution ϵ.

Estimation

We note by the following $\mathrm{ApEn}^{(m)}$ and $\mathrm{SampEn}^{(m)}$ the estimated values of ApEn and SampEn using our own MATLAB implementation, based on Physio-Net packages. We used the commonly-accepted value, $\epsilon = 0.2\sigma$, with σ the standard deviation of X, and $m = 2$. For all quantities, we used $\tau = 5 = f_s/f_{\max}$ with $f_{\max} = 2$ Hz the cutoff frequency above which FHR times series essentially contain no relevant information [55]; this time delay corresponds to 0.5 s.

9.2.2 Connecting Complexity Theory and Information Theory

We consider here for clarity only the relation between ApEn and m-order Shannon entropy rate, although the very same relation holds between SampEn and the m-order Rényi order-two entropy rate. In information theory terms, ApEn appears as a particular estimator of the m-order Shannon entropy rate that computes the probability density by counting, in the m-dimensional embedded space, the number of neighbors in a hypersphere of radius ϵ, which can be interpreted as a particular kernel estimation of the probability density.

Limit of Large Datasets and Vanishing ϵ: Exact Relation

When the size ϵ of the spheres tends to 0, the expected value of ApEn for a stochastic signal X with any smooth probability density is related, in the limit $N \to \infty$, to the m-order Shannon entropy rate [122]:

$$\mathrm{ApEn}(m, \epsilon) \underset{\epsilon \to 0}{\simeq} h^{(m,\tau)}(X) - \log(2\epsilon)\,. \tag{9.10}$$

Both terms involve m-points correlations of the process X. This relation allows a quantitative comparison of ApEn with the m-order Shannon entropy rate $h^{(m,\tau)}$. The $\log(2\epsilon)$ difference corresponds to the paving, with hyperspheres of radius ϵ, of the continuous m-dimensional space over which the probability $p(\mathbf{x}_t^{(m,\tau)})$ involved in eq.(1.7) is defined and, thus, $h^{(m,\tau)}$. This paving defines a discrete phase space, over which eq.(9.1), eq.(9.2) and eq.(9.6) operate to define ApEn [102]. This illustrates that, for a stochastic signal, ApEn diverges logarithmically as the size ϵ approaches 0, as expected for h_{KS}. Fortunately, ϵ is fixed in the definition of ApEn, which allows one in practice to compute it for any signal/process.

New Features

Having recognized the success of ApEn and remembering its relation to $h^{(m,\tau)}$, it seems interesting to probe other m-order Shannon entropy rate estimators. A straightforward improvement would be to consider a smooth, e.g., Gaussian, kernel of width ϵ instead of the step function used in eq.(9.1). We prefer to reverse the perspective and use a k-nearest-neighbor (k-NN) estimate. Instead of counting the number of neighbors in a sphere of size ϵ, this approach searches for the size of the sphere that accommodates k neighbors. In practice, we compute the entropy H with the Kozachenko and Leonenko estimator [114, 184], which we denote $\hat{H}$. We compute the auto-mutual information $I^{(m,p,\tau)}$ with the Kraskov et al. estimator [116], which we denote $\hat{I}^{(m,p)}$. We then combine the two according to eq.(1.25) to get an estimator $\hat{h}^{(m)}$ of the m-order Shannon entropy rate. We use $k = 5$ neighbors and set $\tau = 5$ (see section 9.2.1).

We report in the next section our results for the five features when setting $m = 2, p = 1$ and compare their performances in detecting acidosis. The dependance of the m-order entropy rate (and its estimators) on m is expected to give some insight into the dimension of the attractor of the underlying dynamical system, but as we have pointed out, the dynamics is indeed contained in the AMI part of the entropy rate. This is why we further explore the effect of varying the embedding dimensions m and p on the AMI estimator $\hat{I}^{(m,p)}$.

9.3 Results: Acidosis Detection Performance

9.3.1 Comparison of Features' Performance, Using a Single Time Window, Just Before Delivery

Average features value for normal and abnormal subjects: We compute the five features: $\mathrm{ApEn}^{(m)}$, $\mathrm{SampEn}^{(m)}$, $\hat{h}^{(m)}$, $\hat{H}$ and $\hat{I}^{(m,p)}$ for normal and acidotic (abnormal) subjects in Datasets I and II using data from the last $T = 20$ mn before delivery, which are

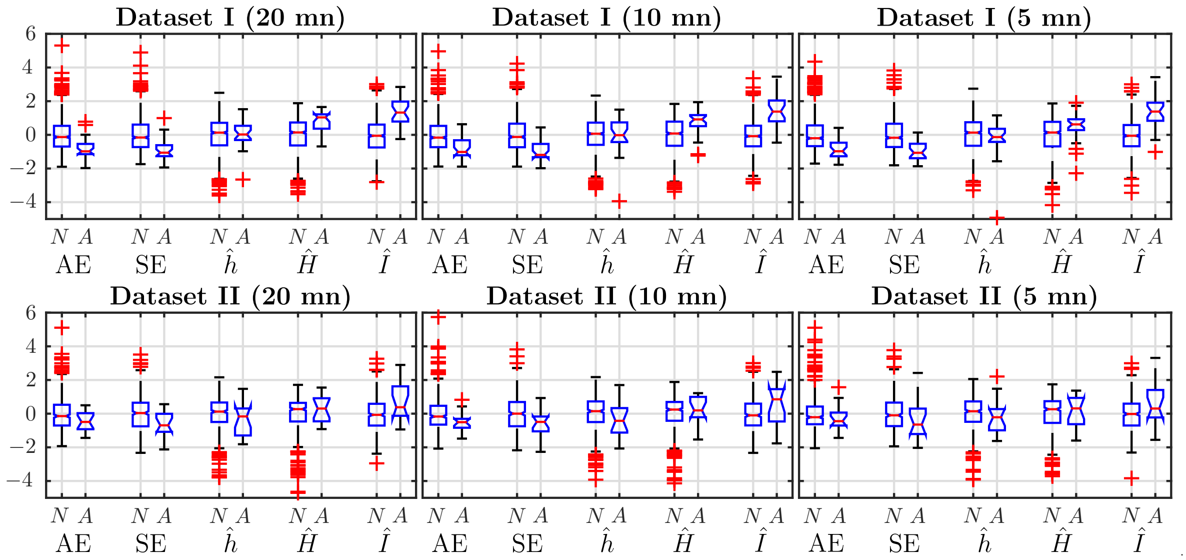

Figure 9.1: Box plot-based comparisons of the five different (normalized) estimates ($\text{ApEn}^{(m)}$ (AE), $\text{SampEn}^{(m)}$ (SE), Shannon entropy rate $h^{(m)}$ ($\hat{h}$), Shannon entropy ($\hat{H}$) and auto-mutual information (AMI) $I^{(m,p)}$ ($\hat{I}$)) for normal (N) and pathological (A for "abnormal") subjects, for Dataset I (top) and Dataset II (bottom). Each column corresponds to a different window size: $T = 20$ mn, 10 mn and 5 mn. All features are computed with $m = 2$, $p = 1$.

the most crucial. We use the classical values $m = 2$ and $p = 1$ for embedding dimensions. To compare the performance, we present the box plots of the five normalized (zero-mean, unit-variance) estimates in the left column of figure 9.1. For Dataset I, the average of ApEn and SampEn for acidotic subjects is smaller than for normal subjects, while the average of the Shannon entropy rate does not show any tendency. This is surprising as one might have expected for $\hat{h}^{(m)}$ a behavior similar to ApEn and SampEn (see section 9.2.2). Average values of $\hat{H}$ and $\hat{I}^{(2,1)}$ are larger for acidotic subjects. The larger value of Shannon entropy H indicates that the acidotic FHR signals contain more information. The larger value of AMI indicates a stronger dependence structure in the dynamics of abnormal subjects.

For subjects in Dataset II, it is harder to find any tendency by looking at the average values.

Features performance: Fetal acidosis detection performance is assessed with the p-value given by the classical Wilcoxon rank sum test. This non-parametric test of the null hypothesis, which corresponds to identical medians of the distributions of estimates in the normal and abnormal classes, is reported in table 9.1. We have added one $\star$ symbol when the p-value is less than 0.05, two $\star\star$ when less than 0.01. We see that for Dataset I, $\text{ApEn}^{(m)}$, $\text{SampEn}^{(m)}$, $\hat{H}$ and $\hat{I}^{(m,p)}$ for $m = 2$ discriminate normal and acidotic subjects, while $\hat{h}^{(m)}$ does not. Out of the three estimates ($\text{ApEn}^{(m)}$, $\text{SampEn}^{(m)}$, $\hat{h}^{(m)}$) based on entropy rates, the nearest-neighbors one for Shannon entropy rate is the poorest, although its decomposition into Shannon entropy (static one-point information) and AMI (which

	Dataset I		**Dataset II**	
	AUC	**p-Value**	**AUC**	**p-Value**
ApEn$^{(2)}$	0.76	4.08e-06 ⋆⋆	0.61	1.33e-01
SampEn$^{(2)}$	0.79	5.92e-07 ⋆⋆	0.67	2.35e-02 ⋆
$\hat{h}^{(2)}$	0.50	9.75e-01	0.39	1.36e-01
$\hat{H}$	0.76	8.36e-06 ⋆⋆	0.56	4.23e-01
$\hat{I}^{(2,1)}$	**0.84**	2.00e-09 ⋆⋆	**0.68**	1.69e-02 ⋆

Table 9.1: Area under Receiver Operational Characteristics curves (as shown in figure 9.2) and p-value obtained from the Wilcoxon rank-sum test, for each of the five different estimates, all with $m = 2$.

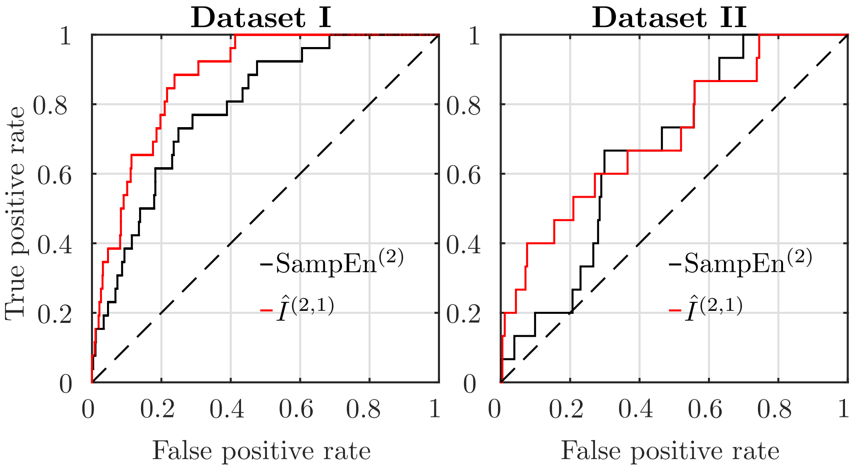

Figure 9.2: Receiver operating characteristics (ROC) curves for SampEn$^{(m)}$ (black) and AMI estimator $\hat{I}^{(m,p)}$ (red), for subjects in Datasets I (left) and II (right). $m = 2$, $p = 1$.

includes dynamic information) leads to two satisfying estimates figure 9.1 and table 9.1 both show that the best performing estimators are SampEn$^{(m)}$ and $\hat{I}^{(m,p)}$. In Dataset II, although all features performs more poorly than in Dataset I, SampEn and AMI are again the best ones, with a p-value lower than 0.05. We focus on these two features in the following.

