Ph. D. Thesis submitted to the
Physics Department – University of Trento
in fulfilment of the requirements for the
Degree of Philosophiæ Doctor in Physics
Under the Supervision of
Prof. Leonardo Ricci

ADDRESSING NONLINEAR SYSTEMS WITH INFORMATION-THEORETICAL TECHNIQUES

MICHELE CASTELLUZZO
XXXV cycle

Abstract

The study of experimental recording of dynamical systems often consists in the analysis of signals produced by that system. Time series analysis consists of a wide range of methodologies ultimately aiming at characterizing the signals and, eventually, gaining insights on the underlying processes that govern the evolution of the system.

A standard way to tackle this issue is spectrum analysis, which uses Fourier or Laplace transforms to convert time-domain data into a more useful frequency space. These analytical methods allow to highlight periodic patterns in the signal and to reveal essential characteristics of linear systems. Most experimental signals, however, exhibit strange and apparently unpredictable behavior which require more sophisticated analytical tools in order to gain insights into the nature of the underlying processes generating those signals. This is the case when nonlinearity enters into the dynamics of a system. Nonlinearity gives rise to unexpected and fascinating behavior, among which the emergence of deterministic chaos. In the last decades, chaos theory has become a thriving field of research for its potential to explain complex and seemingly inexplicable natural phenomena.

The peculiarity of chaotic systems is that, despite being created by deterministic principles, their evolution shows unpredictable behavior and a lack of regularity. These characteristics make standard techniques, like spectrum analysis, ineffective when trying to study said systems. Furthermore, the irregular behavior gives the appearance of these signals being governed by stochastic processes, even more so when dealing with experimental signals that are inevitably affected by noise.

Nonlinear time series analysis comprises a set of methods which aim at overcoming the strange and irregular evolution of these systems, by measuring some characteristic invariant quantities that describe the nature of the underlying dynamics. Among those quantities, the most notable are possibly the Lyapunov exponents, that quantify the unpredictability of the system, and measure of dimension, like correlation dimension, that unravel the peculiar geometry of a chaotic

system's state space. These methods are ultimately analytical techniques, which can often be exactly estimated in the case of simulated systems, where the differential equations governing the system's evolution are known, but can nonetheless prove difficult or even impossible to compute on experimental recordings.

A different approach to signal analysis is provided by information theory. Despite being initially developed in the context of communication theory, by the seminal work of Claude Shannon in 1948, information theory has since become a multidisciplinary field, finding applications in biology and neuroscience, as well as in social sciences and economics. From the physical point of view, the most phenomenal contribution from Shannon's work was to discover that entropy is a measure of information and that computing the entropy of a sequence, or a signal, can answer to the question of how much information is contained in the sequence. Or, alternatively, considering the source, i.e. the system, that generates the sequence, entropy gives an estimate of how much information the source is able to produce. Information theory comprehends a set of techniques which can be applied to study, among others, dynamical systems, offering a complementary framework to the standard signal analysis techniques. The concept of entropy, however, was not new in physics, since it had actually been defined first in the deeply physical context of heat exchange in thermodynamics in the 19th century. Half a century later, in the context of statistical mechanics, Boltzmann reveals the probabilistic nature of entropy, expressing it in terms of statistical properties of the particles' motion in a thermodynamic system. A first link between entropy and the dynamical evolution of a system is made. In the coming years, following Shannon's works, the concept of entropy has been further developed through the works of, to only cite a few, Von Neumann and Kolmogorov, being used as a tool for computer science and complexity theory. It is in particular in Kolmogorov's work, that information theory and entropy are revisited from an algorithmic perspective: given an input sequence and a universal Turing machine, Kolmogorov found that the length of the shortest set of instructions, i.e. the program, that enables the machine to compute the input sequence was related to the sequence's entropy. This definition of the complexity of a sequence already gives

hint of the differences between random and deterministic signals, in the fact that a truly random sequence would require as many instructions for the machine as the size of the input sequence to compute, as there is no other option than programming the machine to copy the sequence point by point. On the other hand, a sequence generated by a deterministic system would simply require knowing the rules governing its evolution, for example the equations of motion in the case of a dynamical system. It is therefore through the work of Kolmogorov, and also independently by Sinai, that entropy is directly applied to the study of dynamical systems and, in particular, deterministic chaos. The so-called Kolmogorov-Sinai entropy, in fact, is a well-established measure of how complex and unpredictable a dynamical system can be, based on the analysis of trajectories in its state space.

In the last decades, the use of information theory on signal analysis has contributed to the elaboration of many entropy-based measures, such as sample entropy, transfer entropy, mutual information and permutation entropy, among others. These quantities allow to characterize not only single dynamical systems, but also highlight the correlations between systems and even more complex interactions like synchronization and chaos transfer. The wide spectrum of applications of these methods, as well as the need for theoretical studies to provide them a sound mathematical background, make information theory still a thriving topic of research.

In this thesis, I will approach the use of information theory on dynamical systems starting from fundamental issues, such as estimating the uncertainty of Shannon's entropy measures on a sequence of data, in the case of an underlying memoryless stochastic process. This result, beside giving insights on sensitive and still-unsolved aspects when using entropy-based measures, provides a relation between the maximum uncertainty on Shannon's entropy estimations and the size of the available sequences, thus serving as a practical rule for experiment design. Furthermore, I will investigate the relation between entropy and some characteristic quantities in nonlinear time series analysis, namely Lyapunov exponents. Some examples of this analysis on recordings of a nonlinear chaotic system are also provided. Finally, I will discuss other entropy-based measures,

among them mutual information, and how they compare to analytical techniques aimed at characterizing nonlinear correlations between experimental recordings. In particular, the complementarity between information-theoretical tools and analytical ones is shown on experimental data from the field of neuroscience, namely magnetoencefalography and electroencephalography recordings, as well as meteorological data.

Contents

Chapter 1

Introduction

The motion of objects is among the first topics studied in Physics, from observation of celestial objects in ancient Greece and later by Galileo and Newton up to modern physics with the theory of quantum mechanics. Beside the mechanical aspects, the concept of a dynamical system broadened to involve any system that "moves", in the sense that it evolves with time. This eventually surpassed the limits of physical systems, and the theory of dynamical systems now involves a multitude of fields, including physiology, climate science and even social studies and economics.

The techniques used to study dynamical systems are therefore general and do not necessarily make assumptions on the nature of the underlying process generating the dynamics. These techniques stem not only from Physics, but from several branches of mathematics, including statistics and information theory.

Studying a dynamical system ultimately consists in trying to characterize its evolution, recognizing possible patterns and most importantly estimating fundamental invariant quantities that govern its behaviour. In other words, understanding the essential properties of the dynamics governing the system.

Since many dynamical systems have been observed to exhibit irregular behaviour, a particular interest has been taken into nonlinear phenomena. Among these, the emergence of chaos has gained a particular interest, as a possible way to describe the underlying dynamics of the systems under study when the assumption of a purely stochastic behaviour was found insufficient.

The evolution of theoretical dynamical system often unfolds in a multi-dimensional space, known as the *state space*. For theoretical systems, when differential equations governing the evolution are known and can be solved by numerical integra-

tion, finding the evolution invariants in order to characterize the evolution is often possible by means of analytical methods proper of dynamical systems theory.

Alas, in experiments, knowledge of the evolution differential equations is missing and dedicated tools are required to recover information about the underlying nature of the systems under study. Also, full access to a system's state space is not always available, but just a scalar signal originating from the system may be available, after measuring sampled points of the signal. Therefore signal analysis is critical for experimental research of such systems.

Time series analysis is possibly the most widespread approach to deal with experimental signals, and comprises a set of methods that ultimately try to grasp fundamental properties of the system's dynamics. These properties will make up the starting point for the theoretical modeling, with the ultimate goal of recovering, if possible, the equations governing the system's evolution.

Among the most widespread methods of time series analysis, spectral analysis by using Fourier or Laplace transforms are often used to convert time-domain data into the frequency space, which allows for important properties of linear systems to be shown. Unfortunately, these methods prove to be inefficient in the case of nonlinear, and chaotic, systems. The basic requirement of linearity, i.e. that small differences in the cause lead to small differences in their effects, is not valid in the case of chaotic regimes.

Although being created by deterministic principles, chaos is a dynamical state marked by unpredictability and lack of regularity. Signals created by chaotic dynamical systems are irregular, sensible to initial conditions, difficult to predict and their evolution within the system's state space can exhibit strange properties. Due to these characteristics, misclassification of chaotic signals as stochastic occurs when inappropriate tools, like linear methods, are used. The recognition and classification of chaos need the evaluation of dynamical invariants that quantify the complicated aspects of the evolution of a chaotic system. Lyapunov exponents, for example, serve as a predictability indicators, while quantities like attractor's dimensions reveal the strange geometry of a chaotic system's state space.

The identification of chaos as a possible dynamical domain for nonlinear sys-

tems gave way to search for chaotic behavior in experimental recordings. Chaotic behavior, therefore, is often searched in several fields like climate science, social science, neurology, and even economics. Detection of a chaotic regime, as opposed to a stochastic one, would reveal the presence of determinism in a system, as well as provide a starting point for predicting the system's future states.

In the next sections, a mathematical definition of a dynamical system will be given, as well as a brief overview of the most widely used method of nonlinear time series analysis.

From dynamical systems to time series

A dynamical system $\mathcal{D}$ evolving in time can be described by a set of vectors $\mathbf{s}(t)$ in a state space $S \subset \mathbb{R}^d$, where d is the dimension of the system. We can define a time-evolution operator $\mathcal{T}$ such that the state of the system at a later time can be written as:

$$\mathbf{s}(t + dt) = \mathcal{T}_{dt}[\mathbf{s}(t)].$$

Typically, the trajectories of a chaotic system, provided long enough observations, will converge to the subset S, in which the evolution is contained, and that is called an attractor, meaning that it "attracts" the trajectories.

In theoretical systems, the description of the evolution operator $\mathcal{T}$ is known, i.e. a set of differential equations (or discrete difference equations) is provided. As an example, we will use the equations of the Röessler attractor [1], a widely used theoretical chaotic model. The Röessler attractor is a a three-dimensional systems, so the vector $\mathbf{s}(t) = \{x(t), y(t), z(t)\}$. The differential equations describing its dynamics are:

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= -y - z \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= x + a\,y \\ \frac{\mathrm{d}z}{\mathrm{d}t} &= b + z\,(x - c)\ . \end{aligned} \tag{1.1}$$

A common choice of parameters is $a = 0.3$, $b = 0.2$, $c = 5.7$, which we will use as a standard to provide examples of applications of nonlinear analysis tools.

Figure 1.1 shows the state space of the Rössler attractor, obtained by integration of Eqs. (1.1) by Runge–Kutta Prince–Dormand (8, 9) method [2].

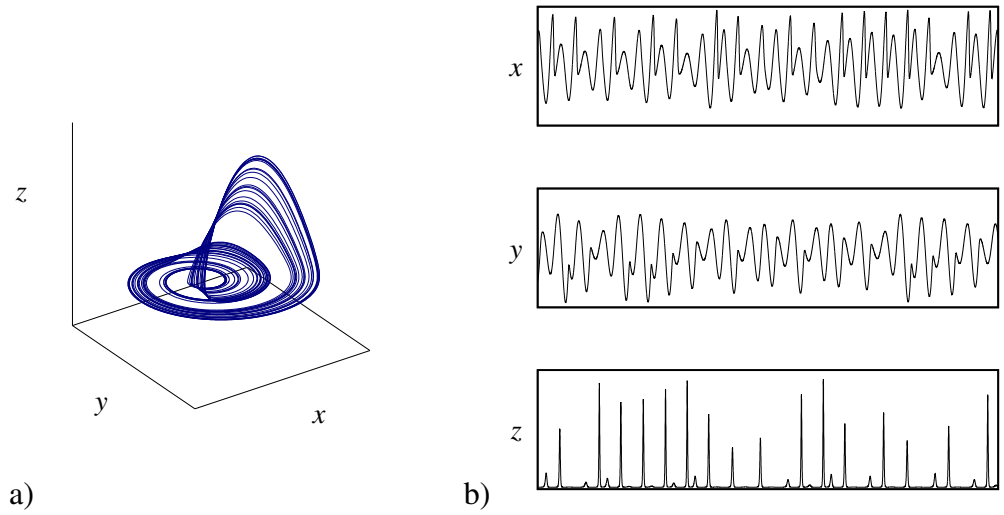

Figure 1.1: State space evolution (a) of a Rössler attractor and time series of its components (b)

In the case of experimental systems, the space state vectors $\mathbf{s}(t)$ are most often not available, but a single scalar sequence $s(t)$ is available. This signal is the result of a measuring process that can be described as $s(t) = \mathcal{M}[\mathbf{s}(t)]$, where $\mathcal{M}$ is the measuring function. The signal $s(t)$ is then digitally sampled so that only a discrete sequence of sampled points remains available. By defining as T the sampling period, we can write the sampled sequence as $x[n] = s(nT)$.

We are going now to describe tools from nonlinear time series analysis that, starting from the sequence $x[n]$ defined above, try to recover dynamical invariants that can characterize the underlying nature of the system's evolution. In particular, we are interested in tools that may give evidence of chaotic behaviour.

Complexity in nonlinear time series analysis

The methods we are going to describe are indicators of a system degree of complexity. In general, having a higher-degree of complexity indicates that the system has a more disordered and irregular dynamic. Chaotic systems usually show an

intermediate level of complexity: despite their evolution is irregular and unpredictable, they are still controlled by deterministic rules. Periodic systems have low-level complexity, while a higher-level complexity represents less regular rules governing the dynamics, as in the case of purely stochastic processes.

Quantify the complexity of time series data using methods derived from nonlinear dynamics, usually requires to reconstruct the dynamic system states based on the phase-space. Reconstruction of the phase-space from a scalar sequence is performed by means of embedding, as proven by Taken. One example of embedding, namely time-delay embedding, consists in constructing vectors $\mathbf{x}$ so that:

$$\mathbf{x}(i) = \{x(i), x(i-\tau), x(i-2\tau), \ldots, x(i-(m-1)\tau)\},$$

where (m, τ) are respectively the dimension and the time-lag parameters of the embedding. An example of time-delayed embedding of the x-component from the Rössler attractor is shown in Figure1.2. The choice of optimal parameters for the embedding procedure is not a trivial subject: despite some general criteria are present in literature, new methods addressing this issue are still an important topic of research [3].

Provided that phase-state reconstruction has been properly obtained, the exploration of the attractor properties can be performed by measuring the invariants of the system's evolution. In terms of exploring attractor properties, the major categories of attractor invariants commonly used to test complexity out of the system states are Lyapunov exponents and fractal dimensions.

Lyapunov exponent investigates how system states change over time in terms of the exponential divergence (or convergence) of initially close trajectories. The growing rate of separation between nearby trajectories is a clear evidence of the system's sensitivity to initial conditions. A positive Lyapunov exponent, for example, indicates the exponential divergence of nearby trajectories; a negative one indicates the exponential convergence of trajectories; and a zero value indicates that expanding and contracting dynamics are in balance. Given a trajectory $\mathbf{x}_i(t)$ and its nearest neighbour $\tilde{\mathbf{x}}_i(t)$, so that the two trajectories are arbitrary close at

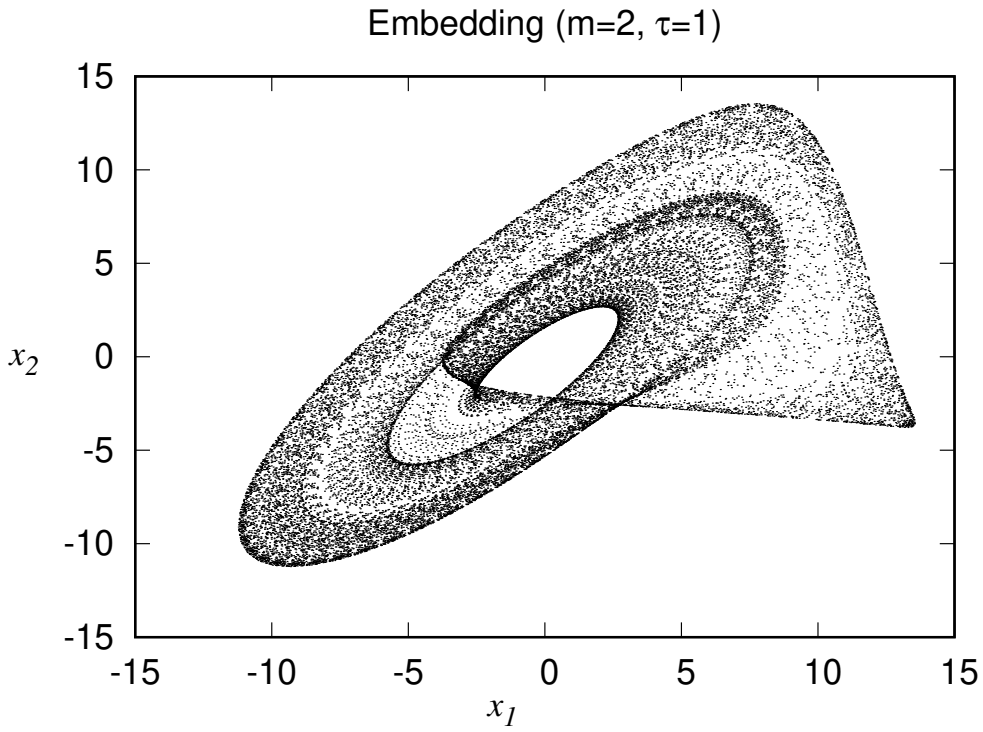

Figure 1.2: Two-dimensional time-delay embedding of a Rössler attractor.

$t = t_0$, the Lyapunov exponent λ_i can be described by:

$$\lambda_i = \lim_{t \to \infty} \frac{1}{t} \ln \frac{d(t)}{d(t_0)}, \tag{1.2}$$

where $d(t) = \|\mathbf{x}_i(t) - \tilde{\mathbf{x}}_i(t)\|$ is the distance between the trajectories at a given time t.

In theoretical systems, Lyapunov exponents can be derived by analytical methods out of the differential equations governing the evolution. One commonly used method, therefore known as the *standard* method was proposed by Benettin [4, 5]. In general, we talk about the Lyapunov spectrum, since the number of exponents is equal to the dimension of the system space state. For experimental data, given an m-dimensional phase-space reconstruction, m Lyapunov exponents exist, but the main variations are described by the largest exponent: this means not only that the largest exponent is the one bearing the most information regarding the system's evolution, but also that it's the one most easily measurable out of the reconstructed phase-space, particularly when the size of the available samples se-

quence is limited. Some algorithms also exist to compute the whole Lyapunov spectrum out of nonlinear experimental signals[6, 7], but while they coincide on evaluation of the mximum Lyapunov exponent (MLE), the results on the others exponents may vary as they depend on the algorithms' parameters [8].

A theoretical definition of Lyapunov exponents came with Pesin, but a mathematical formulation suitable for computation came later with Wolf et al. [9]. Numerous modification and adjustments have been made since, for instance, Rosenstein et al. [10] created a much faster and simpler version to decrease computing complexity and successfully guarantee trustworthy values even for small-sized datasets and later on a more reliable version was provided by Gao and Zheng [11, 12, 13] to compute MLE out of embedded time series. Finally, Kowalik and Elbert [14] proposed a modified version to compute the MLE in a time-dependent way.

The dimensional complexity of phase space can also be estimated using different fractal dimensions from fractal theory. Hausdorff [15] first offered a precise theoretic definition of dimension. A geometric object's dimension indicates how large it is in space, and the dimension of the attractor in phase space may indicate how many degrees of freedom (or how complex) the system is. Particularly, a larger dimension of the attractor denotes a system's increased spatial extensiveness, or complexity of the time series under study. Different fractal dimensions have been introduced to study dimensional complexity, among these correlation dimension [16], box-counting dimension [17] and Rényi dimensions [18] are worth mentioning.

As an example, we will give the definition of correlation dimension, possibly the most used among fractal dimensions. First let us define as $d_{i,j} = \|\mathbf{x}_i - \mathbf{x}_j\|$ the distance between two arbitrary vectors in the reconstructed m-dimensional phase space. Then the correlation integral $C(m, r)$ can be written as:

$$C(m, r) = \frac{2 \sum_{i<j} \Theta(r - d_{i,j})}{(N - (m-1)\tau)\,(N - (m-1)\tau - 1)}\,, \tag{1.3}$$

where Θ is the Heaviside function and r a threshold distance. The correlation

dimension is then:

$$C_m = \lim_{r \to 0} \frac{\ln C(m, r)}{\ln r} .$$

From this definition, we can see how correlation dimension, and fractal dimensions in general, try to capture the spatial complexity of the system in terms of scaling exponents of the distribution of distances between phase space vectors. Some alternative methods have been proposed to compute correlation dimension, which do not require to compute directly the correlation integral in Eq.(1.3). Among these, one method exploits the asymptotic behaviour of the divergence exponent $\Lambda(t, d_0)$ for $t \to \infty$ [19]. The divergence exponent, as we will discuss more in detail in Chapter 4, can be written as $\Lambda(t, d_0) = \langle \ln \left(d_{i,j}(t)/d_0 \right) \rangle$ computed at different values of the initial distance d_0.

Finally, a qualitative analysis of a system complexity can be performed by several useful diagrams, which offer a vivid representation of the system states. Some examples are phase portraits [20], Poincare sections [21], and recurrence plots (RP) [22].

Poincarè sections are made by selecting a plane in the state space, or in the reconstructed phase space, and plotting the points in which the attractor's trajectories cross into the selected plane. Usual choices of planes consist in putting one of the variables equal to a constant values. A different, but still widely used version of Poincarè sections, is shown in Figure 1.3. This diagram was made by plotting consecutive maxima of the x-component (X_n, X_{n+1}) of the Rössler attractor, that ultimately corresponds to selecting the plane where $\mathrm{d}x/\mathrm{d}t = 0$ and the second derivative is negative. This example shows clearly the deterministic nature of the Rössler attractor: despite the apparently irregular evolution of the trajectories and the lack of periodicity, the consecutive maxima of the x-component follow a regular pattern.

Applications of nonlinear analysis methods

As we discussed in the previous sections, nonlinear systems have gained a lot of interest particularly because of their irregular behaviour, which so often is found in real systems.

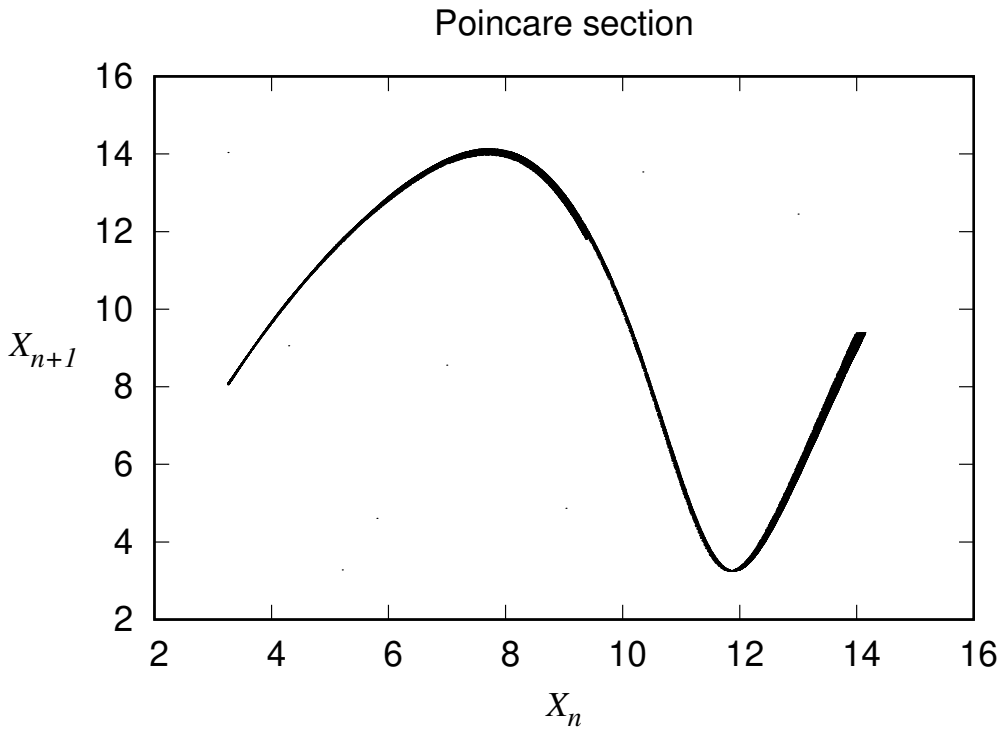

Figure 1.3: Poincarè diagram of consecutive maxima of the x-component of the Rössler attractor.

One of the first experimental evidences of chaotic behaviour were found in optical systems, in particular concerning lasers. Ikeda [23] presented one of the earliest articles on chaos in optical systems in 1979, discussing a model of light in ring cavities, which showed evidence of "optical turbulence" [24] that eventually leads to chaotic behavior. Many experiments using pulsed lasers [25], as well as frequency-modulated [26], and gain-modulated [27] lasers, corroborated these theoretical predictions. Evidence of chaotic behaviour were also found in lasers with feedback mechanisms [28, 29], among which it is worth mentioning the work of Arecchi et al. [30, 31] on CO_2 lasers. From the latter, a set of equations was recently derived that makes up a Minimal Universal Model for chaos, which will be described in more detail in Chapter 4. Among other evidences of chaos in lasers, some systems [32, 33] were found to be described by equations similar to a Lorenz system [34], thus providing an experimentally proven equivalence.

Beside optical systems, electronic circuits were identified as ideal systems to study chaotic regimes. The first observations of chaos in electronics were made in the early 1980s, when one of the most famous chaotic circuits was implemented by Chua [35, 36]. The relative easiness for electronic circuits to be customized makes them an efficient instrument for generating different chaotic regimes [37, 38], as well as simulating theoretical models [39, 40]. Recently, an implementation of the Minimal Universal Model was made on an electronic platform, using analog components as well as digital ones to tune the model's parameters in a fast and repeatable way [41]. Furthermore, the cheap cost and scalability of electronic setups allows to investigate chaotic behavior arising from interaction of multiple systems, as in the case of network structures [42] and for the study of synchronization phenomena [43, 44, 45].

Beside experimental evidence of chaos on controlled systems like the ones mentioned above, there are several fields, beside physics, where the methods discussed in the previous section have been applied to evaluate complexity. Among these, economics, life science, geophysics and climate science, may be the most common application areas of nonlinear dynamics approaches.

The study of the application of methods derived from nonlinear dynamics to economics has become a popular topic over the last 20 years; a seminal work is provided by Brock and Sayers [46], in which tests are conducted for the presence of low-dimensional deterministic chaos in U.S. macroeconomic data, such as the unemployment rate, GNP, and industrial production. More recently, recurrence plots were used to analyze various stock indexes in global stock markets [47], and a complex behaviour was discovered in Hong Kong's stock market by means of correlation dimension and Lyapunov exponents. Barkoulas et al. [48] examined the chaotic properties of oil market indeces, using the correlation dimension, Lyapunov exponents, and recurrence plots.

