AN ENTROPIC APPROACH TO THE ANALYSIS OF TIME SERIES

Nicola Scafetta, B.S., M.S.

Dissertation Prepared for the Degree of
DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS
December 2001

APPROVED:
Paolo Grigolini, Major Professor
Patti Hamilton, Committee Member
James A. Roberts, Committee Member
William D. Deering, Committee Member
Sam Matteson, Chair of the Department of Physics
C. Neal Tate, Dean of the Robert B. Toulouse School of Graduate Studies

Scafetta Nicola, An entropic approach to the analysis of time series. Doctor of Philosophy (Physics), December 2001; 170 pp.; 8 tables; 50 figures; references, 171 titles.

Statistical analysis of time series. With compelling arguments we show that the Diffusion Entropy Analysis (DEA) is the only method of the literature of the Science of Complexity that correctly determines the scaling hidden within a time series reflecting a Complex Process.

The time series is thought of as a source of fluctuations, and the DEA is based on the Shannon entropy of the diffusion process generated by these fluctuations. All traditional methods of scaling analysis, instead, are based on the variance of this diffusion process. The variance methods detect the real scaling only if the Gaussian assumption holds true. We call H the scaling exponent detected by the variance methods and δ the real scaling exponent. If the time series is characterized by Fractional Brownian Motion, we have $H = \delta$ and the scaling can be safely determined, in this case, by using the variance methods. If, on the contrary, the time series is characterized, for example, by Lévy statistics, $H \neq \delta$ and the variance methods cannot be used to detect the true scaling. Lévy walk yields the relation $\delta = 1/(3 - 2H)$. In the case of Lévy flights, the variance diverges and the exponent H cannot be determined, whereas the scaling δ exists and can be established by using the DEA. Therefore, only the joint use of two different scaling analysis methods, the variance scaling analysis and the DEA, can assess the real nature, Gauss or Lévy or something else, of a time series. Moreover, the DEA determines the information content, under the form of Shannon entropy, or of any other convenient entropic indicator, at each time step of the process that, given a sufficiently large number of data, is expected to become diffusion with scaling. This makes it possible to study the regime of transition from dynamics to thermodynamics, non-stationary regimes, and the saturation regime as well.

First of all, the efficiency of the DEA is proved with theoretical arguments and with

numerical work on artificial sequences. Then we apply the DEA to three different sets of real data, Genome sequences, hard x-ray solar flare waiting times and sequences of sociological interest. In all these cases the DEA makes new properties, overlooked by the standard method of analysis, emerge.

ACKNOWLEDGEMENTS

I thank all my professors and my parents.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The goal of this dissertation is to introduce a new method of statistical analysis of time series. We show with compelling arguments that DIFFUSION ENTROPY ANALYSIS is the only method available in the literature of the Science of Complexity determining the correct scaling of a time series.

For the reader to appreciate the importance of this result, we have to outline very quickly what the Science of Complexity is all about. After the discovery of quantum mechanics and of the theories of relativity, humanity seemed to have nothing else to discover about the fundamental laws of nature. This was not true. The world around us, from a mountain to a cloud, from a coastline to a tree, shows a complexity that hardly can be described by applying the basic laws of Nature. The Mandelbrot's masterpiece "The Fractal Geometry of Nature" [1] proved that many natural patterns, characterized by an irregular and fragmented behavior, may be studied scientifically. Since Euclid, the scientists have left aside this research because it was thought that to investigate what looked as "formless" or "amorphous" was not interesting or, even, was impossible. Maldelbrot showed that a revolutionary geometry, the "Fractal Geometry," is able to evidence and analyze the different levels of complexity of a phenomenon. Fractal Geometry is able to make intelligible what looks "formless" and "amorphous."

Today, fractals, chaos and power laws have become common in modern science [2]. Self-similarity is the unifying concept underlying this new frontier of physics. Self-similarity means invariance against changes in scale or size. Self-similarity is like a set of Chinese boxes or Russian dolls - a doll hides a similar smaller one inside its "body" and this is repeated for many generations of dolls. In other words, we say that an object is self-similar, or invariant with scaling, if magnifying some portion of it reproduces itself. Innumerable natural phenomena, from the distribution of atoms in matter to that of the galaxies in the universe or that of stock market trends, show self-similarity attributes. "Chaos," a word used for describing a state of

utter confusion and disorder, may be related to self-similarity and its inherent lack of "smoothness." Of course, it is impossible to find a natural phenomenon characterized by a mathematical self-similar behavior, which implies invariance in scaling ad infinitum. Galileo Galilei, who discovered the scaling law for falling objects, noted that some laws of nature are not unchanged under changes of scale. For example, he noted that it is not true that if an animal is two times longer, wider and taller than another animal, its bones scale with the same factor 2. Its bones must scale with a factor that is higher than 2 because an eight times heavier animal cannot be supported by bones with a four times larger cross section: an anomalous scaling may be involved. This means that in Nature, self-similarity reflects always a range of validity. It is true that the shape of a cloud is self-similar, but the invariance in scaling is obviously lost at the atomic scale. However, self-similarity is a fundamental concept for understanding innumerable natural phenomena.

From a mathematical viewpoint, a function $\Phi(r_1, r_2, \ldots)$ is scaling invariant, if it fulfills the following property:

$$\Phi(r_1, r_2, \ldots) = \gamma^a \ \Phi(\gamma^b r_1, \gamma^c r_2, \ldots) \ . \tag{1.1}$$

Eq. (1.1) means that if we scale all coordinates $\{r\}$ by particular factors, the resulting values of the function remain similar to the original ones; the only change is a change of scales. This dissertation focuses upon the scaling properties of a time series. By summing the terms of a time series we get a trajectory and the trajectory can be used to generate a diffusion process. There is scaling if, in the stationary condition, a diffusion process can be described by the following probability density function (pdf):

$$p(x,t) = \frac{1}{t^\delta} \ F(\frac{x}{t^\delta}), \tag{1.2}$$

where x denotes the diffusion variable and $p(x,t)$ is its pdf at time t. The coefficient δ is called the scaling exponent. We define the scaling of a time series as the scaling exponent of a diffusion process generated by that time series.

Diffusion Entropy Analysis detects the correct scaling exponent, δ, of a time series.

Let us see why. This method of analysis is based upon the evaluation of the Shannon entropy of the pdf of the diffusion process that reads

$$S(t) = -\int_{-\infty}^{+\infty} dx\ p(x,t)\ \ln\left[p(x,t)\right]. \tag{1.3}$$

If the scaling equation (1.2) is fulfilled, the time evolution of the entropy, S(t), is linear with respect to the logarithmic time, $\tau \equiv ln(t/t_0)$, which makes Eq. (2) read

$$S(\tau) = A + \delta\ \tau. \tag{1.4}$$

Eq. (1.4) states that the scaling exponent δ is determined by the asymptotic slope of the entropy $S(\tau)$. The term, t_0, is the unit diffusion "time." Because the exponent δ does not depend on the value of t_0, it will be conventionally put equal to 1 in the entire dissertation. The "time" t will be always measured in unit of t_0.

The Diffusion Entropy Analysis is the only method establishing δ correctly. It is so because Diffusion Entropy Analysis analyzes directly the pdf of the diffusion processes, without using the moments of the distribution. Instead, all the other methods used for detecting scaling –Variance Scaling Analysis, Hurst R/S Analysis, Detrended Fluctuation Analysis, Relative Dispersion Analysis, Spectral Analysis, Spectral Wavelet Analysis– are subtly based on the Gaussian assumption and, so, upon a variance that can be used to monitor scaling. The problem is that the scaling detected by the "variance" methods, usually called H in honor of Hurst [3], may not exist or may not coincide with the correct scaling, δ. If the time series is characterized by what Mandelbrot called Fractional Brownian Motion, we have $H = \delta$. Consequently, the scaling of this type of noise can be detected by using the variance methods. If, on the contrary, the time series is characterized, for example, by Lévy properties [4, 5], $H \neq \delta$ and the variance methods cannot be used to detect the true scaling. A diffusion process generated by Lévy walk is characterized by the relation

$$\delta = \frac{1}{3-2H}\ . \tag{1.5}$$

In the case of Lévy flights, the exponent H cannot be determined because the variance

diverges, whereas the scaling δ exists and can be determined by using the diffusion entropy analysis. The above conclusions suggest that to determine the real statistical properties of a time series it is not enough to study the scaling with only one type of analysis. Only the joint use of two scaling analysis methods, the variance scaling analysis and the diffusion entropy analysis, can determine the real nature, Gauss or Lévy or something else, of a time series. We have to determine H and δ. Then, if $H = \delta$ we can conclude that Fractional Brownian Noise may characterize the signal. If, instead, $H \neq \delta$ we have to look for a different type of noise. If we find that the relation (1.5) holds true, we can have good reasons to conclude that the noise is characterized by Lévy statistics. Moreover, Diffusion Entropy Analysis may be used for studying the transition from the dynamics to the thermodynamics of the dissusion process. This is the transition region before that the Central Limit Theorem or the Generalized Central Limit Theorem hold true and a thermodynamical scaling of the diffusion process may be defined.

The dissertation is organized in two parts followed by a conclusion. The first part addresses the theory of Diffusion Entropy Analysis and the numerical simulations that verify the theoretical results. The second part shows three different applications to real data. Each application addresses a particular way to work with the Diffusion Entropy Analysis for studying the scaling of a time series or statistical properties of the transition region.

Chapter 2 reviews the theory of a diffusion process. Continuous Random Walk theory is used to derive the main scaling properties of Fractional Brownian and Lévy diffusion. Different ways to realize a walk are studied. Chapter 3 reviews the most traditional analysis methods used in literature of the Science of Complexity for detecting the scaling of a time series. All these methods have in common the fact that they are based upon the Gaussian assumption, that is, upon the fact that the real scaling can be detected by studying the variance of the generated diffusion process.