Receiver operating characteristics: To compare the two best performing features, SampEn and AMI, we plot receiver operating characteristics (ROC) curves in figure 9.2, both for Datasets I and II, using data from the last $T = 20$ mn before delivery. For Dataset I, AMI better discriminates acidotic subjects from normal ones. For Dataset II, AMI discrimination is only slightly better than SampEn. The area under the curve (AUC) of the ROC of the features is reported in table 9.1, with bold font indicating the estimator with the largest AUC. Performance is much worse in Dataset II than in Dataset I: the AUC is reduced. Nevertheless, AMI is always the better performing estimator (its AUC reduces from 0.84 to 0.68), followed by SampEn (AUC reducing from 0.79 to 0.67).

		$\mathbf{SampEn^{(2)}}$		$\mathbf{\hat{I}^{(2,1)}}$		$\mathbf{\hat{I}^{(3,3)}}$	
		AUC	**p-value**	**AUC**	**p-value**	**AUC**	**p-value**
Dataset I	20 mn	0.79	5.92e-07 ⋆⋆	0.84	2.00e-09 ⋆⋆	**0.88**	5.46e-11 ⋆⋆
	10 mn	0.76	1.22e-07 ⋆⋆	0.84	2.22e-09 ⋆⋆	**0.87**	1.40e-10 ⋆⋆
	5 mn	0.72	1.97e-07 ⋆⋆	0.83	7.47e-09 ⋆⋆	**0.86**	6.26e-10 ⋆⋆
Dataset II	20 mn	0.67	2.35e-02 ⋆	0.68	1.69e-02 ⋆	**0.71**	5.36e-03 ⋆⋆
	10 mn	0.62	1.56e-02 ⋆	0.64	5.87e-02	**0.68**	1.66e-02 ⋆
	5 mn	0.62	5.16e-02	0.60	1.70e-01	**0.64**	7.29e-02

Table 9.2: AUC and p-value of Wilcoxon test of SampEn and AMI in datasets I and II using data from the last 20, 10 or 5mn before delivery.

9.3.2 Effect of Window Size on Performance

We investigate the robustness of the detection performance when the window size T is reduced, using data from $T = 20, 10$ and five minutes. Results are reported in figure 9.1 and table 9.2. p-values and AUC both indicate that $\hat{I}^{(2,1)}$ and SampEn$^{(2)}$ provide robust discrimination in Dataset I even when the observation length is reduced. Again, $\hat{I}^{(2,1)}$ performs better: its AUC is reduced from 0.84 to 0.83 when T is reduced from 20 mn to 5 mn, where the AUC of SampEn is reduced from 0.79 to 0.72. In Dataset II, once again, performance degrades, but AMI is still better at discriminating acidotic from normal subjects. In the following, we focus on AMI estimates only.

9.3.3 Effect of the Embedding Dimensions on the (Fetal Acidosis Detection) Performance of AMI

In order to improve the acidosis detection performance of the AMI, especially in Dataset II, we increase the embedding dimensions m and p used in computing $\hat{I}^{(m,p)}$. This way, we probe a higher order dependence structure in the dynamics. Because of the symmetry of AMI (eq.(1.27)) and aiming at probing the effect of increasing either m or p, we plot the AUC of ROC as a function of $m + p$ only. The dependence of AMI on $m - p$ is much smaller and not reported here. These computations have been done with a larger value $k = 15$ in the k-NN algorithms, in order to accommodate the possibly large embedding dimensions ($m + p$ up to 12). Results are presented in figure 9.3.

For a fixed window size, the AUC increases when $m + p$ increases and reaches a maximum; it then remains constant or decreases slightly. This behavior is observed in both Datasets I and II and for any window size $T \in [5, 10, 20]$ mn. Varying T dos not seem to change the location of the maximum of the AUC in a given dataset. The optimal embedding dimension is $m + p = 6$ in Dataset I and $m + p = 10$ in Dataset II. This hints at a difference in the dynamics of the FHR in the two datasets. Because both bias and computation time increase with the total dimensionality [69], the maximal embedding is restricted to $m = p = 3$. A reduction of the AUC is observed when the analysis window is reduced, but this is only significant for Dataset II.

We reported in table 9.2 the AUC and p-values of the AMI for two embedding dimensions: $\hat{I}^{(2,1)}$ and $\hat{I}^{(3,3)}$, for Datasets I and II and several window sizes. The best performing

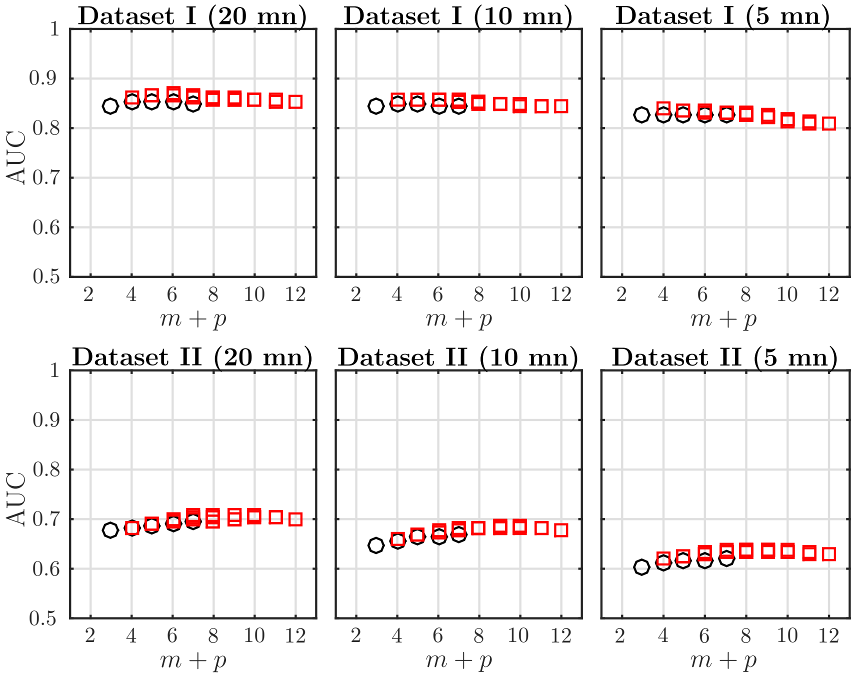

Figure 9.3: AUC of ROC for $\hat{I}^{(m,p)}$ as a function of the total embedding dimension $m+p$, with $m \geq p$, for time windows of size $T = 20, 10, 5$ mn, for data in Dataset I (first line) and Dataset II (second line). Black circles indicate the special case $p = 1$ corresponding to the classical definition of Shannon entropy rate (see eq.(1.25)). Red squares correspond to $p \geq 2$.

estimator is indicated in bold. For all observation windows and for the two datasets, $\hat{I}^{(3,3)}$ achieves the best performance. Their AUC is always larger than the one obtained using SampEn or AMI with $m = 2$ and $p = 1$.

9.3.4 Dynamical Analysis

We now explore how long before delivery the AMI can diagnose fetal acidosis on an FHR signal. To do so, we do not restrict our analysis to the last data points before delivery, but we apply it to an ensemble of windows scanning the first and second stages of labor. We examine the dynamics of $\hat{I}^{(3,3)}$, the best performing feature, for both normal and abnormal subjects.

Dataset I: Rapid Delivery

In this first section, we focus on Dataset I and probe Stage 1, including early labor, active labor and transition. Using the time at which Stage 1 ends as a reference (setting it at $t_0 = 0$), we compute for each subject $\hat{I}^{(3,3)}$ in a set of time windows $[t_i - T, t_i]$, $0 \leq i \leq 50$ of fixed size T ending at $t_i = t_0 - i * 2$ mn, so separated by 2 mn. We perform this analysis for three window sizes $T \in [20, 10, 5]$ mn. The value of AMI computed in the i-th window is then assigned to time $t_i - T/2$, at the center of the interval. By construction of Dataset I, delivery occurs less than 15 min after pushing started and can be as short as 1 mn, so we completely discard data from Stage 2. We then average the values of AMI over the population of normal subjects and over the population of acidotic subjects, respectively. Results, including p-values, are presented in figure 9.4.

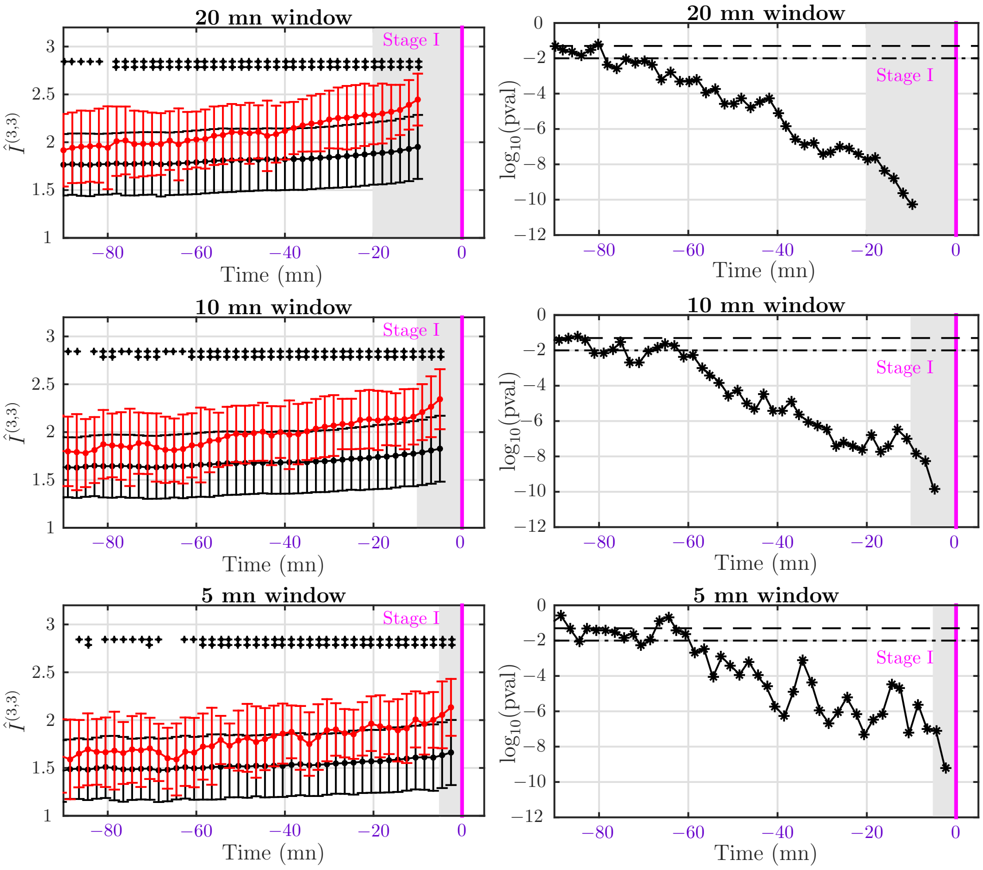

Figure 9.4: Left: Average AMI $\hat{I}^{(3,3)}$ for normal (black) and abnormal (red) subjects in Dataset I (delivery occurring less then 15 mn after pushing started). AMI is computed in windows of size $T = 20$ mn (first line), 10 mn (second line) and 5 mn (third line) shifted by 2 mn. The vertical magenta line indicates the beginning of Stage 2 (pushing). Right: Corresponding p-value. A single black + symbol in the AMI plot indicates a p-value lower than 0.05, two ++ indicate a p-value lower than 0.01. The gray-shaded region represents the time window used to compute the last value of AMI.