In earth science, Ponyavin [49] applied nonlinear analysis to study geomagnetic and climate records for sunspot counts, revealing a more intense solar activity in the last 60 years. Other studies suggested that irregular oscillations in the number of sunspots, may be created by chaotic dynamics [50] and have a self-

similar structure [51]. The sunspot number, on the other hand, was demonstrated to be well represented by a relaxation oscillator model with stochastic perturbations [52], defying the assertion of chaos.

In geophysics, the idea of an underlying deterministic nature prompted the use of nonlinear time series analysis techniques in this field [53], with the aim of finding a deterministic, and therefore predictable, behaviour: chaos was claimed to be present in critical events like seismic activity [54, 55, 56] and volcanic eruptions [57], as well as sea level oscillations [58]. Nonetheless, several studies showed that earthquakes may be represented by stochastic processes [59], proving the difficulty of reaching a definitive result on this topic.

Climate science is another central field where researches expect to find chaotic regimes at certain conditions and dimensional scales. After all, the well-known Lorenz system is a model of atmospheric convection [60]. Nonlinear analytical approaches have thus been used to investigate atmospheric data [61, 62], with the goal of improving forecasts. Analysis of wind recordings [63, 64] yielded promising results for chaos identification in meteorological data. Likewise, several questions have been raised about the proper application of nonlinear analytical techniques to climate data, as well as the interpretation of the results [65, 66]. In many circumstances, such as for daily temperature sequences, using stochastic modeling proves to be sufficient to correctly extract system features [67].

For what regards life sciences, nonlinear analysis methods find a wide range of applications in the study of the human heart and the brain. As an example, Yilmaz and Güler [68] were able to find two indices useful in diagnosing internal carotid artery stenosis, through the estimation of correlation dimension and MLE on the blood flow of healthy subjects and patients. Some groundbreaking works have been conducted by Rapp et al. [69], who addressed spontaneous brain activity in the motor cortex of a monkey using chaotic modeling.

For what regards human brain signals, electroencephalography (EEG) or magnetoencephalography (MEG) are frequently used to record electromagnetic signals from the brain cortex. Brain recordings display irregular evolutions, prompting interest from neuroscientist to find evidence of chaotic regimes [70]. Babloy-

antz et al. [71] were the first to measure chaos in human brain EEG sleep recordings. Thomasson et al. [72] later analyzed epileptic EEG activity and the results showed substantial transients in pre-ictal activity recordings. Other studies are present on the possibility of predicting epileptic seizures using nonlinear approaches [73]. Some evidences were found of "brain attractor" increasing its dimensionality at the occurrence of epileptic seizures [74, 75].

Nonetheless, some studies compared nonlinear techniques with other measures, including linear ones, finding that linear measures can have performances similar to nonlinear ones in EEG data analysis [76]. Similarly, McSharry et al. [77] proposed that some linear approaches are able to identify pre-ictal transitions, as efficiently as nonlinear methods, suggesting that a combination of linear and nonlinear methods would be a useful way for accurate seizure prediction. In conclusion, despite several evidences of a chaotic nature in the brain systems [78, 79] as well as to model activity of single neurons [80, 81], it has been pointed out that the algorithms used to identify chaos were frequently misused, showing the necessity for the creation of new nonlinear metrics to estimate the complexity of those systems, rather than necessarily proving the presence of deterministic chaos [82].

In the next Chapter 2, we are going to present the framework of Information Theory and discuss Shannon entropy, as well as other information-theoretical tools applied to dynamical systems and time series in general.

Despite it's wide use, entropy estimations, in particular Shannon entropy, still has some unresolved issues in what regards estimating it's variance when an entropy estimator is used on a set of real data. This will be the topic of Chapter 3.

Since Information Theory focuses on estimating the amount of information in a given set of possible outcomes, methods stemming from this framework provide a different approach with respect to the nonlinear analytical tools we discussed in this chapter, which evidently stem from dynamical systems theory. In Chapter 4, based on a fundamental theorem establishing a link between dynamical invariants and entropy, we are going to employ some novel methods to study this relation. These methods will be applied to a newly developed model of chaos.

Nonetheless, some similarities, or even equivalence, between entropy and dynamical invariants of chaos theory are known and will be addressed in Chapter 4.

Finally, in Chapter 5 we are going to provide a comparison between a metric from Information Theory, namely mutual information, and a recently developed measure that combines linear correlation metrics with nonlinear analytical elements provided by surrogate data generation for significance estimation. These methods will be used to address correlation between brain signal recordings as well as wind recordings.

Chapter 2

Entropies and dynamical systems

After a brief introduction on some basic concepts regarding dynamical systems theory and time series analysis, in this Chapter, the topic of entropy and the connection between Information Theory and dynamical systems is discussed. The seminal work of Claude Shannon, in 1948[83], addresses for the first time the problem of communication in mathematical terms, trying to answer to the questions: how much information is contained in a message; how can the message be efficiently encoded; what is the maximal communication rate over a channel. After this, many scientist have further expanded the fields of Information Theory. Just to mention a few, Von Neumann, Turing and Kolmogorov, famous for their work in computer science and algorithmic theory.

In this Chapter, I will first start by discussing Shannon's entropy, as it was described in his paper *A mathematical theory of communication*. Then I will proceed with some fundamental definitions in Information Theory, that will be necessary for the next chapters. Later, I will give a brief overlook of how, from Shannon entropy, other measures of entropy have been developed up to Kolmogorov-Sinai entropy, possibly the most important measure of complexity for dynamical systems. I will also present the main problem of Kolmogorov-Sinai entropy that makes it impractical to compute directly, especially on experimental data.

Finally, I will give some definitions of entropies commonly used in real signal analysis, namely Sample Entropy, Approximate Entropy and Permutation Entropy, as well as Mutual Information when information among multiple systems is studied.

2.1. Information

Before going into formal definitions and equations, let's spend a few words on the core concepts discussed in this chapter, in a more informal and intuitive way. The topics we are going to discuss are the basis of Information Theory and, later on, some *tools* that originate from Information Theory that can be used in a wide range of other applications, among which the analysis of nonlinear systems. We should therefore face upfront some fundamental questions: the first questions will be what information is; the second, closely related to the first one, is what entropy is; the third and final question will be why those concepts are so important for our purpose, namely the study of nonlinear systems.

Information is something that we deal with everyday, we can see many manifestation of it, especially if we think about technological applications. Yet if we had to give a clear definition of it, the task would not be so trivial at all. A popular ways to describe information is as removed uncertainty, meaning that gaining information about something, for example the outcome of some event, removes some or all of the uncertainty that we may have about that outcome. Let's use this last description and assume that information does in fact do what we just said: removing uncertainty about the outcome of an event. We'll use this to try and outline some of the properties of information that are required for such description to be true.

Let's suppose the event we are talking about is the toss of a coin. First of all we notice that we need a set of possible outcomes. In our case, the possible outcomes are just two, either head or tail. We also require to represent the information, possibly being able to measure it and quantify it in some units. The universal unit of measure for information is the binary digit, or bit, widely known for its use in every electronical or informatic device. If we use the binary system representation of a bit, this can assume either the value 0 or 1.

Let's go back to our coin. It is of course straightforward to see that either we call it head or tail, or 0 or 1, both cases have the same uncertainty. Therefore reducing the uncertainty over one coin toss, corresponds to gaining an amount

of information of one bit. We will go back on the mathematical calculations of this later in the chapter, but let's focus now on another important aspect of this last step we made: for all practical purposes, we just made an encoding of the outcome of our coin toss into a binary code of 0s and 1s. This may seem trivial, but it gives us the opportunity to notice something else of what information is, or actually what it isn't. Information is not knowledge. If instead of a coin toss, we used the outcome of a football match, or even some statistical distributions of elements in a certain set, we could just as easily encode those into a binary code and then compute its information content in bits. Information concerns data, the representation of some events, or statistical distributions, but it doesn't necessarily corresponds to knowledge of the physical process of coin tossing, or football games and so on.

We got some understanding of what is information, we can therefore move on to the second question and try to understand what entropy is. For this, we can use what we just discussed about information and specifically the fact that information can be measured. As we are going to see more in detail later, entropy is the expectation value of the amount of information of some given process. As before, the nature of said process can be as general as possible, either the statistical distribution of some property over a set of elements, or the probability distribution of outcomes of some event, or the set of words in a transmitted message. Furthermore, entropy can be seen as a measure of variability of the elements in a given set. But we will come back on this later, in a more proper way.

Lastly, let's try to see why information is so useful and important for the study of physical systems. An intuitive way to answer this, is provided by the german physicist Rolf Landauer, who worked at IBM, who wrote that "Information is physical". We can try to look at this sentence in two ways and I believe they provide the answer to our question. The first way is to notice that information requires physical systems, namely bits. The unit blocks of information are in the end physical systems: either the state of some electronic support, or the emission or not of a photon, or even the firing or rest state of a neuron, bits of information ultimately require a physical support. The second way, that answers our question, is

looking at Landauer's quote in the opposite direction: everything physical contains information. Be it again the emission of a photon, or the switching on and off of an electronical component, neurons firing or the sequence of a DNA string, all physical processes do contain information that can be encoded in bits. Those physical processes, those bits, can be studied with the use of Information Theory.

2.2. Shannon entropy

With the introduction to Information Theory given in the previous section, we can now approach Shannon's definition of entropy in a more rigorous way. Claude Shannon first published this results in 1948, in his work *A mathematical theory of communication*. The title already clears out what kind of applications he had in mind when he gave the definition of entropy and laid the ground for what became Information Theory.

In this section, we are going to describe Shannon's Entropy in the same context of communication, as long with some theorems and other quantities that were present in Shannon's work. Nonetheless, as mentioned in the previous section, these concepts are quite general and not at all tied to the field of communication. But we will discuss the generality of his results and the wide range of applications that Shannon's work found in the following years in the next sections, where we will finally introduce the link between Information Theory and Dynamical Systems Theory.

Let's then suppose that we have a message and we want to transmit it. As we mentioned in the previous section, it is convenient to encode this message into bits that can be implemented on physical supports made up of two possible states for each bit, namely 0 and 1. One of the fundamental question that Shannon tried to answer to is, given a message, what is the most efficient way to encode it, i.e. how can we encode it using the least amount of bits necessary. First of all, we should decompose the message in its most basic elements, namely letters. Let's make clear that an optimal encoding will depend on the specific message we want to trasmit: we will have to examine the frequency, or the a priori probability (if

that is known), in which letters appear in the message and try to optimize the encoding on these premises. Of course nothing prevents us from wanting to find an optimal encoding for the entire English language, which is in fact one of the example Shannon examines in his paper, so that we already have a way of coding whatever message in English we may want to. For example, a possible way would be the standard ASCII coding: each letter, or any character, can be mapped into a sequence of 7 bits.

But let's focus just on our simple message and try to describe how to get an optimal encoding for this message only. As we mentioned, we need to find the frequencies p_i with which each letter appears. Intuitively, we can see that an efficient encoding would use short sequences of bits for those letters that more often occur and will need to be transmitted many times over, and instead use longer sequences for letters that more rarely come up. We will henceforth refer to sequences of bits encoding a letter as *words*, which should not be confused with English words making up our message.

Then the average amount of bits per letter used in our code can be easily computed as an expectation value, summing the length of each word weighted by the frequency of the corresponding encoded letter, namely $\sum_i p_i l_i$, where with l_i we express the length of the word encoding the i-th letter. The most striking result of Shannon was that an optimal encoding for a given message can be found. Furthermore, the average amount of bits/letter for such encoding is exactly the Shannon's entropy of the message, defined as:

$$H_2 = \sum_i p_i \log_2 \frac{1}{p_i} = -\sum_i p_i \log_2 p_i \,,$$

where the 2-subscript means that we are encoding in sequences of binary digits, or bits. We will henceforth omit the subscripts to ease the notation.

It is straightforward to see that the lengths of the words used to encode each letter depend on the frequencies with which they occur, namely as $-\log p_i$. This quantity is also known as the **Shannon information**.

To understand more about this, it is best to move from the case of transmitting message to the more general and helpful scenario of a random variable X and the

set of its possible n outcomes $\{x_1, x_2, \ldots, x_n\}$. Just as with the letters of a message, we can define as $p(x_i)$ the probability that the random variable $X = x_i$ and consequently compute the Shannon entropy of X as $H(X) = -\sum_{i=1}^{N} p(x_i) \log p(x_i)$. Than the Shannon information $-\log p(x_i)$ of a certain outcome gives us a sense of how much uncertainty we remove by discovering the outcome: the lower the probability $p(x_i)$, the more surprised we are to find that such a rare event has occurred.

In this sense, as we also mentioned in the previous section, the entropy is a measure of the average information of a given process.

It is useful to mention the case of a uniformly distributed random variable: given n possible outcomes it follows that $p(x_i) = \frac{1}{n}$. The Shannon information of a certain outcome is simply $\log(n)$ and, as expected, it's the same for all outcomes, meaning that witnessing any outcome carries the same amount of information. The Shannon entropy in this case is easily computable as $H(X) = \frac{1}{n} \sum_i^n \log(n) = \log(n)$, which corresponds with the Shannon information.

2.3. Joint entropy, conditional entropy and other definitions

Now that we've established Shannon's entropy for a given random variable, let's look at some additional relevant entropies that can be simply calculated using probability theory. Some of these entropies will be important in the following sections regarding application on time series, while others are included to provide a comprehensive picture of the principles of information theory.

As shown in the previous section, Shannon Entropy can be easily applied to the case of a random variable X, where its probability distribution $p(x)$ is known. I am going to expand this concept to the case where multiple random variables are available, each with its own probability distribution.

Let the Joint Entropy be defined as:

$$H(X_1, X_2, \ldots X_n) = -\sum_{x_1, x_2 \ldots x_n} p(x_1, x_2, \ldots x_n) \log p(x_1, x_2, \ldots x_n)\,,$$

where $p(x_1, x_2, \dots x_n)$ is the joint probability distribution of the variables $\{X_1, \dots, X_n\}$, and the sum is carried on all possible joint outcomes of the n random variables. For clarity, let's point out that the subscripts on x_i here do not enumerate the possible outcomes of a single random variable X, but generically represent the outcomes of the i-th random variable X_i.

Using as a test the case of $X_1, X_2 \dots, X_n$ being independent and identically distributed random variables, it's easy to show that the joint entropy of n random variables will just be equal to $nH(x)$, where $H(x)$ is the Shannon entropy of one single random variable.

The dependence of one random variable from the outcome of another is described by the conditional probability distribution. More precisely, given two random variables X_1 and X_2, the probability of observing the value x_1 for the first variable, once the outcome x_2 of the second variable has been measured, can be written as $p(x_1|x_2)$. We can write:

$$p(x_1, x_2) = p(x_1|x_2)\, p(x_2), \tag{2.1}$$

which links the joint probability to the conditional probability: the probability of the joint events x_1, x_2 to occur is the product of the probability of x_1 to occur knowing that x_2 occurred, times the probability of x_2 occurring in first place. Considering again the case where the random variables are independent, it is clear that $p(x_1|x_2) = p(x_1)$, meaning that the probability of getting the outcome x_1 has no dependence on the outcome of the random variable X_2.

Let's now define the Conditional Entropy as:

$$H(X_1|X_2) = -\sum_{x_2} p(x_2) H(X_1|X_2 = x_2)\,.$$

The last term can be explicitly rewritten as:

$$H(X_1|X_2 = x_2) = \sum_{x_1} p(x_1|x_2) \log p(x_1|x_2)\,,$$

which allows to write Conditional Entropy in terms of the joint probability distribution, as:

$$H(X_1|X_2) = -\sum_{x_1, x_2} p(x_1, x_2) \log \frac{p(x_1, x_2)}{p(x_2)} = H(x_1, x_2) - H(x_2) \tag{2.2}$$

where we made use two times of Eq.(2.1). In other words, the conditional entropy $H(X_1|X_2)$ tells us how much uncertainty we have left on X_1 after the uncertainty of X_2 has been removed.

It is worth noticing that if X_1 and X_2 are independent random variables, then $H(X_1, X_2) = H(X_1) + H(X_2)$ and $H(X_1|X_2) = H(X_1)$, meaning that finding the outcome of X_2 does not tell us anything about X_1.

In order to quantify a sort of distance between two probability density functions, we can use the Kullback-Leibler divergence that can be written as:

$$D_{KL}(p_1|p_2) = \sum_{x_1,x_2} p_1(x) \log \frac{p_1(x)}{p_2(x)} .$$

The Kullback-Leibler divergence is always non-negative, but it's not technically a distance metric since it is not symmetryc, i.e. $D_{KL}(p_1|p_2) \neq D_{KL}(p_2|p_1)$. One way to understand it's meaning is to consider it as a measure of how many extra bits we would require if we were to use a coding optimized on the random variable X_2 to code samples from X_1. It may be clear from this description that in general D_{KL} would not be symmetric.

Another useful measure of information between two random variables is Mutual Information (MI). This metric is symmetric and can be written as:

$$I(X_1, X_2) = \sum_{x_1,x_2} p(x_1, x_2) \log \frac{p(x_1, x_2)}{p(x_1)p(x_2)} .$$

From this definition, it is clear to find that MI corresponds to the sum of the Shannon entropies of both individual variables minus their joint entropy, namely $I(X_1, X_2) = H(X_1) + H(X_2) - H(X_1, X_2)$ and the symmetry of MI is evident. The meaning of MI can better be understood by writing it in another form: using the definition of conditional entropy we gave in Eq.(2.2), we can write:

$$I(X_1, X_2) = H(X_1) - H(X_1|X_2) ,$$

which, due to the symmetry of MI, is equal if we swap X_1 and X_2. It is clear from this last form that mutual information counts the amount of information gained on one variable by knowing the other. Mutual information may be employed as

a measure of correlation between two signals, as we will see in further detail in Chapter 5.

Finally, we are going to introduce the concept of entropy rate. This concept can be explained in two ways. The first is to consider entropy rate as an average of joint entropy per symbol: if we consider a stochastic process $\{X_1, X_2, \ldots, X_n\}$, then the entropy rate of this process can be written as:

$$H(X) = \lim_{n\to\infty} \frac{H(X_1, X_2, \ldots, X_n)}{n} . \tag{2.3}$$

An alternative way, is to consider entropy rate as the amount of information we get by sampling the variable X_n once all the previous variables have been sampled, i.e. the asymptotic value of condition entropy, namely:

$$H'(X) = \lim_{n\to\infty} H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1) .$$

Considering again the case when the X_i are independent identical random variables, the two forms of entropy rate correspond and are equal to the Shannon entropy of a generic variable $H(X_i)$.

Beside stochastic processed, entropy rate is widely used for dynamical systems as it measures the amount of information that the system generates at each new step of its evolution. In the next section, we are going to discuss what is possibly the most fundamental entropy measure on dynamical systems, which is actually a dynamical invariant and corresponds to the system entropy rate.

2.4. Kolmogorov-Sinai entropy and the partitioning problem

Kolmogorov-Sinai entropy (KSE) is a fundamental concept in the field of dynamical systems theory, providing a quantitative measure of the complexity and unpredictability of dynamical behavior. It was introduced by the Russian mathematician Andrey Kolmogorov and Israeli mathematician Yakov G. Sinai in the 1950s as a way to quantify the rate of information production or entropy generation in dynamical systems. KSE has since found widespread applications in various

fields, including physics, information theory, computer science, and complex systems analysis. As already mentioned in the abstract of this thesis, several concepts developed throughout the last centuries have converged into the formulation of KSE. Starting from the first evidences of irregular behavior, in the sense of non-periodic trajectories, from the study of celestial mechanics, to stochastic motion of particles in statistical mechanics. The different approaches used to tackle these issues — from the techniques of dynamical systems theory regarding dynamical invariants of systems defying the notion of periodic and regular motion, to statistical tools and ergodic theory to approach the study of gas particles — find a place as necessary steps towards the concept of complexity that KSE attempts to quantify.

The turning point comes with information theory, which addresses a different type of dynamical systems, but most of all uses a different approach in trying to quantify the amount of information that the evolution of such system generates at each new step. Finally, the revisitation of information theory by Von Neumann, Turing and the same Kolmogorov in terms of algorithmic theory, and therefore in the sense of the complexity of outlining a set of instructions needed for a computer to reproduce a dynamical system, put the final mark on the topic.

In this sense, KSE acts as a "conceptual bridge" [84] between different fields that find in complexity a common denominator.

KSE is a measure of the average rate of entropy production in a dynamical system. Its definition is based on the concept of partitions that are subsets of the state space of a dynamical system. Let's consider a dynamical system, evolving in the phase space S and a time evolution map $\mathcal{T} : S \to S$ that maps a state into the next state.

A partition $\mathcal{P}$ of S is a collection of disjoint subsets, or cells, A_i, that cover the entire phase space, i.e. $S = \bigcup A_i$. Each cell A_i in the partition represents a distinct set of initial conditions that lead to similar dynamical behaviors or states under the action of the map $\mathcal{T}$, in other words the points from each cell A_i can generate similar trajectories in the phase space, described by the dynamical invariants of the system.

Based on the partition $\mathcal{P}$, we can define the Kolmogorov-Sinai entropy H_{KS} as the average rate of entropy production per unit time in the system, given by the formula:

$$H_{KS} = lim_{N\to\infty} \frac{1}{N} \log B(N)\,,$$

where N is the number of iterations of the map $\mathcal{T}$ and $B(N)$ is the total number of cells in the partition visited by the system after N iterations.

The crucial issue in evaluating KSE is the partitioning of the phase space. We can describe the partition cells A_i as an hypercube in the phase space, namely with volume ϵ^D, where D is the dimension of the phase space. Then KSE can be expressed in an alternative form that makes clearer the correspondence of KSE with an entropy rate and highlights the issue of finding a suitable partition, especially in the case of evaluation on real data. By defining the set of random variables $\{X_1, X_2, \ldots, X_N\}$, where the outcomes of X_1 corresponds to selecting the cell where the trajectory is at the time step $t = \tau$, the outcome of X_2 to the cell at time $t = 2\tau$ and so on, we can write KSE as:

$$H_{KS} = -\lim_{\tau\to 0}\lim_{\epsilon\to 0}\lim_{N\to\infty} \frac{1}{N\tau} \sum_{x_1,x_2,\ldots x_N} p(X_1, X_2, \ldots X_n) \log(p(X_1, X_2, \ldots X_N))\,,$$

where $p(X_1, X_2, \ldots, X_n)$ is the joint probability of finding the trajectory in the cell x_1 at time $t = \tau$, in x_2 at $t = 2\tau$ and so on.

From this form, and by the definition of entropy rate given in Eq. (2.3), KSE corresponds to the entropy rate of a joint distribution that essentially maps a trajectory of the system. In addition, this entropy rate is computed for finer and finer refinements of the partitioning, namely the hypercube size ϵ, and with finer temporal resolution as $\tau \to \infty$.

In Chapter 1, upon introducing deterministic chaos, we discussed that these systems are crucially dependent on the initial conditions and that arbitrarily small differences in the initial state will eventually evolve in larger and larger distances among the resulting trajectories. This behavior is described by means of the Lyapunov exponents, a dynamical invariant that quantifies the rate at which the two trajectories grow apart. Nonetheless, these systems are deterministic: if we were to know with absolute precision the initial state of a trajectory, for a theoretical

system where no noise perturbations are present, we would be able to recreate that exact trajectory by iterating the time evolution operator $\mathcal{T}$. Given these premises, an interesting point of view to understand the concept of KSE, in terms of the amount of information generated at each time step, is to consider how much new information we acquire on the initial state at every new evolution step.

Let's suppose we have a partitioning of the state space and a trajectory made of the union of the cells containing the states visited by the trajectory itself. We want to know with as much precision as possible the initial state of this trajectory, which we will refer to as the original trajectory. At first, the resolution we can get on determining the initial state is set by the size of the cell that contains it. We now consider a secondary trajectory, with the initial state arbitrarily close to the initial state of the original trajectory so that it falls in the same cell. We evolve the two trajectories in time and, for a certain number of steps, their new states end up in the same cells. Since the distance between the two trajectories grows over time, at some point, the two trajectories will end up in different cells and grow even further apart at each new step. This information translates into knowing always more precisely the initial state of the original trajectory: it is basically a prediction problem where we can tune the secondary trajectory initial state in order to make it *coincide* to the original trajectory for an increasing number of steps. The link between the KSE in terms of rate of information generated by the system and the trajectories divergence rate given by Lyapunov exponents is evident.

Despite how fascinating this concept appears, it is evident how difficult it is to apply KSE estimates on real data. Furthermore, only *generating partitions* allow for a correct estimation of KSE, while generic partitions provide only a lower bound [85].

This is a critical issue in nonlinear systems analysis and, as we are going to discuss in the next sections of this chapter, it is one that has prompted interest in other entropy measures, for example permutation entropy, due to the fact that these may provide an alternative indirect way of estimating Kolmogorov-Sinai entropy.

2.5. Entropies on time series

In this section, we will go over some of the entropy metrics that are widely used to analyze time series. The ones discussed here are only examples of how information-theoretical methods may be used on time series, but the list is far from exhaustive: there is a huge number of metrics derived from information theory that take the name of entropy, a comprehensive list is provided by Ribeiro et al. [86] in a review that, quite understandably, is titled "The entropy universe".

The selection chosen here consists of two common entropy metrics that originate from KSE, namely approximate entropy and its improved version, sample entropy. Finally, we will introduce permutation entropy, which combines an intuitive and easy to compute definition, with an interesting connection to KSE.

2.5.1. Approximate and sample entropy

Due to the impossibility of computing KSE, one of the first alternative measures proposed to give a similar estimation of complexity was approximate entropy. This measure, derived from KSE, was introduced by Pincus in 1991 [87] and it was applied to classify both chaotic and stochastic systems, proving to be effective even with a limited size of data. Its formulation has some similarities with the correlation integral we discussed in Chapter 1: after performing an m-dimensional embedding to reconstruct the phase space from the input scalar sequence, we compute $C_m^r(i)$ as:

$$C_m^r(i) = \frac{1}{N-m+1} \sum_{j=1}^{N-m+1} \Theta(d_{ij} - r)\,,$$

where $\Theta(x)$ is the Heaviside function, r is a threshold value and $d_{i,j}$ is the distance between two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ in the embedded space, defined as $d_{i,j} = max_{k=1,\dots,m}\,|(x(i+k-1) - x(j+k-1)|$. We then take the average of the natural logarithm of $C_m^r(i)$ over all i:

$$\Phi_m^r = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \ln C_m^r(i)\,,$$

and finally compute the approximate entropy as:

$$H_{Ap}(m,r) = \Phi_m^r - \Phi_{m+1}^r\,.$$

Nevertheless, approximate entropy has some drawbacks as lack of consistency, high dependability on data length, overestimation of probabilities due to self-matches and resulting in lower estimation than expected for small-sized datasets.

These shortcomings were highlighted by Richman and Moorman [88] when they introduced an improved version of approximate entropy, namely sample entropy. Sample entropy essentially removes self-matches from approximate entropy, therefore also reducing the computation time as compared to approximate entropy.