The results illustrated from Chapters 4 to the end are original results of this dissertation and have produced several papers, some already published [6, 7], two submitted to Phys. Rev. Lett. and Phys. Rev. E [8, 9], and some others almost ready for submission. Many more publications are expected to emerge from the

results of this dissertation. Chapter 4 focuses on the meaning of the entropic analysis of a time series. Shannon entropy, Tsallis non-extensive entropy and Kolmogorov-Sinai entropy of a time series are reviewed. The theory of Diffusion Entropy Analysis is exposed. Chapter 5 is devoted to the verification of the theory of the previous chapters by using artificial sequences. The main characteristics of Diffusion Entropy Analysis are discussed and shown. The second part of the dissertation begins with Chapter 6 in which Diffusion Entropy Analysis is applied to different type of DNA sequences. The large number of data available for a DNA sequences makes this type of time series suitable for the analysis of asymptotic scaling properties. Chapter 6 proves that DNA sequences are characterized by long-range Lévy correlation. This may have important consequences upon the understanding of the complexity and the origin of the life. Chapter 7 is dedicated to the study of waiting times of hard x-ray solar flares. It is shown how a dynamical entropic fit can be realized and its superiority is compared to the static fit of the waiting time distribution. Moreover, a dynamical entropic fit is sensitive to the time evolution of the data. This result may be important for understanding the dynamics of the solar flare phenomenon and, in general, the turbulent behavior of the Sun and of other similar turbulent phenomena. In fact, a theoretical model for describing turbulence should not simulate only the waiting time distribution, but it should simulate the real temporal dynamics of the turbulence as well. Diffusion Entropy Analysis may be used to determine the waiting time distribution properties and the temporal dynamics of the turbulent phenomenon. The goodness of a theoretical model may be checked by an entropic dynamical fit. Chapter 8 is dedicated to thermodynamics of social processes. Diffusion Entropy Analysis is used for studying two different sets of data regarding the births of babies to married and unmarried teenagers. The Diffusion Entropy Analysis applied to the two groups allows a distinction between them based upon thermodynamic properties. The non-extensive Tsallis entropy is used to measure the thermodynamic difference between the two groups. The applications to study of the transition region and the non-stationary condition are shown.

The dissertation ends with a short conclusion and with a research project concerning further applications of this theory.

Part I

THEORY

CHAPTER 2

CONTINUOUS TIME RANDOM WALK

In this chapter we discuss some aspects of the Continuous Time Random Walk (CTRW) theory . We demostrate how to obtain the probability density $p(x,t)$ of a diffusion process according to different rules of walking. The chapter focuses upon those diffusion processes that are characterized by a probability density function (pdf) of the type

$$p(x,t) = \frac{1}{t^{\delta}} \cdot F\left(\frac{x}{t^{\delta}}\right) \tag{2.1}$$

that shows scaling proprieties. The coefficient δ is the pdf scaling coefficient. Moreover, we compare the pdf scaling coefficient δ with the variance scaling exponent H defined by

$$\Sigma^2(t) \sim t^{2H}, \tag{2.2}$$

where Σ^2 is the variance of the diffusion process. In literature, the exponent H is usually called *Hurst exponent* because of the well know *Hurst's scaled range analysis (R/S analysis)* [3]. We note that the Hurst's analysis does not coincide exactly with the second moment analysis. The exponents measured by the two techniques may be slightly different [10]. However, for the purpose of this dissertation, the slight distinction between the Hurst analysis and the variance scaling analysis is not important and we use the same symbol, H, to indicate the scaling exponent given by the two techniques without further distinctions.

If $< x(t) >= 0$, the variance, $\Sigma^2(t)$, will coincide with the mean squared displacement, and Eq. (2.2) will read

$$< x^2(t) >= \int_{-\infty}^{-\infty} x^2 \; p(x,t) \; dx \sim t^{2H} \; . \tag{2.3}$$

The pdf scaling coefficient δ and the second moment scaling exponent H coincide in all cases in which

$$\int_{-\infty}^{\infty} y^2 F(y) \; dy = const < \infty \; , \tag{2.4}$$

where *const* is a constant that does not depend on the diffusion time. This holds true, for example, in the normal and fractional Brownian diffusion. However, there are cases in which they do not coincide. Possible cases include diffusion processes characterized by Lévy flights where Eq. (2.1) still holds true. However, the variance is not finite and, therefore, the variance scaling exponent, H, cannot be defined. Another interesting situation is when, for example, a diffusion process is characterized by Lévy walks where $\delta = 1/(3-2H)$. In chapter 3, we review the common techniques used to measure the variance scaling exponent H. In chapter 4 we show that the Diffusion Entropy Analysis (DEA), the subject of this dissertation, is the first technique for detecting directly the pdf scaling coefficient δ. This allow us to distinguish anomalous Brownian diffusion from other types of diffusion like Lévy diffusion. In this chapter, we address the basic theory of normal and fractional Brownian motion and of anomalous Lévy diffusion (long flights and walks) [11].

2.1 Brownian and anomalous diffusion: historical remarks.

In 1785, the Dutch physician J. Ingenhousz observed a small flickering of coal dust particles on the surface of alcohol. In 1827, the Scottish botanist R. Brown [12] observed that, when suspended in water, small pollen grains appear to move in a very irregular way. At the beginning, Brown thought that this motion had an organic origin. However, further experiments made by using any type of fine particles –glass, minerals and even a fragment of the sphinx– showed that this motion could not be a manifestation of life. In 1855, by studying Fourier's works about the heat conduction equation, A. Fick discovered the diffusion equation [13]. Fick used the theory of the continuum formulation of fluid dynamics that had already been fully developed at his time. Fick's law states that if $\rho(\mathbf{r},t)$ is the concentration of a chemical substance in the position $\mathbf{r}$ at time t, a diffusion current $\mathbf{j}(\mathbf{r},t)$ exists such that

$$\mathbf{j}(\mathbf{r},t) = -D\nabla\rho(\mathbf{r},t), \tag{2.5}$$

where D is called the diffusion coefficient. If there is no chemical reaction, the current obeys to the following conservation equation

$$\frac{d}{dt}\int\limits_V \rho(\mathbf{r},t)\ d^3\mathbf{r} = -\int\limits_S \mathbf{j}(\mathbf{r},t)\cdot d\mathbf{S} = -\int\limits_V \nabla\cdot\mathbf{j}(\mathbf{r},t)\ d^3\mathbf{r}, \tag{2.6}$$

where V and $\mathbf{S}$ are the volume and the boundary of the substance. Hence, since V is arbitrary, it follows the conservation equation:

$$\frac{\partial}{\partial t}\rho(\mathbf{r},t) + \nabla\cdot\mathbf{j}(\mathbf{r},t) = 0\ . \tag{2.7}$$

By substituting Fick's law (2.5) into the conservation equation (2.7), we get the diffusion equation

$$\frac{\partial}{\partial t}\rho(\mathbf{r},t) = D\nabla^2\rho(\mathbf{r},t), \tag{2.8}$$

that can be easily solved. In the one-dimensional case the solution is given by

$$\rho(x,t) = \frac{1}{\sqrt{4\pi D t}}\exp\left(-\frac{x^2}{4Dt}\right)\ . \tag{2.9}$$

In 1863, C. Weiner attempted a first explanation of Brownian motion based on the kinetic theory. Further developments of the theory based on the collision model were attempted by von Nägeli, W. Strutt and Lord Rayleigh. Finally, in 1905, A. Einstein unified the continuum formulation given by Fick with the stochastic theory based upon the collision model and gave an excellent explanation of the Brownian motion [14]. Einstein's solution to the problem of Brownian motion was based upon the following two assumptions:

(i) The motion is caused by the exceedingly frequent impacts on the pollen grain of the incessantly moving molecules of liquid in which it is suspended.

(ii) The motion of these molecules is so complicated that its effect on the pollen grain can only be described probabilistically in terms of exceedingly frequent statistically independent impacts.

By using these two assumptions, Einstein proved that pollen grains suspended upon a liquid diffuse according to the equation (2.9) with the diffusion coefficient D

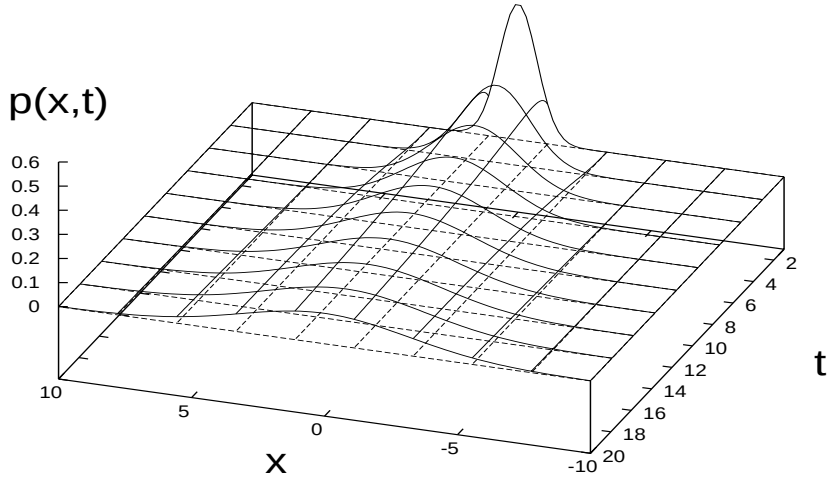

Figure 2.1: Diffusion Equation. Eq.(2.9) with $D = 1/4$.

given by

$$D = \frac{RT}{N} \cdot \frac{1}{6\pi kP}, \tag{2.10}$$

where N is Avogadro's number, k is the coefficient of viscosity, P is the radius of the suspended particles which are supposed spherical, R and T are the gas constant and the temperature respectively. In 1926, J. B. Perrin won the Nobel Prize for measuring the Avogadro's number with a rather high accuracy by using the Einstein's theory.

It is easy to see that Eq. (2.9), a Gaussian function, belongs to the class of the scaling equations (2.1) with

$$\delta = \frac{1}{2} = H \ . \tag{2.11}$$

The Gaussian distribution is important for many reasons. Empirically many diffusion processes are well described by the Gaussian diffusion equation (2.9) because of the Central Limit Theorem (CLT) [15]. In fact, under general conditions, a random variable composed of the sum of many parts, each independent but arbitrarily distributed, is Gaussian.

The CLT states that if X_1, X_2, X_3, ..., X_n are independent random variables

such that

$$< X_i >= 0, \qquad var\{X_i\} = \sigma_i < \infty, \tag{2.12}$$

and $p_i(x_i)$ is the distribution function of X_i, the variable S_n defined by

$$S_n = \sum_{i=1}^{n} X_i \tag{2.13}$$

tends to the Gaussian with zero mean and variance

$$\Sigma_n^2 = var\{S_n\} = \sum_{i=1}^{n} \sigma_i^2 \ . \tag{2.14}$$

This holds true if the Linderberg condition

$$\lim_{n\to\infty} \left[\frac{1}{\Sigma_n^2} \sum_{i=1}^{n} \int\limits_{|x|>t\sigma_n} dx \ x^2 \ p_i(x) \right] = 0 \tag{2.15}$$

is fulfilled for any fixed $t > 0$.