A first observation is that AMI is always larger for acidotic subjects than for normal subjects. As labor progresses, AMI increases in both populations, but the increase is stronger for acidotic subjects. As a consequence, the p-value of the test decreases clearly, so the feature performs better and better when approaching delivery. Detection of acidosis using the AMI feature and $T = 20$ mn can be obtained in Dataset I as early as 80 min before entering the second stage. Using shorter windows, $T = 10$ mn or 5 mn, detection is still reliable as early as one hour before Stage 2. We interpret this reduced forecast of acidosis detection in Dataset I as a direct consequence of the reduction of the statistics when the window size T is reduced.

Dataset II: Delivery after Pushing More Than 15 mn

For Dataset II, we performed the same dynamical analysis as in the previous section, using the end of Stage 1 as the reference time ($t_0 = 0$). Because there is now enough data in the pushing stage, we also perform the analysis of this stage using the delivery time ($t_0 = D$) as the reference. All results are presented in figure 9.5. At the end of Stage 1, we observe again that AMI is larger for acidotic subjects than for normal ones, but the difference is not significant in this group (see the corresponding p-value on the right of figure 9.5). The situation is identical at the end of Stage 2, although we obtain a lower p-value in some windows. The p-value does not decrease clearly when approaching delivery time, as it was in Dataset I, see figure 9.4. For subjects in Dataset II, it is very

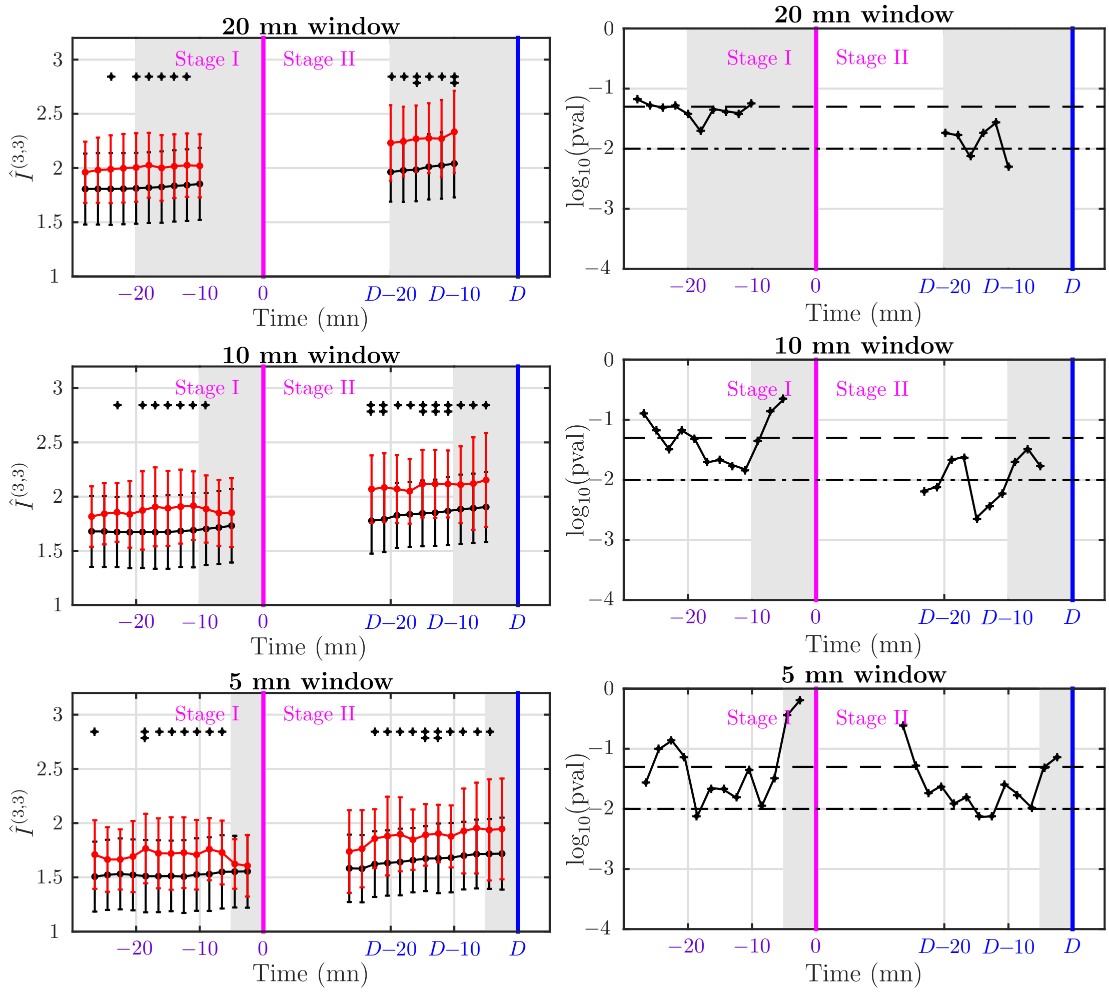

Figure 9.5: Left: Average AMI $\hat{I}^{(3,3)}$ for normal (black) and abnormal (red) subjects in Dataset II (delivery occurring more then 15 mn after pushing started). AMI is computed in windows of size $T = 20$ mn (first line), 10 mn (second line) and 5 mn (third line) shifted by 2 mn. The vertical magenta line indicates the beginning of Stage 2 (pushing), and the vertical blue line indicates delivery. Right: Corresponding p-value. Black + symbols in the AMI plot indicate a p-value lower than 5% (+) or lower than 1% (++). The gray-shaded region represents the time window used to compute the last value of AMI.

difficult to make an early detection of acidosis. However, we observe in figure 9.5 that the average AMI is significantly larger at the end of Stage 2 than at the end of Stage 1. The increase of AMI is larger for abnormal subjects.

To examine more precisely the dynamical increase of AMI, especially when entering Stage 2, we computed $\hat{I}^{(3,3)}$ over an ensemble of windows of size $T = 20$ mn spanning continuously a large time interval that includes the end of the active labor and the beginning of the pushing stage. Results are reported in figure 9.6. We see a continuous increase of AMI values when evolving from Stage 1 to Stage 2. The increase is more important for abnormal subjects, which corroborates the findings in figure 9.5. For smaller window sizes, the situation is less clear.

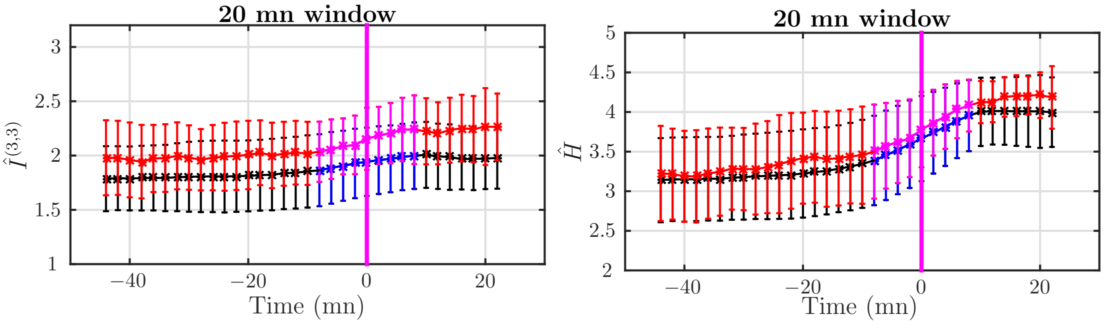

Figure 9.6: Average behavior of AMI $\hat{I}^{(3,3)}$ (left) and Shannon entropy $\hat{H}$ (right) for normal (black) and acidotic (red) subjects in Dataset II (delivery occurring more than 15 mn after pushing started).Quantities are computed in windows of size $T = 20$ mn. The vertical magenta line indicates the beginning of Stage 2 (pushing). We have used a different color code for windows spanning both Stages 1 and 2: blue for normal subjects and magenta for acidotic ones.

We also studied the dynamical evolution of the Shannon entropy estimate $\hat{H}$, which, together with the AMI, combines into the Shannon entropy rate (see eq.(1.25)). Changes in the Shannon entropy H indicate changes in the probability density of the signal. Results are presented in figure 9.6, side by side with the AMI. We observe a dramatic rise of the value of $\hat{H}$ when subjects evolve from Stage 1 to Stage 2. This increase is clearly observed for normal and abnormal subjects. No significant difference between normal and acidotic subjects is observed for this static quantity. The start of pushing implies a strong deformation of the probability density of the FHR, indicating strong perturbations of the FHR, for both normal and acidotic subjects.

9.4 Discussion, Conclusions and Perspectives

We now discuss the interpretation of Shannon entropy and AMI measurements in different stages of labor. The fetuses are classified as normal or acidotic depending on a post-delivery pH measurement, which gives a diagnosis of acidosis at delivery only. There is no information on the health of the fetuses during labor.

The physiological interpretation of a feature, and especially its relation to specific FHR patterns, e.g., like those detailed in [64, 14], is a difficult task that is only scarcely reported in the literature [180, 75]. In this chapter, we have averaged our results over large numbers of (normal or acidotic) subjects, which jeopardizes any precise interpretation in terms of a specific FHR pattern that may appear only intermittently.

9.4.1 Acidosis Detection in the First Stage

We can nevertheless suggest that the value of the Shannon entropy H is related to the frequency of decelerations in the FHR signals. Indeed, Shannon entropy strongly depends on the standard deviation of the signal (e.g., see eq.(1.38)), which in turn depends on the variability in the observation window. A larger number of decelerations in the observation window deforms the PDF of the FHR signal by increasing its lower tail; in particular, this increases the width of the PDF and hence increases the standard deviation and the Shannon entropy. This explains our findings in figure 9.1 (for Dataset I).

When acidosis develops in the first stage of labor, the Shannon auto-mutual information estimator $\hat{I}^{(m,p)}$ significantly outperforms all other quantities both in terms of p-value and AUC. The performance of AMI is robust when tuning either the size of the observation, and hence the number of points in the data, and the embedding dimensions (m, p). In addition, the performance slightly increases when the total embedding dimension $m+p$ increases; although one has to care about the curse of dimensionality.

For abnormal subjects from Dataset II, AMI is not able to detect acidosis using data from Stage 1. This suggests that acidosis develops later, in the second stage of labor.