Sample entropy requires the same parameters r and of course m. We first define the quantities $B_m^r(i)$ as:

$$B_m^r(i) = \frac{1}{N-m-1} \sum_{j=1, j\neq i}^{N-m} \Theta(d_{ij} - r)\,,$$

and its average $B_m^r = \frac{\sum_i^{N-m} B_m^r(i)}{N-m}$, then sample entropy becomes:

$$H_S = -\ln \frac{B_{m+1}^r}{B_m^r}\,.$$

Several other entropies have been developed from approximate and sample entropy, one example are fuzzy entropies where the Heaviside function is replaced with smoother functions, for example sigmoids.

Some applications of these metrics involve studies on stock markets [89], heart beat rate [87, 88] and EEG signals [90], as well as earth science [91], just to mention a few.

2.5.2. Permutation entropy

Permutation entropy (PE) is an entropy measure used to quantify the complexity or irregularity of time series data. PE is a non-linear and non-parametric measure that characterizes the disorder or randomness of a time series by analyzing the ordinal patterns of the data.

Permutation entropy was first introduced by Bandt and Pompe in 2002 [92] as a measure to assess the complexity of time series data without making any assumptions about the underlying dynamics. Since then, it has gained a lot of

interest as a measure of complexity for its intuitiveness, relatively easy computation and invariance with respect to scaling. PE is based on the concept of ordinal patterns, which represent the relative ranking of values within a window of data points in a time series. The idea is to transform the original time series into a sequence of symbols based on the ranking of values of neighbouring points, and then compute the entropy (formally, a Shannon entropy) of the symbols distribution.

Let's denote a sequence as $X = \{x_1, x_2, ..., x_N\}$, where x_i represents the i-th point in the sequence, and N is the length of the sequence. The ordinal pattern of a window of length m is defined as the ranking of the values in that window. In other words, the ordinal pattern is a sequence of m integers that represents the order of the values within the window. It follows that the number of possible patterns corresponds to the number of possible permutations, i.e. $m!$, hence the name of permutation entropy.

An example can help clarify this concept: assuming $m = 3$, the possible ordinal patterns are $\Pi = \{\pi_1 = [1,2,3], \pi_2 = [1,3,2], \pi_3 = [2,1,3], \pi_4 = [2,3,1], \pi_5 = [3,1,2], \pi_6 = [3,2,1]\}$. A possible mapping of these patterns, or symbols, onto the ranking of the sequence values is for example:

$$[1,2,3] \rightarrow x_1 < x_2 < x_3$$
$$[1,3,2] \rightarrow x_1 < x_3 < x_2$$
$$[2,1,3] \rightarrow x_2 < x_1 < x_3$$
$$\dots .$$

Figure 2.1 gives a graphical representation of how a pattern in a sequence ranking is mapped into a symbol.

Based on the occurrence of symbols, we can construct a histogram of the occurrences of each symbol. Finally, permutation entropy is defined as:

$$H_\Pi(m) = -\sum_{i=1}^{m!} N_i \log N_i \,,$$

where N_i is the number of occurrences of the i-th symbol. The equivalence of the formulation of permutation entropy and Shannon entropy is evident.

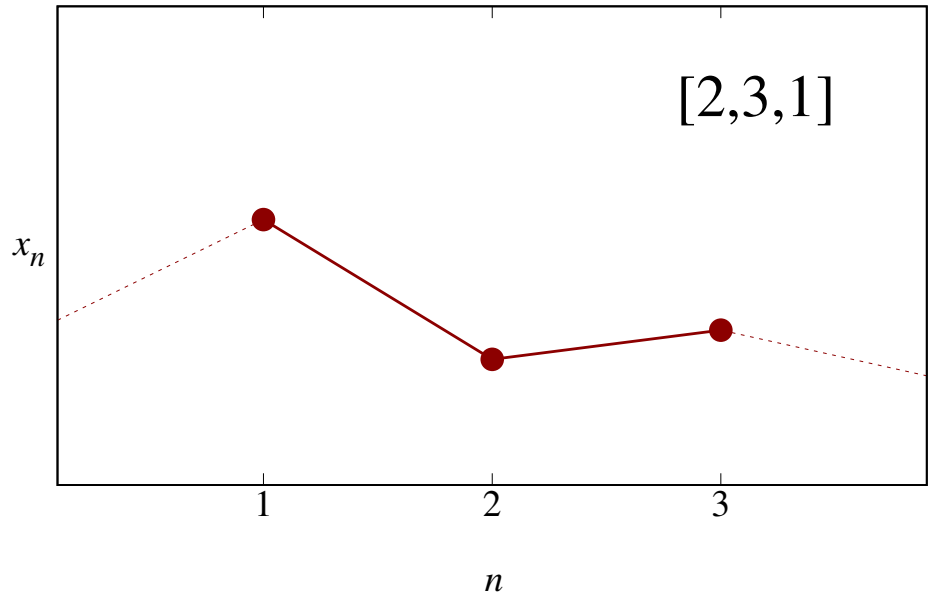

Figure 2.1: Visual representation of a pattern in a sequence and it's corresponding symbol.

As mentioned earlier in this chapter, one of the main reasons for the wide interest in permutation entropy is that the rate of permutation entropy growth is asymptotically equivalent to the Kolmogorov-Sinai entropy. The permutation entropy rate h_Π can be computed as a derivative, namely:

$$h_\Pi = \lim_{m\to\infty} \frac{H_\Pi(m+1) - H_\Pi(m)}{T}, \tag{2.4}$$

where T is the sequence sampling period.

A drawback of this approach is that the number of symbols of permutation entropy increases as $m!$, i.e. the number of possible permutations of m elements. This causes an high computational cost, counterbalancing the easiness of the permutation entropy algorithm, and furthermore a proper estimation of the occurrence histogram on a support as large as $m!$ requires an increasingly higher number of available data points to be correctly estimated.

In Chapter 4, we discuss this matter in more detail, as well as a recently developed method proposing a correction to permutation entropy that makes its entropy rate converge faster to the value of KSE.

Chapter 3

Uncertainty of Shannon Entropy

In the previous chapter, an introduction of Information Theory was given and some of the most commonly used tools for time series analysis have been described. Despite such a widespread application of entropy evaluation in many different fields, a proper analysis of the uncertainty of such estimations is often lacking. Assessing the reliability of an entropy estimator, in terms of its bias and variance, is not a trivial task and in general requires knowledge of the *a priori* probability distribution.

In this chapter, the estimation of the variance of the so-called plug-in estimator of Shannon Entropy will be discussed, in the case of an underlying multinomial distribution. The demonstration of the results obtained will require quite a long and tedious mathematical description. Therefore, in order to make the reading more fluid, the details will be discussed in the Appendix A.

We will instead focus on the calculation of the maximum entropy variance, which in turn can be easily obtained from the number of points available and the support size of the possible outcomes. These results, as it will be argued later, allow to make up a simple rule of thumb to fix the required number of samples in order to obtain entropy estimates within a certain level of confidence. The results discussed in this chapter were published in 2021 [93].

3.1. Evaluation of Shannon entropy on real data: the plug-in estimator and the variance estimator

We will consider to have a set of N discrete samples of an experimental recording of some system's time evolution. We will define as M the size of the support, namely the number of different states visited by the system. The *a priori* probabil-

ity of the number of visits for each state is usually unknown, therefore only sample probabilities can be evaluated as rate of visits, from the N available samples.

Let s_i be the probability with which the i-th state is visited. Then the system Shannon Entropy is:

$$H = -\sum_{i=1}^{M} s_i \ln s_i \,.$$

We consider that the support size M of the probability distribution $\{s_i\}$ is such that $s_i > 0$ for each value of $i = 1, \ldots, M$.

We take into account the N available successive system evolution steps and the M-dimensional set $\hat{j}_i$ of the number of visits to each of the M states, such that $\sum_{i=1}^{M} \hat{j}_i = N$ and $\{j_i\}$ is distributed as a multinomial distribution. Then the observed rate $\hat{p}_i$ is computed as $\hat{p}_i = \hat{j}_i/N$. The plug-in estimator of Shannon Entropy $\widehat{H}$ is defined as:

$$\widehat{H} = -\sum_{i=1}^{M} \hat{p}_i \ln \hat{p}_i \,.$$

The plug-in estimator reliability can be evaluated in terms of its bias and variance. The bias term is known as the Miller-Madow correction [94, 95, 96] and is known to be equal to $-(M-1)/(2N)$. Keeping in mind that the entropy is the expectation value of the Shannon Information $I_i \equiv -\ln s_i$, as discussed in Sec.2.2, it follows that the variance parameter can be written as:

$$\begin{aligned} \Lambda_0 &= \mathbf{E}(I_i^2) - \mathbf{E}(I_i)^2 \\ &= \sum_{i=1}^{M} s_i \ln^2 s_i - \left(\sum_{i=1}^{M} s_i \ln s_i \right)^2 , \end{aligned} \tag{3.1}$$

which can also be rewritten as $\Lambda_0 = \sum_{i=1}^{M} s_i \, (H + \ln s_i)^2$, making clear that $\Lambda_0 \geq 0$, except in the case of s_i being uniformly distributed.

As found by Basharin [94] and Harris [96], the entropy variance for the multinomial case is then:

$$\sigma^2_{\widehat{H}} = \frac{\Lambda_0}{N} + O\left(\frac{1}{N^{\frac{3}{2}}}\right) ,$$

Knowledge of the parameter Λ_0 is therefore needed in order to estimate the Shannon Entropy plug-in estimator variance. Using Eq. (3.1) we can thus define

an estimator $\widehat{\Lambda_0}$ of the variance parameter as:

$$\widehat{\Lambda_0} = \sum_{i=1}^{M} \left(\hat{p}_i \ln^2 \hat{p}_i \right) - \widehat{H^2} \, .$$

For this estimator as well, the corresponding reliability requires the evaluation of its bias and variance. This can be done by considering the asymptotic behaviour of $\widehat{\Lambda_0}$ for $N \to \infty$. It can be proven that asymptotically the variance parameter estimator behaves as a normal distribution:

$$\widehat{\Lambda_0} \sim \mathcal{N}\left(\Lambda_0 + \frac{\gamma}{N}, \frac{\Gamma}{N} \right) \quad \text{as } N \to \infty \, .$$

The demonstration of this property is discussed in Appendix A. Let's notice that the estimator converges to the variance parameter with bias and variance terms of order N^{-1}. The two parameters γ and Γ are defined in Appendix A, but it should be noticed that they do depend on the knowledge of the *a priori* distribution $\{s_i\}$ and higher orders moments of the distribution of single-state Shannon Information I_i.

Evaluating the variance of the Shannon Entropy estimator seems then an inescapable loop where the *a priori* distribution is in any case required to draw any conclusion. Nonetheless, it is possible to circumvent this issue when evaluating the maximum value of Λ_0, as discussed in the next section.

3.2. The maximum Shannon Entropy uncertainty

Let's first define as $\mathbb{D}$ the probability M-simplex such that:

$$\mathbb{D} = \{ \mathbf{r} \in \mathbb{R}^M | \sum_{i=1}^{M} r_i = 1, \; r_i \geq 0 \; \forall i \} \, .$$

Considering $x \ln^n x = 0$ for $x = 0$ and $\forall n \in \mathbb{N}$, both H and Λ_0 are continuous on the simplex $\mathbb{D}$. Therefore, due to Weierstrass extreme value theorem [97] and the theorem of necessary conditions for extreme values [98], the maximums of Λ_0 (or H) can either be on stationary points within $\mathbb{D}$ or on the boundary $\partial\mathbb{D}$.

We will first start considering the case when the stationary point are within $\mathbb{D}$. Since the probabilities $\{s_i\}$ must satisfy the constraint $\sum_{i=1}^{M} s_i = 1$, we only have $M - 1$ independent variables, since we can always write one as a function

of the others. Let's the choose to write $s_M = 1 - \sum_{i=1}^{M-1} s_i$. We now have to write out the partial derivative of Λ_0 with respect to the remaining $M-1$ independent variable and solve for it to be zero. Before going into the derivation, it may be helpful to write out explicitly the derivatives of s_M and H with respect to s_i (with $1 \leq i \leq M-1$) in order to make the reading more clear.

$$\begin{aligned} \frac{\partial s_M}{\partial s_i} &= \frac{\partial}{\partial s_i}\left(1 - \sum_{i=1}^{M-1} s_i\right) = -1\,, \\ \frac{\partial H}{\partial s_i} &= \frac{\partial}{\partial s_i}\left(-\sum_{i=1}^{M-1} s_i \ln s_i - s_M \ln s_M\right) = \ln s_M - \ln s_i\,. \end{aligned} \tag{3.2}$$

Using the definition in Eq. (3.1) and the partial derivatives found above, we can write the partial derivative of Λ_0 as:

$$\frac{\partial \Lambda_0}{\partial s_i} = (\ln s_i - \ln s_M)\,(\ln s_i + \ln s_M + 2 + 2H)\ . \tag{3.3}$$

We now impose the partial derivative to be zero $\forall i$, thus it easily follows that we have two possible conditions for s_i:

$$\begin{aligned} s_i &= p_0 \equiv s_M \\ s_i &= q_0 \equiv \frac{1}{s_M}\exp^{(-2-2H)}\ . \end{aligned} \tag{3.4}$$

These two must be mutually exclusive. If they were simultaneously valid, then all s_i would be equal to each other, furthermore the first condition would require them to be all equal to s_M: we would then have a uniform distribution giving $H = \ln M$. But the second condition being simultaneously valid would give $s_M = \frac{1}{s_M}\exp^{(-2-2H)}$, resulting in $H = 1 + \ln M$.

We must then have a mixed situation where some of the s_i satisfy the first condition while the remaining ones satisfy the second condition. We can define as k the number of $s_i = p_0$, while the remaining $(M-k)$ satisfy $s_i = q_0$. Since, trivially, s_M always satisfies the first condition, then $1 \leq k \leq M-1$, where we exclude the case of a uniform distribution with $k = M$. For the uniform distribution, it is in fact known that $\Lambda_0 = 0$ [96]. It then follows that also $1 \leq (M-k) \leq M-1$. We

thus can write a set of three equations:

$$\begin{aligned} H &= -kp_0 \ln p_0 - (M-k) q_0 \ln q_0 \\ q_0 &= \frac{1-kp_0}{M-k} \\ 2 + 2H &= -\ln p_0 - \ln q_0 \,, \end{aligned} \tag{3.5}$$

where the first is the equation of the Shannon entropy from its definition, the second equation comes from the requirement that the probabilities sum up to one and the third equation corresponds to Eq. (3.4). By substituting the expression of H from the first equation and replacing it in the last equation, with some algebra, we can rewrite this set of equations as:

$$(1-2kp_0) \ln \frac{1-kp_0}{kp_0} = 2 + (1-2kp_0) \ln \frac{M-k}{k} \,. \tag{3.6}$$

We then have a single equation in the variables H, p_0, q_0, while we consider M and k to be fixed parameters.

We can now define the variable $v \equiv 2kp_0 - 1$. Since we are considering the solutions within the simplex $\mathbb{D}$, it must hold that $0 < kp_0 < 1$, therefore we get $|v| < 1$. By defining the function:

$$f(v) \equiv v \ln \frac{1+v}{1-v} \,, \tag{3.7}$$

we can rewrite Eq. (3.6) into:

$$f(v) = 2 - v \ln \frac{M-k}{k} \,, \tag{3.8}$$

where we used the equalities:

$$\begin{aligned} p_0 &= (1+v)/2k \\ q_0 &= (1-v)/2(M-k) \,. \end{aligned} \tag{3.9}$$

By rewriting $f(v) = v \ln(1+v) - v \ln(1-v)$, it is straightforward to notice that $f(v) = f(-v)$. This function is also always positive with the exception of the origin, since $f(0) = 0$. The first derivative is:

$$f'(v) = \frac{f(v)}{v} + \frac{2v}{1-v^2} \,.$$

Since both terms of the derivative have the same sign as v, it follows that $f(v)$ is monotonically increasing for $v > 0$ and decreasing for $v < 0$, resulting in the function diverging to $+\infty$ for $v = \pm 1$ and $v = 0$ being the only minimum. Figure 3.1 shows the behaviour of $f(v)$ in red. The second member of Eq. (3.8), given M and k as fixed parameters, results in a set of straight lines going through the point $(0, 2)$. These lines are also shown in Fig. 3.1, in the case of $M = 7$ with $1 \leq k \leq 6$.

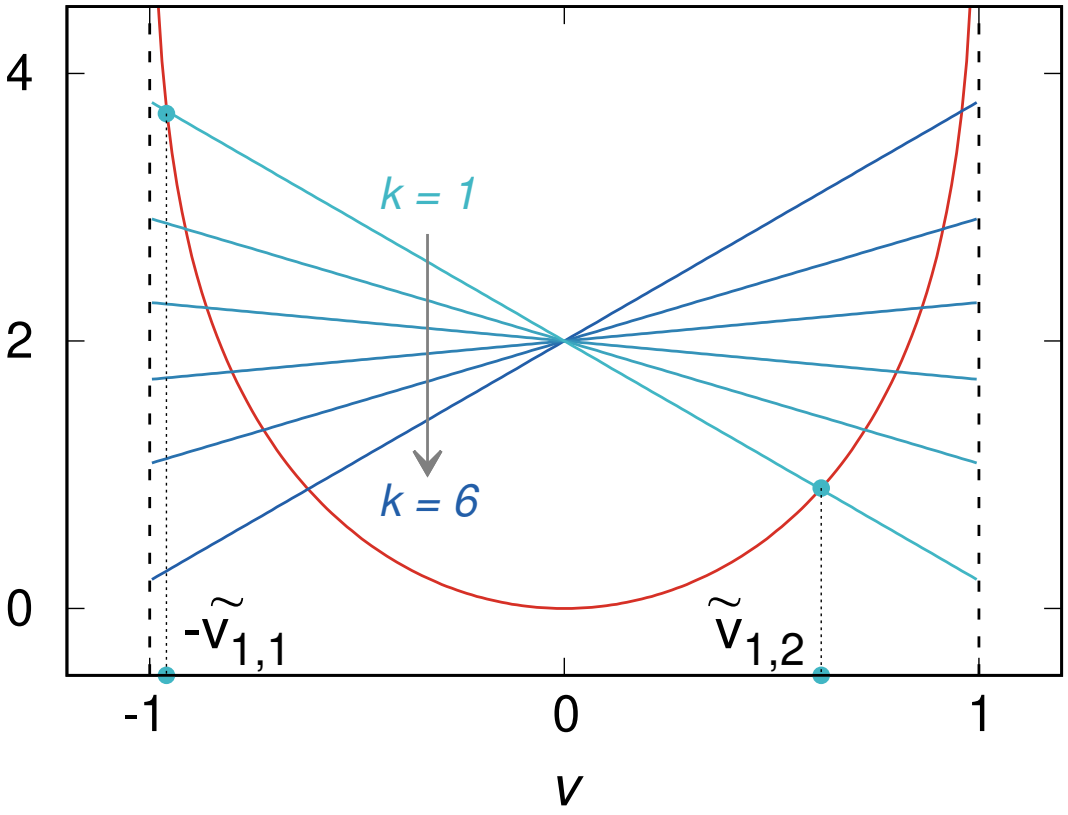

Figure 3.1: Graphical representation of Eq. (3.8). The red, solid line corresponds to the graph of the function $f(v)$ defined in Eq. (3.7). The straight-lines correspond to the right-hand term of Eq. (3.8) in the case $M = 7$ and for different values of the parameter k. The blue dots mark the intersection of $f(v)$ with the straight-line in the case $k = 1$, as well as the two abscissae $-\tilde{v}_{k,1}$, $\tilde{v}_{k,2}$.
Reproduced from Phys. Rev. E 104, 024220 (2021), with the permission of APS Publishing.

For a fixed M, all the intersections of $f(v)$ with the set of straight lines given by the parameter k are therefore roots of Eq. (3.3) and thus stationary points for

Λ_0: we will define these points as $\tilde{v}_k$. By using the expression for Λ_0 in Eq. (3.1) we get:

$$\begin{aligned}\Lambda_0 &= k\left(p_0 \ln^2 p_0\right) + (M-k)\left(q_0 \ln^2 q_0\right) - H^2 \\ &= (kp_0 - k^2 p_0^2)\ln^2 p_0 + \left[(M-k)kq_0 - (M-k)^2 q_0^2\right]\ln^2 q_0 \\ &\quad - 2k(M-k)\ln q_0 \ln p_0\,,\end{aligned} \tag{3.10}$$

where we expanded H as in the first of Eqs. (3.5). Using the fact that $\tilde{v}$ is a root of $f(v)$ and manipulating Eqs. (3.9), we get the following equations:

$$\begin{aligned}\tilde{v}\ln\frac{M-k}{k} &= 2 \\ \tilde{v}\ln\frac{1+\tilde{v}}{1-\tilde{v}} &= 0 \\ \frac{1+\tilde{v}}{2} &= kp_0 \\ \frac{1-\tilde{v}}{2} &= (M-k)q_0\,.\end{aligned}$$

Equation (3.10) then becomes:

$$\begin{aligned}\Lambda_0 &= \frac{1}{4}\left(1-\tilde{v}^2\right)\left(\ln\frac{p_0}{q_0}\right)^2 \\ &= \frac{1}{4}\left(1-\tilde{v}^2\right)\left(\ln\frac{1+\tilde{v}}{1-\tilde{v}} + \ln\frac{M-k}{k}\right)^2 \\ &= \frac{1}{4}\left(1-\tilde{v}^2\right)\left(\frac{2}{\tilde{v}}\right)^2 \\ &= \frac{1}{\tilde{v}_k^2} - 1\,.\end{aligned} \tag{3.11}$$

Therefore, for each value of k we have two possible roots: $\tilde{v}_{k,1}$ and $\tilde{v}_{k,2}$ and in order to find the value of k giving the maximum value Λ_0, we need to select the smallest $\tilde{v}_k$.

We can notice that, for each value of k, the two roots $\tilde{v}_{k,1}$ and $\tilde{v}_{k,2}$ are one negative and one positive. We are going to define $-\tilde{v}_{k,1}$ as the negative solution and $\tilde{v}_{k,2}$ as the positive one. We also observe from Eq. (3.8) that $\tilde{v}_{k,1} = \tilde{v}_{M-k,2}$, so we can limit the analysis to the cases where $k < M/2$ and the term $\ln\frac{M-k}{k}$ is non-negative.

Then by putting $-\tilde{v}_{k,1}$ and $-\tilde{v}_{k,2}$ in Eq. (3.8) and subtracting them, we get:

$$f(-\tilde{v}_{k,1}) - f(-\tilde{v}_{k,2}) = (\tilde{v}_{k,1} + \tilde{v}_{k,2})\ln\frac{M-k}{k} \geq 0\,,$$

which, since $f(v) = f(-v)$ and $f(v)$ has positive derivative for $v > 0$, implies that $\tilde{v}_{k,1} > \tilde{v}_{k,2}$. Therefore, $\tilde{v}_{k,2}$ is the smallest root we were searching, providing the maximum value of Λ_0.

In order to find the probability distribution giving the maximum entropy uncertainty, we must study the behaviour of $\tilde{v}_{k,2}$ at different k. We can do this by considering the variation of $\tilde{v}_{k,2}$ when k increases by one. Since $\frac{M-k}{k} > \frac{M-k-1}{k+1}$, it follows that:

$$\frac{2 - f(\tilde{v}_{k,2})}{\tilde{v}_{k,2}} > \frac{2 - f(\tilde{v}_{k+1,2})}{\tilde{v}_{k+1,2}},$$

and therefore:

$$\frac{2}{\tilde{v}_{k,2}} - \frac{2}{\tilde{v}_{k+1,2}} > \frac{f(\tilde{v}_{k,2})}{\tilde{v}_{k,2}} - \frac{f(\tilde{v}_{k+1,2})}{\tilde{v}_{k+1,2}}.$$

If we assume $\tilde{v}_{k,2} \geqslant \tilde{v}_{k+1,2}$, than the left-hand term of the previous inequality would be non-positive, where instead the right-hand term would be non-negative (due to the fact that the function $f(v)/v = \ln \frac{1+v}{1-v}$ is monotonically increasing if $0 < v < 1$), which leads up to a contradiction. Therefore, it follows that $\tilde{v}_{k,2} < \tilde{v}_{k+1,2}$. Consequently, $\tilde{v}_{1,2}$ provides the minimum absolute value of a root of Eq. (3.8) and thus, via Eq. (3.11), the maximum value of Λ_0 for a given M, when considering the interior of the simplex $\mathbb{D}$.

By the same argument, it follows that increasing M leads to a decrease of $\tilde{v}_{1,2}$ and thus to an increase of $\Lambda_{0,\max}$. This corollary is important in order to discuss the behavior of the variance parameter Λ_0 on the boundary of the probability simplex.

It should finally be noted that for $k = 1$ we have a single *outlier* among the set of probabilities $\{s_i\}$, with value s_M, namely the one that was chosen as dependent on all the others. Since the dependent probability can be selected arbitrarily among the M possible choices, the number of equivalent maxima of Λ_0 is M, namely any state i could be the outlier and result in the same value of Λ_0.

We proceed now to the situation when the maxima of Λ_0 occurr on the boundary of the M-simplex. By noticing that the boundary $\partial\mathbb{D}$ of the M-simplex $\mathbb{D}$ is the union of M boundary facets, with each facet defined by a single s_j, with $1 \leqslant j \leqslant M$, set to zero, it follows that such boundary corresponds to a lower-dimensional sim-

plex, namely a $(M-1)$-simplex.

By the initial assumption that the support size of the probability distribution is M, we could promptly disregard this situation. Nevertheless, assuming that the s_i values can be arbitrarily small, makes it valuable to discuss also this situation.

Because each facet corresponds to a simplex, the procedure applied for the interior of the M-simplex can be iteratively applied. First, in the interior of the j-th facet, the maximum of Λ_0 will be $1/\tilde{v'}_{1,2}^2 - 1$, where $\tilde{v'}_{1,2}$ is the positive root of Eq. (3.8) when $k = 1$ and M is replaced by $M-1$. Second, the boundary of the j-th facet has to be considered as a $(M-2)$-dimensional set of $(M-2)$-simplexes, and so on. By re-iterating the procedure, we end up with a set of two-dimensional simplexes where only a pair $(s_i,\ s_j)$ of probabilities is nonzero: in this case the boundaries are the point-like vertices of a segment, where $\Lambda_0 = 0$.

By means of the argument discussed above, consisting in the proof that increasing M results in an increase of Λ_0, it follows that on a two-dimensional simplex the maximum Λ_0 is obtained on a stationary point in the interior of the two-dimensional simplex $\mathbb{D}$, namely $1/\tilde{v}_{1,2}^2 - 1$, and not on the facet corresponding to either s_i or s_j being zero.