In nature, however, many anomalous diffusion processes are observed as well. These processes are characterized by a variance scaling exponent $H \neq 0.5$. If $0 < H < 0.5$, the process is characterized by subdiffusion. If $H > 0.5$, there is superdiffusion. The Brownian diffusion, $H = 0.5$, is the threshold between sub- and super-diffusion. $H = 1$ is the special case that corresponds to the ballistic motion. As normal diffusion rests on the validity of the CLT, anomalous diffusion rests on the validity of the Lévy-Gnedenko Generalized Central Limit Theorem (GCLT) that can be applied where not all moments of the underlying elementary transport events, the jumps, exist [16, 17, 18, 4, 5, 19].

The first one who studied anomalous diffusion was Richardson in his treatise on turbulent diffusion in 1926 [20]. Today, anomalous diffusion is observed in many systems. Subdiffusion regimes are observed in charge carrier transport in amorphous semiconductors [21, 22, 23, 24, 25, 26, 27], in nuclear magnetic resonance (NMR) diffusometry in percolative [28, 29] and porous systems [30, 31], in reptation dynamics in polymeric systems [32, 33, 34, 35, 36, 37, 38], in transport on fractal geometries

[39, 40], in the diffusion of a scalar tracer in an array of convection rolls [41, 42], and in the dynamics of a bead in a polymeric network [43, 44]. Superdiffusion and/or Lévy statistics are observed in special domains of rotating flows [45], in collective slip diffusion on solid surfaces [46], in layered velocity fields [47, 48], in Richardson turbulent diffusion [20, 49, 50, 51, 52], in bulk-surface exchange controlled dynamics in porous glasses [53, 54, 55], in the transport in micelle systems and in heterogeneous rocks [56, 57, 58], in quantum optics [59, 60], in single molecule spectroscopy [61, 62], in the transport in turbulent plasma [63], in bacterial motion [64, 65, 66, 67, 68] and even for the flight of an albatross [69].

Anomalous diffusion has been modelled in numerous ways: fractional Brownian motion introduced by B. Mandelbrot [1, 70, 71, 72, 73, 74]; generalized diffusion equations [75]; continuous time random walk models [21, 22, 23, 24, 25, 76, 77, 78, 79, 80, 81, 82, 83, 84]; Langevin equations [85, 86, 87, 88, 89]; generalized Langevin equations [90, 91, 92, 93]; generalized master equations [94, 95, 96, 97]; generalized thermostatistics [98, 99, 100, 101, 102]. In the next sections we develop some continuous time random walk models.

2.2 From Discrete to Continuous Time Random Walk: Brownian diffusion.

The problem of a discrete time random walk (DTRW) may be addressed as follows. At time $l = 0$ a walker is in the position $m = 0$. At each temporal step the walker moves with one jump of intensity 1 in either a positive or a negative direction. Let us suppose that the probability of doing a positive jump +1 is equal to the probability of doing a negative jump -1. The problem is to determine the probability $p(m, l)$ for the random walker to be at position m after l jumps. This probability is given by the binomial distribution [103]:

$$p(m,l) = \frac{1}{2^l}\binom{l}{\frac{l+m}{2}}\frac{1+(-1)^{l+m}}{2} \ . \tag{2.16}$$

For large values of l, the binomial distribution (2.16) converges to the Gaussian distribution

$$p(x,t) = \frac{1}{\sqrt{2\pi\sigma^2 t}} \cdot exp\left(-\frac{x^2}{2\sigma^2 t}\right), \tag{2.17}$$

with $\sigma = 0.5$ and where m and l are replaced with the continuous variables x and t respectively. Eq. (2.17) is of the type of Eq. (2.1) with the scaling coefficient $\delta = 0.5$ as expected for a diffution process where the CLT holds true.

Whereas the DTRW model is based on the idea that the walker can jump only by a discrete unit length for each discrete time unit, in the CTRW model the length of a given jump, as well as the waiting time elapsing between two successive jumps, are regulated by a jump pdf, $w(x,t)$. The function $w(x,t)$ determines the jump length pdf

$$\lambda(x) = \int_0^\infty w(x,t)\ dt \tag{2.18}$$

and the waiting time pdf

$$\psi(t) = \int_{-\infty}^\infty w(x,t)\ dx\ . \tag{2.19}$$

The quantity $\lambda(x)\ dx$ is the probability of having a jump length in the interval $(x, x+dx)$, whereas $\psi(t)\ dt$ is the probability for a waiting time between two successive jumps in the interval $(t, t+dt)$. If the jump length and the waiting time are two independent random variables, the jump pdf $w(x,t)$ is decoupled and can be written as $w(x,t) = \lambda(x)\psi(t)$. If the jump length and the waiting time are coupled, it is possible to write the jump pdf $w(x,t)$ by using the conditional probabilities: $w(x,t) = p(x|t)\psi(t)$ or $w(x,t) = p(t|x)\lambda(x)$, that indicate a jump of a certain length needs a time cost or that in a given time interval the walker cannot travel more than a maximum distance.

By using the jump length and waiting time pdfs it is possible to determine the jump length variance

$$\Sigma^2 =< x^2 >= \int_{-\infty}^\infty x^2\ \lambda(x)\ dx \tag{2.20}$$

and the characteristic waiting time

$$T = \int_0^\infty t\ \psi(t)\ dt\ . \tag{2.21}$$

If at $t = 0$ the walker is at position $x = 0$, the pdf $\eta(x,t)$ of just having arrived at position x at time t from the position x' at time t' is determined by

$$\eta(x,t) = \int_{-\infty}^{\infty} dx' \int_0^\infty dt'\ \eta(x',t')\ w(x-x',t-t') + \delta(x)\delta(t)\ . \tag{2.22}$$

Let us now define the cumulative probability

$$\Psi(t) = 1 - \int_0^t dt'\ \psi(t') \tag{2.23}$$

of having no jump events during the time interval $(0,t)$. It is not difficult to prove that the pdf $p(x,t)$ of being in x at time t is given by

$$p(x,t) = \int_0^t dt'\ \eta(x,t')\ \Psi(t-t')\ . \tag{2.24}$$

Because it is easier to work in the Fourier-Laplace space, let us calculate the Fourier-Laplace transform of $p(x,t)$

$$\hat{p}(k,s) = \mathcal{F}\{\mathcal{L}\{p(x,t); t \to s\}; x \to k\}\ . \tag{2.25}$$

We obtain [80]

$$\hat{p}(k,s) = \frac{1-\hat{\psi}(s)}{s} \frac{\hat{p}_0(k)}{1-\hat{w}(k,s)}\ , \tag{2.26}$$

where $\hat{p}_0(k)$ is the Fourier transform of the initial condition $p(x,0)$. If the jump length and waiting time are independent random variable, we have that $\hat{w}(k,s) = \hat{\psi}(s)\hat{\lambda}(k)$.

This means

$$\hat{p}(k,s) = \frac{1-\hat{\psi}(s)}{s}\frac{\hat{p}_0(k)}{1-\hat{\psi}(s)\hat{\lambda}(k)} \ . \tag{2.27}$$

Finally, to obtain the pdf $p(x,t)$ in the diffusion limit, we have to invert the Fourier-Laplace transform in the limit $(k,s) \to (0,0)$, that is

$$p(x,t) = \mathcal{L}^{-1}\{\lim_{s\to 0}\mathcal{F}^{-1}\{\lim_{k\to 0}\hat{p}(k,s)\}\} \ . \tag{2.28}$$

Let us now use the CTRW model for describing the Brownian motion. Let us suppose that the Brownian diffusion is given by a decoupled jump pdf $w(x,t) = \lambda(x)\psi(t)$ with both characteristic waiting time and jump length variance finite. Let the waiting time pdf $\psi(t)$ be a Poissonian of the type

$$\psi(t) = \frac{1}{\tau}\exp\left(-\frac{t}{\tau}\right), \tag{2.29}$$

where $\tau = T < \infty$ is the characteristic waiting time. Let the jump length pdf $\lambda(x)$ be the Gaussian

$$\lambda(x) = \frac{1}{\sqrt{4\pi\sigma^2}}\exp\left(-\frac{x^2}{4\sigma^2}\right), \tag{2.30}$$

where σ^2 is the variance. By doing the Laplace transform of $\psi(t)$ and the Fourier transform of $\lambda(x)$ and taking only the lowest order in s and k, we obtain:

$$\hat{\psi}(s) \sim 1 - \tau s + O(\tau^2), \tag{2.31}$$

$$\hat{\lambda}(k) \sim 1 - \sigma^2 k^2 + O(k^4) \ . \tag{2.32}$$

Finally, by using Eq. (2.27), we get the propagator $\hat{p}(k,s)$ of the Brownian diffusion

$$\hat{p}(k,s) = \frac{1}{s+Dk^2}, \tag{2.33}$$

where $D = \sigma^2/\tau$. Eq. (2.33) is the Fourier-Laplace transform of the well-know Gaussian propagator

$$p(x,t) = \frac{1}{\sqrt{4\pi Dt}}\exp\left(-\frac{x^2}{4Dt}\right) \ . \tag{2.34}$$

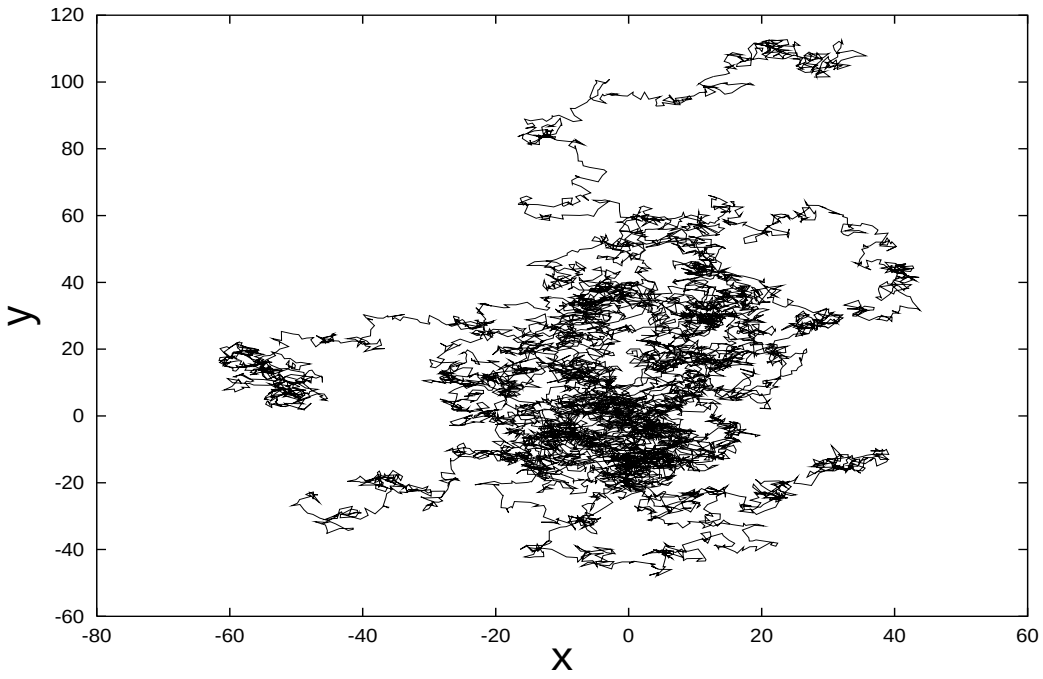

Figure 2.2: Diffusion of 10 Brownian particles with $\sigma^2 = 1$ in two dimensions. The trajectories are statistically self-similar. The walk starts from (0,0) and is drawn for 1000 steps.