For all datasets, AMI computed with $\tau = 0.5$ s is always larger for acidotic subjects than for normal subjects. This is in agreement with results obtained with ApEn and SampEn, which are both lower for acidotic subjects. This shows that FHR classified as abnormal have a stronger dependence structure at a small scale than normal ones. We can relate this increase of the dependence structure of acidotic FHR to the short-term variability and to its coupling with particular large-scale patterns. For example, a sinusoidal FHR pattern [14], especially if its duration is long, should give a larger value of the AMI, because its large-scale dynamics is highly predictable. As another example, we expect variable decelerations (with an asymmetrical V-shape) and late decelerations (with a symmetrical U-shape and/or reduced variability) to impact AMI differently. Of course, the choice of the embedding parameter τ is then crucial, and this is currently under investigation.

AMI and entropy rates depend on the dynamics as they operate on time-embedded vectors. AMI focuses on nonlinear temporal dynamics, while being insensitive to the dominant static information. AMI is independent of the standard deviation, which on the contrary contributes strongly to the Shannon entropy. This explains why AMI performs better than entropy rate estimates, such as $\text{ApEn}^{(m)}$, $\text{SampEn}^{(m)}$ and $\hat{h}^{(m)}$, which depend also on the standard deviation.

9.4.2 Acidosis Detection in Second Stage

The results reported for Stage 2 show a severe decrease in the performance of the five estimated quantities. Analyzing Stage 2 is far more challenging than analyzing Stage 1, which suggests that temporal dynamics in Stage 2 differ notably from those of Stage 1 [189] or simply that our database does not contain enough acidotic subjects in that case.

$\hat{I}^{(m,p)}$ achieves the best performance in terms of p-value and AUC; this clearly underlines that the analysis of nonlinear temporal dynamics is critical for fetal acidosis detection in Stage 2. As in Stage 1, the AMI is always larger in Stage 2 for acidotic subjects than for normal subjects.

Although the Shannon entropy computed from the last 20 mn of Stage 2 before delivery does not show a clear tendency in figure 9.1 for Dataset II, looking at figure 9.6 clearly shows that $\hat{H}$ increases as labor progresses: this is probably related to the average increase of the number of decelerations, which is expected in both the normal and acidotic population.

SampEn$^{(2)}$ is also able to perform discrimination in Stage 2. From these observations, one may envision the definition of a new estimator that would measure the auto-mutual information using the Rényi order-two entropy by applying eq.(1.25). Nevertheless, it should be emphasized that Rényi order-q entropy is lacking the chain rule of conditional probabilities as soon as $q \neq 1$, which may jeopardize any practical use of such an estimator.

9.4.3 Probing the Dynamics

Increasing the total embedding dimensions in AMI improves the performance in the detection of acidotic subjects, in both the first and second stages. The best performance is found for different total embedding dimension in the two datasets. This suggests that FHR dynamics is different in each stage.

As seen in eq.(1.25), the Shannon entropy rate can be split into two contributions: one that depends only on static properties (the Shannon entropy, estimated by $\hat{H}$) and one that involves the signal dynamics (the auto-mutual information, estimated by $\hat{I}^{(m,1)}$). By following the time evolution of these two parts, we were able to relate Shannon entropy $\hat{H}$ with the evolution of the labor and AMI not only with the evolution of the labor, but also with possible acidosis. Looking at subjects for which the pushing phase is longer than 15 mn, it clearly appears that all fetuses are affected by the pushing, as evidenced by a large increase of the Shannon entropy $\hat{H}$ and a small increase of AMI. Additionally, the increase of AMI is steeper for abnormal subjects than for normal ones, which may indicate different reactions to the pushing and can be related to specific pathological FHR patterns. When the pushing stage is long (Dataset II), fetuses reported as acidotic do not show any sign of acidosis until prolonged pushing. These fetuses appear as normal until delivery is near.

When acidosis develops during the first stage of labor, in Dataset I, we observe clearly that while AMI increases steadily till delivery for healthy fetuses, it increases faster for acidotic ones. This suggests that acidotic fetuses in Dataset I react to early labor, as early as one hour before pushing starts. This could not only indicate that some fetuses are prone to acidosis, but also may pave the way for an early detection of acidosis in this case.

Discussion, Conclusions and Perspectives

Most of the time when studying a system, scientists face processes whose properties are *a priori* unknown. Characterising these processes is a major task to describe the studied system. During this thesis, which combines signal processing and physics, we were mainly motivated by the study of complex systems and turbulence, and consequently, we were interested in the study of regularity and self-similarity properties, long range dependence structures and multi-scale behavior. In order to perform this kind of study, we use information theory quantities, which are functions of the probability density function of the analysed process, and so depend on any order statistics of its PDF (chapter 3).

We developed different, but related, data analysis methodologies, based on information theory, to analyse a process across scales τ. These scales are usually identified with the sampling parameter of Takens embedding procedure (section 1.2), but also with the size of the increments of the process (chapters 7 and 8). The methodologies developed during this thesis, can be used to characterize stationnary and non-stationnary processes by analysing time windows of length T of the studied signal (chapter 4).

Characterization of a process

Regularity

In order to characterize the regularity of an unknown process X, we combine entropy rate $h^{(m,\tau)}(X)$ (section 1.4.1) and an adapted Theiler prescription (section 2.5) to perform an analysis across scales. The entropy rate across scales, that quantifies the unknown information of a sample knowing the m previous ones, can be expressed following eq.(1.25). Consequently, entropy rate across scales presents two different contributions: the entropy of the process, which only depends on its PDF, and the auto-mutual information across scales (section 1.4.2) that quantifies the shared information between two consecutive chunks of the signal.

Both, auto-mutual information and entropy rate across scales perform very similar statistical characterizations of a process (see eq.(1.25)). However, entropy rate presents some estimation advantages: it is less biased and has a smaller standard deviation than auto-mutual information, when analysing non-stationnary processes (section 4.3.1). The local slope of entropy rate across scales in a given range of scales, can be related to a measure of the exponent of "global regularity" of the process in this range (chapter 4):

- If the local slope, in a range of scales, is 0, the process, sampled at these scales, corresponds to a white noise. The analysed process can be seen as discontinuous at these scales (figure 9.7, a)). Furthermore, when the scale decreases the slope of entropy rate can only increase (or remains constant), due to the increase of dependences between samples of the process at these scales. This effect indicates the increase of the regularity of the process for the small scales.
- A local slope between 0 and 1 ($\in (0,1)$), in a range of scales, indicates the continuity but non-differentiability of the process at these scales. The value of the slope characterizes the exponent of "global regularity" of the process in this range (figure 9.7, b)).
- A local slope of 1, in a range of scales, indicates that the process is differentiable with a continuous derivative at these scales (figure 9.7, c)).

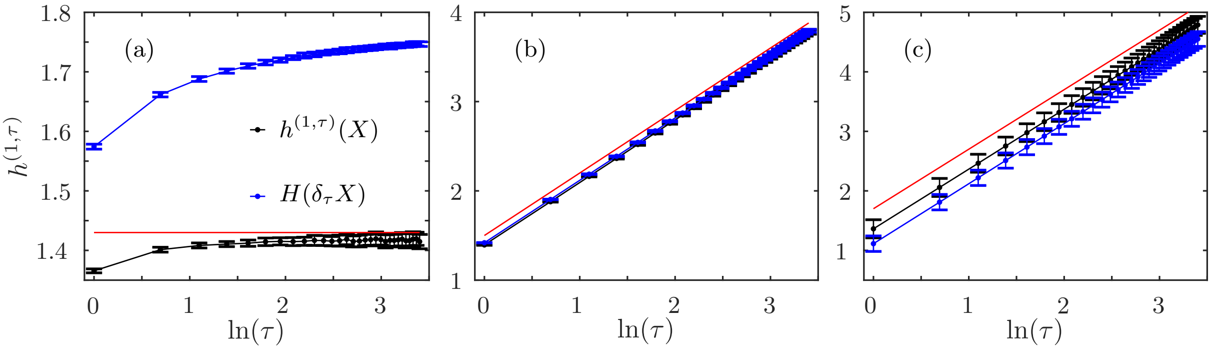

Figure 9.7: Entropy rate across scales (black) and entropy of the increments (blue) for: a) fractional Gaussian noise with $\mathcal{H} = -0.3$, the red line indicates the slope 0, b) fractional Brownian motion with $\mathcal{H} = 0.7$, the red line indicates the slope 0.7 c) process with $\mathcal{H} = 1.7$ generated by integrating a fractional Brownian motion with $\mathcal{H} = 0.7$, the red line indicates the slope 1.

Scale invariance

A rigorous linear behavior of entropy rate in function of the logarithm of the scale indicates the scale invariance (monofractality) of the analysed process, and therefore the non-existence of integral scale. In this case the exponent of global regularity is equal to the Hurst exponent of the process.

The entropy rate across scales of a process and the entropy of the increments of the process are strongly connected (eq.(4.8)). Their relationship is completely general: it is valid for both stationnary and non-stationnary processes and it is independent of the PDF and the regularity properties of the process. In addition, for scale-invariant processes, we show that both measures are almost identical, and both characterize the Hurst exponent of the process (figures 9.7, b) and c)). This equality shows the strong relation between entropy rate across scales of a process and the PDFs of its increments.

Self-similarity: monofractality and multifractality

While for scale-invariant processes the entropy rate is linear in function of the logarithm of the scales, for multifractal processes, the non-linear behaviour of entropy rate across scales allows to describe the distance from monofractality of the process (figure 7.7), and indicates the existence of an integral scale. This non-linear behavior for multifractal processes is explained since the regularity is not defined by a single exponent but by a distribution of exponents (Hölder exponent) characterized by the singularity spectrum of the process (section 5.2). Nevertheless, the entropy rate across scales mainly depends on the second order moment of the analysed process and deviations from linearity, due to higher order moments, are in general very weak, and difficult to study. For multifractal processes, the dominant linear behavior of entropy rate across scales characterizes the exponent of global regularity of the process defined as the Hölder exponent of order 2, $h(p) = \frac{\partial \zeta(p)}{\partial p}$ with $p = 2$ (see section 5.2).

In order to properly characterize self-similarity and discriminate between monofractal and multifractal processes, we used the concept of Kullback-Leibler distance between two PDFs. We define a measure of distance from Gaussianity of a PDF (section 7.1) as the Kullback-Leibler divergence between this PDF and its Gaussian approximation (Gaussian PDF with same standard deviation). Then, we measure the distance from Gaussianity of the increments of the analysed process to measure the deformation of their PDFs across scales and hence the intermittency of the process (chapter 7).

Long-range dependences

Combining Takens embedding procedure with information theory allows to measure dependences by taking into account n-points interactions, with n being defined by the total embedding dimension of the measure (section 3.3.2). This provides the framework to test Markov properties and long-range dependences. In order to identify long range dependences in a process, we can derive it until obtaining a discontinuous process (order of exponent of global regularity less than or equal to 0). Then, we measure auto-mutual information for different embedding dimensions m and p in order to vary the number of points interacting in the measure. For long range dependences, the auto-mutual information measure should increase when the embedding increases, while for short-range dependences, the auto mutual information remains constant when the embedding increases (figure 3.6).