3.3. Maximum entropy uncertainty and its distributions

The results of the analysis in the previous section show that the maximum entropy uncertainty in the multinomial case occurs when the distribution of the probabilities $\{s_i \,|\, i \in [1,\ M]\}$ has a single outlier equal to p_0, while the other probabilities are uniformly distributed and equal to q_0.

We can then write Eq. (3.8) for $k = 1$, as:

$$f(v) = 2 - v \ln(M-1)\,, \tag{3.12}$$

and write as $\tilde{v}$ its root.

The outlier value p_0 is then:

$$p_0 = \frac{1+\tilde{v}}{2}\,. \tag{3.13}$$

By replacing $f(v)$ with the first term of its Taylor expansion, namely $2v^2$, we can approximately solve Equation (3.12) that becomes a second-degree equation with positive solution:

$$\tilde{v} \approx \sqrt{\frac{\ln^2(M-1)}{16} + 1} - \frac{\ln(M-1)}{4}\,. \tag{3.14}$$

.

For large M values, the previous equation can be approximated as $\tilde{v} \sim 2/\ln(M)$. Furthermore, the approximations improve with increasingly higher M, corresponding to smaller $\tilde{v}$. In Figure 3.2, we show the plots of p_0, $\Lambda_{0,\max}$ as a function of the dimension M. It can be noticed that already for $M \gtrsim 10$, the approximations provide a good matching with the analytical solution.

For asymptotical values $M \to \infty$, we obtain:

$$\Lambda_{0,\max} \cong \frac{\ln^2(M)}{4}\,. \tag{3.15}$$

The outlier value p_0 then tends to 1/2. Since the outlier value could be any of the M possible choices for the dependent s_i, we then have M equivalent distributions providing the maximum Λ_0.

In Figure 3.3, we show the case of $M = 3$.

It is important to notice that, in the case of a distribution that maximizes Λ_0, the parameter Γ determining the variance of the estimator $\widehat{\Lambda_0}$, shown in Equation Eq. (3.2), vanishes [93]. A similar situation occurs for the Shannon entropy plug-in estimator in the case of a uniform distribution: it is well-know that in this case entropy reaches its maximum value of $\ln M$, while Λ_0, the parameter quantifying the entropy estimator variance, vanishes.

3.4. Final remarks

The results discussed in this chapter show that a reliable estimator of Shannon entropy variance allows to determine a maximum uncertainty in entropy estimation, in the case of multinomial distributions. This result though has three requirements.

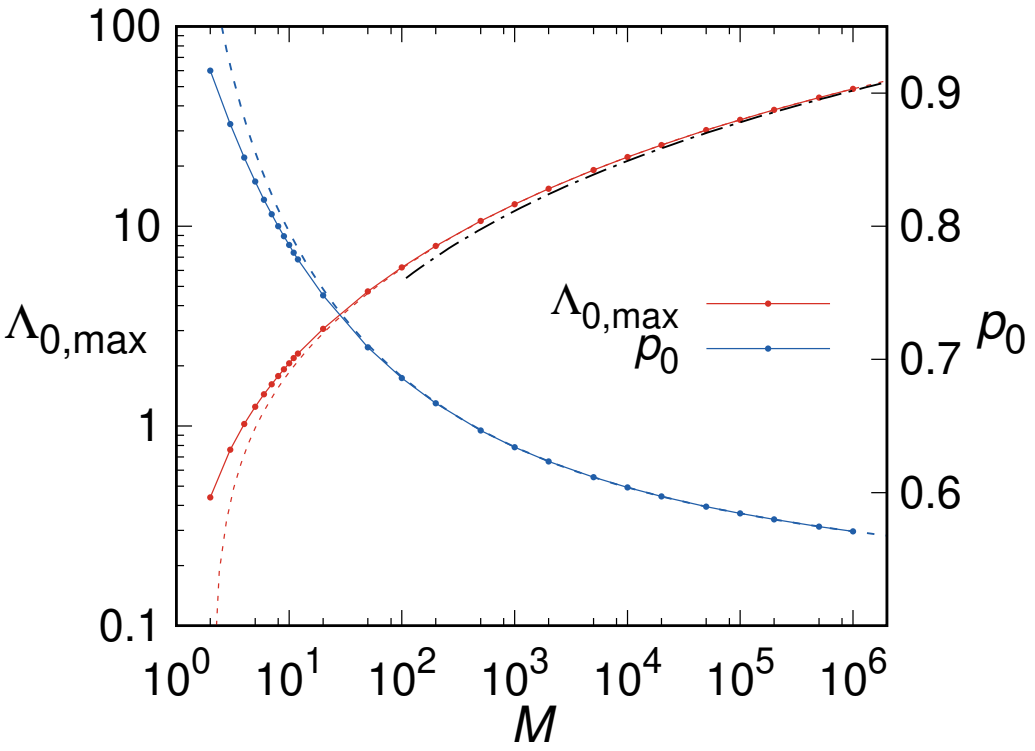

Figure 3.2: Maximum variance parameter $\Lambda_{0,\max}$ as a function of the support size M: (red, solid line) exact computation; (red, dashed line) evaluation via Eq. (3.11) by using the approximated value for $\tilde{v}$ given by Eq. (3.14); (black, dash-dotted line) approximation described by Eq. (3.15) and valid for $M \gtrsim 100$. *Outlier* probability p_0 as a function of the support size M: (blue, solid line) exact computation; (blue, dashed line) evaluation via Eq. (3.13) by using the approximated value for $\tilde{v}$ given by Eq. (3.14).

Reproduced from Phys. Rev. E 104, 024220 (2021), with the permission of APS Publishing.

The first is to know the support size M. In the case of experimental data, the value of M is not usually known but ultimately depends in the method used to partition the values of the data being analyzed. The issue of choosing a method for partitioning is a non-trivial one and can drastically influence the outcomes of entropy estimates. For this reason, some improved estimators of entropy have been developed, avoiding the partitioning problem of the plug-in estimator. More details on this are discussed in Chapter 5, regarding the estimation of mutual

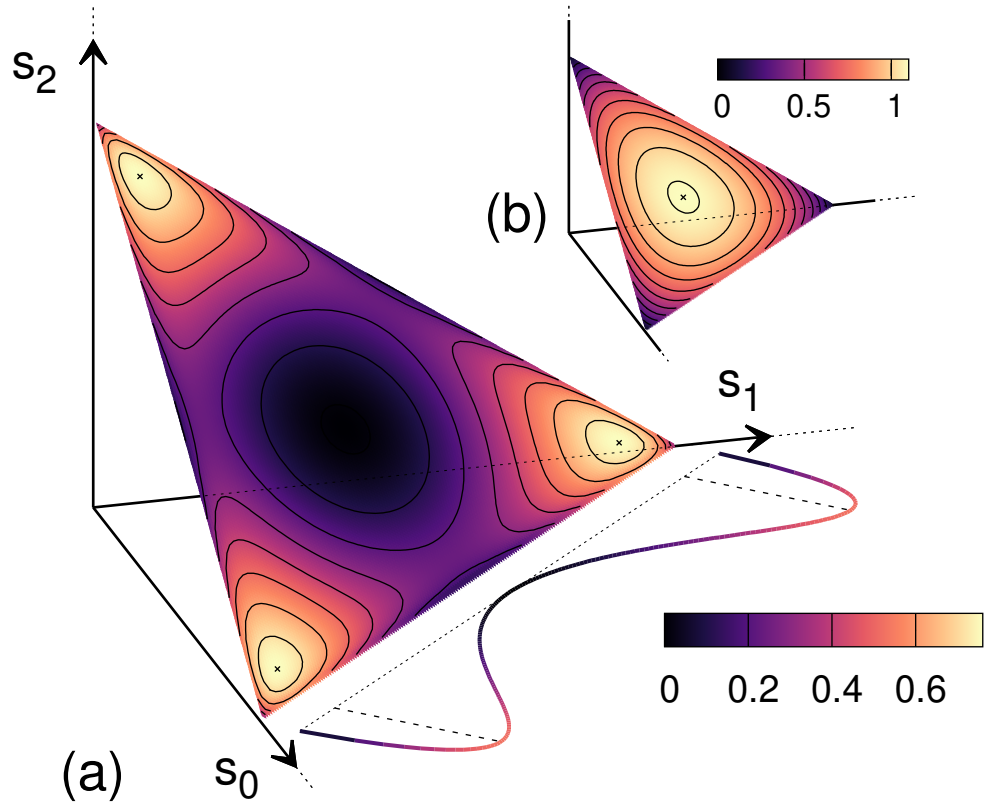

Figure 3.3: (a) Color map of the variance parameter Λ_0 on the probability simplex in the case $M = 3$. Three points, located at one coordinate being equal to $p_0 \cong 0.88$ and the other two equal to $q_0 \cong 0.06$, provide the maximum value of Λ_0, namely $\Lambda_{0,\max} \cong 0.762$. The curve placed on the plane π_{s_0,s_1} represents the plot of Λ_0 on the simplex facet corresponding to $s_2 = 0$. (b) Color map of Shannon entropy H on the probability simplex in the case $M = 3$. The maximum value occurs in the center, where $s_0 = s_1 = s_2 = 1/3$ and Λ_0 vanishes. No additional stationary points are present (which is true at any dimension).
Reproduced from Phys. Rev. E 104, 024220 (2021), with the permission of APS Publishing.

information.

The second requirement, as mentioned several times throughout this chapter, is that data are sampled from a memoryless multinomial distribution. This means that the outcome of each state is independent from the previous state. It appears clear that, in the context of dynamical systems, moreso for deterministic chaos,

this is not a valid assumption. As a first approach, when no insight is available about the underlying nature of the process generating the data under study, assuming the data to follow a multinomial distribution is a reasonable, despite raw, initial assumption. Nonetheless, a proper analysis, in terms of Shannon entropy, of data from a deterministic source eludes the case discussed in this chapter. It is worth noticing that Markov chains provide a valuable approximation of deterministic system's evolution. A recent work [99], based on a revisited form of the central theorem for Markov chains [100], discusses a method to compute the uncertainty of entropy estimations in the case of an underlying regular Markov chain process, using a similar approach to the one used in this chapter, and described in more detail in Appendix A. Nonetheless, this latter method requires the knowledge of the Markov chain process, or at least an estimate of the matrix of transition rates from one state to the other.

The third and weaker requirement is that in order to visit all the M states making up the distribution support, the number N of trials has to exceed M, so that $N > M > \sqrt{M}$. Provided this condition, the upper bound to entropy variance described in this chapter turns out to be smaller, and thus more accurate, than the one derived by Antos and Kontoyiannis [101], namely $(\ln N)/\sqrt{N}$.

Chapter 4

Lyapunov Exponents and Entropy

In Chapter 1 we already gave the definition of Lyapunov exponents: in simple words, they estimate the rate of growth of the distance between two initially close trajectories. Lyapunov exponents are the mathematical version of what is commonly known as the "butterfly effect", a concept first used by Lorenz in 1962, which highlights the critical dependency of chaotic dynamics on initial conditions.

By considering the state space of a dynamic system, or the reconstructed phase space of an experimental signal, as divided into arbitrary small cells by some partitioning, we can imagine two trajectories that at some initial time are close enough to be contained in the same cell. Lyapunov exponents quantify the rate at which, after a certain amount of time, the two trajectories will start to diverge and therefore be contained in separate and further apart cells.

From this point of view, it is not so difficult to grasp that Lyapunov exponents may have some similarities with what we defined as the Kolmogorov-Sinai entropy (or metric entropy, or topological entropy) in Chapter 2. It was in fact proven by Pesin [102], that the two concepts are equivalent. More precisely, he found that:

$$H_{KS} = \sum_{i \,|\, \Lambda_i > 0} \Lambda_i \,,$$

so that the KSE is equal to the sum of positive Lyapunov exponents of the system.

Nonetheless, as we discussed in Chapter 2, direct computations of KSE are difficult to perform on experimental data. We could therefore use the Lyapunov spectrum to get an estimate of KSE, as we will do in this chapter, by using the so-called standard method [4, 5, 103] to compute the Lyapunov exponents. Unfortunately, the complete Lyapunov spectrum is equally difficult to obtain when no knowledge of the differential equations describing the system's evolution is

available. In addition, methods to estimate Lyapunov spectrum out of scalar sequences, using for example embedding for phase space reconstruction, are reliable for the computation of the maximum exponent but can produce unexpected results when computing the whole spectrum.

An interesting approach would be to use other entropy metrics, for example permutation entropy rate, that asymptotically is said to coincide to KSE and therefore to the sum of positive Lyapunov exponents [104].

Not surprisingly, there are some drawbacks also with this approach, since it requires to compute PE asymptotically, i.e. for an infinite window m. This poses some computational problems as well as finite-size issues when applied on a limited real set of data.

In this Chapter, we will try some of this approaches using as a test-bench the Minimal Universal Model (μ-Model) for chaos [105].

4.1. Minimal Universal Model for chaos

The μ-Model was developed by Meucci et al. [105] by revisiting the model of a laser with feedback in order to provide the minimal amount of nonlinearities needed to exhibit a chaotic behaviour. The model clearly originates from laser physics, specifically from CO_2 lasers [30, 31], but tuning the values of its parameters, it can be generalized to produce different nonlinear regimes that can model different phenomena, like neuron dynamics [70, 106, 107], electronics, opto-electronics and possibly population dynamics [108, 109, 110, 111] and epidemiological models [112]. In this sense, the model is referred to as a *Minimal Universal Model* for chaos.

The μ-Model is a three-dimensional model, where each variable corresponds to a physical quantity in its originating context of laser physics: the variable x corresponds to laser intensity, y corresponds to population inversion and z represents the feedback strength.

The system of differential equations, in its dimensionless form, of the μ-Model

is:

$$\begin{aligned}\frac{\mathrm{d}x}{\mathrm{d}\tau} &= -\varepsilon_1 x\left(1 + k_1 z^2 - p_0 y\right) ,\\ \frac{\mathrm{d}y}{\mathrm{d}\tau} &= -y - xy + 1 ,\\ \frac{\mathrm{d}z}{\mathrm{d}\tau} &= -\varepsilon_2 \left(z - B_0 + B_1 x\right) .\end{aligned} \tag{4.1}$$

Some of these parameters are considered fixed and take values given by the time scales of the processes involved in CO_2 laser, namely the decay rates of laser pulses and population inversion, and the feedback usual time scale [105]. We therefore set $\varepsilon_1 = 200$, $\varepsilon_2 = 6$, $k_1 = 12$, $B_1 = 0.555$. For this choice of parameters, the dynamics of the system is given by the interaction of three variables that vary with very different time scales: namely the laser intensity x is the fastest variable, the population inversion y is the slowest and the feedback z operates at intermediate frequencies between the former two.

The two controlling parameters p_0 and B_0, correspond to the pump and bias and can be tuned to change the system dynamics [105, 41]. Here we will set $p_0 = 1.208$ and use two values of B_0, which, as discussed in Ref. [105], are both chaotic but show different kinds of evolution. The values are namely $B_0 = 0.124$ (for what will be referred as system 1) and $B_0 = 0.1246$ (system 2).

We computed time series of the μ-Model by integration of the differential equations (4.1) via Runge–Kutta Prince–Dormand (8, 9) method [2], with sampling period of $T = 0.1$ ($\approx 1/10$ of the main oscillations period) and a total size $N \approx 10^6$, so that finite-size corrections in entropy evaluations are negligible. Figures 4.1 and 4.2 show the state space of the two attractors for the parameters chosen above.

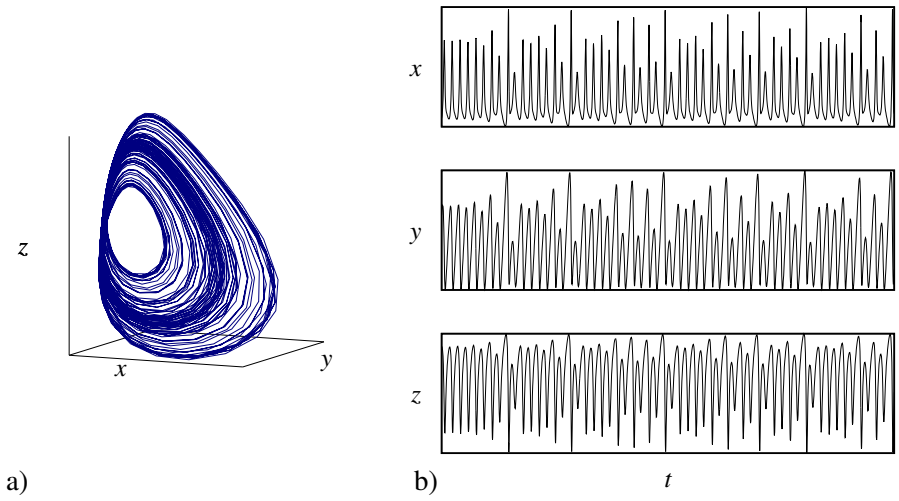

Figure 4.1: State space evolution (a) of the μ-Model attractor, system 1, and time series of its components (b). The parameter B_0 is here set to 0.124.

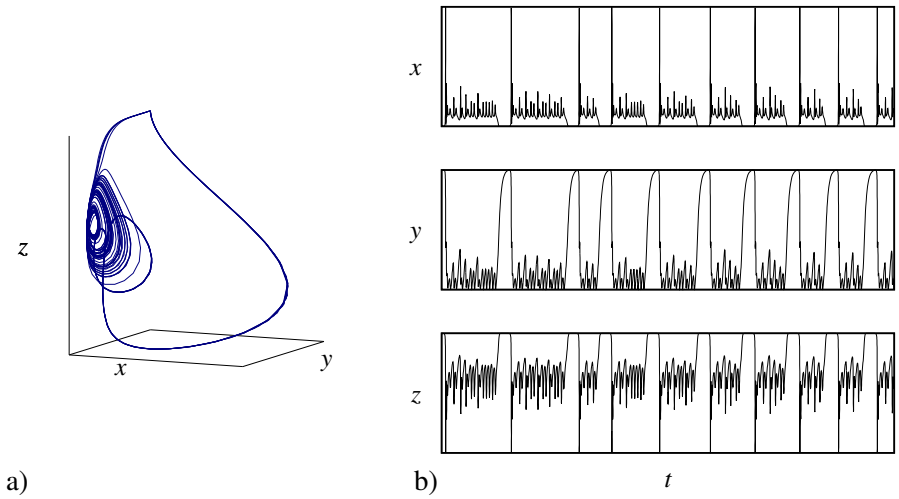

Figure 4.2: State space evolution (a) of the μ-Model attractor, system 2, and time series of its components (b). The parameter B_0 is here set to 0.1246.

4.2. Permutation entropy rate

We already discussed in Chapter 2 that the permutation entropy rate converges to Kolmogorov-Sinai entropy in the asymptotic limit of the window m going to infinity. Permutation entropy rate h_Π can be computed according to Eq. (2.4). In

this section, we are also proposing one other possible method of computing the rate of permutation entropy. Namely, we use a corrected version of permutation entropy, recently proposed by Politi [113], whose rate has been shown to converge faster to KSE.

In this recent work, the entropy K_P is computed as:

$$K_P(m) = H_\Pi(m) + D\,\langle \ln \sigma(m) \rangle\,, \tag{4.2}$$

where $H_\Pi(m)$ is the standard permutation entropy of window m, D is the system dimension (here computed by means of Kaplan-Yorke conjecture [114]). The meaning of $\sigma(m)$ requires a longer explanation.

Given the permutation entropy window m, for each symbol π_i and each occurrence of the given symbol (if any), it will correspond a given segment $\{x_{N_i}(j)\}$, where j is the position of each point of the segment, going from 1 to the segment length m, and N_i is the number of occurrences of the symbol π_i. Then $\sigma_i(j)$ is the standard deviation of the points $\{x_{N_i}(j)\}$ over the N_i segments of the given symbol. The reason for the additional term $D\,\langle \ln \sigma(m) \rangle$ is explained in detail in Ref. [113], but to give a simple explanation we can consider the symbols of window m as a partitioning of the system into the permutation "cells", where each cell is a symbol π_i. Then $\langle \ln \sigma(m) \rangle$ corresponds to the average cell size, being a measure of the dispersion among the different trajectories falling into a given symbol/cell π_i. Furthermore, we take the average only of the term $j = m$ since in this case we have the maximum value of $\sigma(j)$, and therefore of its logarithm [113]. In Ref. [113], it is argued that the trajectories dispersion scales as $\sigma(m) \approx m^{-\gamma}$, and the term $D\,\langle \ln \sigma(m) \rangle$ in Eq.(4.2) consists in a finite-size correction that improves the asymptotic convergence of the rate k_P to KSE, where the rate is computed as $k_P = \lim_{m\to\infty} \left(K_P(m+1) - K_P(m)\right)/T$, similarly to Eq. (2.4).

In Figure 4.4, we show the behaviour of $\langle \ln \sigma(m) \rangle$ for the two μ-Model systems under study, which partially confirms what expected from what we discussed above.

In particular, the power-law decrease $\sigma(m) \approx m^{-\gamma}$ seems to only appear for values of $m \geqslant 15$.

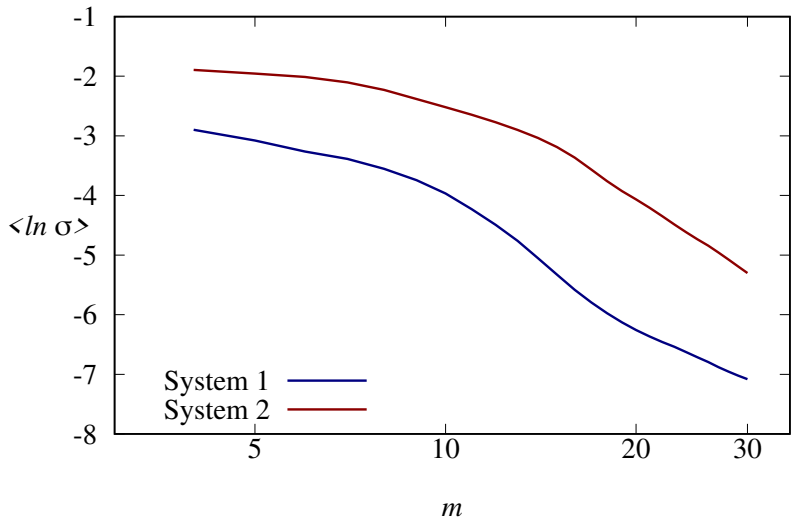

Figure 4.3: $\langle \ln \sigma(m) \rangle$ as a function of the permutation entropy window m, for system 1 (red) and system 2 (blue). In both cases, the logarithmic scale shows a linear behaviour as described in Ref. [113], but only for higher values of m, namely for $m \geqslant 15$.

4.3. Permutation entropy rate of the μ-Model

We show here the results of computing the standard permutation entropy rate h_Π and the corrected version k_P, discussed in the previous section.

In system 1, we can see that the standard permutation entropy rate converges quite slowly and does not reach the asymptotic value of h_{KS} within the considered range of window m. The corrected permutation entropy, instead, despite having more abrupt fluctuations, does converge more rapidly to h_{KS}.

This approach shows to work more effectively in the case of the first system, while in the second system, the standard permutation entropy rate outperforms the corrected permutation entropy, reaching the value of h_{KS} with an overestimation of $\approx 23\%$ within the considered window m.

The poorer performance of the corrected permutation entropy rate k_P on the second system could be due to the nature of the system evolution, which as shown in Fig. 4.2, exhibits abrupt spikes in all three of its coordinates, while system 1 has a more *smooth* behaviour. For the sake of comparison with Ref. [113], it should be mentioned that the chaotic attractor used in that work is a Rössler

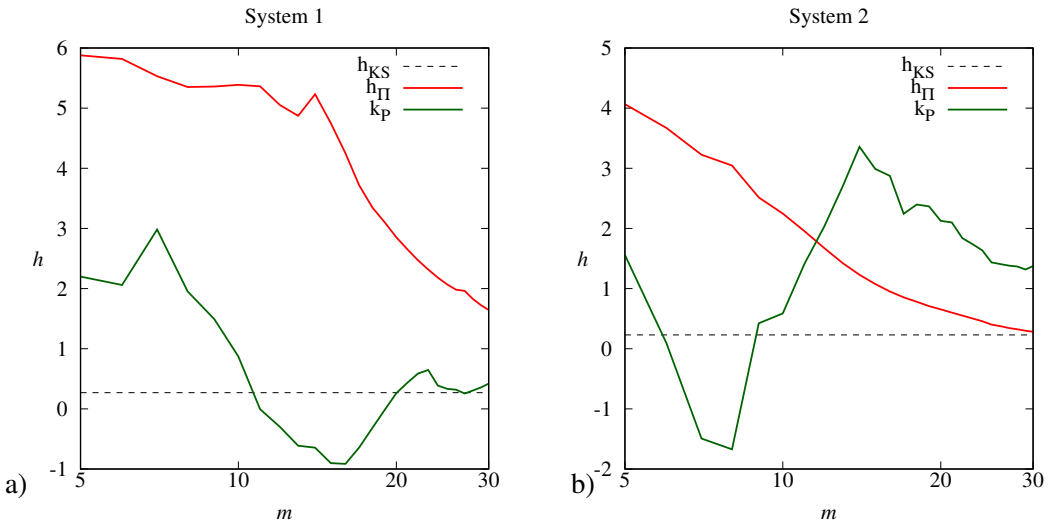

Figure 4.4: Entropy rates as a function of the window m, for system 1 (a) and system 2 (b). The dotted black line corresponds to the Kolmogorov-Sinai entropy; red lines correspond to the standard permutation entropy rate and green lines to the corrected permutation entropy rate [113].

attractor, namely the x-coordinate of the attractor. As we showed in Fig. 1.1 of Chapter 1, the Rössler system x-component does have a smooth evolution, while the z-component shows a spiking behaviour. This may suggest that the method used in this chapter may be less efficient on spiking time series, where dedicated methods for testing nonlinearity are recommended [115, 116].

4.4. Final remarks

In this chapter, we have discussed the relation between the Lyapunov spectrum of a chaotic system and its information evaluated by means of dynamical entropies. While in theory the identity of Pesin clearly reveals how Lyapunov exponents and Kolmogorov-Sinai entropy are two strictly connected concepts, and actually equivalent, we argued in previous chapters that KSE is hardly directly computed. In most cases, as well as in this chapter, the most efficient approach to estimate KSE is through the Lyapunov spectrum.

Nevertheless, knowledge of the Lyapunov spectrum is not generally available for real time series, therefore the need to find a reliable alternative method, pos-

sibly based on entropy, to evaluate KSE is a fundamental issue. Here we showed the application of two approaches, based on permutation entropy, on a newly developed, and therefore unconventional, chaotic model.