2.3 Anomalous Diffusion: Fractional Brownian Motion (FBM).

For describing anomalous diffusion, Mandelbrot introduced Fractional Brownian Motion (FBM). FBM of index η may be easily defined as a simple generalization of the Brownian motion. For definition, FBM of index η is described by the fractional Gaussian propagator

$$p(x,t) = \frac{1}{\sqrt{4\pi D t^{\eta}}} \exp\left(-\frac{x^2}{4Dt^{\eta}}\right) \quad . \tag{2.35}$$

The second moment is given by

$$< x^2 >= \int_{-\infty}^{\infty} dx\ x^2\ p(x,t) = 2D\ t^{\eta}\ . \tag{2.36}$$

For $\eta = 1$ the normal Brownian motion is recovered. The case $0 < \eta < 1$ corresponds to subdiffusion and the case $\eta > 1$ corresponds to superdiffusion. In this case, the pdf scaling coefficient, δ, coincides with the variance scaling exponent, H. Precisely:

$$\delta = \frac{\eta}{2} = H \ . \tag{2.37}$$

Fractional Brownian motion is characterized by long range correlation. In fact, the correlation function of future increments, $\xi_\eta(t)$, with past increments, $-\xi_\eta(t)$, is given by

$$C(t) = \frac{< -\xi_\eta(-t)\xi_\eta(t) >}{< \xi_\eta^2(t) >} \propto 2^{\eta-1} - 1 \ . \tag{2.38}$$

The correlation function, $C(t)$, is calculated with the assumption that $\xi_\eta(0) = 0$. Eq. (2.38) justifies the Mandelbrot notation [1] of *persistent* fractional Brownian motion for $\eta > 1$ that corresponds to a positive correlation function, and of *antipersistent* fractional Brownian motion for $0 < \eta < 1$ that corresponds to a negative correlation function.

In the nice book of Feder [70], there is an algorithm to generate fractional Brownian noise. Let $\{\xi_i\}$ be a set of Gaussian random variables with unit variance and zero mean. The discrete fractional Brownian increment is given by

$$\begin{aligned} x_H(t) - x_H(t-1) \ &= \ \frac{n^{-H}}{\Gamma(H+0.5)} \left\{ \sum_{i=1}^{n} i^{H-0.5} \ \xi_{1+n(M+t)-i} + \right. \\ &\left. \sum_{i=1}^{n(M-1)} \left((n+i)^{H-0.5} - i^{H-0.5} \right) \ \xi_{1+n(M-1+t)-i} \right\} , \end{aligned}$$

where M is an integer that should be moderately large, and n indicates the number of the fractional steps for each unit time. In the simulation, good results are obtained with $M = 1000$ and $n = 10$.

2.4 Waiting time distributions with extended tails.

In this section, we will discuss the scaling properties of two diffusion processes characterized by a waiting time pdf $\psi(t)$ with long tails , that is, an asymptotic behavior

given by

$$\psi(t) \sim \left(\frac{\tau}{t}\right)^{\mu}, \tag{2.39}$$

where $\mu > 1$. If $1 < \mu < 2$, the characteristic waiting time T diverges and the process is non-stationary. For $\mu > 2$, T is finite and the process is stationary. A waiting time pdf with the asymptotic behavior given by (2.39) may be produced by a Lévy distribution if $2 < \mu < 3$ because in this region the second moment of the waiting time pdf $\psi(t)$ diverges whereas T is finite. The jump length variance is kept finite. However, there are two different ways of jumping. We use *Symmetric Jump Model* (SJM) to refer to a diffusion process where the walker may randomly make a jump in both positive and negative directions. Instead, we call *Asymmetric Jump Model* (AJM) that diffusion model where the walker makes a jump always in the same direction. For simplicity, it may be supposed that the length of the jump is fixed.

2.4.1 Symmetric Jump Model (SJM).

In the Symmetric Jump Model a walker remains in a state of rest for a time t according to a long-tail waiting time pdf with the asymptotic behavior given by the Eq. (2.39). Then the walker makes a jump randomly forward or backward with a length whose variance is kept finite. We also suppose that the jump length is kept fixed. The scaling corresponding to this process has been studied by Shlesinger in the pioneer paper of ref. [104].

Let us first suppose that $1 < \mu < 2$. Under this condition, the characteristic waiting time T diverges. In Laplace space at the lowest order, the waiting time pdf is

$$\hat{\psi}(s) \sim 1 - (\tau s)^{\mu-1}. \tag{2.40}$$

By using a Gaussian jump length pdf, the jump pdf in the Fourier-Laplace space becomes

$$\hat{p}(k,s) = \frac{1}{s}\ \frac{\hat{p}_0(k)}{1 + D_\mu s^{1-\mu} k^2}, \tag{2.41}$$

where $D_\mu = \sigma^2/\tau^{\mu-1}$. By doing the calculation as shown in [11], we see that the

second moment is given by

$$< x^2(t) >= \frac{2D_\mu}{\Gamma(\mu)}\ t^{\mu-1}, \tag{2.42}$$

where $\Gamma(\mu)$ is the Gamma function. Eq. (2.42) implies that

$$H = \frac{\mu - 1}{2}\ . \tag{2.43}$$

A closed-form solution for the jump pdf can be found in terms of Fox functions [11], that is,

$$p(x,t) = \frac{1}{\sqrt{4\pi D_\mu t^{\mu-1}}}\ H_{1,2}^{2,0}\left[\frac{x^2}{4D_\mu t^{\mu-1}}\ \middle|\ \begin{matrix}((3-\mu)/2, \mu-1)\\ (0,1),(0.5,1)\end{matrix}\right]. \tag{2.44}$$

Eq. (2.44) is scaling equation of the type of Eq. (2.1) with the scaling coefficient δ given by

$$\delta = \frac{\mu - 1}{2}\ . \tag{2.45}$$

In this case the pdf scaling coefficient δ is equal to the second moment scaling exponent H. For $1 < \mu < 2$, $\delta < 0.5$. This means that the SJM yields a type of subdiffusion. This is due to the fact that the characteristic waiting time T is not finite. For $\mu > 2$, the characteristic waiting time T is finite and the diffusion process obeys to the CLT, therefore yielding

$$\delta = \frac{1}{2} = H\ . \tag{2.46}$$

2.4.2 Asymmetric Jump Model (AJM).

In the Asymmetric Jump Model [105], the walker can make a jump only in one direction. The waiting time distribution is still given by Eq. (2.39). For simplicity, let us suppose that the length of the jump is fixed to a unit length, **1**. Under this condition the jump length pdf is given by

$$\lambda(x) = \delta(x - \mathbf{1}), \tag{2.47}$$

where $\delta(x)$ is the usual Dirac delta function. The Fourier transform of $\lambda(x)$ is given by

$$\hat{\lambda}(k) = \exp(ik\mathbf{1}) \ . \tag{2.48}$$

By plugging Eq. (2.48) in Eq. (2.27), we obtain the jump pdf

$$\hat{p}(k,s) = \frac{1-\hat{\psi}(s)}{s} \ \frac{1}{1-\hat{\psi}(s)\exp(ik\mathbf{1})} = \frac{1-\hat{\psi}(s)}{s} \sum_{n=0}^{\infty} \hat{\psi}(s)^n \ e^{ink\mathbf{1}} \ . \tag{2.49}$$

By evaluating the inverse Fourier transform of Eq. (2.49), we arrive at

$$\hat{p}(x,s) = \frac{1-\hat{\psi}(s)}{s} \sum_{n=0}^{\infty} \hat{\psi}(s)^n \ \delta(x-n\mathbf{1}) \ . \tag{2.50}$$

For large distances we can adopt the following expression

$$\hat{p}(x,s) = \frac{1-\hat{\psi}(s)}{s} \ [\hat{\psi}(s)]^x \ . \tag{2.51}$$

Let us now use the waiting time pdf in Laplace space given by Eq. (2.40). We obtain

$$\hat{p}(x,s) = \frac{(\tau s)^{\mu-1}}{s} \ [1-(\tau s)^{\mu-1}]^x \ . \tag{2.52}$$

By inverting the Laplace transform with the method of Ref. [106], we obtain for $1 < \mu < 2$

$$p(x,t) \approx \frac{1}{t^{\mu-1}} \ \left(\frac{x}{t^{\mu-1}}\right)^{-\frac{\mu}{\mu-1}} \ L\left[\left(\frac{x}{t^{\mu-1}}\right)^{-\frac{1}{\mu-1}}, \mu-1, 1-\mu\right], \tag{2.53}$$

and for $2 < \mu < 3$

$$p(x,t) \approx \frac{1}{t^{\frac{1}{\mu-1}}} \ L\left[\frac{x}{t^{\frac{1}{\mu-1}}}, \mu-1, 1-\mu\right], \tag{2.54}$$

where $L(y,\alpha,-\alpha)$ denotes a fully asymmetric Lévy stable law of index α [5]. From Eqs. (2.53) and (2.54) we obtain the following scaling prescriptions

$$\delta = \mu - 1, \qquad 1 < \mu < 2, \tag{2.55}$$

and

$$\delta = \frac{1}{\mu - 1} \ , \qquad 2 < \mu < 3 \ . \tag{2.56}$$

For $\mu > 3$ the CLT holds true because the second moment of the waiting time pdf $\psi(t)$ is finite, therefore, we have $\delta = 0.5$. We observe that for $1 < \mu < 1.5$ there is subdiffusion, whereas for $1.5 < \mu < 3$ there is superdiffusion. Moreover, for $0.5 < \delta < 1$ there are two possible values for μ smaller or greater than 2. The value $\mu = 2$ is important because is the border between the non-stationary ($T = \infty$ for $\mu < 2$) and the stationary ($T < \infty$ for $\mu > 2$) region.