Information flows

We also looked for causality relationships between the different scales of a process (chapter 8). We used transfer entropy (section 1.6.2) as a Wiener causality measure between two time processes. However, the generation of the scales of a time process has a strong influence on the causality measures, and renders their interpretation very difficult, *i.e.* we are not able to distinguish which contribution of the measure is due to the properties of the system and which one is due to the way of generating the scales (section 8.2). The results we obtained are therefore negative results, considering that no knowledge from the analysed process was obtained. However, all the studied systems presented a flow of information from small to large scales for any way of generating the scales, which can suggest the generality of this result for dynamical systems.

We used the above methodologies mainly in the study of turbulence (chapters 6, 7 and 8), but also in a bio-medical application (chapter 9), and discrimination between processes (chapters 3 and 4).

Perspectives

The methodologies presented in this manuscript have been used to characterize synthetic processes or high quality experimental velocity measures. Only in the biomedical applications, (chapter 9), these methodologies have been tested on low quality data. As we explained throughout the thesis, these methodologies are of general application and can be used to study generic systems. New applications of the above methodologies on real world datasets, and specially on low quality datasets could reveal some flaws of the methods, or some aspects that need to be improved in order to keep the measures instructive when dealing with this kind of data.

Throughout the manuscript, we analysed stationnary and non-stationnary, stochastic or chaotic processes and we give an interpretation to our measures for all these kinds of processes. In the case of non-stationnary processes, we did that in two different frameworks. The question that naturally arises is if we can use the methodologies developed during the thesis to characterize non-chaotic deterministic processes (for exemple, a sine wave), and what are the interpretation of the measures in this case. This study can be interesting to analyse independently the stochastic and deterministic components of a process; for exemple, a noise on a sine wave.

In the analysis of information flows and causality relationships is, perhaps, where more questions remain opened. We concluded that the causality measures between the scales of a time series are strongly influenced by the way of generating the scales. However, for any scale function we found that the small scales cause the large ones. This general result for any process and any scale function needs to be studied carefully. On the other hand, to avoid the generation of the scales of a time series, we propose, as future research, to analyse the causality between experimental signals with a meaning of scale, *i.e.* not generating the different scales from an only signal, but experimentally measuring several signals associated with different scales. This way, with the scales measured experimentally and turned into different time series, the causality measures would be only due to the properties of the system.

Bibliography

[1] Abramson, I. (1982). On bandwidth variation in kernel estimates- a square root law. *The annals of Statistics*, 10:1217–1223.

[2] Ahn, Y., Bagrow, J., and Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. *Nature*, 466:761–4.

[3] Akselrod, S., Gordon, D., Ubel, F. A., Shannon, D. C., Berger, A. C., and Cohen, R. J. (1981). Power spectrum analysis of heart rate fluctuation: a quantitative probe of beat-to-beat cardiovascular control. *Science*, 213(4504):220–222.

[4] Albers, D. and Hripcsak, G. (2012a). Estimation of time-delayed mutual information and bias for irregularly and sparsely sampled time-series. *Chaos, Solitons & Fractals*, 45(6):853 – 860.

[5] Albers, D. and Hripcsak, G. (2012b). Using time-delayed mutual information to discover and interpret temporal correlation structure in complex populations. *Chaos*, 22:013111.

[6] Alfirevic, Z., Devane, D., and Gyte, G. M. (2006). Continuous cardiotocography (CTG) as a form of electronic fetal monitoring (EFM) for fetal assessment during labour. *Cochrane Database Syst Rev*, 3(3):CD006066.

[7] Amblard, P. and Michel, O. (2013). The relation between Granger causality and directed information theory: A review. *Entropy*, 15:113–143.

[8] Anselmet, F., Gagne, Y., Hopfinger, E., and Antonia, R. (1984). High-order velocity structure functions in turbulent shear flow. *Journal of Fluid Mechanics*, 140:63–89.

[9] Arneodo, A., Bacry, E., and Muzy, J. F. (1998). Random cascades on wavelet dyadic trees. *Journal of Mathematical Physics*, 39(8):4142–4164.

[10] Arneodo, A., Manneville, S., Muzy, J., and Roux, S. (1999). Revealing a lognormal cascading process in turbulent velocity statistics with wavelet analysis. *Philosophical Transactions Of The Royal Society Of London Series A-Mathematical Physical And Engineering Sciences*, 357(1760):2415–2438.

[11] Ash, R. B. (1972). *Real Analysis and Probability.* Academic Press.

[12] Astola, J. T., Egiazarian, K. O., Khlopov, G. I., Khomenko, S. I., Kurbatov, I. V., Morozov, V. Y., and Totsky, A. V. (2008). Application of bispectrum estimation for time-frequency analysis of ground surveillance doppler radar echo signals. *IEEE Transactions on Instrumentation and Measurement*, 57(9):1949–1957.

[13] Atta, C. V. and Park, J. (1972). Statistical self-similarity and inertial subrange turbulence. In *Statistical models and turbulence.*

[14] Ayres-de Campos, D., Spong, C. Y., Chandraharan, E., and Intrapartum Fetal Monitoring Expert Consensus Panel, F. I. G. O. (2015). Figo consensus guidelines on intrapartum fetal monitoring: Cardiotocography. *Int J Gynaecol Obstet*, 131(1):13–24.

[15] Bacry, E., Delour, J., and Muzy, J. F. (2001). Multifractal random walk. *Phys. Rev. E*, 64:026103.

[16] Bair, W., Koch, C., Newsome, W., and Britten, K. (1994). Power spectrum analysis of bursting cellss in area mt in the behaving monkey. *Journal of Neuroscience*, 14:2870–2892.

[17] Bar-Yam, Y. (2004). Multiscale variety in complex systems. *Complexity*, 9(4):37–45.

[18] Bar-Yam, Y. (2006a). *Engineering Complex Systems: Multiscale Analysis and Evolutionary Engineering*, pages 22–39. Springer Berlin Heidelberg.

[19] Bar-Yam, Y. (2006b). Improving the effectiveness of health care and public health: a multiscale complex systems analysis. *American Journal of Public Health*, 96(3):459–466.

[20] Batchelor, G. K. (1951). Pressure fluctuations in isotropic turbulence. *Mathematical Proceedings of the Cambridge Philosophical Society*, 47:359–374.

[21] Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour.* Princeton University Press.

[22] Benzi, R., Paladin, G., Parisi, G., and Vulpiani, A. (1985). Characterisation of intermittency in chaotic systems. *J. Phys. A*, 18:2157–2165.

[23] Bollt, E., Stanford, T., Lai, Y., and Życzkowski, K. (2000). Validity of threshold-crossing analysis of symbolic dynamics from chaotic time series. *Phys. Rev. Lett.*, 85:3524–3527.

[24] Borland, L., Bouchaud, J., Muzy, J., and Zumbach, G. (2005). The dynamics of financial markets- mandelbrot's multifractal cascades and beyond. *Wilmott Magazine.*

[25] Boubekeur, L. and Lyth, D. H. (2006). Detecting a small perturbation through its non-gaussianity. *Phys. Rev. D*, 73:021301.

[26] Bouchaud, J. and Potters, M. (2003). *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management.* Cambridge University Press.

[27] Breiman, L., Meisel, W., and Purcell, E. (1977). variable kernel estimates of multivariate densities. *Technometrics*, 19:135–144.

[28] Bressler, S. and Seth, A. (2011). Wiener-Granger causality: A well established methodology. *NeuroImage*, 58:323–329.

[29] Brown, T. (1982). Information theory and the spectrum of isotropic turbulence. *Journal of Physics A: Mathematical and General*, 15(7):2285.

[30] Cacoullos, T. (1964). Estimation of a multivariate density. *Technical Report*, 40.

[31] Carnot, S. (1824). *Reflexions sur la puissance motrice du feu et sur les machines propres a developper cette puissance.* Bachelier Libraire.

[32] Castaing, B., Gagne, Y., and Hopfinger, E. (1990). Velocity probability density functions of higher reynolds number turbulence. *Physica D*, 46:177–200.

[33] Cerbus, R. and Goldburg, W. (2013). Information content of turbulence. *Phys. Rev. E*, 88:053012.

[34] Chanal, O., Chabaud, B., Castaing, B., and Hebral, B. (2000). Intermittency in a turbulent low temperature gaseous helium jet. *European Physical Journal B*, 17:309–317.

[35] Chandraharan, E. and Arulkumaran, S. (2007). Prevention of birth asphyxia: responding appropriately to cardiotocograph (CTG) traces. *Best Pract. Res. Clin. Obstet. Gynaecol.*, 21(4):609–624.

[36] Chandran, V., Elgar, S., and Vanhoff, B. (1995). Statistics of tricoherence. *IEEE Transactions on Signal Processing*, 42:3430–3440.

[37] Chevillard, L. (2004). *Description multifractale unifieé du phénomène d'intermittence en turbulence Eulérienne et Lagrangienne.* PhD thesis, Université Sciences et Technologies – Bordeaux I.

[38] Chevillard, L., Castaing, B., Arneodo, A., Lévêque, E., Pinton, J., and Roux, S. (2012). A phenomenological theory of eulerian and lagrangian velocity fluctuations in turbulent flows. *Comptes Rendus Physique*, 13(9):899 – 928.

[39] Chevillard, L., Castaing, B., and Lévêque, E. (2005). On the rapid increase of intermittency in the near-dissipation range of fully developed turbulence. *The European Physical Journal B - Condensed Matter and Complex Systems*, 45(4):561–567.

[40] Chicharro, D. and Panzeri, S. (2017). Synergy and redundancy in dual decompositions of mutual information gain and information loss. *Entropy*, 19:71.

[41] Chodrow, P. (2017). Divergence, entropy, information: An opinionated introduction to information theory. *CoRR*, abs/1708.07459.

[42] Chudáček, V., Anden, J., Mallat, S., Abry, P., and Doret, M. (2014). Scattering transform for intrapartum fetal heart rate variability fractal analysis: A case-control study. *IEEE Transactions on Biomedical Engineering*, 61(4):1100–1108.

[43] Clausius, R. (1854). Ueber eine veranderte form des zweiten haupsatzes der mechanischen wamertheoriein. *Annalen der Physik und Chemie*, 93(12):481–506.

[44] Cohen, M. and Kohn, A. (2011). Measuring and interpreting neuronal correlations. *Nature Neuroscience*, 14(7):811–819.

[45] Collis, W., White, P., and Hammond, J. (1998). Higher-order spectra: The bispectrum and trispectrum. *Mechanical Systems and Signal Processing*, 12(3):375 – 394.

[46] Costa, A., Ayres-de Campos, D., Costa, F., Santos, C., and Bernardes, J. (2009). Prediction of neonatal acidemia by computer analysis of fetal heart rate and ST event signals. *Am J Obstet Gynecol*, 201(5):464.e1–464.e6.