This model has the advantage of being tunable by means of few parameters to exhibit different behaviours, which make it a promising candidate to model several natural phenomena, from neural spiking to social dynamics [105]. Furthermore, an electronic implementation was developed recently [41], by making use of operational amplifiers, signal multipliers and digitally controlled potentiometer, to solve the system's differential equations. The whole setup was made on a custom printed circuit board and using integrated circuit components, allowing reduced electronic noise and repeatable conditions. The board also had dedicated ports to allow remote control of the potentiometer, necessary to tune the parameters B_0 and p_0, via an FPGA, which in turn was controlled by a computer through LAN network. An oscilloscope, connected to the computer via USB, was employed to measure the output signals. This setup allows for a fast and completely remote-controlled acquisition procedure, making it easy to automate and perform data collection and analysis in one single pipeline. Furthermore, the reduced dimension and relatively low cost, allow to easily scale up this implementation in order to make multiple units of the MUM model and have them coupled. This could then prompt the experimental study of interactions in chaotic network structures, including for example synchronization phenomena.

Nonetheless, the results shown in this chapter are far from satisfactory. The problem of finding a reliable proxy for KSE, especially when studying continuous-time chaotic systems, is still an open and fundamental issue in nonlinear time series analysis.

Chapter 5

Mutual information and CTO for correlation analysis

The discovery and characterisation of network structures that arise inside a particular system is a critical subject in nonlinear research. Taking on this problem requires assessing the relationships between system nodes, a process that may be accomplished using a variety of metrics, each with its own set of benefits and drawbacks. A common strategy is to discover significant values of correlation estimators between experimental time series generated by elements, or nodes, in order to assess the existence of an underlying network structure [117]. This approach is common in a variety of fields, like climate science [118, 119, 120] and brain connectivity [121, 122]. Functional networks are studied in this latter context using functional magnetic resonance imaging (fMRI) [123], electroencephalography (EEG) [124], or magnetoencephalography (MEG)[125, 126], which produce multivariate recordings that allow reconstruction of brain activity in different brain regions. Assessing the degree of connectivity is also important for investigating the nonlinear dynamics of synchronization processes that contribute to the formation of a network [127].

Different connection measures might be used depending on the type of interaction that is theorized to generate the link. The sample Pearson correlation coefficient between two sequences, for example, assesses the mutual degree of linear, non-directed correlation [117]. In more general circumstances, more sophisticated connection measures, such as Spearman's and Kendall's correlation coefficients can detect nonlinear relationships. Furthermore, measurements like transfer entropy [128] and Granger causality [129, 130, 131] are intended to discern the direction of connectivity as well as its strength. Finally, metrics that op-

erate in the frequency domain, such as coherence [132, 133] and methods derived from this last one [134, 135, 132], phase synchronization [136], and wavelet-based envelope correlation [136], are frequently used [121], particularly in the field of neuroimaging [137].

Mutual information, initially suggested by Shannon in 1948, is regarded as a valuable technique for detecting linear, and possibly nonlinear [138, 139], correlations among the metrics that measure amplitude-related correlations, particularly due to the relative simplicity of implementation.

Recently, a method for identifying links was presented [140] that evaluates the relevance of zero-delay cross-correlation between pairs of time series via surrogate data creation [141]. The approach determines the shortest observation period, known as the correlation time scale of observability (CTO), required to detect a significant cross-correlation.

In this chapter, we examine the relationship between this last metric and mutual information by evaluating both metrics on large sets of data extracted from three experimental contexts: human brain magnetoencephalography, human brain electroencephalography, and small-scale regional surface wind measurements. A power-law describes this connection effectively. A basic noise and signal model is used in a theoretical model that explains these findings, yielding the equation:

$$\widehat{I}W^{1-\gamma} \approx \eta^* \cdot c \, \ln(N)\,, \tag{5.1}$$

where $\widehat{I}$ and W are the sample Mutual Information and CTO, respectively; η^* is an efficiency parameter required for the CTO algorithm (described in the next section), N is the size of the available sequences and γ and c are costants to be fitted on the experimental data. The demonstration of this last equation, and a more detailed description of its parameters, is discussed in Appendix A.

These results were published in 2021 [142] and demonstrate that the combination of this metric and mutual information may be used as a powerful tool to detect and quantify connectivity relationships in a variety of experimental scenarios.

5.1. Correlation time-scale of observability (CTO)

The CTO method computes a cross-correlation time scale of observability W from a pair of sequences. The approach is thoroughly discussed in the study where it was initially introduced [140], as well as in two papers where it was applied to MEG recordings [143, 144]. A C++ implementation has also been released as an open-source software package [145, 146].

The first step of the CTO method consists in computing the Pearson cross-correlation coefficient on running windows of varying widths w. Given a pair of nodes and their corresponding sequences of size N and sampled with period T, the sample Pearson cross-correlation coefficient $r(k, w)$ is computed on a moving window of width w and centered at time t_k. The window width takes the values mw_0, where w_0 is a "base window" and m is an integer number running between 1 and M. The moving windows's center point moves between $t_1 = \frac{Mw_0}{2}$ and $t_K = NT - \frac{Mw_0}{2}$, incremented at each step by the base window w_0. It should be noted that the number of windows, K, is the same independently of the window width. Therefore, at small values of window width, for example at the minimum value w_0, the initial and last segments of the time series, each long $\frac{(M-1)w_0}{2}$, are discarded. This should be considered when choosing the values of M and w_0, which ultimately set the minimum and maximum window width to be used. A reasonable choice is to ensure that the segments discarded using short windows are at least one order of magnitude less than the whole time series. Furthermore, the base window width w_0 should be set based on the time series sampling period, making sure that an adequate number of sampled points is present in a window of width w_0. To simplify the notation, we will henceforth refer to the time index k instead of the explicit time value t_k, with k taking all the integers values from 1 to K.

Given k and w, the corresponding element of the *cross-correlation diagram*, $r(k, w)$, is then linked to a p-value using a surrogate-based technique [141]. Surrogate generation is an efficient strategy for estimating statistical significance when the underlying null-hypothesis distribution is unknown and must be derived from

data. To calculate the p-value, an iterative amplitude-adjusted Fourier transform (IAAFT) technique is used to construct a collection of m pairs of surrogate sequences from a given original pair. A surrogate sequence generated through IAAFT essentially consists in a sequence with approximately the same power spectrum (or autocorrelation, by means of the Wiener-Kinchin theorem) and amplitude distribution. Nonetheless, a surrogate sequence has a purely stochastic nature: IAAFT ultimately performs a random reshuffle of the input data, with the power spectrum and amplitude constraints mentioned earlier. This method therefore provides a set of synthetic stochastic data that can be used as a comparison to test the null hypothesis of the two original sequences being uncorrelated. From the original pair of sequences, out of each sequence independently the IAAFT algorithm is used to generate a set of surrogate sequences. These are then paired, one surrogate sequence from each of the two original ones, in order to make a surrogate pair.

Out of each pair of surrogate sequences, a surrogate cross-correlation diagram $r_i(k, w)$ is computed, where $i = 1, \ldots, m$ is the index enumerating each surrogate pair. Finally, the original cross-correlation diagram element $r(k, w)$ is ranked in ascending order among the m corresponding surrogate elements $r_i(k, w)$: at this point, a p-value can be obtained simply by dividing the rank number by the total number of pairs m [140]. The resulting collection of $p(k, w)$ p-values is henceforth referred to as the p-value diagram.

The CTO method's second step consists in reducing the two-dimensional information contained inside a p-value diagram to a one-dimensional efficiency function $\eta(w)$ of the window width w. This function is defined as the fraction of windows of width w whose p-value is smaller than a certain significance threshold α, here set to 0.05. Considering that the total number of windows, at a given width, is K, we can write the efficiency function as:

$$\eta(w) = \frac{\#\left(k \in 1, \ldots, K \mid p(k, w) \leqslant \alpha\right)}{K},$$

where with # we denote counting every k that meets the p-value condition.

The last step consists in selecting the minimum window width W, if any, such

that $\eta(w) > \eta^*$, where η^* is a threshold value. If such condition is fulfilled, the value W will correspond to the *time scale of observability* of the correlation between the two input sequences, or in other words to the link between the two selected nodes generating the sequences.

Figure 5.1 shows a graphic summary of the method's results: (a) and (b) show the p-value diagram in the case where a link is not found (a) and in the case where a link is found (b). For the latter case, the dotted green line marks the value of w such that a fraction η^* of windows shows a significant correlation, i.e. p-value less than α. Finally, in (c), the efficiency functions are shown: by increasing the window width w, the event in red eventually reaches the threshold value η^*, meaning that a link between the two considered nodes exists, while in blue the spurious event is shown, where the efficiency function never reaches the threshold and therefore the link is deemed not to exist.

In Perinelli et al. [140], a discussion is present about how this procedure skims out spurious events: if a correlation is found at small windows due to a spurious event, once the window width is increased the spurious correlation will be averaged over larger and larger portions of uncorrelated noise, effectively reducing the overall correlation and the p-value given by surrogate-based significance estimation. Therefore, the p-value diagram, as shown in Figure 5.1(a), will not have a funnel-like structure and reach the efficiency threshold necessary for the link to be assessed. This previous case assumes that the spurious events have low signal-to-noise ratio. As discussed in Appendix A, a single delta-like event with high signal-to-noise ratio would in fact generate a funnel-like structure in the p-value diagram. Nonetheless, a single spurious event would trigger a slowly increasing efficiency function, resulting in a high value of time scale of observability W. Therefore, suitably setting the efficiency threshold could prove useful to skim this kind of events.

To further understand the meaning of W, consider that its value is an estimate of the window width such that the chance of detecting a statistically significant cross-correlation along the sequences in a window of that width equals η^*. In the limit situation $\eta^* = 1$, for example, W refers to the window width for which all

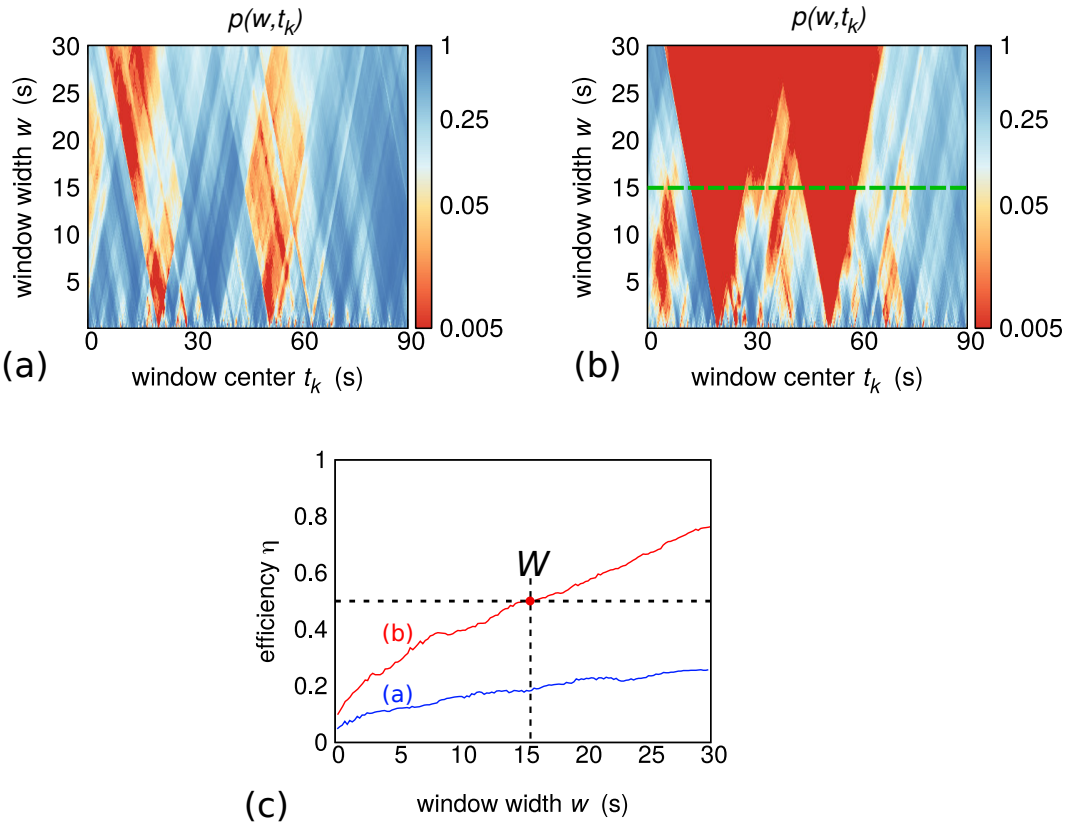

Figure 5.1: Scheme of the CTO algorithm: in (a) the p-value diagram is shown in the case of spurious correlation events; (b) shows the p-value diagram in case of significantly correlated nodes, with the green dotted line showing the window width value w where a number of windows η^* shows significant correlation; in (c) the efficiency function $\eta(w)$ for the detected link (red) and the spurious correlation event (blue) are shown. The black lines highlight the value W for which the threshold efficiency η^* is reached.

windows along the sequences display a significant cross-correlation. Crucially, the higher the threshold is set, the more constringent the requirement for the existence of a link. Therefore, the threshold can be used to tune the sensitivity of the method: when the threshold is raised, only stronger links are detected, while lowering it allows for the detection of weaker ones.

Aside from indicating the presence of a link, the time scale W measures its strength: the more correlated the two sequences, the shorter the window width W required to detect a significant cross-correlation. Therefore the inverse of the

time scale of observability, W^{-1} makes up a measure of connectivity strength [144], which will be used as a comparison with MI.

It is worth noting that, while the Pearson cross-correlation coefficient is a linear measure, the running-window approach combined with the surrogate-based significance evaluation can prove useful in detecting behaviors typical of nonlinear dynamics, such as intermittency or nonstationarity. In the case of intermittent behavior, differently from a forward application of the cross-correlation coefficient on the two whole sequences, the CTO algorithm can be more efficient when intermittency causes shared events to take place in a limited fraction of the time spanned by the time series: as shown in Fig. 5.1(b), a funnel-like structure can be easily triggered by significant events taking place in a short amount of time. Nonetheless, the intermittent behaviour may cause an increased value of W, thus a weaker connectivity, which should be taken into account when setting the efficiency threshold. In general, the flexibility given by the CTO algorithm in terms of temporal resolution (by tuning the base window width w_0 and the maximum width Mw_0) and connectivity strength (by tuning efficiency threshold) can prove useful in extracting valuable information in a wide range of situations, involving those given by the presence of nonlinear effects on the time series.

5.2. Mutual information

The sequences $\{x_t, y_t\}$ can be seen as a realization of a bivariate random variable (X, Y), with joint density $p_{X,Y}(x, y)$ and marginal densities $p_X(x)$, $p_Y(y)$, respectively. We already gave the definition of mutual information in Chapter 2 and we discussed that MI can be rewritten in a more practical form as:

$$I(X, Y) = -H(X, Y) + H(X) + H(Y)\,,$$

where $H(X, Y)$ is the joint Shannon entropy and $H(X)$ and $H(Y)$ the Shannon entropies of the two single random variables X and Y.

Computing MI therefore consists in evaluating the marginal sample Shannon entropies $\widehat{H(X)}$, $\widehat{H(Y)}$ and the joint sample Shannon entropy $\widehat{H(X, Y)}$, where the

hat denotes that these are sample estimations and therefore affected by uncertainties due to noise in the sequences $\{x_t, y_t\}$ and finite-size effects.

The method used here to evaluate the sample Shannon entropies is the Kozachenko-Leonenko k-nearest-neighbor estimator [147], which is consistently used as a benchmark [148, 149], despite being computationally cumbersome.

The Kozachenko-Leonenko estimator for the marginal sample entropies $\widehat{H}(X)$, $\widehat{H}(Y)$ can be written as:

$$\widehat{H}(X) = \psi(N) - \psi(k) + \ln(2) + \frac{1}{N}\sum_{i=1}^{N} \ln(\varepsilon_i) \ ,$$

where ψ is the digamma function, k is the selected nearest neighbor's rank, and ε_i is the k-th distance in ascending order within the set $\{|x_i - x_j|\}$, where i is fixed and j takes on all the values between 1 and N, except $j = i$.

Similarly, the estimator for the sample joint entropy becomes:

$$\widehat{H}(X, Y) = \psi(N) - \psi(k) + \ln(\pi) + \frac{2}{N}\sum_{i=1}^{N} \ln(\varepsilon_i) \ ,$$

where ε_i is k-th distance now in ascending order within the set $\{d_{ij}\}$, with d_{ij} being the euclidean distance between the points (x_i, y_i) and (x_j, y_j), again with i being fixed and j ranging from 1 to N, with the exception of $j = i$.

For sufficiently smooth distributions, the estimator value is independent of the nearest neighbor's rank k. Here, k is set to be 20 [142]. A more detailed discussion on the Kozachenko-Leonenko method is present in several works, including Lombardi & Pant [150], and Delattre & Fournier [151]).

Another standard approach for MI estimation [138] consists in the plug-in estimator:

$$\widehat{I}(X, Y) = \sum_{i=1}^{M_x}\sum_{j=1}^{M_y} P_{xy}(i, j) \ln\left[\frac{P_{xy}(i, j)}{P_x(i)P_y(j)}\right] . \tag{5.2}$$

where the sample joint probability distribution is evaluated as a two-dimensional histogram $P_{xy}(i, j)$ out of the two sequences $\{x_t\}$, $\{y_t\}$. The indexes i, j identify the rectangular bins used to partition the joint distribution support [152, 153]: Concurrently, two marginal distributions histograms $P_x(i)$, $P_y(j)$ are made to estimate the marginal distributions $p_X(x)$, $p_Y(y)$, respectively. The estimate $\widehat{I}(X, Y)$

converges to $I(X, Y)$ in the limit $N \to \infty$ [138]. This estimator would be relatively easy to compute but its accuracy strongly depends on the bin partitioning, which proves to be a non-trivial choice [154].

5.3. Experimental dataset

Both methods considered in the present work, namely MI and CTO, are applied to pairs of scalar sequences, henceforth referred to as $\{x_t\}$, $\{y_t\}$, where t is an integer between 1 and N. Each sequence, sampled with a sampling period T, is produced by a node, i.e. a point-like location within the space region encompassed by the system of interest. If the system contains ν nodes, the number of possible links to analyze is $\nu(\nu - 1)/2$.

Three sets of experimental recordings are analyzed to investigate the relationship between MI and CTO. The first two datasets correspond to electromagnetic recordings of human brain activity in resting state, one acquired by means of magnetoencephalography and one by means of electroencephalography. In both cases, power sequences are reconstructed for a given set of source locations, or nodes, within the brain. The third dataset concerns surface wind speed data recorded by meteorological stations—and thus nodes—in the Province of Trento.

5.3.1. MEG recordings

Magnetoencephalography (MEG) is a powerful non-invasive tool that allows to capture the dynamic electrical activity of the brain with unparalleled precision. MEG is a neuroimaging technique that measures the magnetic fields generated by the electrical activity of the neurons in the brain. MEG recordings are collected via superconducting quantum interference devices (SQUIDs) that are able to detect the weak magnetic fields generated by brain activity. SQUIDs are capable of measuring magnetic fields at the order of femto-Tesla, making MEG an extremely sensitive technique for capturing electrical dynamics of the brain.

The use of MEG in neuroscience research has grown exponentially in recent years due to its unique advantages. One of the key strengths of MEG is the

ability to provide millisecond-level temporal resolution, allowing to capture rapid changes in brain activity that occur during various cognitive processes, such as perception, attention and memory. This temporal precision is crucial for understanding real-time dynamics of the brain and unraveling the sequence of events that occur during different cognitive tasks.

Another advantage of MEG is its ability to directly measure the magnetic fields produced by neural currents that are highly localized and specific to the underlying brain activity. Unlike other neuroimaging techniques such as functional magnetic resonance imaging (fMRI) that rely on changes in blood flow as a proxy for brain activity, MEG provides a direct measure of the electrical activity of neurons. This makes MEG highly sensitive to the actual neural processes that occur in the brain and allows for more precise localization of brain activity.

Due to the possibility of mapping the brain activity during the execution of functional tasks, it's possible to identify the brain regions that are involved in specific cognitive processes. This can be used to gain insight about the functional organization of the brain and establish the existence of neural networks involved in language processing, motor control, and sensory perception.

MEG recording typically involves placing the subject's head inside a highly shielded room to minimize interference from external magnetic fields. The participant wears a helmet-like device that contains the SQUIDs, connected to a data acquisition system. During the acquisition, the participant may be asked to perform a specific task or engage in a cognitive process, or stay in a relaxed state, commonly referred as resting state. Processing the recorded data involves sophisticated algorithms and statistical techniques in order to reconstruct the spatio-temporal dynamics of brain activity. The final outcome of this preprocessing consists in temporal recordings associated with *voxels*, small volumes that make up a mapping of the brain cortex.

Here we use MEG recordings collected from the CamCan public database [155, 156]. The dataset is made up of recordings from 20 healthy subjects with age of 18-28 years. The recordings were aquired with the subjects in resting state and with eyes closed. For each subject, brain activity is reconstructed at 72 brain

locations uniformly distributed on the cortex. The details of the preprocessing procedure to reconstruct brain activity is described in Castelluzzo et al. [144]. For each node, a recording sampled at 250 Hz with a duration between 180 and 240 seconds is available. The final dataset therefore consists in $72 \cdot 71/2 = 2556$ pairs of sequences available for each subjects, making up a total of 51120 pairs.

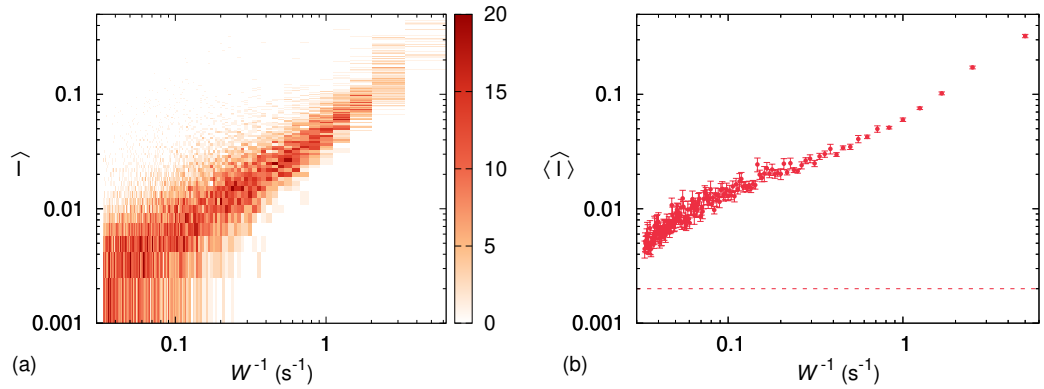

Figure 5.2: (a) Sample joint histogram of $W^{-1}, \widehat{I}$ built out of MEG recordings. The color scale identifies the bin occupation. (b) Average MI $\langle \widehat{I} \rangle$ as a function of W^{-1} computed along each column of the joint histogram in (a), i.e. for fixed values of W^{-1}. The errorbars correspond to the standard error $s_{\langle \widehat{I} \rangle}$ of $\langle \widehat{I} \rangle$. The dashed line corresponds to the value of $\langle \widehat{I} \rangle$ computed for all the pairs of sequences that do not provide a finite W value.

Reproduced from Chaos 31, 073106 (2021), with the permission of AIP Publishing.

In Figure 5.2 we show the joint histogram of $\widehat{I}$ and W^{-1}, evaluated by setting the threshold efficiency η^* to 0.5. The number of bins for the continuous variable $\widehat{I}$ was computed by using the Freedman-Diaconis rule [157, 158]. The binning of W^{-1} is instead determined by discrete values of W, that as mentioned in Sec. 5.1 can take on values being multiples of the base window width w_0.

Due to a total of 36462 pairs that did not exhibit a link, according to the CTO algorithm, and therefore did not produce a finite value of W, the histogram is made up of 14704 points. The existence of a relationship betweeen W^{-1} and $\widehat{I}$ appears evident. The average value of the MI estimator $\langle \widehat{I} \rangle$, and the related uncertainty estimated as the standard error $s_{\langle \widehat{I} \rangle}$, was then considered for each available

discrete value of W, as shown in Figure 5.2 (b).

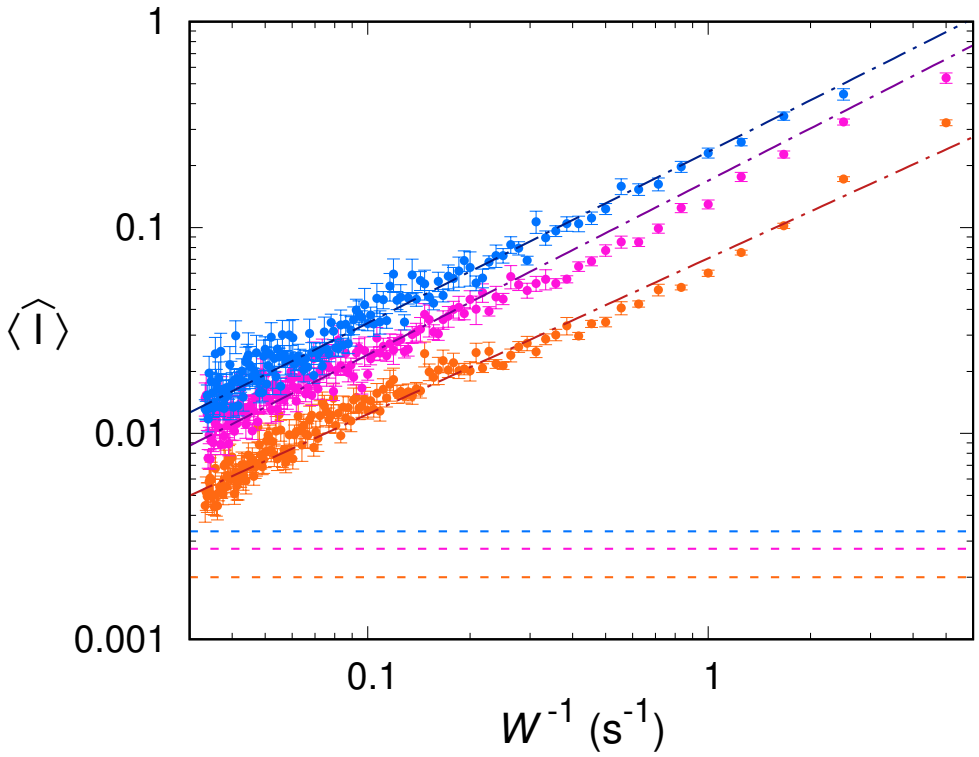

Figure 5.3: Average MI $\langle \widehat{I} \rangle$ as a function of W^{-1} evaluated out of MEG recordings. The errorbars correspond to the standard error $s_{\langle \widehat{I} \rangle}$ of $\langle \widehat{I} \rangle$. The orange, magenta and blue points correspond to evaluations of W carried out by setting the efficiency threshold η^* to 0.5, 0.75 and 0.9, respectively. The dashed, horizontal lines correspond to the values of $\langle \widehat{I} \rangle$ computed for all the pairs of sequences that do not provide a finite W (the color key is the same as for the data points). The three dash-dotted lines correspond to the fits of the linear parts of the log-log plots. The fit procedure is discussed in Ref. [142].

Reproduced from Chaos 31, 073106 (2021), with the permission of AIP Publishing.