2.5 Lévy Flights: Long Jumps Model (LJM).

In this section, we study a diffusion process generated by a waiting time pdf with a finite characteristic time T that may be modeled by a Poissonian distribution, and a jump length pdf $\lambda(x)$ given by a Lévy distribution with index $0 < \alpha < 2$. By definition, the Fourier transform of $\lambda(x)$ is

$$\hat{\lambda}(k) = \exp(-\sigma^\alpha |k|^\alpha) \sim 1 - \sigma^\alpha |k|^\alpha \ . \tag{2.57}$$

$\lambda(x)$ has the asymptotic behavior given by

$$\lambda(x) \sim A_\alpha \ \sigma^\alpha |x|^{-1-\alpha} = A_\alpha \ \sigma^{1-\mu} |x|^{-\mu} \tag{2.58}$$

for $|x| \gg \sigma$ and $\mu = 1 + \alpha$. Substituting the asymptotic expansion of the jump length pdf $\hat{\lambda}(k)$ in the Fourier space and the waiting time pdf of Eq. (2.29) in the Laplace space into Eq. (2.27), we obtain the following jump pdf in the Fourier-Laplace space

$$\hat{p}(k, s) = \frac{1}{s + K^\alpha \ |k|^\alpha}, \tag{2.59}$$

where $K^\alpha = \sigma^\alpha/\tau$ is the generalized diffusion constant. Eq. (2.59) is the solution of the generalized diffusion equation

$$\frac{\partial p(x,t)}{\partial t} = K^\alpha \ {}_{-\infty}D_x^\alpha \ p(x,t), \tag{2.60}$$

where ${}_{-\infty}D_x^\alpha$ is the fractional Weyl operator [107, 108]. Upon Laplace inversion of Eq. (2.59), we get the characteristic function of the jump pdf

$$\hat{p}(k,t) = \exp\left(-K^\alpha \ t \ |k|^\alpha\right) \ . \tag{2.61}$$

Eq. (2.61) is the characteristic function of a centered and symmetric Lévy distribution. The Fourier inversion of (2.61) can be obtained analytically by making use of the Fox function [109, 110]

$$p(x,t) = \frac{1}{\mu-1} \ \frac{1}{t^{1/(\mu-1)}} \ \left(\frac{|x|}{t^{1/(\mu-1)}}\right)^{-1} \ H_{2,2}^{1,1}\left[\frac{|x|}{(K^\alpha \ t)^{1/(\mu-1)}} \ | \ \begin{matrix}(1,1/\alpha),(1,1/2)\\(1,1),(1,1/2)\end{matrix}\right] . \tag{2.62}$$

The pdf scaling coefficient δ for the LFM with $1 < \mu < 3$ is

$$\delta = \frac{1}{\mu-1} \ . \tag{2.63}$$

We observe that Eq. (2.62) has an power-law asymptotic of the type

$$p(x,t) \sim \frac{1}{t^{1/(\mu-1)}} \ \left(\frac{|x|}{t^{1/(\mu-1)}}\right)^{-\mu} \qquad \mu < 3 \ . \tag{2.64}$$

Due to this propriety, the mean squared displacement, $< x^2(t) >$, diverges. For this reason, for the LJM, the variance scaling exponent, H, cannot be defined.

2.6 Lévy Walk: Symmetric Velocity Model (SVM).

In this section, we address a dynamic derivation of a Lévy walk diffusion [111]. This type of diffusion is characterized by the fact that the walker travels with a constant

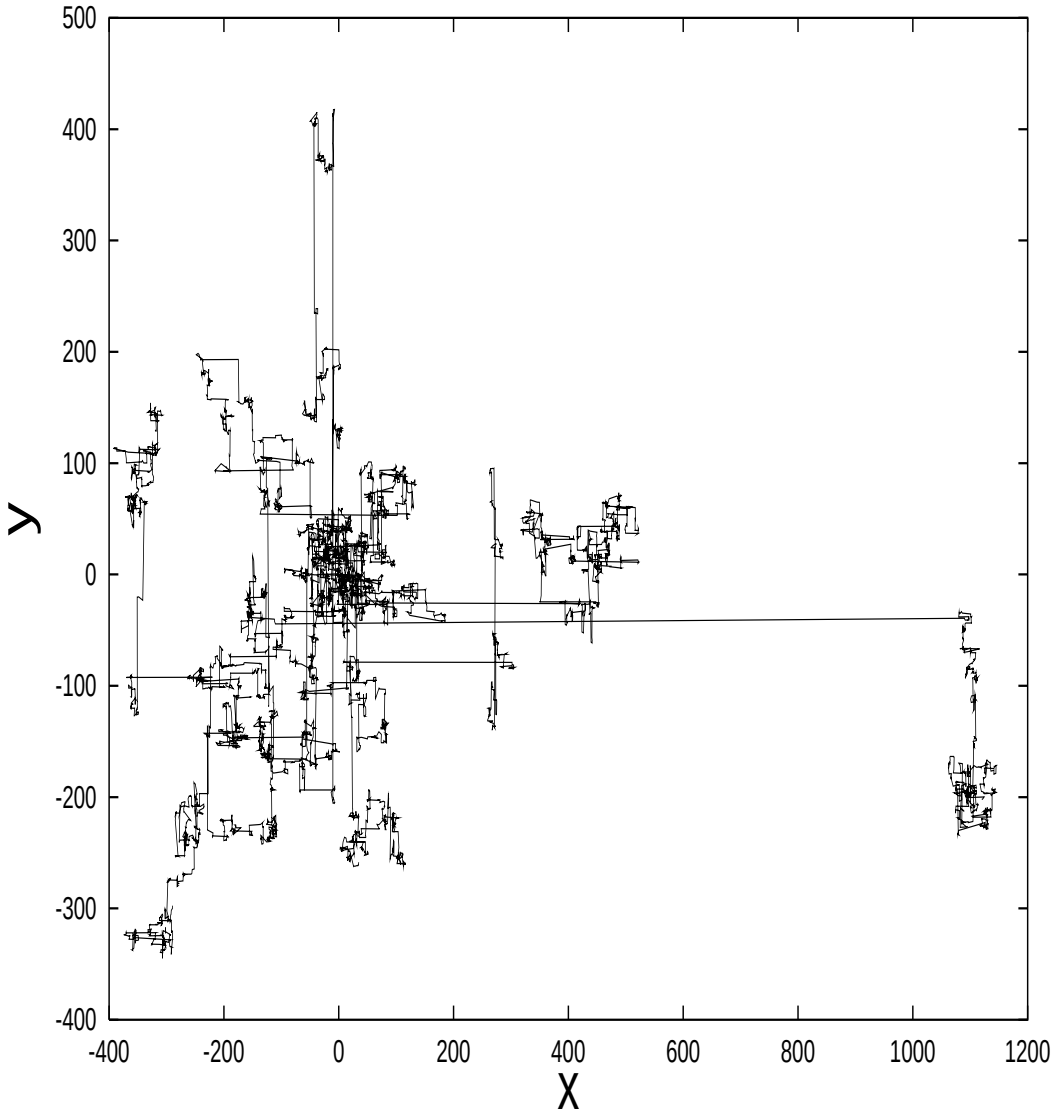

Figure 2.3: Long jump diffusion of 10 Lévy particles with $\mu = 2.5$ and $T = 1$ in two dimensions. The typical island structure of clusters of smaller steps connected by a long step is evident. The trajectories are statistically self-similar. The walk starts from (0,0) and is drawn for 1000 steps.

velocity throughout a whole time interval t chosen from a long- tail waiting time pdf of the kind

$$\psi(t) = (\mu - 1)\frac{T^{\mu-1}}{(T+t)^{\mu}} \ . \tag{2.65}$$

At the end of the time interval, the walker may or may not change direction and travel through the next time interval with the same or the opposite constant velocity. We focus our attention upon the interval $2 < \mu < 3$ that is characterized by a finite characteristic waiting time T and by an divergent second moment: a fact that is required for a Lévy statistics.

A diffusion process concerning the variable $x(t)$ is dynamically described by the equation of motion

$$\dot{x}\ (t) = \xi(t)\ , \tag{2.66}$$

where $\xi(t)$ are the fluctuations that the variable $x(t)$ collects. We make the simplifying assumption that ξ is a dichotomous variable: $\xi = \pm\mathbf{1}$, where $\mathbf{1}$ is a unit of length. The solution of (2.66) is given by

$$x(t) = x(0) + \int_0^t dt'\ \xi(t')\ . \tag{2.67}$$

Let us assume stationarity, namely, that the normalized correlation function

$$\Phi_y(t_1, t_2) = \frac{<\xi(t_1)\xi(t_2)>}{<\xi^2>} \tag{2.68}$$

depends only on $t = |t_2 - t_1|$, that is, $\Phi_y(t)$. Moreover, we observe that due to the dichotomous nature of $\xi(t) = \pm\mathbf{1}$, we have $<\xi^2>= 1$. It is easy to prove that the mean squared displacement $<x^2(t)>$ is given by

$$\frac{d}{dt} <x^2(t)>= 2<\xi^2> \int_0^t dt'\ \Phi_y(t-t')\ . \tag{2.69}$$

The normalized correlation function, $\Phi_y(t)$, is related to the waiting time pdf, $\psi(t)$,

through the relation

$$\psi(t) = \frac{d^2}{dt^2}\ \Phi_y(t)\ . \tag{2.70}$$

By using Eqs. (2.65, 2.67 and 2.70) it is easy to prove that

$$\lim_{t\to\infty} < x^2(t) > \propto t^{2H}, \tag{2.71}$$

with

$$H = \frac{4-\mu}{2}\ . \tag{2.72}$$

On the other hand, by using the result of [111], it is possible to prove that the jump pdf, $p(x,t)$, obeys the following equation

$$\frac{\partial}{\partial t}p(x,t) = \frac{1}{2}\int_{-\infty}^{\infty} dx'\ \psi\left(\frac{|x-x'|}{\mathbf{1}}\right)\ p(x',t)\ . \tag{2.73}$$

By substituting (2.65) in (2.73) and making the plausible assumption that the short-range region $|x-x'| \approx T$ does not contribute to the long-time process, we get the integro-differential equation

$$\frac{\partial}{\partial t}p(x,t) \propto \int_{-\infty}^{\infty} dx'\ \frac{p(x',t)}{|x-x'|^{\mu}} \tag{2.74}$$

that describes a centro-symmetric Lévy stable process [85]. Because we expect that the scaling given by Eq. (2.1) takes place for large t, with a simple dimensional analysis it is possible to get the right pdf scaling coefficient δ:

$$\delta = \frac{1}{\mu-1}\ . \tag{2.75}$$

Eq. (2.72) and (2.75) state that for the SVM both the second moment scaling exponent, H, and the pdf scaling coefficient, δ, exist. However, they coincide only for

$\mu = 2$ and $\mu = 3$. In general we have

$$\delta = \frac{1}{3 - 2H} \ . \tag{2.76}$$

2.7 Conclusion: we need both H and δ!

The relation between H and δ given by Eq. (2.76) is very important because it may be used to distinguish a fractional Brownian diffusion, that is characterized by $H = \delta$, from a diffusion with Lévy proprieties. The usual methods for detecting H, as the Hurst analysis, are able only to measure H. This is not enough if we want to study the statistical proprieties of the diffusion propagator $p(x, t)$. We need to measure the pdf scaling coefficient δ, as well. The Diffusion Entropy Analysis (DEA), developed in Chapter 4, is the first technique that determines δ.