[47] Costa, M., Goldberger, A. L., and Peng, C. K. (2002). Multiscale entropy analysis of complex physiologic time series. *Physical Review Letters*, 89(6):068102.

[48] Courtney, C., Neild, S., Wilcox, P., and Drinkwater, B. (2010). Application of the bispectrum for detection of small nonlinearities excited sinusoidally. *Journal of Sound and Vibration*, 329(20):4279 – 4293.

[49] Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. John Wiley & Sons, Inc.

[50] Czabanski, R., Jezewski, J., Matonia, A., and Jezewski, M. (2012). Computerized analysis of fetal heart rate signals as the predictor of neonatal acidemia. *Expert Systems with Applications*, 39(15):11846–11860.

[51] Dawes, G., Moulden, M., Sheil, O., and Redman, C. (1992). Approximate entropy, a statistic of regularity, applied to fetal heart rate data before and during labor. *Obstet Gynecol*, 80(5):763–8.

[52] Delour, J., Muzy, J., and Arnéodo, A. (2001). Intermittency of 1d velocity spatial profiles in turbulence: a magnitude cumulant analysis. *The European Physical Journal B*, 23(2):243–248.

[53] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L1 view*. Wiley Series in probability and Mathematical Statistics. John Wiley & Sons, Inc.

[54] Donner, R., Barbosa, S., Kurths, J., and Marwan, N. (2009). Understanding the Earth as a complex system – recent advances in data analysis and modelling Earth sciences. *The european Physical Journal Special Topics*, 174(1):1–9.

[55] Doret, M., Helgason, H., Abry, P., Gonçalvés, P., Gharib, C., and Gaucherand, P. (2011a). Multifractal analysis of fetal heart rate variability in fetuses with and without severe acidosis during labor. *American Journal of Perinatology*, 28(4):259–266.

[56] Doret, M., Massoud, M., Constans, A., and Gaucherand, P. (2011b). Use of peripartum ST analysis of fetal electrocardiogram without blood sampling: a large prospective cohort study. *Eur J Obstet Gynecol Reprod Biol*, 156(1):35–40.

[57] Doret, M., Spilka, J., Chudáček, V., Gonçalves, P., and Abry, P. (2015). Fractal Analysis and Hurst Parameter for intrapartum fetal heart rate variability analysis: A versatile alternative to Frequency bands and LF/HF ratio. *PLoS ONE*, 10(8):e0136661.

[58] Ebeling, W. and Pöschel, T. (1994). Entropy and long-range correlations in literary english. *EPL (Europhysics Letters)*, 26(4):241.

[59] Echeverria, J. C., Hayes-Gill, B. R., Crowe, J. A., Woolfson, M. S., and Croaker, G. D. H. (2004). Detrended fluctuation analysis: a suitable method for studying fetal heart rate variability? *Physiol Meas*, 25(3):763–774.

[60] Eckmann, J. P. and Ruelle, D. (1985). Ergodic-theory of chaos and strange attractors. *Reviews of Modern Physics*, 57(3):617–656.

[61] Faes, L., Porta, A., and Nollo, G. (2015). Information decomposition in bivariate systems: Theory and application to cardiorespiratory dynamics. *Entropy*, 17(1):277–303.

[62] Falcioni, M., Loreto, V., and Vulpiani, A. (2003). *Kolmogorov's Legacy about Entropy, Chaos, and Complexity*, pages 85–108. Springer Berlin Heidelberg, Berlin, Heidelberg.

[63] Francis, D. P., Willson, K., Georgiadou, P., Wensel, R., Davies, L., Coats, A., and Piepoli, M. (2002). Physiological basis of fractal complexity properties of heart rate variability in man. *J Physiol*, 542(Pt 2):619–629.

[64] Freeman, R. (2002). Problems with intrapartum fetal heart rate monitoring interpretation and patient management. *Obstet Gynecol*, 100:813–26.

[65] Frenzel, S. and Pompe, B. (2007). Partial mutual information for coupling analysis of multivariate time series. *Phys. Rev. Lett.*, 99:204101.

[66] Frisch, U. (1995). *Turbulence: The Legacy of A.N. Kolmogorov.* Cambridge University Press.

[67] Frisch, U. and Parisi, G. (1985). On the singularity structure of fully developed turbulence. *Turbulence and Predictability in Geophysical Fluid Dynamics and Climate Dynamics*, 01:71–88.

[68] Frisch, U., Sulem, P., and Nelkin, M. (1978). A simple dynamical model of intermittent fully developed turbulence. *J. Fluid Mech*, 87:719–736.

[69] Gao, W., Oh, S., and Viswanath, P. (2016). Demystifying fixed k-nearest neighbor information estimators. *CoRR*, abs/1604.03006.

[70] Georgieva, A., Papageorghiou, A. T., Payne, S. J., Moulden, M., and Redman, C. W. G. (2014). Phase-rectified signal averaging for intrapartum electronic fetal heart rate monitoring is related to acidaemia at birth. *BJOG*, 121(7):889–894.

[71] Georgieva, A., Payne, S. J., Moulden, M., and Redman, C. W. G. (2013). Artificial neural networks applied to fetal monitoring in labour. *Neural Computing and Applications*, 22(1):85–93.

[72] Geweke, J. (1982). Measurement of linear dependence and feedback between multiple time series. *Journal of the american statistical association*, 77(378):304–313.

[73] Gomez, C., Lizier, J., Schaum, M., Wollstadt, P., Grutzner, C., Uhlhaas, P., Freitag, C., Schlitt, S., Bolte, S., Hornero, R., and Wibral, M. (2014). Reduced predictable information in brain signals in autism spectrum disorder. *Frontiers in neuroinformatics*, 8:9.

[74] Gómez Herrero, G., Wu, W., Rutanen, K., Soriano, M. C., Pipa, G., and Vicente, R. (2015). Assessing coupling dynamics from an ensemble of time series. *entropy*, 17:1958–1970.

[75] Gonçalves, H., Henriques-Coelho, T., Bernardes, J., Rocha, A., Nogueira, A., and Leite-Moreira, A. (2008). Linear and nonlinear heart-rate analysis in a rat model of acute anoxia. *Physiol Meas*, 29(9):1133–43.

[76] Gonçalves, H., Rocha, A. P., de Campos, D. A., and Bernardes, J. (2006). Linear and nonlinear fetal heart rate analysis of normal and acidemic fetuses in the minutes preceding delivery. *Med Biol Eng Comput*, 44(10):847–855.

[77] Granero-Belinchon, C., Roux, S., Abry, P., Doret, M., and Garnier, N. (2017). Information theory to probe intrapartum fetal heart rate dynamics. *Entropy*, 19(12):640.

[78] Granero-Belinchón, C., Roux, S., and Garnier, N. (2017). Un estimateur du taux d'entropie basé sur l'information mutuelle. In *Colloque Gretsi*.

[79] Granero-Belinchon, C., Roux, S., Garnier, N., Abry, P., and Doret, M. (2017). Mutual information for intrapartum fetal heart rate analysis. In *Con Proc IEEE Eng Med Biol Soc (EMBC)*.

[80] Granero-Belinchon, C., Roux, S. G., and Garnier, N. B. (2016). Scaling of information in turbulence. *Epl*, 115(5):58003.

[81] Granero-Belinchón, C., Roux, S. G., and Garnier, N. B. (2018). Kullback-leibler divergence measure of intermittency: Application to turbulence. *Phys. Rev. E*, 97:013107.

[82] Granger, C. (1969). Investigating causal relations by econometric models and cross-spectral mehods. *Econometrica*, 37(3):424–438.

[83] Grassberger, P. and Procaccia, I. (1983). Estimation of the kolmogorov-entropy from a chaotic signal. *Physical Review A*, 28(4):2591–2593.

[84] Gray, C. and McCormick, D. (1996). Chattering cells: Superficial pyramidal neurons contributing to the generation of synchronous oscillations in the visual cortex. *Science*, 274:109–113.

[85] Gray, R. M. (2009). *Probability, Random Processes, and Ergodic Properties*. Springer US, 2 edition.

[86] Greb, U. and Rusbridge, M. G. (1988). The interpretation of the bispectrum and bicoherence for non-linear interactions of continuous spectra. *Plasma Physics and Controlled Fusion*, 30(5):537.

[87] Grünwald, P. and Vitányistard, P. (2003). Kolmogorov complexity and information theory. with an interpretation in terms of questions and answers. *Journal of Logic, Language and Information*, 12(4):497–529.

[88] Gurvitch, A. (1960). Eksperimentalnye issledovania chastotnykh spektrov. *Izvest. Akad. Nauk SSSR*, 7.

[89] Haritopoulos, M., Illanes, A., and Nandi, A. (2016). *Survey on Cardiotocography Feature Extraction Algorithms for Fetal Welfare Assessment*, pages 1187–1192. Springer International Publishing, Cham.

[90] Helgason, H., Pipiras, V., and Abry, P. (2011a). Fast and exact synthesis of stationary multivariate gaussian time series using circulant embedding. *Signal Process.*, 91(5):1123–1133.

[91] Helgason, H., Pipiras, V., and Abry, P. (2011b). Synthesis of multivariate stationary series with prescribed marginal distributions and covariance using circulant matrix embedding. *Signal Processing*, 91:1741–1758.

[92] Herzel, H., Ebeling, W., and Schmitt, A. O. (1994). Entropies of biosequences: The role of repeats. *Phys. Rev. E*, 50:5061–5071.

[93] Horowitz, J. M. and Esposito, M. (2014). Thermodynamics with continuous information flow. *Phys. Rev. X*, 4:031015.

[94] Hruban, L., Spilka, J., Chudek, V., Jank, P., Huptych, M., Bura, M., Hudec, A., Kacerovsk, M., Kouck, M., Prochzka, M., Koreko, V., Segetá, J., imetka, O., Mchurov, A., and Lhotsk, L. (2015). Agreement on intrapartum cardiotocogram recordings between expert obstetricians. *Journal of Evaluation in Clinical Practice*, 21(4):694–702.

[95] Husar, P. and Henning, G. (1997). Bispectrum analusis of visually evoked potentials. *IEEE Engineering in medicine and biology magazine*, 16:57–63.

[96] Ikeda, K. and Matsumoto, K. (1989). Information theoretical characterization of turbulence. *Physical Review Letters*, 62(19):2265–2268.

[97] Izenman, A. (1991). Recent developments in nonparametric density estimation. *Journal of the american statistical association*, 86(413):205–224.

[98] J. Ladyman, J. Lambert, K. W. (2012). What is a complex system? *European Journal of Philosophy of Science*.

[99] James, R. G., Barnett, N., and Crutchfield, J. P. (2016). Information flows? a critique of transfer entropies. *Phys. Rev. Lett.*, 116:238701.

[100] Jaynes, E. T. (1957a). Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630.

[101] Jaynes, E. T. (1957b). Information theory and statistical mechanics. ii. *Phys. Rev.*, 108:171–190.

[102] Jaynes, E. T. (1962). Information theory and statistical mechanics. In *Brandeis Lectures in theoretical Physics*, volume 3. Brandeis Summer Institute.

[103] Jaynes, E. T. (1965). Gibbs vs Boltzmann entropies. *American Journal of Physics*, 33(5):391–398.