In Fig. 5.3, we show $\langle \widehat{I} \rangle$ vs. W^{-1} at different values of the threshold η^*, namely $(0.5, 0.75, 0.9)$. Using more conservative threshold values causes an increased number of discarded pairs due to the absence of a link, namely 41831 and 44347 for the threholds $\eta^* = 0.75$ and $\eta^* = 0.9$ respectively.

Nonetheless, the resulting plots show the existence of a power-law relation-

ship between the two quantities estimating the connectivity strenght. Furthermore, the effects of changing the efficiency threshold seems to cause a parallel shift in the log-log scale, which gives hints of the parameter η^* acting as a multiplying factor in the function binding $\langle \widehat{I} \rangle$ to W^{-1}.

5.3.2. EEG recordings

Electroencephalography (EEG) is a non-invasive neurophysiological technique that has been widely used to study brain activity. It measures the electrical activity of the brain providing valuable insights into brain function and dysfunction. EEG has been a fundamental tool in neuroscience research, clinical practice, and various applications, such as diagnosing epilepsy, studying sleep patterns, and assessing cognitive processes.

EEG recordings essentially measures electrical signals produced by synchronous firing of neurons, by means of electrodes placed on the scalp. The electrical activity is measured as a series of waves with specific frequency bands associated with different brain states and can provide insights into various aspects of brain functioning. The source of the electrical brain signals is the postsynaptic potentials of cortical neurons. When a large number of neurons fire synchronously, they generate an electrical field that can be detected by the electrodes. EEG measures the voltage difference between two or more electrodes, resulting in time series known as an EEG waveform.

The dataset used here was collected from the LEMON public database [159, 160]. A total of 30 healthy subjects between 20-35 years of age has been selected, with recordings carried out in resting state with closed eyes. For each subject, we reconstructed the brain activity within the so-called alpha band, namely between 8 Hz and 14 Hz, from 30 brain locations uniformly distributed on the cortex. A more detailed description of the required EEG preprocessing is discussed in Perinelli et al. [142].

For each of the 30 brain locations, three segments are sampled at 250 Hz, with a duration of 60 s each. The total amount of pairs of sequences analyzed therefore corresponds to $30 \cdot 29/2 = 435$ for each subject, for a total of 39150 pairs.

The average $\langle \widehat{I} \rangle$, and the related standard error $s_{\langle \widehat{I} \rangle}$, for each value of W^{-1}, computed following a similar procedure as the one discussed in the previous section, is shown in Fig. 5.4. Again, we performed the analysis on three values of η^*, namely 0.5, 0.75, 0.9. For each threshold the number of links that were discarded due to the absence of a detectable link is 10895, 17917, 23087, respectively.

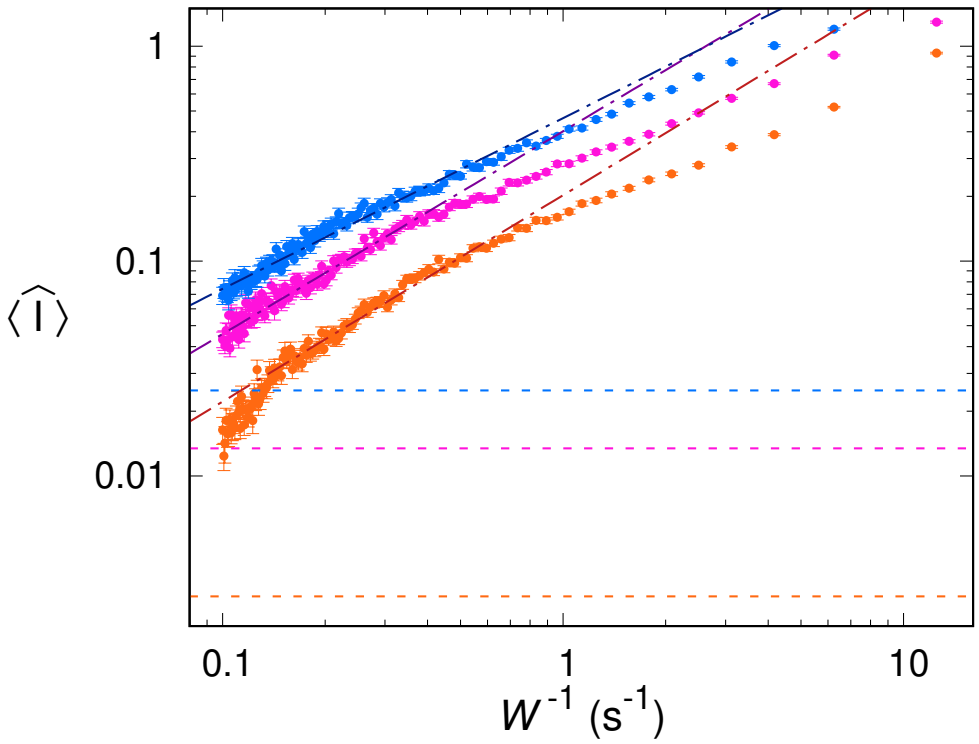

Figure 5.4: Average MI $\langle \widehat{I} \rangle$ as a function of W^{-1} evaluated out of EEG recordings in the alpha frequency band. The errorbars correspond to the standard error $s_{\langle \widehat{I} \rangle}$ of $\langle \widehat{I} \rangle$. The orange, magenta and blue points correspond to evaluations of W carried out by setting the efficiency threshold η^* to 0.5, 0.75 and 0.9, respectively. The dashed, horizontal lines correspond to the values of $\langle \widehat{I} \rangle$ computed for all the pairs of sequences that do not provide a finite W (the color key is the same as for the data points). The three dash-dotted lines correspond to the fits of the linear parts of the log-log plots. The fit procedure is discussed in Ref. [142].
Reproduced from Chaos 31, 073106 (2021), with the permission of AIP Publishing.

Figure 5.4 shows a behaviour similar to the one found in the previous section

for MEG data, with the data hinting at a power-law relationship with multiplying factor depending on the selected efficiency threshold.

5.3.3. Surface wind speed recordings

The wind recordings dataset was collected from the public weather database of the Province of Trento [161], Italy. The selected data refer to 30 weather stations and, for each station, 4 wind recordings were selected. Each recording covers 120 days, namely the first 120 days of each year from 2015 to 2018, sampled with a period of 600 s. Due to occasional sensors malfunctioning and environmental issues affecting the weather stations, a preprocessing has been performed to recover missing data by means of linear interpolation of the neighboring available points. In any case, the occurrence rate of missing data amounts to less than 1% and, occasionally, occuring for consecutive points in the recordings. The final dataset then amounts to 435 pairs of nodes and, considering 4 recordings for each weather station, to a total of 1740 pairs in the whole dataset.

Following the precedure applied to MEG and EEG recordings, Figure 5.5 shows the average $\langle \widehat{I} \rangle$ and the related standard error $s_{\langle \widehat{I} \rangle}$ for each value of W^{-1}. The same values for the threshold have used, namely 0.5, 0.75, 0.9 with a number of discarded pairs equal to 499, 1150 and 1513, respectively.

The plot in Fig. 5.5 shows an increased dispersion in the values of $\langle \widehat{I} \rangle$ with respect to MEG and EEG recordings, particularly to those points corresponding to lower values of W^{-1}. Nevertheless, the same features emerging from the analysis on the two brain activity recordings are evident from this analysis.

5.4. Final remarks

The analysis shown in this chapter allow to draw some meaningful conclusions from the comparison of MI with the CTO method.

The first interesting aspect is validating further the reliability of the CTO method. This method was only recently developed and has been mainly applied to the study on brain networks, specifically out of EEG and MEG recordings. Pro-

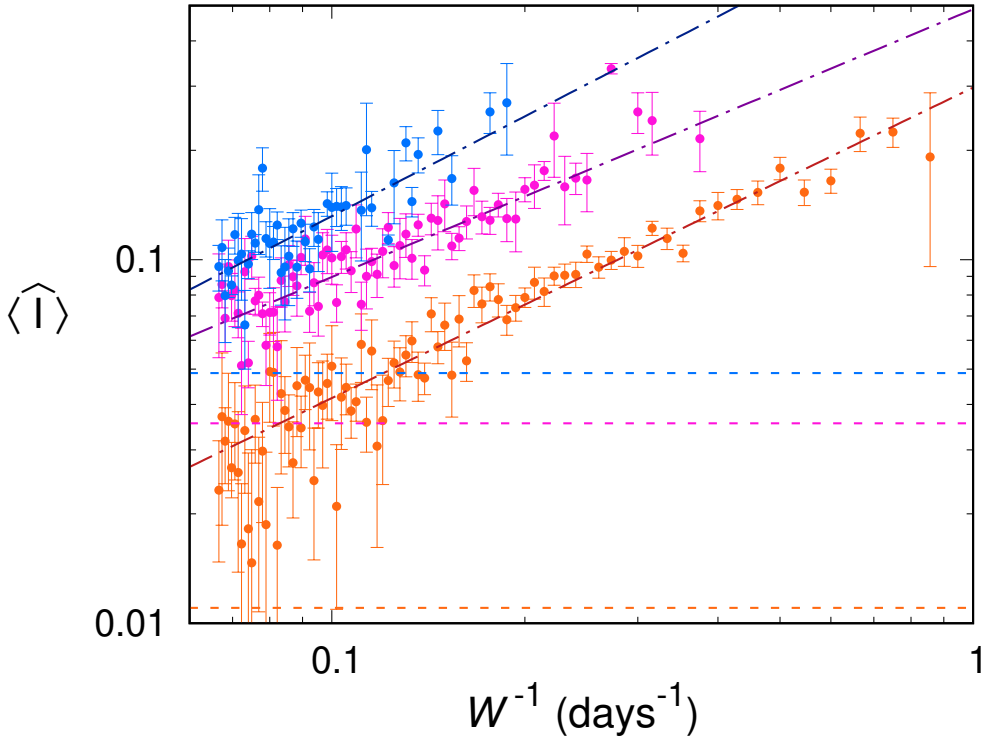

Figure 5.5: Average MI $\langle \widehat{I} \rangle$ as a function of W^{-1} evaluated out of surface wind speed recordings. The errorbars correspond to the standard error $s_{\langle \widehat{I} \rangle}$ of $\langle \widehat{I} \rangle$. The orange, magenta and blue points correspond to evaluations of W carried out by setting the efficiency threshold η^* to 0.5, 0.75 and 0.9, respectively. The dashed, horizontal lines correspond to the values of $\langle \widehat{I} \rangle$ computed for all the pairs of sequences that do not provide a finite W (the color key is the same as for the data points). The three dash-dotted lines correspond to the fits of the linear parts of the log-log plots. The fit procedure is discussed in Ref. [142].

Reproduced from Chaos 31, 073106 (2021), with the permission of AIP Publishing.

viding a comparison with an established and widely used metric as mutual information, and the evident and repeated occurrence of a power-law relationship between these two metrics on heterogeneous experimental data, allows to reliably apply the CTO method to contexts other than neuroscience. Furthermore, the theoretical model proposed to explain this empirical power-law relationship, discussed in Appendix A, allows to gain some insights into the nature of the cor-

relation of the system under study, in terms of the parameters making up the theoretical model. This model describes the signals as a sum of delta-like correlation events summed to a noise component. Therefore, it should be noted that such a model works in describing the experimental evidences shown in this chapter with the assumption that correlation events are independent and sparse, so that the funnel-like structures occurring in the cross-correlation diagram and the p-value diagrams (as shown in Fig. 5.1) do not overlap.

The second conclusion is the complementarity between MI and CTO. First in terms of resolution: by construction, the CTO method provides a higher resolution at small values of connectivity strength, due to the uniformly distributed values of W that, as already discussed, take on values being multiples of the base window width w_0. By the evidence of the analysis shown here, MI appears instead to provide a more limited range of outcomes for weakly-linked nodes. Considering also the issue of spurious events, CTO algorithm's parameter could be tuned to detect spurious correlations and, due to its higher resolution at weaker connectivity, it could prove useful in characterizing these events which, as previously discussed in Section 5.1, are expected to result in weaker connections in terms of W^{-1}. Conversely, while MI would prove more robust against few spurious events (provided their duration is short compared to the time series), it does not provide any additional information on these events. In addition, MI lacks an absolute interpretation when the outcome is significantly different from 0. A comparison between multiple pairs by means of MI allows to draw conclusions in terms of which pairs are more correlated than other pairs. This makes MI more suitable to quantify differences in connectivity, more than absolute connectivity on a single pair. CTO instead, providing an outcome with the dimension of time, is more adapt to be employed as an absolute connectivity metric.

Finally, it should be mentioned that CTO appears to be more conservative than MI. In Perinelli et al. [142], the two metrics are in fact tested on synthetic surrogate data that are therefore non-significantly correlated by means of the CTO method (see Section 5.1). Nonetheless, the spurious correlations occurring solely due to the stochastic nature of surrogate data, produce significantly non-zero values of

MI.

Conclusions

The use of techniques stemming from Information Theory on the analysis of time series has gained a lot of interest in the last decades. Several metrics based on the concept of entropy have been developed and adapted to a wide range of fields, from economics, neuroscience, physiology and earth sciences.

As we discussed in Chapter 2, the work of several scientist in the last century, starting from the seminal work by Shannon, has eventually led to find the link between entropy and the intrinsic complexity of dynamical systems. Therefore, beside being able to quantify the variability of a sequence of data, entropy metrics provide interesting insights into recognizing evidences of deterministic chaotic behaviour in a system's evolution.

Despite their wide use, entropy measures still require a solid mathematical background to properly characterize their behaviour on real data. For example, in Chapter 3 we discuss a recent result regarding the estimation of the Shannon entropy variance. A proper evaluation of this last quantity is well-known in some situations, but even for the relatively simple case of data sampled from an underlying multinomial memoryless distribution, such an estimation was lacking. Our work proves the reliability of a plug-in entropy variance estimator. Despite a proper evaluation, in general, requires to know the *a priori* distribution of the sampled data, namely higher-order moments of the Shannon Information, we were able to study the case in which a distribution gives out the maximum entropy variance and thus make up a practical rule of thumb to link the number of required samples to the wanted accuracy on entropy estimations.

In Chapter 4, we showed how permutation entropy has gained significant in-

terest for its potential to estimate Kolmogorov-Sinai entropy, despite being simple and intuitive. Nonetheless, most of the works on this topic involve time-discrete dynamics, while few attempts have been tried on continuous systems. Several recent works have studied the reliability of estimating KSE rate through permutation entropy by means of some corrective terms or modifications of the original formula of PE from Bandt and Pompe. Nonetheless, at least for the approach we considered, some issues are still present. As an example, we have found some evidences suggesting that in the case of signals with abrupt spikes one of these methods performs worst than the standard permutation entropy.

Finally, in the last chapter, we discussed an interesting comparison of mutual information with a recently developed correlation algorithm. This result does not aim at proving the validity of mutual information to study signal correlations: this in fact is already well-established and finds several applications, especially in neuroscience. Rather, it gives an interesting insight on the complementarity of the CTO method and MI, suggesting that the use of both techniques on the same dataset can uncover more interesting properties than each of the two methods alone.

Information-theoretical techniques then provide an alternative and complementary set of tools to other approaches from nonlinear time series analysis. The main advantage of the former, being that no further knowledge of the system's properties are required, since entropy metrics are usually non-parametric. On the contrary, as we discussed in Chapter 1, the application of nonlinear time series analysis methods can easily turn out to produce conflicting results when these are not applied in a conscious way.

Bibliography

[1] O. E. Rössler. An equation for continuous chaos. *Phys. Lett.*, 57A:397–398, 1976.

[2] P.J. Prince and J.R. Dormand. High order embedded runge-kutta formulae. *Journal of Computational and Applied Mathematics*, 7(1):67–75, 1981.

[3] A. Perinelli and L. Ricci. Identification of suitable embedding dimensions and lags for time series generated by chaotic, finite-dimensional systems. *Phys. Rev. E*, 98:052226, 2018.

[4] G. Benettin, L. Galgani, A. Giorgilli, and J.M. Strelcyn. Lyapunov characteristic exponents for smooth dynamical systems and for Hamiltonian systems; a method for computing all of them. Part 1: Theory. *Meccanica*, 15:9–20, 1980.

[5] G. Benettin, L. Galgani, A. Giorgilli, and J.M. Strelcyn. Lyapunov characteristic exponents for smooth dynamical systems and for Hamiltonian systems; a method for computing all of them. Part 2: Numerical application. *Meccanica*, 15:21–30, 1980.

[6] M. Sano and Y. Sawada. Measurement of the Lyapunov spectrum from a chaotic time series. *Phys. Rev. Lett.*, 55:1082–1085, Sep 1985.

[7] J. P. Eckmann, S. Oliffson Kamphorst, D. Ruelle, and S. Ciliberto. Lyapunov exponents from time series. *Phys. Rev. A*, 34:4971–4979, Dec 1986.

[8] A. Karantonis and M. Pagitsas. Comparative study for the calculation of the lyapunov spectrum from nonlinear experimental signals. *Phys. Rev. E*, 53:5428–5444, May 1996.

[9] A. Wolf, J. B. Swift, H. L. Swinney, and J. A. Vastano. Determining Lyapunov exponents from a time-series. *Physica D*, 16(3):285–317, 1985.

[10] M.T. Rosenstein, J.J. Collins, and C.J. De Luca. A practical method for calculating largest lyapunov exponents from small data sets. *Physica D*, 65:117–134, 1993.

[11] J. Gao and Z. Zheng. Local exponential divergence plot and optimal embedding of a chaotic time-series. *Phys. Lett. A*, 181:153–158, 1993.

[12] J. Gao and Z. Zheng. Direct dynamical test for deterministic chaos and optimal embedding of a chaotic time-series. *Phys. Rev. E*, 49:3807–3814, 1994.

[13] J. Gao and Z. Zheng. Direct dynamical test for deterministic chaos. *Europhys. Lett.*, 25:485–490, 1994.

[14] Z. J. Kowalik and T. Elbert. A practical method for the measurements of the chaoticity of electric and magnetic brain activity. *International Journal of Bifurcation and Chaos*, 5(2):475–490, 1995.

[15] F. Hausdorff. Dimension und äußeres maß. *athematische Annalen*, 79:157–179, 1918.

[16] P. Grassberger and I. Procaccia. Estimation of the Kolmogorov entropy from a chaotic signal. *Phys. Rev. A*, 28:2591–2593, Oct 1983.

[17] L. S. Liebovitch and T. Toth. A fast algorithm to determine fractal dimensions by box counting. *Physics Letters A*, 141(8):386–390, 1989.

[18] H.G.E. Hentschel and I. Procaccia. The infinite number of generalized dimensions of fractals and strange attractors. *Physica D: Nonlinear Phenomena*, 8(3):435–444, 1983.

[19] L. Ricci, A. Perinelli, and M. Franchi. Asymptotic behavior of the time-dependent divergence exponent. *Phys. Rev. E*, 101:042211, 2020.

[20] K. Pyragas. Continuous control of chaos by self-controlling feedback. *Physics Letters A*, 170(6):421–428, 1992.

[21] A. Brandstäter, J. Swift, Harry L. Swinney, A. Wolf, J. Doyne Farmer, Erica Jen, and P. J. Crutchfield. Low-dimensional chaos in a hydrodynamic system. *Phys. Rev. Lett.*, 51:1442–1445, Oct 1983.

[22] J.-P. Eckmann, S. Oliffson Kamphorst, and D. Ruelle. Recurrence plots of dynamical systems. *Europhysics Letters*, 4(9):973, nov 1987.

[23] K. Ikeda. Multiple-valued stationary state and its instability of the transmitted light by a ring cavity system. *Optics Communications*, 30(2):257–261, 1979.

[24] K. Ikeda, H. Daido, and O. Akimoto. Optical turbulence: Chaotic behavior of transmitted light from a ring cavity. *Phys. Rev. Lett.*, 45:709–712, Sep 1980.

[25] F. T. Arecchi, R. Meucci, G. Puccioni, and J. Tredicce. Experimental evidence of subharmonic bifurcations, multistability, and turbulence in a q-switched gas laser. *Phys. Rev. Lett.*, 49:1217–1220, Oct 1982.

[26] T. Midavaine, D. Dangoisse, and P. Glorieux. Observation of chaos in a frequency-modulated co_2 laser. *Phys. Rev. Lett.*, 55:1989–1992, Nov 1985.

[27] Dhruba J. Biswas, Vas Dev, and U. K. Chatterjee. Experimental observation of oscillatory instabilities and chaos in a gain-modulated single-mode cw co_2 laser. *Phys. Rev. A*, 35:456–458, Jan 1987.

[28] F. A. Hopf, D. L. Kaplan, H. M. Gibbs, and R. L. Shoemaker. Bifurcations to chaos in optical bistability. *Phys. Rev. A*, 25:2172–2182, Apr 1982.

[29] Y. Cho and T. Umeda. Observation of chaos in a semiconductor laser with delayed feedback. *Optics Communications*, 59(2):131–136, 1986.

[30] F. T. Arecchi, W. Gadomski, and R. Meucci. Generation of chaotic dynamics by feedback on a laser. *Physical Review A*, 34:1617–1620, 1986.

[31] F. T. Arecchi, R. Meucci, and W. Gadomski. Laser dynamics with competing instabilities. *Physical Review Letters*, 58:2205–2208, 1986.

[32] E. H. M. Hogenboom, W. Klische, C. O. Weiss, and A. Godone. Instabilities of a homogeneously broadened laser. *Phys. Rev. Lett.*, 55:2571–2574, Dec 1985.

[33] U. Hübner, N. B. Abraham, and C. O. Weiss. Dimensions and entropies of chaotic intensity pulsations in a single-mode far-infrared nh_3 laser. *Phys. Rev. A*, 40:6354–6365, Dec 1989.

[34] H Haken. Analogy between higher instabilities in fluids and lasers. *Physics Letters A*, 53(1):77–78, 1975.

[35] T. Matsumoto, L. O. Chua, and S. Tanaka. Simplest chaotic nonautonomous circuit. *Phys. Rev. A*, 30:1155–1157, Aug 1984.

[36] T. Matsumoto. A chaotic attractor from chua's circuit. *IEEE Transactions on Circuits and Systems*, 31(12):1055–1058, 1984.

[37] B. Mithuswamy. Implementing memristor based chaotic circuits. *International Journal of Bifurcation and Chaos*, 20(05):1335–1350, 2010.

[38] L. Minati, M. Frasca, N. Yoshimura, L. Ricci, P. Oświecimka, Y. Koike, Kazuya Masu, and H. Ito. Current-starved cross-coupled cmos inverter rings as versatile generators of chaotic and neural-like dynamics over multiple frequency decades. *IEEE Access*, 7:54638–54657, 2019.

[39] A. Namajūnas, K. Pyragas, and A. Tamaševičius. An electronic analog of the mackey-glass system. *Physics Letters A*, 201(1):42–46, 1995.

[40] M. S. Patel, U. Patel, A. Sen, G. C. Sethia, C. Hens, S. K. Dana, U. Feudel, K. Showalter, C. N. Ngonghala, and R. E. Amritkar. Experimental observa-

tion of extreme multistability in an electronic system of two coupled rössler oscillators. *Phys. Rev. E*, 89:022918, 2014.

[41] L. Ricci, A. Perinelli, M. Castelluzzo, S. Euzzor, and R. Meucci. Experimental evidence of chaos generated by a minimal universal oscillator model. *International Journal of Bifurcation and Chaos*, 31(12):2150205, 2021.

[42] L. Minati. Experimental synchronization of chaos in a large ring of mutually coupled single-transistor oscillators: Phase, amplitude, and clustering effects. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 24(4), 10 2014. 043108.

[43] H. D. I. Abarbanel. *Analysis of Observed Chaotic Data*. Springer Verlag, 1996.

[44] N. F. Rulkov. Images of synchronized chaos: Experiments with circuits. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 6(3):262–279, 1996.

[45] E. Padmanaban, S. Saha, M. Vigneshwaran, and S. K. Dana. Amplified response in coupled chaotic oscillators by induced heterogeneity. *Phys. Rev. E*, 92:062916, 2015.

[46] W. A. Brock and C. L. Sayers. Is the business cycle characterized by deterministic chaos? *Journal of Monetary Economics*, 22(1):71–90, 1988.

[47] J. A. Bastos and J. Caiado. Recurrence quantification analysis of global stock markets. *Physica A: Statistical Mechanics and its Applications*, 390(7):1315–1325, 2011.

[48] J. T. Barkoulas, A. Chakraborty, and A. Ouandlous. A metric and topological analysis of determinism in the crude oil spot market. *Energy Economics*, 34(2):584–591, 2012.

[49] D. I. Ponyavin. Solar cycle signal in geomagnetic activity and climate. *Solar Physics*, 224:465–471, 2004.

[50] M. Lei and G. Meng. Detecting nonlinearity of sunspot number. *International Journal of Nonlinear Sciences and Numerical Simulation*, 5(4):321–326, 2004.

[51] N. Jevti´c, J. S. Schweitzer, and C. J. Cellucci. Research note: Nonlinear time series analysis of northern and southern solar hemisphere daily sunspot numbers in search of short-term chaotic behavior. *A&A*, 379(2):611–615, 2001.

[52] P. D. Mininni, D. O. Gómez, and G. B. Mindlin. Stochastic relaxation oscillator model for the solar cycle. *Phys. Rev. Lett.*, 85:5476–5479, Dec 2000.

[53] B. Sivakumar. Chaos theory in geophysics: past, present and future. *Chaos, Solitons Fractals*, 19(2):441–462, 2004. Fractals in Geophysics.

[54] J. Huang and D. L. Turcotte. Are earthquakes an example of deterministic chaos? *Geophysical Research Letters*, 17(3):223–226, 1990.

[55] F.B. Pelap, L.Y. Kagho, and C.F. Fogang. Chaotic behavior of earthquakes induced by a nonlinear magma up flow. *Chaos, Solitons Fractals*, 87:71–83, 2016.

[56] A. Gualandi, J.-P. Avouac, S. Michel, and D. Faranda. The predictable chaos of slow earthquakes. *Science Advances*, 6(27):eaaz5548, 2020.

[57] A. Sornette, J. Dubois, J. L. Cheminée, and D. Sornette. Are sequences of volcanic eruptions deterministically chaotic? *Journal of Geophysical Research: Solid Earth*, 96(B7):11931–11945, 1991.

[58] M. De Domenico, M. A. Ghorbani, O. Makarynskyy, D. Makarynska, and H. Asadi. Chaos and reproduction in sea level. *Applied Mathematical Modelling*, 37(6):3687–3697, 2013.

[59] V. Cuomo, V. Lapenna, M. Macchiato, C. Serio, and L. Telesca. Stochastic behaviour and scaling laws in geoelectrical signals measured in a seismic

area of southern Italy. *Geophysical Journal International*, 139(3):889–894, 12 1999.

[60] E. N. Lorenz. Deterministic Nonperiodic Flow. *J. Atmos. Sci.*, 20:130–141, 1963.

[61] A. A. Tsonis and J. B. Elsner. Chaos, strange attractors, and weather. *Bulletin of the American Meteorological Society*, 70(1):14 – 23, 1989.

[62] J. Slingo and T. Palmer. Uncertainty in weather and climate prediction. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 369(1956):4751–4767, 2011.