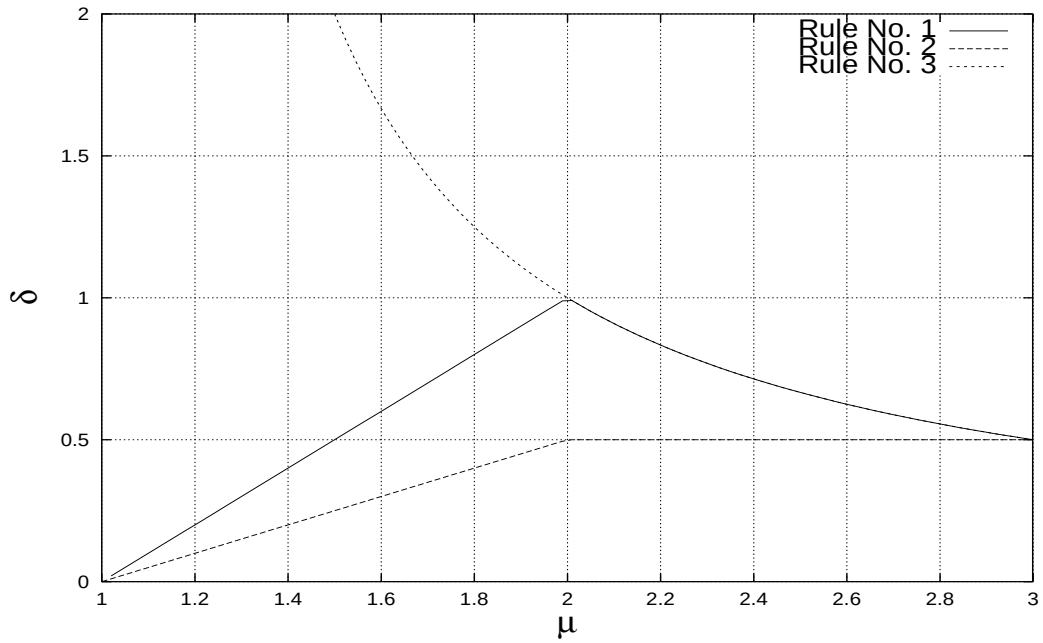

Figure 2.4: δ as a function of μ according to three rules. The solid, dashed and dotted lines denote AJM (rules No. 1), SJM (No. 2) and LJM (No. 3), respectively.

Summary. Let us give a short summary of the results of this chapter.

Fractional Brownian Motion (FBM):

$$\delta = H = \frac{\eta}{2} \ . \tag{2.77}$$

$\eta = 1$ implies normal Brownian motion.

Symmetric Jump Model (SJM):

$$\delta = H = \begin{cases} (\mu - 1)/2 & 1 < \mu < 2 \\ 1/2 & \mu > 2 \ . \end{cases} \tag{2.78}$$

Asymmetric Jump Model (AJM):

$$\delta = \begin{cases} (\mu - 1) & 1 < \mu < 2 \\ 1/(\mu - 1) & 2 < \mu < 3 \ . \end{cases} \tag{2.79}$$

$$\delta = H = 1/2 \qquad \mu > 3 \ . \tag{2.80}$$

Long Jump Model (LJM):

$$\delta = \frac{1}{\mu - 1} \qquad 1 < \mu < 3 \ . \tag{2.81}$$

$$\delta = H = 1/2 \qquad \mu > 3 \ . \tag{2.82}$$

The second moment scaling exponent, H, cannot be defined for $\mu < 3$ because the mean squared displacement diverges.

Symmetric Velocity Model (SVM):

$$\delta = \frac{1}{\mu - 1} \ . \tag{2.83}$$

$$H = \frac{4 - \mu}{2} \ . \tag{2.84}$$

$$\delta = \frac{1}{3 - 2H} \ . \tag{2.85}$$

These relations are valid for $2 < \mu < 3$. $\delta = H = 1/2$ for $\mu > 3$.

CHAPTER 3

VARIANCE SCALING ANALYSIS.

In this chapter, we review the most common methods used to measure the variance scaling exponent, H, associated with a set of data $\{\xi_i\}$. The variance scaling analysis studies the fractal properties of the variance (3.6). In Chapter 2, we saw that the variance scaling exponent, H, is defined by

$$\Sigma^2(t) \sim t^{2H}, \tag{3.1}$$

where Σ^2 is the variance of the diffusion process. If $< x(t) >= 0$, the variance $\Sigma^2(t)$ coincides with the mean squared displacement:

$$< x^2(t) > \sim t^{2H}, \qquad < x(t) >= 0 \ . \tag{3.2}$$

In chapter 2, we saw that Brownian noise is characterized by $H = 0.5$. The correlation function of Brownian noise is zero. If $0 < H < 0.5$, the noise shows antipersistent properties, that is, a negative correlation. If $0.5 < H < 1$, the noise shows persistent properties, that is a positive correlation.

In this chapter, after the analysis of the basic algorithm for studying the variance, we review the Hurst's analysis, detrended fluctuation analysis, relative dispersion analysis, spectral analysis and wavelet spectral analysis. All these methods are related to the variance scaling analysis also if they do not coincide and may give a slightly different value for the scaling exponent H, see Ref. [10]. We use the same symbol H to indicate the scaling coefficient obtained with all methods.

3.1 Basic algorithm.

The basic algorithm for the variance scaling analysis of a set of data is the following. Let us suppose that a temporal series $\{\xi_i\}$ of N data is given:

$$\xi_1,\ \xi_2,\ \xi_3,\ \xi_4,\ \xi_5,\ \xi_6,\ \xi_7,\ \xi_8,\ \xi_9,\ \xi_{10},\ \xi_{11},\ \xi_{12},\ ...,\ \xi_{N-1},\ \xi_N\ . \tag{3.3}$$

We use the set of data $\{\xi_i\}$ to build a trajectory

$$x(t) \equiv \sum_{i=1}^{t} \xi_i\ , \qquad n = 1,\ 2,\ 3,\ ...,\ N\ . \tag{3.4}$$

The trajectory (3.4) is then used to build a series of sub-trajectories $\{x_n(t)\}$ according to the following algorithm

$$x_n(t) \equiv \sum_{i=1}^{t} \xi_{i+n-1}\ , \qquad n = 1,\ 2,\ 3,\ ...,\ N-t\ . \tag{3.5}$$

where $x_n(t)$ indicates the position of the n^{th} sub-trajectory at time t. For each t, there are only $N-t$ available sub-trajectories because the last available sub-trajectory is made by the last t data, that is, by $\xi_{N-t+1},\ \xi_{N-t+2},\ ...,\ \xi_{N-1},\ \xi_N$. All trajectories start from the origin $x(t=0)=0$. At the increasing of the time t, the sub-trajectories generate a diffusion process. At each time t, it is possible to calculate the variance of the position of the $N-t$ available sub-trajectories according to the well known variance equation:

$$\Sigma^2(t) = var(x(t)) = \frac{\sum_{n=1}^{N-t} \left(x_n(t) - \overline{x}(t)\right)^2}{N-t-1}, \tag{3.6}$$

where $\overline{x}(t)$ is the average of the positions of the $N-t$ sub-trajectories at time t. We note that the way to build sub-trajectories shown in (3.5) is not unique. We can also adopt a non-overlapping window method. In this case the original trajectory (3.4) is divided in $M = int(N/t)$ non-overlapping available sectors or sub-trajectories of size t; $int(x)$ is the integer part of x. We can then use the M sub-trajectories to calculate

the variance. The non-overlapping method has the advantage of using independent sub-trajectories, but the disadvantage of statistics poorer than those obtained by using the overlapping window method.