[104] Kahalerras, H., Malecot, Y., Gagne, Y., and Castaing, B. (1998). Intermittency and Reynolds number. *Physics of Fluids*, 10:910–921.

[105] Kantz, H. and Schreiber, T. (2004). *Nonlinear time series analysis.* Cambridge University Press.

[106] Kertai, M., Whitlock, E., and Avidan, M. (2012). Brain monitoring with electroencephalography and the electroencephalogram-derived bispectral index during cardiac surgery. *Anesthesia and Analgesia*, 114:533–546.

[107] Kolmogorov, A. N. (1941a). Dissipation of energy in locally isotropic turbulence. *Dokl. Akad. Nauk SSSR*, 32:16–18.

[108] Kolmogorov, A. N. (1941b). The local structure of turbulence incompressible viscous fluid for very large reynolds number. *Dokl. Akad. Nauk SSSR*, 30:299–303.

[109] Kolmogorov, A. N. (1941c). On degeneration (decay) of isotropic turbulence in an incompressible viscous liquid. *Dokl. Akad. Nauk SSSR*, 31:538–540.

[110] Kolmogorov, A. N. (1956). On the shannon theory of information transmission in the case of continuous signals. *IRE Transactions on Information Theory*, 2:102–108.

[111] Kolmogorov, A. N. (1958). New metric invariant of transitive dynamical systems and automorphism of lebesgue spaces. *Dokl. Akad. Nauk SSSR*, 119.

[112] Kolmogorov, A. N. (1962). A refinement of previous hypotheses concerning the local structure of turbulence in a viscous incompressible fluid at high reynolds number. *Journal of Fluid Mechanics*, 13:82–85.

[113] Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1):1–7.

[114] Kozachenko, L. and Leonenko, N. (1987). Sample estimate of entropy of a random vector. *Problems of Information Transmission*, 23:95–100.

[115] Kramer, G. (1998). *Directed Information for Channels with Feedback.* PhD thesis, Swiss Federal Institute of Technology Zurich.

[116] Kraskov, A., Stöbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, 69:066138–1/066138–16.

[117] Krieger, W. (1970). On entropy and generators of measure-preserving transformations. *Transactions of the American Mathematical Society*, 149(2):453–464.

[118] Krieger, W. (1972). On unique ergodicity. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, pages 327–346, Berkeley, Calif. University of California Press.

[119] Krstacic, G., Krstacic, A., Martinis, M., Vargovic, E., Knezevic, A., Smalcelj, A., Jembrek-Gostovic, M., Gamberger, D., and Smuc, T. (2002). Non-linear analysis of heart rate variability in patients with coronary heart disease. *Computers in Cardiology*, 29:673–675.

[120] Kullback, S. (1968). *Information Theory and Statistics.* Dover Publications.

[121] Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86.

[122] Lake, D. E. (2006). Renyi entropy measures of heart rate gaussianity. *IEEE Transactions on Bio-Medical Engineering*, 53(1):21–27.

[123] Lake, D. E. (2011). Improved entropy rate estimation in physiological data. *Conf. Proc. of the IEEE Eng Med Biol Soc (EMBC)*, pages 1463–1466.

[124] Lake, D. E., Richman, J. S., Griffin, M. P., and Moorman, J. R. (2002). Sample entropy analysis of neonatal heart rate variability. *Am J Physiol Regul Integr Comp Physiol*, 283(3):R789–R797.

[125] Landau, L. and Lifshitz, E. (1987). *Fluid Mechanics: Volume 6 of course of theoretical physics.* Pergamon Press.

[126] Leonarduzzi, R. (2014). *Análisis multifractal basado en coeficientes ondita líderes: selección automática del rango de escalamiento, formalismo multifractal basado en p-líderes y aplicación a señales biomedicas.* PhD thesis, Universidad Nacional del Litoral.

[127] Leonenko, N., Pronzato, L., and Savani, V. (2008). A class of rnyi information estimators for multidimensional densities. *Ann. Statist.*, 36(5):2153–2182.

[128] Lesne, A. (2014). Shannon entropy: a rigorous notion at the crossroads between probability, information theory, dynamical systems and statistical physics. *Mathematical Structures in Computer Science*, 24(3).

[129] Lesne, A., Blanc, J., and Pezard, L. (2009). Entropy estimation of very short symbolic sequences. *Phys. Rev. E*, 79:046208.

[130] Li, J., Zhang, J., W.Ge, and Liu, X. (2004). Multi-scale methodology for complex systems. *Chemical engineering science*, 59(8-9):1687–1700.

[131] Lim, J., Kwon, J. Y., Song, J., Choi, H., Shin, J. C., and Park, I. Y. (2014). Quantitative comparison of entropy analysis of fetal heart rate variability related to the different stages of labor. *Early Hum Dev*, 90(2):81–85.

[132] Lizier, J., Harré, M., Barnett, L., and Bossomaier, T. (2016). *An introduction to Transfer entropy: Information flows in complex systems.* Springer.

[133] Loftsgaarden, D. and Quesenberry, C. (1965). A nonparametric estimate of a multivariate density function. *The annals of Mathematical Statistics*, 36:1049–1051.

[134] Lohse, D. and Muller-Groeling, A. (1995). Bottleneck effects in turbulence: Scaling phenomena in r versus p space. *Physical Review Letters*, 74:1747–1750.

[135] Mack, Y. and Rosenblatt, M. (1979). Multivariate k-nearest neighbor density estimates. *Journal of multivariate analysis*, 9:1–15.

[136] Magenes, G., Signorini, M. G., and Arduini, D. (2000). Classification of cardiotocographic records by neural networks. *Proc. IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN)*, 3:637–641.

[137] Mandelbrot, B. and Van Ness, J. (1968). Fractional brownian motions fractional noises and applications. *SIAM Review*, 10(4):422–437.

[138] Markovitch, N. and Krieger, U. (2000). Nonparametric estimation of long-tailed density functions and its application to the analysis of World Wide Web traffic. *Performance Evaluation*, 42:205–222.

[139] Matsubara, T. (2010). Analytic minkowski functionals of the cosmic microwave background: Second-order non-gaussianity with bispectrum and trispectrum. *Phys. Rev. D*, 81:083505.

[140] Mauboussin, M. (2002). Revisiting market efficiency: the stock market as a complex adaptive system. *Journal of applied corporative finance*, 14(4):47–55.

[141] McGarigal, K., Wan, H., Zeller, K., Timm, B., and Cushman, S. (2016). Multi-scale habitat selection modeling: a review and outlook. *Landscape ecology*, 31(6):1161–1175.

[142] Miller, K. J., Zanos, S., Fetz, E. E., den Nijs, M., and Ojemann, J. G. (2009). Decoupling the cortical power spectrum reveals real-time representation of individual finger movements in humans. *Journal of Neuroscience*, 29(10):3132–3137.

[143] Miramontes, O. (1995). Order-disorder transitions in the behavior of ant societies. *Complexity*, 1(3):56–60.

[144] Mitchell, M. (2009). *Complexity: A guided tour.* Oxford University Press.

[145] Moon, Y., Rajagopalan, B., and Lall, U. (1995). Estimation of mutual information using kernel density estimators. *Physical Review E*, 52:2318–2321.

[146] Moore, D. and Yackel, J. (1977). Consistency properties of nearest neighbor density function estimators. *The annals of Statistics*, 5(1):143–154.

[147] Moore, G. P., Segundo, J. P., Perkel, D. H., and Levitan, H. (1970). Statistical signs of synaptic interaction in neurons. *Biophysical Journal*, 10(9):876 – 900.

[148] Ning, T. and Bronzino, J. (1990). Autoregressive and bispectral analysis techniques: Eeg applications. *IEEE Engineering in Medicine and Biology Magazine*, 9:47–50.

[149] Noullez, A., Wallace, G., Lempert, W., Miles, R. B., and Frisch, U. (1997). Transverse velocity increments in turbulent flow using the relief technique. *Journal of Fluid Mechanics*, 339:287307.

[150] Oboukhov, A. (1962). Some specific features of atmospheric turbulence. *Journal of Fluid Mechanics*, 13:77–81.

[151] Ou, Q., Jin, Y.-D., Zhou, T., Wang, B.-H., and Yin, B.-Q. (2007). Power-law strength-degree correlation from resource-allocation dynamics on weighted networks. *Phys. Rev. E*, 75:021102.

[152] Paluš, M. (2014). Multiscale atmospheric dynamics: Cross-frequency phase amplitude coupling in the air temperature. *Physical Review Letters*, 112:078702.

[153] Paluš, M., Komárek, V., Hrnčíř, Z., and Štěrbová, K. (2001). Synchronization as adjustement of information rates: Detection from bivariate time series. *Physical Review E*, 63(4):046211.

[154] Papana, A. and Kugiumtzis, D. (2008). Evaluation of mutual information estimators on nonlinear dynamic systems. *Nonlinear phenomena in complex systems*, 11(2):225–232.

[155] Papoulis, A. (1991). *Probability, Random Variables, ans Stochastic Processes.* McGraw-Hill International Editions, third edition edition.

[156] Parrondo, J. M. R., M. Horowitz, J., and Sagawa, T. (2015). Thermodynamics of information. *Nature Physics*, 11:131–139.

[157] Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33:1065–1076.

[158] Pincus, S. M. (1991). Approximate entropy as a measure of system-complexity. *Proc. Natl. Acad. Sci. U.S.A.*, 88(6):2297–2301.

[159] Pincus, S. M. (1995). Approximate entropy (ApEn) as a complexity measure. *Chaos*, 5 (1):110–117.

[160] Pincus, S. M. and Viscarello, R. R. (1992). Approximate entropy: a regularity measure for fetal heart rate analysis. *Obstet Gynecol*, 79(2):249–255.

[161] Porta, A., Bari, V., Bassani, T., Marchi, A., Tassin, S., Canesi, M., Barbic, F., and Furlan, R. (2013). Entropy-based complexity of the cardiovascular control in Parkinson disease: Comparison between binning and k-nearest-neighbor approaches. *Conf Proc IEEE Eng Med Biol Soc*, 2013:5045–5048.

[162] Priemer, R. (1991). *Introductory Signal Processing.* World Scientific.

[163] Prokopenko, M., Boschetti, F., and Ryan, A. (2008). An information-theoretic primer on complexity, self-organization, and emergence. *Complexity*, 15(1):11–28.

[164] Qian, H. (2003). *Fractional Brownian Motion and Fractional Gaussian Noise*, pages 22–33. Springer Berlin Heidelberg, Berlin, Heidelberg.

[165] Ray, A. and Chowdhury, A. R. (2010). On the characterization of non-stationary chaotic systems: Autonomous and non-autonomous cases. *Physica A*, 389:5077–5083.

[166] Rényi, A. (1961). On measures of information and entropy. In *Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability*, pages 547–561.

[167] Richardson, L. F. (1921). Some measurements of atmospheric turbulence. *Philosophical Transactions of the Royal Society of London. Series A, Containing papers of a Mathematical or Physical Character*, 221:1–28.