[63] H. N. Shirer, C. J. Fosmire, R. Wells, and L. Suciu. Estimating the correlation dimension of atmospheric time series. *Journal of the Atmospheric Sciences*, 54(1):211 – 230, 1997.

[64] T. E. Karakasidis and A. Charakopoulos. Detection of low-dimensional chaos in wind time series. *Chaos, Solitons Fractals*, 41(4):1723–1732, 2009.

[65] A. A. Tsonis, J. B. Elsner, and K. P. Georgakakos. Estimating the dimension of weather and climate attractors: Important issues about the procedure and interpretation. *Journal of Atmospheric Sciences*, 50(15):2549 – 2555, 1993.

[66] I. Procaccia. Complex or just complicated? *Nature*, 333:498 – 499, 1988.

[67] A. Király and I. M. Jánosi. Stochastic modeling of daily temperature fluctuations. *Phys. Rev. E*, 65:051102, Apr 2002.

[68] D. Yılmaz and N. F. Güler. Analysis of the doppler signals using largest lyapunov exponent and correlation dimension in healthy and stenosed internal carotid artery patients. *Digital Signal Processing*, 20(2):401–409, 2010.

[69] P.E. Rapp, I.D. Zimmerman, A.M. Albano, G.C. Deguzman, and N.N. Greenbaun. Dynamics of spontaneous neural activity in the simian motor cortex: The dimension of chaotic neurons. *Physics Letters A*, 110(6):335–338, 1985.

[70] M. R. Guevara, L. Glass, M. C. Mackey, and A. Shrier. Chaos in neurobiology. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(5):790–798, 1983.

[71] A. Babloyantz, J.M. Salazar, and C. Nicolis. Evidence of chaotic dynamics of brain activity during the sleep cycle. *Physics Letters A*, 111(3):152–156, 1985.

[72] N. Thomasson, T. J. Hoeppner, C. L. Webber, and J. P. Zbilut. Recurrence quantification in epileptic eegs. *Physics Letters A*, 279(1):94–101, 2001.

[73] J. P. Pijn, J. Van Neerven, A. N., and F. H. Lopes da Silva. Chaos or noise in eeg signals; dependence on state and brain site. *Electroencephalography and Clinical Neurophysiology*, 79(5):371–381, 1991.

[74] A Babloyantz and A Destexhe. Low-dimensional chaos in an instance of epilepsy. *Proceedings of the National Academy of Sciences*, 83(10):3513–3517, 1986.

[75] K. Lehnertz and C.E. Elger. Spatio-temporal dynamics of the primary epileptogenic area in temporal lobe epilepsy characterized by neuronal complexity loss. *Electroencephalography and Clinical Neurophysiology*, 95(2):108–117, 1995.

[76] F. Mormann, T. Kreuz, C. Rieke, R. G. Andrzejak, A. Kraskov, P. D., C. E. Elger, and K. Lehnertz. On the predictability of epileptic seizures. *Clinical Neurophysiology*, 116(3):569–587, 2005.

[77] P. E. McSharry, L. A. Smith, and L. Tarassenko. Prediction of epileptic seizures: are nonlinear methods relevant? *Nature Medicine*, 9:241–242, 2003.

[78] M.I Rabinovich and H.D.I Abarbanel. The role of chaos in neural systems. *Neuroscience*, 87(1):5–14, 1998.

[79] H. Korn and P. Faure. Is there chaos in the brain? ii. experimental evidence and related models. *Comptes Rendus Biologies*, 326(9):787–840, 2003.

[80] H. Hayashi, M. Nakao, and K. Hirakawa. Chaos in the self-sustained oscillation of an excitable biological membrane under sinusoidal stimulation. *Physics Letters A*, 88(5):265–266, 1982.

[81] Hatsuo Hayashi and Satoru Ishizuka. Chaotic nature of bursting discharges in the onchidium pacemaker neuron. *Journal of Theoretical Biology*, 156(3):269–291, 1992.

[82] M. J. Brookes, E. L. Hall, S. E. Robson, D. Price, L. Palaniyappan, E. B. Liddle, P. F. Liddle, S. E. Robinson, and P. G. Morris. Complexity measures in magnetoencephalography: Measuring "disorder" in schizophrenia. *PLOS ONE*, 10(4):1–23, 04 2015.

[83] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Techn. J.*, 27:379–423,623–656, 1948.

[84] M. Badino. Bridging conceptual gaps: The kolmogorov-sinai entropy. *Isonomía. Revista de Teoría y Filosofía Del Derecho*, forthcoming.

[85] J. P. Eckmann and D. Ruelle. Ergodic theory of chaos and strange attractors. *Rev. Mod. Phys.*, 57:617–656, Jul 1985.

[86] M. Ribeiro, T. Henriques, L. Castro, A. Souto, L. Antunes, C. Costa-Santos, and A. Teixeira. The entropy universe. *Entropy*, 23(2), 2021.

[87] S. M. Pincus. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, 88(6):2297–2301, 1991.

[88] J. S. Richman and J. R. Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, 278(6):H2039–H2049, 2000. PMID: 10843903.

[89] S. Pincus and R. E. Kalman. Irregularity, volatility, risk, and financial market time series. *Proceedings of the National Academy of Sciences*, 101(38):13709–13714, 2004.

[90] R. Alcaraz and J.J. Rieta. A novel application of sample entropy to the electrocardiogram of atrial fibrillation. *Nonlinear Analysis: Real World Applications*, 11(2):1026–1035, 2010.

[91] D.T. Mihailović, E. Nikolić-Đorić, N. Drešković, and G. Mimić. Complexity analysis of the turbulent environmental fluid flow time series. *Physica A: Statistical Mechanics and its Applications*, 395:96–104, 2014.

[92] C. Bandt and B. Pompe. Permutation Entropy: A Natural Complexity Measure for Time Series. *Phys. Rev. Lett.*, 88(17):174102, 2002.

[93] L. Ricci, A. Perinelli, and M. Castelluzzo. Estimating the variance of shannon entropy. *Phys. Rev. E*, 104:024220, Aug 2021.

[94] G. P. Basharin. On a statistical estimate for the entropy of a sequence of independent random variables. *Theor. Probab. Appl.*, 4:333–336, 1959.

[95] G. Miller. *Note on the bias of information estimates*, pages 95–100. Free Press, Glencoe, IL, 1955.

[96] B. Harris. *The statistical estimation of entropy in the non-parametric case*, pages 323–355. North-Holland, Amsterdam, 1975.

[97] R. Courant and D. Hilbert. *Methods of Mathematical Physics, vol. I.* Interscience Publishers, Inc., New York, 1953.

[98] R. A. Adams and C. Essex. *Calculus: a complete course. 7th ed.* Pearson Canada, Toronto, 1953.

[99] L. Ricci. Asymptotic distribution of sample Shannon entropy in the case of an underlying finite, regular Markov chain. *Phys. Rev. E*, 103:022215, 2021.

[100] L. Ricci. A quantum-mechanical derivation of the multivariate central limit theorem for Markov chains. *Chaos Soliton. Fract.*, 142:110450, 2021.

[101] A. Antos and I. Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Struct. Algor.*, 19:163–193, 2001.

[102] Y. B. Pesin. Characteristic lyapunov exponents and smooth ergodic theory. *Russian Mathematical Surveys*, 32(4):55, aug 1977.

[103] M. Franchi and L. Ricci. Statistical properties of the maximum Lyapunov exponent calculated via the divergence rate method. *Phys. Rev. E*, 90:062920, 2014.

[104] Jose Amigó. *Permutation Complexity in Dynamical Systems*. Springer Series in Synergetics. Springer, Dordrecht, 2010.

[105] Riccardo Meucci, S. Euzzor, T. F. Arecchi, and J. Ginoux. Minimal universal model for chaos in laser with feedback. *International Journal of Bifurcation and Chaos*, 31(04):2130013, 2021.

[106] Bhavin J. Shastri, Mitchell A. Nahmias, Alexander N. Tait, Ben Wu, and Paul R. Prucnal. SIMPEL: Circuit model for photonic spike processing laser neurons. *Opt. Express*, 23:8029–8044, 2015.

[107] J. Tiana-Alsina, C. Quintero-Quiroz, and C. Masoller. Comparing the dynamics of periodically forced lasers and neurons. *New J. Phys.*, 21:103039, 2019.

[108] V. Volterra. Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. *Mem. Acad. Lincei Roma*, 2:31–113, 1926.

[109] V. Volterra. *Variations and Fluctuations of the Number of Individuals in Animal Species living together*, pages 409–448. McGraw-Hill, New York, USA, 1931.

[110] A. J. Lotka. Contribution to the theory of periodic reaction. *J. Phys. Chem.*, 14:271–274, 1910.

[111] A. J. Lotka. Analytical note on certain rhythmic relations in organic systems. *Proc. Natl. Acad. Sci. USA*, 6:410–415, 1920.

[112] I. B. Schwartz and H. Smith. Infinite subharmonic bifurcation in an SEIR epidemic model. *J. Math. Biol.*, 18:233–253, 1983.

[113] A. Politi. Quantifying the dynamical complexity of chaotic time series. *Phys. Rev. Lett.*, 118:144101, 2017.

[114] P. Frederickson, J. L. Kaplan, E. D. Yorke, and J. A. Yorke. The Liapunov dimension of strange attractors. *J. Differ. Equations*, 49:185–207, 1983.

[115] L. Ricci, M. Castelluzzo, L. Minati, and A. Perinelli. Generation of surrogate event sequences via joint distribution of successive inter-event intervals. *Chaos*, 29:121102, 2019.

[116] A. Perinelli, M. Castelluzzo, L. Minati, and L. Ricci. SpiSeMe: A multi-language package for spike train surrogate generation. *Chaos*, 30:073120, 2020.

[117] M. Timme and J. Casadiego. Revealing networks from dynamics: an introduction. *J. Phys. A: Math. Theor.*, 47:343001, 2014.

[118] F. Arizmendi, A. C. Martí, and M. Barreiro. Evolution of atmospheric connectivity in the 20th century. *Nonlin. Processes Geophys.*, 21:825–839, 2014.

[119] N. Ekhtiari, A. Agarwal, N. Marwan, and R. V. Donner. Disentangling the multi-scale effects of sea-surface temperatures on global precipitation: A coupled networks approach. *Chaos*, 29:063116, 2019.

[120] N. Boers, B. Goswami, A. Rheinwalt, B. Bookhagen, B. Hoskins, and J. Kurths. Complex networks reveal global pattern of extreme-rainfall teleconnections. *Nature*, 566:373–377, 2019.

[121] A. Bastos and title = A Tutorial Review of Functional Connectivity Analysis Methods and Their Interpretational Pitfalls journal = Front. Sys. Neurosci. volume = 9 pages = 175 year = 2016 doi = 10.3389/fnsys.2015.00175 Schoffelen, J.

[122] P. M. Rossini, R. Di Iorio, M. Bentivoglio, G. Bertini, F. Ferreri, C. Gerloff, R. J. Ilmoniemi, F. Miraglia, M. A. Nitsche, F. Pestilli, M. Rosanova, Y. Shirota, C. Tesoriero, Y. Ugawa, F. Vecchio, U. Ziemann, and M. Hallett. Meth-

ods for analysis of brain connectivity: An IFCN-sponsored review. *Clin. Neurophisiol.*, 130:1833–1858, 2019.

[123] M. E. Raichle, A. M. MacLeod, A. Z. Snyder, W. J. Powers, D. A. Gusnard, and G. L. Shulman. A default mode of brain function. *Proc. Natl. Acad. Sci.*, 98(2):676–682, 2001.

[124] E. Niedermeyer and F. H. L. da Silva. *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Lippincott Williams & Wilkins, Philadelphia, 5th edition, 2005.

[125] S. Supek and C. J. Aine. *Magnetoencephalography: From Signals to Dynamic Cortical Networks*. Springer-Verlag, Berlin Heidelberg, 1st edition, 2014.

[126] M. Hämäläinen, R. Hari, R. J. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Rev. Mod. Phys.*, 65:413–497, 1993.

[127] L. Minati, H. Ito, A. Perinelli, L. Ricci, L. Faes, N. Yoshimura, Y. Koike, and M. Frasca. Connectivity influences on nonlinear dynamics in weakly-synchronized networks: Insights from rössler systems, electronic chaotic oscillators, model and biological neurons. *IEEE Access*, 7:174793–174821, 2019.

[128] T. Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85:461–464, 2000.

[129] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37:424–438, 1969.

[130] A. K. Seth, A. B. Barrett, and L Barnett. Granger causality analysis in neuroscience and neuroimaging. *J. Neurosc.*, 35:3293–3297, 2015.

[131] L. Faes, G. Nollo, S. Stramaglia, and D. Marinazzo. Multiscale granger causality. *Phys. Rev. E*, 96:042150, 2017.

[132] R. D. Pascual-Marqui, D. Lehmann, M. Koukkou, K Kochi, P. Anderer, B. Saletu, H. Tanaka, K. Hirata, E. R. John, L. Prichep, R. Biscay-Lirio, and T. Kinoshita. Assessing interactions in the brain with exact low-resolution electromagnetic tomography. *Phil. Trans. R. Soc. A*, 369:3768–3784, 2011.

[133] S. M. Bowyer. Coherence a measure of the brain networks: past and present. *Neuropsych. Electrophysiol.*, 2:1, 2016.

[134] G. Nolte, O. Bai, L. Wheaton, Z. Mari, S. Vorbach, and M. Hallett. Identifying true brain interaction from EEG data using the imaginary part of coherency. *Clin. Neurophisiol.*, 115:2292–2307, 2004.

[135] L. Baccalá and K. Sameshima. Partial directed coherence: a new concept in neural structure determination. *Biol. Cybern.*, 84:463–474, 2001.

[136] A. Bruns. Fourier-, Hilbert- and wavelet-based signal analysis: are they really different approaches? *J. Neurosci. Methods*, 137:321–332, 2004.

[137] D. Tabarelli, C. Keitel, J. Gross, and D. Baldauf. Spatial attention enhances cortical tracking of quasi-rhythmic visual stimuli. *NeuroImage*, 208:116444, 2020.

[138] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, 2004.

[139] W. Xiong, L. Faes, and P. Ch. Ivanov. Entropy measures, entropy estimators, and their performance in quantifying complex dynamics: Effects of artifacts, nonstationarity, and long-range correlations. *Phys. Rev. E*, 95:062114, 2017.

[140] A. Perinelli, D. E. Chiari, and L. Ricci. Correlation in brain networks at different time scale resolution. *Chaos*, 28:063127, 2018.

[141] T. Schreiber and A. Schmitz. Surrogate time series. *Physica D*, 142:346 – 382, 2000.

[142] A. Perinelli, M. Castelluzzo, D. Tabarelli, V. Mazza, and L. Ricci. Relationship between mutual information and cross-correlation time scale of observability as measures of connectivity strength. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(7), 07 2021. 073106.

[143] A. Perinelli, D. Tabarelli, C. Miniussi, and L. Ricci. Dependence of connectivity on geometric distance in brain networks. *Sci. Rep.*, 9:13412, 2019.

[144] M. Castelluzzo, A. Perinelli, D. Tabarelli, and L. Ricci. Dependence of Connectivity on the Logarithm of Geometric Distance in Brain Networks. *Front. Physiol.*, 11:611125, 2021.

[145] A. Perinelli and L. Ricci. NetOnZeroDXC: A package for the identification of networks out of multivariate time series via zero-delay cross-correlation. *SoftwareX*, 10:100316, 2019.

[146] See https://github.com/LeonardoRicci/netOnZeroDXC for the NetOnZeroDXC software package.

[147] L. F. Kozachenko and N. N. Leonenko. Sample Estimate of the Entropy of a Random Vector. *Probl. Peredachi Inf.*, 23:9–16, 1987.

[148] D. P. Shorten, R. E. Spinney, and J. T. Lizier. Estimating Transfer Entropy in Continuous Time Between Neural Spike Trains or Other Event-Based Data. *PLOS Comput. Biol.*, 17:1–45, 2021.

[149] G. Mijatovic, Y. Antonacci, T. Loncar Turukalo, L. Minati, and L. Faes. An Information-Theoretic Framework to Measure the Dynamic Interaction between Neural Spike Trains. *IEEE Trans. Biomed. Eng.*, pages 1–1, 2021.

[150] D. Lombardi and S. Pant. Nonparametric k-nearest-neighbor entropy estimator. *Phys. Rev. E*, 93:013310, 2016.

[151] S. Delattre and N. Fournier. On the Kozachenko-Leonenko entropy estimator. *J. Stat. Plan. Infer.*, 185:69–93, 2017.

[152] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Hoboken, New Jersey, 2nd edition, 2006.

[153] A. Galka, T. Ozaki, J. B. Bayard, and O. Yamashita. Whitening as a tool for estimating mutual information in spatiotemporal data sets. *J. Stat. Phys.*, 124:1275–1315, 2006.

[154] G. A. Darbellay and I. Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45:1315–1321, 1999.

[155] M. A. Shafto, L. K. Tyler, M. Dixon, J. R. Taylor, J. B. Rowe, R. Cusack, A. J. Calder, W. D. Marslen-Wilson, J. Duncan, T. Dalgleish, R. N. Henson, C. Brayne, F. E. Matthews, and Cam-CAN. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC Neurology*, 14, 2014.

[156] J. R. Taylor, N. Williams, R. Cusack, T. Auer, N. A. Shafto, M. Dixon, L. K. Tyler, Cam-CAN, and R. N. Henson. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. *Neuroimage*, 144:262–269, 2016.

[157] D. Freedman and P. Diaconis. On the Histogram as a Density Estimator: L_2 Theory. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 57:453–476, 1981.

[158] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes. The Art of Scientific Computing*. Cambridge University Press, Cambridge, UK, 3rd edition, 2007.

[159] See http://fcon_1000.projects.nitrc.org/indi/retro/MPI_LEMON.html for the LEMON public database.

[160] A. Babayan, M. Erbey, and et al. A mind-brain-body dataset of MRI, EEG, cognition, emotion, and peripheral physiology in young and old adults. *Sci. Data*, 6:180308, 2019.

[161] See http://storico.meteotrentino.it/web.htm for public climatological data., accessed April 2021.

[162] L. Paninski. Estimation of Entropy and Mutual Information. *Neural Computation*, 15:1191–1253, 2003.

List of publications

- **Generation of surrogate event sequences via joint distribution of successive inter-event intervals**, L. Ricci, M. Castelluzzo, L. Minati and A. Perinelli, Chaos 29 (2019), 121102, doi:10.1063/1.5138250

  The study of many dynamical systems relies on the analysis of experimentally-recorded sequences of events for which information is encoded in the sequence of interevent intervals. A correct interpretation of the results of the application of analytical techniques to these sequences requires the assessment of statistical significance. In most cases, the corresponding null-hypothesis distribution is unknown, thus forbidding an evaluation of the significance. An alternative solution, which is efficient in the case of continuous signals, is provided by the generation of surrogate data that share statistical and spectral properties with the original dataset. However, in the case of event sequences, the available algorithms for the generation of surrogate data can become cumbersome and computationally demanding. In this work, we present a new method for the generation of surrogate event sequences that relies on the joint distribution of successive interevent intervals. Our method, which was tested on both synthetic and experimental sequences, performs equally well or even better than conventional methods in terms of interevent interval distribution and autocorrelation while abating the computational time by at least one order of magnitude.

- **SpiSeMe: A multi-language package for spike train surrogate generation**, A. Perinelli, M. Castelluzzo, L. Minati and L. Ricci, Chaos 30 (2020), 073120, doi: 10.1063/5.0011328

  Many studies in nonlinear science heavily rely on surrogate-based hypothesis testing to provide significance estimations of analysis results. Among the complex data produced

by nonlinear systems, spike trains are a class of sequences requiring algorithms for surrogate generation that are typically more sophisticated and computationally demanding than methods developed for continuous signals. Although algorithms to specifically generate surrogate spike trains exist, the availability of open-source, portable implementations is still incomplete. In this paper, we introduce the SpiSeMe (Spike Sequence Mime) software package that implements four algorithms for the generation of surrogate data out of spike trains and more generally out of any sequence of discrete events. The purpose of the package is to provide a unified and portable toolbox to carry out surrogate generation on point-process data. Code is provided in three languages, namely, C++, Matlab, and Python, thus allowing straightforward integration of package functions into most analysis pipelines.

- **MiRNA-QC-and-Diagnosis: An R package for diagnosis based on MiRNA expression**, M. Castelluzzo, A. Perinelli, S. Detassis, M. A. Denti and L. Ricci, SoftwareX 12 (2020), 100569, doi:10.1016/j.softx.2020.100569

  The possibility of using microRNA (miRNA) levels as diagnostic and prognostic tools to detect different pathologies requires the implementation of reliable classifiers, whose training and use call for quality control of data corresponding to miRNA expression. In this work we present the MiRNA-QC-and-Diagnosis package. The package provides a set of functions for the R environment that implement the required quality control steps and thereupon allow to train, use and optimize a Bayesian classifier for diagnosis based on the measured miRNA expressions. The package thus makes up a complete and dedicated analytical toolbox for miRNA-based diagnosis.

- **Dependence of Connectivity on the Logarithm of Geometric Distance in Brain Networks**, M. Castelluzzo, A. Perinelli, D. Tabarelli and L. Ricci, Frontiers in Physiology 11 (2021), 611125, doi:10.3389/fphys.2020.611125

  Physical connections between nodes in a complex network are constrained by limiting factors, such as the cost of establishing links and maintaining them, which can hinder network capability in terms of signal propagation speed and processing power. Trade-off mechanisms between cost constraints and performance requirements are reflected in the topology of a network and, ultimately, on the dependence of connectivity

on geometric distance. This issue, though rarely addressed, is crucial in neuroscience, where physical links between brain regions are associated with a metabolic cost. In this work we investigate brain connectivity—estimated by means of a recently developed method that evaluates time scales of cross-correlation observability—and its dependence on geometric distance by analyzing resting state magnetoencephalographic recordings collected from a large set of healthy subjects. We identify three regimes of distance each showing a specific behavior of connectivity. This identification makes up a new tool to study the mechanisms underlying network formation and sustainment, with possible applications to the investigation of neuroscientific issues, such as aging and neurodegenerative diseases.

- **A Novel Hybrid Microdosimeter for Radiation Field Characterization Based on the Tissue Equivalent Proportional Counter Detector and Low Gain Avalanche Detectors Tracker: A Feasibility Study**, M. Missiaggia, E. Pierobon, M. Castelluzzo, A. Perinelli, F. Cordoni, M. Centis Vignali, G. Borghi, E. V. Bellinzona„ E. Scifoni, F. Tommasino, V. Monaco, L. Ricci, M. Boscardin, and C. La Tessa, Frontiers in Physics 8 (2021), 472, doi:10.3389/fphy.2020.578444

  In microdosimetry, lineal energies y are calculated from energy depositions ϵ inside the microdosimeter divided by the mean chord length, whose value is based on geometrical assumptions on both the detector and the radiation field. This work presents an innovative two-stages hybrid detector (HDM: hybrid detector for microdosimetry) composed by a tissue equivalent proportional counter and a silicon tracker made of 4 low gain avalanche diode. This design provides a direct measurement of energy deposition in tissue as well as particles tracking with a submillimeter lateral spatial resolution. The data collected by the detector allow to obtain the real track length traversed by each particle in the tissue equivalent proportional counter and thus estimates microdosimetry spectra without the mean chord length approximation. Using Geant4 toolkit, we investigated HDM performances in terms of detection and tracking efficiencies when placed in water and exposed to protons and carbon ions in the therapeutic energy range. The results indicate that the mean chord length approximation underestimate

particles with short track, which often are characterized by a high energy deposition and thus can be biologically relevant. Tracking efficiency depends on the low gain avalanche diode configurations: 34 strips sensors have a higher detection efficiency but lower spatial resolution than 71 strips sensors. Further studies will be performed both with Geant4 and experimentally to optimize the detector design on the bases of the radiation field of interest.The main purpose of HDM is to improve the assessment of the radiation biological effectiveness via microdosimetric measurements, exploiting a new definition of the lineal energy (y_T), defined as the energy deposition ϵ inside the microdosimeter divided by the real track length of the particle.

- **Relationship between mutual information and cross-correlation time scale of observability as measures of connectivity strength**, A. Perinelli, M. Castelluzzo, D. Tabarelli, V. Mazza and L. Ricci, Chaos 31 (2021), 073106, doi:10.1063/5.0053857

  The task of identifying and characterizing network structures out of experimentally observed time series is tackled by implementing different solutions, ranging from entropy-based techniques to the evaluation of the significance of observed correlation estimators. Among the metrics that belong to the first class, mutual information is of major importance due to the relative simplicity of implementation and its relying on the crucial concept of entropy. With regard to the second class, a method that allows us to assess the connectivity strength of a link in terms of a time scale of its observability via the significance estimate of measured cross correlation was recently shown to provide a reliable tool to study network structures. In this paper, we investigate the relationship between this last metric and mutual information by simultaneously assessing both metrics on large sets of data extracted from three experimental contexts, human brain magnetoencephalography, human brain electroencephalography, and surface wind measurements carried out on a small regional scale, as well as on simulated coupled, auto-regressive processes. We show that the relationship is well described by a power law and provide a theoretical explanation based on a simple noise and signal model. Besides further upholding the reliability of cross-correlation time scale of observability, the results show that the combined use of this metric and

mutual information can be used as a valuable tool to identify and characterize connectivity links in a wide range of experimental contexts.

- **Estimating the variance of Shannon entropy**, L. Ricci, A. Perinelli and M. Castelluzzo, Phys. Rev. E. 104 (2021), 024220, doi:10.1103/PhysRevE.104.024220

  The statistical analysis of data stemming from dynamical systems, including, but not limited to, time series, routinely relies on the estimation of information theoretical quantities, most notably Shannon entropy. To this purpose, possibly the most widespread tool is provided by the so-called plug-in estimator, whose statistical properties in terms of bias and variance were investigated since the first decade after the publication of Shannon's seminal works. In the case of an underlying multinomial distribution, while the bias can be evaluated by knowing support and data set size, variance is far more elusive. The aim of the present work is to investigate, in the multinomial case, the statistical properties of an estimator of a parameter that describes the variance of the plug-in estimator of Shannon entropy. We then exactly determine the probability distributions that maximize that parameter. The results presented here allow one to set upper limits to the uncertainty of entropy assessments under the hypothesis of memoryless underlying stochastic processes.