3.2 Hurst's Rescaled Range Analysis (R/S analysis).

In 1965, in the book, *Long-Term Storage: An Experimental Study* [3], Hurst introduced a method for studying the fractal properties of a time series. Hurst developed his method for studying the water storage of the Nile river. The problem was to design a reservoir, which never overflows or empties, based upon the given record of observed discharges from a lake. Let us suppose that ξ_i is the amount of water flowing from a lake to a reservoir for each year. The problem is to determine the needed capacity of the reservoir under the condition that each year the reservoir releases a volume of water equal to the mean influx. In τ years, the average influx is

$$< \xi >_\tau = \frac{1}{\tau} \sum_{i=1}^{\tau} \xi_i \ . \tag{3.7}$$

The amount of water accumulated in the reservoir in t years is

$$x(t,\tau) = \sum_{i=1}^{t} \{\xi_i - < \xi >_\tau\} \ . \tag{3.8}$$

The reservoir neither overflows nor empties during the period of τ years if its storage capacity is larger than the difference, $R(\tau)$, between the maximum and minimum amounts of water contained in the reservoir. $R(\tau)$ is

$$R(\tau) = \max_{1 \le t \le \tau} x(t,\tau) - \min_{1 \le t \le \tau} x(t,\tau) \ . \tag{3.9}$$

For getting a dimensionless value, Hurst divided $R(\tau)$ by the standard deviation $S(\tau)$ of the data during the τ years:

$$S(\tau) = \sqrt{\frac{1}{\tau} \sum_{i=1}^{\tau} \{\xi_i - < \xi >_\tau\}^2} \ . \tag{3.10}$$

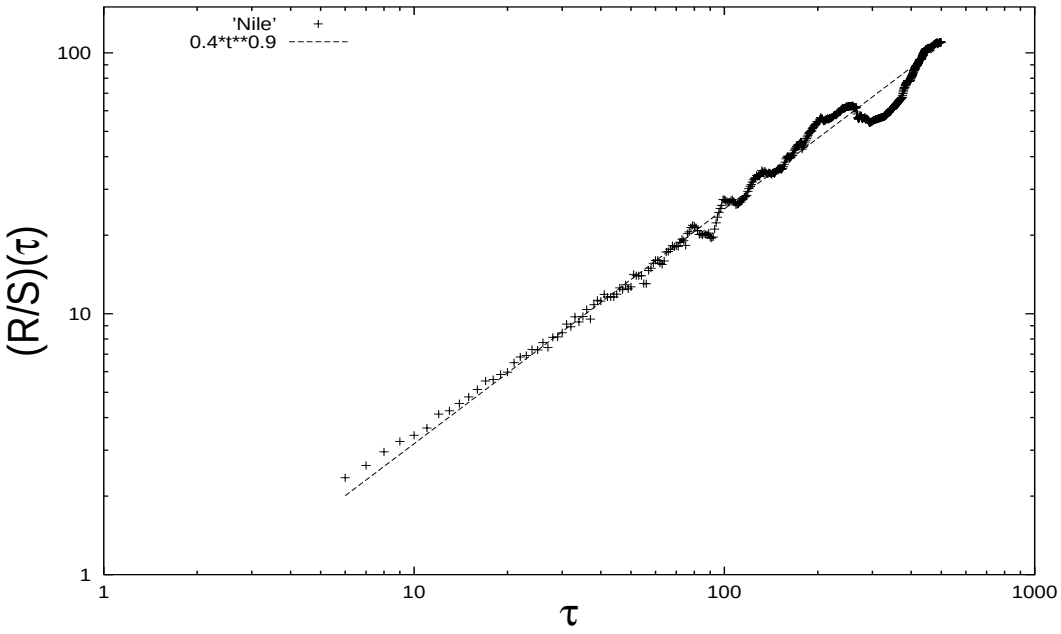

Figure 3.1: Hurst R/S analysis of the measured annual discharge of the Nile river (years 622-1284). The scaling exponent is $H = 0.90 \pm 0.02$.

Hurst observed that many phenomenon are very well described by the following scaling relation:

$$\frac{R(\tau)}{S(\tau)} \propto \tau^H \ . \tag{3.11}$$

The exponent H (called K by Hurst) was called the Hurst exponent, H, by Mandelbrot [1], who makes it a famous scaling analysis. The exponent H is directly related to the Lipschitz-Hölder exponent α [70] and to the variance scaling exponent also if they do not coincide [10].

In the case of the Nile river, Fig. 3.1, Hurst measured an exponent $H = 0.9$. This means that the Nile is characterized by a long range persistence that requires unusually high barriers, such as the Aswân High Dam, to contain damage and rein in the floods.

3.3 Detrended Fluctuation Analysis.

In 1994, Stanley and others [112] introduced a new method called Detrended Fluctuation Analysis (DFA). Since 1994, hundreds of papers, which analyze fractal properties of time series with the DFA, have been published.

Given a time sequence $\{\xi_i\}$ $(i = 1,\ ...,\ N)$, the DFA is based upon the following steps. First, the entire sequence of length N is integrate

$$x(t) = \sum_{i=1}^{t} (\xi_i - < \xi >) , \tag{3.12}$$

where

$$< \xi >= \frac{1}{N} \sum_{i=1}^{N} \xi_i \ . \tag{3.13}$$

Second, the time series is divided into $int(N/n)$ non-overlapping boxes. The number n, which indicates the size of the box, is an integer smaller than N. A local trend is defined for each box by fitting the data in the box. The linear least-squares fit may be done with a polynomial function of order $l \geq 0$ [113]. Let $x_n(t)$ be the local trend built with boxes of size n. Third, a detrended walk is defined as the difference between the original walk and the local trend given by the linear least-squares fit according to the following relation

$$X(t) = x(t) - x_n(t) \ . \tag{3.14}$$

Finally, the mean squared displacement of the detrended walk is calculated, that is,

$$F_D^2(n) = \frac{1}{N} \sum_{t=1}^{N} [X(t)]^2 \ . \tag{3.15}$$

Stanley and many others have proved that many time series are characterized by the following scaling relation

$$F_D(n) \propto n^H, \tag{3.16}$$

where the scaling exponent is $H = 0.5$ for a random walk or $H \neq 0.5$ for fractal noise.

3.4 Relative Dispersion Analysis.

Relative Dispersion Analysis (RDA) [114] is another technique that makes use of the variance. The purpose of the RDA is to study the fractal properties of a time series by measuring an dimensionless value called Relative Dispersion (RD). The RD is obtained by dividing the standard deviation, calculated from the variance of the basic algorithm method explained in section 3.1 by using non-overlapping aggregations of nearest neighbors, by the average among all aggregations of the sum of the data in each aggregation.

The algorithm is the following. Let $\{\xi_i\}$ be a set of N data with $i = 1,\ 2,\ 3,\ ...,\ N$. Let r be an integer smaller than N. Let us divide the original sequence in $int(N/r)$ contiguous aggregations of r nearest data. $int(y)$ is the integer part of y. Let $x_n(r)$ be the integral of the n^{th} aggregation of size r,

$$x_n(r) = \sum_{i=1}^{r} \xi_{n*r-r+i} \ . \tag{3.17}$$

Let $\overline{x}(r)$ be the average of all $x_n(r)$,

$$\overline{x}_n(r) = \frac{\sum_{n=1}^{int(N/r)} x_n(r)}{int(N/r)} \ . \tag{3.18}$$

Let $SD(r)$ be the standard deviation of all $x_n(r)$,

$$SD(r) = \sqrt{\frac{\sum_{n=1}^{int(N/r)} \left[x_n(r) - \overline{x}(r)\right]^2}{int(N/r) - 1}} \ . \tag{3.19}$$

Finally, the Relative Dispersion of r nearest neighbors is given by

$$RD(r) = \frac{SD(r)}{\overline{x}_n(r)} \ . \tag{3.20}$$

If the time series is a simple fractal process, then the Relative Dispersion scales as

$$RD(r) \propto r^{\beta}, \tag{3.21}$$

where $\beta = H - 1$ because SD(r) scales as r^H and $\overline{x}(r)$ scales as r.

3.5 Spectral Density Analysis.

Spectral analysis can be used for detecting the scaling in a stationary long memory process because of the validity of the Parseval's theorem that connects the variance of a time series to the spectral density function $S(f)$, where f is the frequency. In fact, noise $\{\xi_i\}$ with power law correlation is characterized by a spectral density function of the type

$$S(f) \propto \frac{1}{f^\alpha}, \tag{3.22}$$

where α is the spectral density scaling exponent. For $\alpha < 0$ the noise shows anti-persistent properties because the spectral density increases with the frequency. For $\alpha > 0$, the noise shows persistent properties because the spectral density is higher at low frequencies. The case $\alpha = 0$ corresponds to a white noise, that is, Brownian noise.

In this dissertation we are interested in studying time series. Therefore, we focus our attention upon the Discrete Fourier Transform (DFT). Given a time sequence $\{\xi_t\}$ $(t = 0,\ ...,\ N-1)$, the DFT of the data is the sequence $\{\hat{\xi}_k\}$ of N variable given by

$$\hat{\xi}_k = \sum_{t=0}^{N-1} \xi_t \exp\left(-\frac{i2\pi tk}{N}\right), \qquad k = 0,\ 1,\ ...,\ N-1\ . \tag{3.23}$$

It is well known that the Fourier transform can be inverted. It is possible to reconstruct the original set of data $\{\xi_t\}$ from its DFT $\{\hat{\xi}_k\}$ using the following equation

$$\xi_t = \frac{1}{N} \sum_{k=0}^{N-1} \hat{\xi}_k \exp\left(\frac{i2\pi tk}{N}\right), \qquad t = 0,\ 1,\ ...,\ N-1\ . \tag{3.24}$$

The Parseval's theorem is based on the validity of the following identity. Let $\{a_t\} \leftrightarrow \{\hat{a}_k\}$ and $\{b_t\} \leftrightarrow \{\hat{b}_k\}$ two series of data with their DFT, then we have

$$\sum_{t=0}^{N-1} a_t b_t^* = \frac{1}{N} \sum_{k=0}^{N-1} \hat{a}_k \hat{b}_k^*, \tag{3.25}$$

where b_t^* is the complex conjugate of b_t. Now, by letting $a_t = b_t = \xi_t$, we obtain the Parseval's theorem for finite sequences, namely,

$$\sum_{t=0}^{N-1} |\xi_t|^2 = \frac{1}{N} \sum_{k=0}^{N-1} |\hat{\xi}_k|^2 \ . \tag{3.26}$$

In the case in which $< \xi >= 0$, the term of (3.26) is equal to $N\sigma^2$, where σ^2 is the variance of the noise $\{\xi_t\}$. Therefore, we have

$$\sigma^2 = \frac{1}{N^2} \sum_{k=0}^{N-1} |\hat{\xi}_k|^2 = \sum_{k=0}^{N-1} S_\xi(f_k), \qquad f_k = \frac{2\pi k}{N}, \tag{3.27}$$

where f_k is the value of the k^{th} frequency and $S_\xi(f_k)$ is the spectral density function at the frequency f_k. Eq. (3.27) shows that the variance of the noise σ^2 can be decomposed into parts that depend upon frequency, $S_\xi(f_k)$. This allows us to find the following relation between the spectral density scaling exponent, α, and the variance scaling exponent H [74],

$$\alpha = 2H - 1 \ . \tag{3.28}$$

Instead of studying the spectral density function of the fluctuations ξ_i, it is possible to study the spectral density function of the integral of the fluctuations. As done before, let us define

$$x(t) = \sum_{i=1}^{t} (\xi_i - < \xi >) \, , \tag{3.29}$$

where

$$< \xi >= \frac{1}{N} \sum_{i=1}^{N} \xi_i \ . \tag{3.30}$$

Let $S_x(f)$ be the spectral density function of the data $x(t)$, Mandelbrot and van Ness [74] proved that, in this case, the spectral density scaling exponent, α, is related to the variance scaling exponent H through the following equation

$$\alpha = 2H + 1 \ . \tag{3.31}$$

The case $\alpha = 1$ is generated by pink noise, Brownian noise corresponds to $\alpha = 2$, and

for $\alpha > 2$ we have black noise.