[168] Richman, J. S. and Moorman, J. R. (2000a). Physiological time-series analysis using approximate entropy and sample entropy. *Am J Physiol Heart Circ Physiol*, 278(6):H2039–H2049.

[169] Richman, J. S. and Moorman, J. R. (2000b). Time series analysis using approximate entropy and sample entropy. *Biophysical Journal*, 78(1):218A–218A.

[170] Rissanen, J. and Wax, M. (1987). Measures of mutual and causal dependence between two time series. *IEEE Transactions on Information Theory*, 33(4):598–601.

[171] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The annals of Mathematical Statistics*, 27:832–837.

[172] Schreiber, T. (2000). Measuring information transfer. *Physical Review Letters*, 85(2):461–464.

[173] Scott, D. (1977). Kernel density estimation revisited. *Nonlinear Analysis, Theory, Methods and Applications*, 1:339–372.

[174] Scott, D. (1985). Averaged shifted histograms: Effective nonparametric density estimators in several dimensions. *The annals of Statistics*, 13:1024–1040.

[175] Scott, D. W. (2015). *Multivariate Density Estimation: Theory, practice and Visualization*. Wiley Series in Probability and Statistics. John Wiley and Sons, second edition edition.

[176] Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, XXVII:388–427.

[177] Shannon, C. and Weaver, W. (1949). *The Mathematical Theory of Information*. University of Illinois Press.

[178] She, Z. and Lévêque, E. (1994). Universal scaling laws in fully developed turbulence. *Phys. Rev. Lett.*, 72:336–339.

[179] Siegelmann, H. (2010). Complex systems science and brain dynamics. *Frontiers in Computational Neuroscience*, 4:7.

[180] Signorini, M., Magenes, G., Cerutti, S., and Arduini, D. (2003). Linear and non-linear parameters for the analysis of fetal heart rate signal from cardiotocographic recordings. *IEEE Transactions on Biomedical Engineering*, 50:365–374.

[181] Siira, S. M., Ojala, T. H., Vahlberg, T. J., Rosén, K. G., and Ekholm, E. M. (2013). Do spectral bands of fetal heart rate variability associate with concomitant fetal scalp pH? *Early Hum Dev*, 89(9):739–742.

[182] Silverman, B. (1986). Monographs on statistics and applied probability. In *Density Estimation for Statistics and Data Analysis*, volume 26. Chapman and Hall/CRC.

[183] Sinai, Y. G. (1959). On the notion of entropy of a dynamical system. *Doklady of Russian Academy of Sciences*, 124:768–771.

[184] Singh, H., Misra, N., Hnizdo, V., Fedorowicz, A., and Demchuk, E. (2003). Nearest neighbor estimates of entropy. *American Journal of Mathematical and Management Sciences*, 23:301–321.

[185] Slepian, Z. and Eisenstein, D. (2015). Computing the three-point correlation function of galaxies in $\mathcal{O}(\mathrm{n}^2)$. *Monthly Notices of the Royal Astronomical Society*, 454(4):4142–4158.

[186] Spilka, J., Abry, P., Goncalves, P., and Doret, M. (2014a). Impacts of first and second labour stages on hurst parameter based intrapartum fetal heart rate analysis. In *Computing in Cardiology Conference (CinC)*, pages 777–780.

[187] Spilka, J., Chudáček, V., Koucký, M., Lhotská, L., Huptych, M., Janků, P., Georgoulas, G., and Stylios, C. (2012). Using nonlinear features for fetal heart rate classification. *Biomedical Signal Processing and Control*, 7(4):350–357.

[188] Spilka, J., Frecon, J., Leonarduzzi, R., Pustelnik, N., Abry, P., and Doret, M. (2016a). Sparse support vector machine for intrapartum fetal heart rate classification. *IEEE Journal of Biomedical and Health Informatics*, PP:1–1.

[189] Spilka, J., Leonarduzzi, R., Chudáček, V., Abry, P., and Doret, M. (2016b). Fetal heart rate classification: First vs second stage of labor. In *Proc. Int. Workshop Biosignal Interpretation (BSI)*, Osaka, Japan.

[190] Spilka, J., Roux, S., Garnier, N., Abry, P., Goncalves, P., and Doret, M. (2014b). Nearest-neighbor based wavelet entropy rate measures for intrapartum fetal heart rate variability. *Conf. Proc. of the IEEE Eng. Med. Biol. Soc. (EMBC)*, pages 2813–2816.

[191] Staicu, A. (2002). *Intermittency in turbulence.* PhD thesis, University of technology Eindhoven.

[192] Steuer, R., Kurths, J., Daub, C., Weise, J., and Selbig, J. (2002). The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics*, 18:S231–S240.

[193] Strogatz, S. H. (1994). *Nonlinear Dynamics and Chaos: with applications to Physics, Biology, Chemistry and Engineering.* Perseus Books.

[194] Subcommittee on Standards in Perinatal Medicine, F. (1987). Guidelines for the Use of Fetal Monitoring. *Int J Gynaecol Obstet*, 25:159–167.

[195] Takada, M. and Jain, B. (2002). The three-point correlation function in cosmology. *Monthly Notices of the Royal Astronomical Society*, 340(2).

[196] Takahira, R., Tanaka-Ishii, K., and Debowski, L. (2016). Entropy rate estimates for natural language —a new extrapolation of compressed large-scale corpora. *Entropy*, 18:364.

[197] Takens, F. (1981). *Dynamical Systems and Turbulence, Warwick 1980: Proceedings of a Symposium Held at the University of Warwick 1979/80*, chapter Detecting strange attractors in turbulence, pages 366–381. Springer Berlin Heidelberg, Berlin, Heidelberg.

[198] Tarter, M. and Kronmal, R. (1976). An introduction to the implementation and theory of nonparametric density estimation. *The American Statistician*, 30:105–112.

[199] Tennekes, H. and Wyngaard, J. (1972). The intermittent small-scale structure of turbulence: data-processing hazards. *J. Fluid Mech*, 55:93–103.

[200] Terrel, G. and Scott, D. (1980). On improving convergence rates for nonnegative kernel density estimator. *The annals of Statistics*, 8:1160–1163.

[201] Terrel, G. and Scott, D. (1992). Variable kernel density estimation. *The annals of Statistics*, 20:1236–1265.

[202] Theiler, J. (1986). Spurious dimension from correlation algorithms applied to limited time-series data. *Physical Review A*, 34:2427–2432.

[203] Tobin, R. and Houghton, C. (2013). A kernel-based calculation of information on a metric space. *Entropy*, 15:4540–4552.

[204] Tourassi, G., Frederick, E., Markey, M., and Jr, C. F. (2001). Application of the mutual information criterion for feature selection in computer-aided diagnosis. *Medical Physics*, 28:2394–2402.

[205] Trivedi, P., Seshadri, T. R., and Subramanian, K. (2012). Cosmic microwave background trispectrum and primordial magnetic field limits. *Phys. Rev. Lett.*, 108:231301.

[206] Van den Broeck, C. (2014). Stochastic thermodynamics: A brief introduction. In Bechinger, C., Sciortino, F., and Ziherl, P., editors, *Proceedings of the International School of Physics "Enrico Fermi": Physics of Complex Colloids*, pages 155–193.

[207] Van Laar, J., Porath, M., Peters, C., and Oei, S. (2008). Spectral analysis of fetal heart rate variability for fetal surveillance: Review of the literature. *Acta Obstet. Gynecol. Scand.*, 87(3):300–306.

[208] Venugopal, V., Roux, S. G., Foufoula-Georgiou, E., and Arneodo, A. (2006). Revisiting multifractality of high-resolution temporal rainfall using a wavelet-based formalism. *Water Resources Research*, 42:W06D14.

[209] Vicente, R., Wibral, M., Lindner, M., and Pipa, G. (2010). Transfer entropy- a model-free measure of effective connectivity for the neurosciences. *Journal of Computational Neuroscience*, 30:45–67.

[210] Victor, J. (2002). Binless strategies for estimation of information from neural data. *Phys. Rev. E*, 66:051903.

[211] Vu, V. Q., Yu, B., and Kass, R. E. (2009). Information In The Non-Stationary Case. *Neural Computation*, 21(3):688–703.

[212] Walter, G. and Blum, J. (1979). Probability density estimation using delta sequences. *The annals of Statistics*, 7:328–340.

[213] Walters-Williams, J. and Li, Y. (2009). Estimation of mutual information: A survey. In Wen, P., Li, Y., Polkowski, L., Yao, Y., Tsumoto, S., and Wang, G., editors, *Rough Sets and Knowledge Technology*, pages 389–396, Berlin, Heidelberg. Springer Berlin Heidelberg.

[214] Warrick, P., Hamilton, E., Precup, D., and Kearney, R. (2010). Classification of normal and hypoxic fetuses from systems modeling of intrapartum cardiotocography. *IEEE Transactions on Biomedical Engineering*, 57(4):771–779.

[215] Wendt, H., Abry, P., and Jaffard, S. (2007). Bootstrap for empirical multifractal analysis. *IEEE Signal Proces. Mag.*, 24(4):38–48.

[216] Wendt, H., Roux, S., Jaffard, S., and Abry, P. (2009). Wavelets leaders and bootstrap for multifractal analysis of images. *Signal Processing*, 89:1100–1114.

[217] Wibral, M., Pampu, N., Priesemann, V., Siebenhhner, F., Seiwert, H., Lindner, M., Lizier, J. T., and Vicente, R. (2013). Measuring information-transfer delays. *PLOS ONE*, 8(2):1–19.

[218] Wiener, N. (1956). *Modern Mathematics for engineers*, volume 1, chapter 8 *The theory of Prediction.* McGraw-Hill, New York.

[219] Williams, P. L. and Beer, R. D. (2010). Nonnegative decomposition of multivariate information. *CoRR*, abs/1004.2515.

[220] Williams, P. L. and Beer, R. D. (2011). Generalized Measures of Information Transfer. *ArXiv e-prints.*

[221] Wollstadt, P., Martínez-Zarzuela, M., Vicente, R., Díaz-Pernas, F. J., and Wibral, M. (2014). Efficient transfer entropy analysis of non-stationary neural time series. *PLOS ONE*, 9(7):1–21.

[222] Wyner, A. (1978). A definition of conditional mutual information for arbitrary ensembles. *Information and Control*, 38(1):51 – 59.

[223] Xiong, W.and Faes, L. and Ivanov, P. C. (2017). Entropy measures, entropy estimators, and their performance in quantifying complex dynamics: Effects of artifacts, nonstationarity, and long-range correlations. *Phys. Rev. E*, 95:062114.

[224] Yamasaki, K., Gozolchiani, A., and Havlin, S. (2008). Climate networks around the globe are significantly affected by el niño. *Phys. Rev. Lett.*, 100:228501.

[225] Zhang, W. (2011). Constructing ecological interaction networks by correlation analysis: hints from community sampling. *Network Biology*, 1(2):81–98.

[226] Zografos, K. and Nadarajah, S. (2005). Expressions for renyi and shannon entropies for multivariate distributions. *Statistics & Probability Letters*, 71(1):71–84.