- **Experimental evidence of chaos generated by a Minimal Universal Oscillator Model**, L. Ricci, A. Perinelli, M. Castelluzzo, S. Euzzor, R. Meucci, Int. J. Bifurcation and Chaos 31 (2021), 2150205, doi:10.1142/S0218127421502059

  Detection of chaos in experimental data is a crucial issue in nonlinear science. Historically, one of the first evidences of a chaotic behavior in experimental recordings came from laser physics. In a recent work, a Minimal Universal Model of chaos was developed by revisiting the model of laser with feedback, and a first electronic implementation was discussed. Here, we propose an upgraded electronic implementation of the Minimal Universal Model, which allows for a precise and reproducible analysis of the model's parameters space. As a marker of a possible chaotic behavior the variability of the spiking activity that characterizes one of the system's coordinates was used. Relying on a numerical characterization of the relationship between spiking

activity and maximum Lyapunov exponent at different parameter combinations, several potentially chaotic settings were selected. The analysis via divergence exponent method of experimental time series acquired by using those settings confirmed a robust chaotic behavior and provided values of the maximum Lyapunov exponent that are in very good agreement with the theoretical predictions. The results of this work further uphold the reliability of the Minimal Universal Model. In addition, the upgraded electronic implementation provides an easily controllable setup that allows for further developments aiming at coupling multiple chaotic systems and investigating synchronization processes.

- **Experimental Implementation of a Laser Model with Cavity Loss Modulation**, M. Castelluzzo, M. Cescato, L. Ricci, R. Meucci, A. Perinelli, 2022 IEEE Workshop on Complexity in Engineering (COMPENG) (2022), COMPENG2022

  Experimental evidence of chaos and generalized multistability in a laser model with cavity loss modulation was first found in 1982, and recently revisited in a theoretical framework. Here we propose and investigate an electronic implementation of such laser model. Although a chaotic behavior has not been observed yet, the analysis provides insights in the difficulties of an experimental implementation of the model. These difficulties appear to be mainly due to the laser intensity, indeed a non-negative quantity, being simulated by a voltage, which can take on both positive and negative values.

- **Identification of miRNAs regulating MAPT expression and their analysis in plasma of patients with dementia**, P. Piscopo, M. Grasso, V. Manzini, A. Zeni, M. Castelluzzo, F. Fontana, G. Talarico, A. E. Castellano, R. Rivabene, A. Crestini, G. Bruno, L. Ricci, M. A. Denti, Frontiers in Molecular Neuroscience 16 (2023), doi:10.3389/fnmol.2023.1127163

  Background: Dementia is one of the most common diseases in elderly people and hundreds of thousand new cases per year of Alzheimer's disease (AD) are estimated. While the recent decade has seen significant advances in the development of novel biomarkers to identify dementias at their early stage, a great effort has been recently made to identify biomarkers able to improve differential diagnosis. However, only

few potential candidates, mainly detectable in cerebrospinal fluid (CSF), have been described so far. Methods: We searched for miRNAs regulating MAPT translation. We employed a capture technology able to find the miRNAs directly bound to the MAPT transcript in cell lines. Afterwards, we evaluated the levels of these miRNAs in plasma samples from FTD ($n = 42$) and AD patients ($n = 33$) and relative healthy controls (HCs) ($n = 42$) by using qRT-PCR. Results: Firstly, we found all miRNAs that interact with the MAPT transcript. Ten miRNAs have been selected to verify their effect on Tau levels increasing or reducing miRNA levels by using cell transfections with plasmids expressing the miRNAs genes or LNA antagomiRs. Following the results obtained, miR-92a-3p, miR-320a and miR-320b were selected to analyse their levels in plasma samples of patients with FTD and AD respect to HCs. The analysis showed that the miR-92a-1-3p was under-expressed in both AD and FTD compared to HCs. Moreover, miR-320a was upregulated in FTD vs. AD patients, particularly in men when we stratified by sex. Respect to HC, the only difference is showed in men with AD who have reduced levels of this miRNA. Instead, miR-320b is up-regulated in both dementias, but only patients with FTD maintain this trend in both genders. Conclusions: Our results seem to identify miR-92a-3p and miR-320a as possible good biomarkers to discriminate AD from HC, while miR-320b to discriminate FTD from HC, particularly in males. Combining three miRNAs improves the accuracy only in females, particularly for differential diagnosis (FTD vs. AD) and to distinguish FTD from HC.

Appendix A

Appendices

Statistical properties of the plug-in estimator of the Shannon entropy variance parameter

Let's write out the plug-in estimator of the variance parameter Λ_0, evaluated out of the rate histogram $\{\widehat{p_i}\}$ as

$$\widehat{\Lambda_0}\left(\widehat{p_1}, \ldots \widehat{p_M}\right) = \sum_{i=1}^{M} \left(\widehat{p_i} \ln^2 \widehat{p_i}\right) - \widehat{H^2} \,. \tag{1}$$

The quantity $\widehat{\Lambda_0}$ is a sample statistic, therefore it is necessary to assess its reliability in terms of bias and variance. This can be done by determining the asymptotic behavior of $\widehat{\Lambda_0}$ as $N \to \infty$. We define, for each i, $\zeta_i \equiv \left(\widehat{p_i} - s_i\right) \sqrt{N}$, then the rates can be written as:

$$\widehat{p_i} = s_i + \frac{\zeta_i}{\sqrt{N}} \,.$$

It follows:

$$\begin{aligned} \widehat{\Lambda_0} &= \sum_{i=1}^{M} \left[\left(s_i + \frac{\zeta_i}{\sqrt{N}} \right) \ln^2 \left(s_i + \frac{\zeta_i}{\sqrt{N}} \right) \right] \\ &- \left[\sum_{i=1}^{M} \left(s_i + \frac{\zeta_i}{\sqrt{N}} \right) \ln \left(s_i + \frac{\zeta_i}{\sqrt{N}} \right) \right]^2 \,. \end{aligned}$$

The previous expression can be expanded in terms of order $N^{-n/2}$, with $n \in \mathbb{N}$, as:

$$\begin{aligned}\widehat{\Lambda_0} &= \Lambda_0 + \frac{1}{\sqrt{N}} \sum_{i=1}^{M} \zeta_i \left[\ell_i^2 + 2(1+H)\ell_i\right] \\ &+ \frac{1}{N}\left[\sum_{i=1}^{M}\left(\frac{1+H+\ell_i}{s_i}\zeta_i^2\right) - \sum_{i=1}^{M}\sum_{j=1}^{M}\left(\zeta_i\zeta_j\ell_i\ell_j\right)\right] \\ &+ O\left(\frac{1}{N^{3/2}}\right),\end{aligned}$$

where we defined $\ell_i \equiv \ln s_i$ and the identity $\sum_{i=1}^{M} \zeta_i = 0$, following from the constraint $\sum_{i=1}^{M} \widehat{p_i} = 1$.

We write as $G_{\widehat{\delta\Lambda_0}}(t)$ the moment generating function of the residual $\widehat{\delta\Lambda_0} \equiv \widehat{\Lambda_0} - \Lambda_0$, where t is a real variable defined in a neighborhood of the origin. By the definition of moment generating function, it holds that:

$$G_{\widehat{\delta\Lambda_0}}(t) = \mathbf{E}\left(e^{t\,\widehat{\delta\Lambda_0}}\right).$$

We can then expand the exponential $e^{t\,\widehat{\delta\Lambda_0}}$ as:

$$\begin{aligned}e^{t\,\widehat{\delta\Lambda_0}} &= 1 + \frac{t}{\sqrt{N}} \sum_{i=1}^{M} \zeta_i \left[\ell_i^2 + 2(1+H)\ell_i\right] \\ &+ \frac{t}{N}\left[\sum_{i=1}^{M}\left(\frac{1+H+\ell_i}{s_i}\zeta_i^2\right) - \sum_{i=1}^{M}\sum_{j=1}^{M}\left(\zeta_i\zeta_j\ell_i\ell_j\right)\right] \\ &+ \frac{t^2}{2N}\sum_{i=1}^{M}\sum_{j=1}^{M}\zeta_i\zeta_j\left[\ell_i^2 + 2(1+H)\ell_i\right]\left[\ell_j^2 + 2(1+H)\ell_j\right] \\ &+ O\left(\frac{1}{N^{3/2}}\right).\end{aligned}$$

Knowing that the starting vector of the system's evolution is chosen randomly out of the distribution $\{s_i\}$, the expected value of each ζ_i is given by $\mathbf{E}(\zeta_i) = O(1/N)$ [99], therefore the term proportional to $t/\sqrt{N}$ is absorbed within the term $O\left(N^{-3/2}\right)$. In addition [99], one has $\mathbf{E}\left(\zeta_i\zeta_j\right) = \chi_{i,j} + O\left(N^{-1/2}\right)$, where $\chi_{i,j} = \delta_{i,j}s_i - s_i s_j$ is the covariance matrix of rates being distributed according to a multinomial distribution,

with $\delta_{i,j}$ being the Kronecker delta. It follows:

$$\begin{aligned} G_{\delta\widehat{\Lambda_0}}(t) &= \frac{t}{N}\left[\sum_{i=1}^{M}\left(\frac{1+H+\ell_i}{s_i}\chi_{i,i}\right)-\sum_{i=1}^{M}\sum_{j=1}^{M}\left(\chi_{i,j}\ell_i\ell_j\right)\right] \\ &+ \frac{t^2}{2N}\sum_{i=1}^{M}\sum_{j=1}^{M}\chi_{i,j}\left[\ell_i^2+2(1+H)\ell_i\right]\left[\ell_j^2+2(1+H)\ell_j\right] \\ &+ O\left(\frac{1}{N^{3/2}}\right). \end{aligned} \tag{2}$$

We now define the coefficients of t/N and $t^2/(2N)$ in Eq. (2) as γ and Γ, respectively:

$$\begin{aligned} \gamma &\equiv \sum_{i=1}^{M}\left(\frac{1+H+\ell_i}{s_i}\chi_{i,i}\right)-\sum_{i=1}^{M}\sum_{j=1}^{M}\left(\chi_{i,j}\ell_i\ell_j\right), \\ \Gamma &\equiv \sum_{i=1}^{M}\sum_{j=1}^{M}\chi_{i,j}\left[\ell_i^2+2(1+H)\ell_i\right]\left[\ell_j^2+2(1+H)\ell_j\right]. \end{aligned}$$

We can then rewrite Equation (2) as:

$$G_{\delta\widehat{\Lambda_0}}(t) = \frac{t}{N}\gamma + \frac{t^2}{2N}\Gamma + O\left(\frac{1}{N^{3/2}}\right). \tag{3}$$

It is worth noting that, being χ positive semi-definite, from the definition of Γ we get $\Gamma \geqslant 0$.

From the properties of the moment generating function, we get that the coefficient of t in this last equation corresponds to the expected value of $\delta\widehat{\Lambda_0}$, i.e. of $\widehat{\Lambda_0} - \Lambda_0$:

$$\mu_{\delta\widehat{\Lambda_0}} = \mu_{\widehat{\Lambda_0}} - \Lambda_0 = \frac{\gamma}{N} + O\left(\frac{1}{N^{3/2}}\right). \tag{4}$$

Also, because the square of this last term is of order N^{-2}, the coefficient of $t^2/2$ in Eq. (3) directly provides the variance of $\delta\widehat{\Lambda_0}$ and thus of $\widehat{\Lambda_0}$:

$$\sigma^2_{\widehat{\Lambda_0}} = \sigma^2_{\delta\widehat{\Lambda_0}} = \frac{\Gamma}{N} + O\left(\frac{1}{N^{3/2}}\right). \tag{5}$$

Equations (4), (5) provide the bias and the variance of the plug-in estimator $\widehat{\Lambda_0}$ of the variance parameter Λ_0, respectively.

Besides these results, Eq. (3) can be further exploited: starting from it, the moment generating function of the random variable $w \equiv \left(\delta\widehat{\Lambda_0} - \frac{\gamma}{N}\right)\sqrt{N}$ is given by:

$$G_w(t) = \exp\left(\frac{t^2}{2}\Gamma\right) + O\left(\frac{1}{N^{1/2}}\right),$$

where t is a real variable defined in a neighborhood of the origin. This expression shows that, provided that Γ is non-vanishing, w is asymptotically normally distributed with zero mean and variance Γ. It is therefore possible to formulate a central limit theorem for the plug-in estimator $\widehat{\Lambda_0}$ of the variance parameter Λ_0:

$$\widehat{\Lambda_0} \sim \mathcal{N}\left(\Lambda_0 + \frac{\gamma}{N}, \frac{\Gamma}{N}\right) \qquad \text{as } N \to \infty\,.$$

We then found that the plug-in estimator $\widehat{\Lambda_0}$ from Eq. (1) is asymptotically normal and a consistent estimator of the variance parameter Λ_0, given that its population mean and variance both depend on N^{-1}.

A theoretical signal and noise model for MI and CTO relationship

The results discussed in Chapter 5 prove that the inverse time scale of observability W^{-1} and the MI $\widehat{I}$ are proportional to each other and their relationship follows a power law. Remarkably, this relationship does not seem to depend on the system of interest, at least for the three cases discussed (and an additional case of synthetic auto-regressive data, discussed in Ref. [142]. In this section, we will try to provide a theoretic model able to explain the empirical results in Chapter 5.

A possible explanation of the relationship between W^{-1} and $\widehat{I}$ relies on the interpretation of each one of these two quantities in terms of effective number of shared events occurring in the two sequences. We consider a simple noise-and-signal model where two sequences $\{x_t\}$, $\{y_t\}$ are given by two independent white noise sources and share a single delta-like signal, as follows:

$$\begin{aligned} x_t &= u_t + \mathcal{S}\delta_{t,t_0}\,, \\ y_t &= v_t + \mathcal{S}\delta_{t,t_0}\,, \end{aligned} \tag{6}$$

where u_t, v_t are independent and identically distributed standard normal random variables, with equal power distribution, and δ_{t,t_0} is the Kronecker delta. The positive real number $\mathcal{S}$ is the amplitude of the shared delta-like signal. We define as N the length of the two sequences. Although the model relies on delta-like signals, its validity can be extended to more realistic shared events, such as in the case of oscillatory activity typical of brain recordings, by expressing the finite duration of these events as a sum of delta-like signals, as discussed below.

Modeling of $\widehat{I}$

We first consider the expected MI $I_0(X,Y)$ in the case of no shared event. Assuming the underlying variables to be normally distributed, $I_0(X,Y)$ can be directly estimated by using the expression [154] $I = -\frac{1}{2}\ln\left(1-r^2\right)$, with r being the coefficient of correlation between the two variables that, for sufficiently large values of N, is given by $(N-1)^{-1}$. We thus get:

$$\widehat{I_0}(X,Y) \approx \frac{1}{2N}\,.$$

To study now the effects of delta-like events, it is convenient to consider the expression of the plug-in estimator of Eq. (5.2), assuming that a suitable partition of the support was carried out. Then the definition of MI from Eq. (5.2) can be rewritten as:

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

where $\widehat{H}(X,Y)$ is the sample joint Shannon entropy, and $\widehat{H}(X)$, $\widehat{H}(Y)$ are the sample Shannon entropies of the marginal distributions $P_x(i)$, $P_y(j)$. These entropies can then be written as:

$$\widehat{H}(X) = -\sum_{i=1}^{M_x} P_x(i)\ln\left[P_x(i)\right]\,,$$

$$\widehat{H}(Y) = -\sum_{j=1}^{M_y} P_y(j)\ln\left[P_y(j)\right]\,,$$

$$\widehat{H}(X,Y) = -\sum_{i=1}^{M_x}\sum_{j=1}^{M_y} P_{xy}(i,j)\ln\left[P_{xy}(i,j)\right]\,.$$

We will consider the datasets to be sufficiently large, so that bias corrections, like Miller-Madow [95, 94, 96, 162] can be neglected. Furthermore, we can neglect other bias terms assuming that the access to each bin is essentially memoryless, i.e. the observed histograms are essentially realizations of multinomial distributions [99, 100].

We suppose that, in the absence of the delta-like signal, the amplitude pair (u_{t_0}, v_{t_0}) falls in the joint histogram bin (i_1, j_1). If $\mathcal{S}$ is so small that the perturbed amplitude pair $(u_{t_0}+\mathcal{S}, v_{t_0}+\mathcal{S})$ still belongs to the same bin, nothing changes. However, for sufficiently large $\mathcal{S}$, the occupation of the former bin (i_1, j_1) is decreased

by $1/N$, also considering the normalization of each bin occupation by N. Instead, considering the amplitude pair $(u_{t_0} + \mathcal{S}, v_{t_0} + \mathcal{S})$ to fall in bin (i_2, j_2), the normalized occupation of this latter bin is increased by $1/N$. We call $N_1 = NP_{xy}(i_1, j_1)$ the occupation of the bin (i_1, j_1) and $N_2 = NP_{xy}(i_2, j_2)$ the occupation of the bin (i_2, j_2), both evaluated in the case of no delta-like event.

We now consider the contribution to MI from the joint entropy $-\widehat{H}(X, Y)$ when a transition from bin (i_1, j_1) to bin (i_2, j_2) occurs due to an event. The change in MI can be written as:

$$-\delta\widehat{H}(X, Y) = \left[\frac{N_1 - 1}{N} \ln\left(\frac{N_1 - 1}{N}\right) + \frac{N_2 + 1}{N} \ln\left(\frac{N_2 + 1}{N}\right)\right] - \left[\frac{N_1}{N} \ln\left(\frac{N_1}{N}\right) + \frac{N_2}{N} \ln\left(\frac{N_2}{N}\right)\right].$$

We can now consider two regimes, based on the values of N_1 and N_2.

In the first one we consider N_1 and N_2 to be sufficiently larger than 1. Then the Taylor expansion of the last expression provides:

$$-\delta\widehat{H}(X, Y) \approx \frac{1}{N} \ln\left(\frac{N_2}{N_1}\right).$$

Similarly, for the two marginal entropies we get:

$$\delta\widehat{H}(X) \approx \frac{1}{N} \ln\left(\frac{N_{x;1}}{N_{x;2}}\right), \qquad \delta\widehat{H}(Y) \approx \frac{1}{N} \ln\left(\frac{N_{y;1}}{N_{y;2}}\right),$$

where $N_{x;1} = NP_x(i_1)$, $N_{x;2} = NP_x(i_2)$ and $N_{y;1} = NP_y(j_1)$, $N_{y;2} = NP_y(j_2)$. Summing the three contributions, we get the resulting MI of a single delta-like event shared by two noisy sequences:

$$\delta\widehat{I}(X, Y) \approx \frac{1}{N} \left[\ln\left(\frac{N_{x;1}N_{y;1}}{N_1}\right) - \ln\left(\frac{N_{x;2}N_{y;2}}{N_2}\right)\right]. \tag{7}$$

Assuming that the delta-like event will lead, on average, to transitions to less occupied bins, the term $\delta\widehat{I}(X, Y)$ is on average positive. By also assuming the bins to be roughly uniformly populated, we can write $N_1 \approx N/(M_x M_y)$, $N_{x;1} \approx N/(M_x)$, $N_{y;1} \approx N/(M_y)$. Therefore we get:

$$\ln\left(\frac{N_{x;1}N_{y;1}}{N_1}\right) \approx \ln N.$$

Similarly, also the term $\ln\left(\frac{N_{x;2}N_{y;2}}{N_2}\right)$ is of order $\ln N$. Therefore, the total variation of MI due to the delta-like event as given by Eq. (7) is expected to be of order $\ln(N)/N$.

The second case instead corresponds to the signal amplitude $\mathcal{S}$ overcoming the noise level. Therefore, N_2 can be considered to be void without the occurrence of a delta-like event. Instead, we can still assume that N_1 is still sufficiently larger than 1. Then Equation (7) becomes:

$$\delta\widehat{I}(X,Y) \approx \frac{1}{N}\left[\ln\left(\frac{N_{x;1}N_{y;1}}{N_1}\right) - \ln(N)\right].$$

By the same reasoning applied to the first case, we conclude that also for the second case the variation of MI due to the delta-like event is of order $\ln(N)/N$.

By writing $\widehat{I} = \widehat{I_0} + \delta\widehat{I}$, and considering sufficiently large values of N, the expected MI between two sequences sharing a single, delta-like event, with signal amplitude overcoming the noise level, can be written as:

$$\widehat{I}(X,Y) \approx a\frac{\ln(N)}{N}, \tag{8}$$

where a is a constant [142].

We will now address more realistic events that cover a finite time τ. These can be written as the sum of time-correlated delta-like events. The time correlation, nonetheless, does not play an important role. What matters is the fact that the amplitudes of the single delta-like components fall within different bins of the P_{xy} joint histogram, so that they can be assumed to be amplitude-uncorrelated. Then, since an event of duration τ corresponds to the sum of τ/T delta-like events, with T the sampling period of the sequence, also the expected MI for such event will be the right-hand term of Eq. (8) multiplied times a factor τ/T. Finally, we consider a number n_0 of independent and effective shared events occurring in the time span NT of the sequences, where by effective we mean that the signal amplitude overcomes the noise level. Than MI is given by:

$$\widehat{I}(X,Y) \approx a\frac{n_0\tau\,\ln(N)}{NT}. \tag{9}$$

Modeling of W^{-1}

We consider a delta-like event having a sufficiently large signal-to-noise ratio to generate a "funnel-like structure" [140] in the cross-correlation and p-value diagrams, as shown in Fig. 1. Such structures are not instead generated by spurious correlations, as we also mentioned in Section 5.1.

Given a window width w, the number of available running windows of that width that contain the shared peak is w/w_0, where w_0 is the duration of the base window, as explained in Sec. 5.1.

We can then write the efficiency $\eta(w) = w/Kw_0$, where we define as K the number of running windows having the same w. The efficiency $\eta(w)$ reaches the threshold η^* at the window width $W = Kw_0\eta^*$, namely the time scale of observability. In the case of two shared delta-like events with non-overlapping funnel structures, as we show in Fig. 1 (c), the threshold η^* is reached two times faster by increasing the window width, thus halving the time scale of observability W. Therefore, for a number n of independent events generating non-overlapping funnel-like structures, we can write $nW \approx Kw_o\eta^*$.

The number of events n is proportional to n_0, i.e. the number of independent and effective shared events of Eq. (9). Furthermore, n is dependent on W: as shown in Figs. 1 (a),(b), we can observe that, the lower w, the larger the number of significant cross-correlation values that originate from funnel-like structures and eventually fade out due to a lower signal-to-noise ratio [140].

We can assume a power-law distribution for $n = n(W)$, namely $n = n_0 \, (W/W_0)^{-\gamma}$, where W_0 is a time constant that corresponds to the window width at which $n = n_0$. We can thus write:

$$W^{1-\gamma} \approx \frac{Kw_o\eta^*}{n_0 W_0^{\gamma}} \, ,$$

where the exponent γ has to be positive and less than one to ensure a decrease of n in the case of an increasing W.

From the properties of the CTO algorithm [140], we also have that $Kw_0 = NT - w_{\max}$, where $w_{\max}$ is the maximum window width analyzed. Typically, $w_{\max}$ is one order of magnitude smaller than NT. Therefore the ratio $Kw_0/(NT)$ is ap-

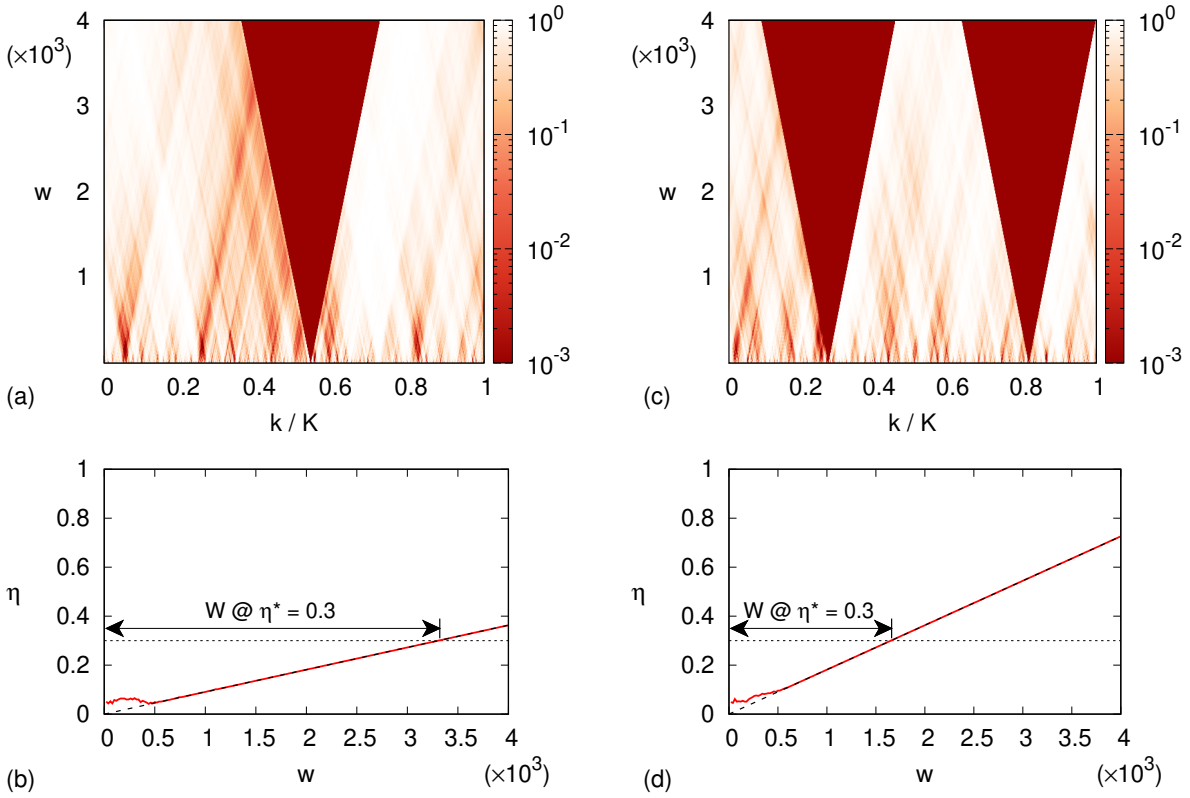

Figure 1: (a) p-value diagram built out of a pair of Gaussian white noise sequences sharing a single delta-like event. The window width w is reported in units of the sampling period. (b) The red line corresponds to the efficiency function $\eta(w)$ built out of the diagram in (a). For sufficiently large w values, the function $\eta(w)$ grows linearly with w as $\eta(w) = w/K$ (black, dashed line). The window width W at which $\eta(w)$ overcomes the efficiency threshold $\eta^* = 0.3$ (dotted line) is also highlighted. (c) p value diagram built out of a pair of Gaussian white noise sequences sharing two delta-like events. (d) Efficiency function $\eta(w)$ built out of the diagram in (c). The slope of the linear part is twice the slope of the linear section of efficiency vs. window width in panel (b).

Reproduced from Chaos 31, 073106 (2021), with the permission of AIP Publishing.

proximately one, which allows to rewrite the previous equation as:

$$W^{1-\gamma} \approx \eta^* \frac{NT}{n_0 W_0^\gamma} \,. \tag{10}$$

By combining this latter equation with Eq. (9), we can finally write:

$$\widehat{IW}^{1-\gamma} \approx \eta^* \cdot c \ \ln(N)\,, \tag{11}$$

where $c = a\tau/W_0^{\gamma}$ is a constant.

This last expression is the same we reported at the beginning of Chapter 5. The values of c and γ can be found on experimental data by fitting Equation (5.1). In Ref. [142], the values found for the experimental data considered in Chapter 5 are reported.