Spectral analysis is widely used because of its simplicity. Moreover, for very large sets of data the spectral density function may be calculated with a fast algorithm, the Fast Fourier Transform (FFT), that works for $N = 2^j$ data, where j is an integer. Having a fast algorithm may be important for studying a large set of data. The FFT is a $O(N \log(N))$ algorithm.

3.6 Spectral Wavelet Analysis.

Spectral Wavelet Analysis (SWA) is a new and powerful method for studying the fractal properties of the variance [115]. Similar to SDA, SWA is able to decompose the sample variance of a time series on a *scale-by-scale* basis. In SDA, as well known, sine and cosines wave functions are used for the spectral analysis, whereas SWA makes use of scaling wavelets that have the characteristics of being localized in space and in frequencies. If $S_W(\tau)$ is the wavelet spectral density function at the scale τ, for a noise $\{\xi_i\}$ with power law correlation we have

$$S_W(\tau) \propto \tau^{\alpha}, \tag{3.32}$$

where the exponent α is the same exponent used for the SDA in Eq. (3.22). The relation between α and the variance scaling exponent H is the same that that valid for the Fourier Analysis. Therefore, $\alpha = 2H - 1$ for the wavelet spectral analysis of the noise, and $\alpha = 2H$ for the wavelet spectral analysis of the integral of the noise.

The essence of the wavelet analysis is the following. Let $\tilde{\psi}(u)$ be a real-valued function defined over the real axis $(-\infty,\ \infty)$ such that the two basic properties

$$\int_{-\infty}^{\infty} \tilde{\psi}(u)\ du = 0 \tag{3.33}$$

and

$$\int_{-\infty}^{\infty} [\tilde{\psi}(u)]^2\ du = 1 \tag{3.34}$$

are satisfied. The wavelets are characterized by the fact that they can be localized in the space and depend upon a scaling coefficient that gives the width of the wavelet. Two typical wavelets widely used are the Haar wavelet localized in t with a width coefficient τ defined by

$$ {}^{(H)}\tilde{\psi}_{\tau,t}(u) \equiv \begin{cases} -1/\tau, & t-\tau < u < t \\ 1/\tau, & t < u < t+\tau \\ 0, & otherwise \end{cases} , \tag{3.35} $$

and the Mexican hat wavelet localized in t and with a width coefficient τ defined by

$$ {}^{(Mh)}\tilde{\psi}_{\tau,t}(u) \equiv \frac{2\left[1-(u-t)^2/\tau^2)\right]\ e^{-(u-t)^2/2\tau^2}}{\pi^{1/4}\sqrt{3\tau}} . \tag{3.36} $$

The Mexican hat wavelet is the second derivative of a Gaussian. The width coefficient τ defines a kind of scale analyzed by the wavelet. Given a signal $\xi(u)$, the Continuous Wavelet Transform (CWT) is defined by

$$ W(\tau,t) = \int_{-\infty}^{\infty} \tilde{\psi}_{\tau,t}(u)\ \xi(u)\ du\ . \tag{3.37} $$

The CWT of a signal depends on two variables, τ that is a width coefficient and looks like the sub-trajectory size used in the basic algorithm in paragraph 3.1 or the DFA box size discussed in paragraph 3.3, and t that is the position in the sequence that indicate the region analyzed by the wavelet. If the wavelet $\tilde{\psi}(u)$ satisfies the admissibility condition, that is, if it is such that

$$ C_{\tilde{\Psi}} = \int_{0}^{\infty} \frac{|\tilde{\Psi}(f)|^2}{f}\ df < \infty, \tag{3.38} $$

where $\tilde{\Psi}(f)$ is the wavelet Fourier transform, namely,

$$ \tilde{\Psi}(f) = \int_{-\infty}^{\infty} \tilde{\psi}(u)\ e^{-i2\pi f u}\ du, \tag{3.39} $$

and the signal $\xi(u)$ is such that

$$\int_{-\infty}^{\infty} \xi^2(u)\ du < \infty, \tag{3.40}$$

then, the original signal can be recovered from its CWT via

$$\xi(u) = \frac{1}{C_{\tilde{\psi}}} \int_0^{\infty} \left[\int_{-\infty}^{\infty} W(\tau, t)\ \tilde{\psi}_{\tau,t}(u)\ dt \right]\ \frac{d\tau}{\tau^2}\ . \tag{3.41}$$

Finally, it is possible to prove that

$$\int_{-\infty}^{\infty} \xi^2(u)\ du = \frac{1}{C_{\tilde{\psi}}} \int_0^{\infty} \left[\int_{-\infty}^{\infty} W^2(\tau, t)\ dt \right] \frac{d\tau}{\tau^2} = \int_0^{\infty} S_W(\tau)\ d\tau\ . \tag{3.42}$$

The function $W^2(\tau, t)/\tau^2$ defines an energy density function that decomposes the energy across different scales and times. Eq. (3.42) is the wavelet equivalent to the Fourier Parseval's theorem. The function $S_W(\tau)$ is the wavelet spectral density function that gives the contribution to the energy at the scale τ.

As for the Fourier Analysis, when we have to study a time series, it is better to adopt a discrete version of the wavelet spectral analysis. The Discrete Wavelet Transform (DWT) and its generalization, the Maximal Overlap Discrete Wavelet Transform (MODWT), both work well. In the book of Percival [115], it is possible to find complete details. Due to the existence of a special algorithm, called *Pyramid Algorithm*, the DWT is the fastest algorithm for a spectral analysis; a $O(N)$ algorithm against the $O(N \log(N))$ algorithm for the Fast Fourier Transform. This makes the Wavelet Analysis the fastest analysis for very large set of data. However, the results depend slightly upon the particular wavelet chosen for the analysis.

CHAPTER 4

DIFFUSION ENTROPY ANALYSIS.

In this chapter, we introduce a new method, Diffusion Entropy Analysis (DEA), (the topic of this dissertation) that has the propriety of detecting the correct pdf scaling coefficient, δ, of a diffusion process generated by a time series. The pdf scaling coefficient δ is defined by the equation

$$p(x,t) = \frac{1}{t^{\delta}} \cdot F\left(\frac{x}{t^{\delta}}\right), \tag{4.1}$$

where $p(x,t)$ is the pdf of a diffusion process with fractal scaling. DEA allows us to detect the correct scaling coefficient also in the case in which a time series is characterized by Lévy properties. None of the methods available in the literature of the Science of Complexity can do that. In fact, the variance of an ideal Lévy flight process is divergent as it was shown in section 2.5, and all the methods discussed in Chapter 2 related to the variance (Hurst R/S Analysis, Detrended Fluctuation Analysis, Relative Dispersion Analysis, Spectral Analysis, Spectral Wavelet Analysis) are subtly based on the Gaussian assumption and, so, upon a variance that can be used to monitor scaling. The variance scaling coefficient, H, coincides with the pdf scaling, δ, of a time series only in particular cases such as for Fractional Brownian Motion. In Section 2.6, we proved that a diffusion process generated by Lévy walk is characterized by a variance scaling coefficient H and a pdf scaling coefficient δ that do not coincide and are related one to the other by the following equation

$$\delta = \frac{1}{3-2H} \ . \tag{4.2}$$

Eq. (4.2) suggests a new powerful method for distinguishing a Fractional Brownian Motion from a Lévy Walk Motion. Given a time series, we measure the variance scaling coefficient, H, by using, for example, the basic algorithm of Variance Scaling Analysis, (sec. 3.1). Then, we measure the pdf scaling coefficient δ with the Diffusion

Entropy Analysis. Finally, we compare δ and H. If $\delta = H$, the time series may be characterized by Fractional Brownian Motion. If $\delta \neq H$, the time series cannot be Fractional Brownian Motion. If δ and H are related each the other by Eq. (4.2), the time series is characterized by a Lévy distribution.

Diffusion Entropy Analysis is based upon the Shannon Entropy. Moreover, because we want to study the statistical properties of a time series, we need to introduce the Kolmogorov-Sinai entropy that studies the statistics of a dynamical symbol sequence. After reviewing the Kolmogorov-Sinai entropy, we briefly discuss the computational limits of it. These limits justify the adoption of the Diffusion Entropy Analysis that bypasses those limits.

4.1 The Shannon entropy and the Khinchin axioms.

Let $\{\xi_n\}$ be a long series of N observations with R different possible events. Let N_i be the number of times that the i-th event is observed. Therefore, we have

$$N = \sum_{i=1}^{R} N_i \tag{4.3}$$

and

$$p_i = \frac{N_i}{N}, \tag{4.4}$$

where p_i is the probability with which the i-th event occurs. We assume that the set of data is large enough for frequencies to coincide with the probability.

The problem is to measure the information or the entropy –the indicator of the lack of information– about the measure of an event that occurs with a probability p. Given a probability distribution $\{p_i\}$ of R events, where $i = 1,\ 2,\ ...,\ R$, the entropy indicator $S(p)$ of the probability distribution should fulfill the following conditions that are known as the Khinchin axioms:

(I) $S(p)$ is a function of the probabilities p_i only, that is

$$S(p) = S(p_1,\ p_2,\ ...,\ p_R)\ . \tag{4.5}$$

(II) The entropy $S(p)$ takes its maximum value in the correspondence of the uniform distribution, that is,

$$S(p) \leq S\left(\frac{1}{R}, \frac{1}{R},\ ..., \frac{1}{R}\right)\ . \tag{4.6}$$

(III) The entropy $S(p)$ should remain unchanged if the sample set is enlarged by a new event with probability $p_{R+1} = 0$, that is,

$$S(p_1,\ p_2,\ ...,\ p_R) = S(p_1,\ p_2,\ ...,\ p_R,\ 0)\ . \tag{4.7}$$

(IV) If a system Σ is divided in two subsystems Σ^I and Σ^{II}, the entropy $S(p)$ should fulfill the following condition

$$S(p) = S(p^I) + \sum_i p_i^I \sum_j Q(j|i) \ln Q(j|i), \tag{4.8}$$

where $S(p^I)$ is the entropy of the subsystem Σ^I and $Q(j|i)$ is the conditional probability that the subsystem Σ^{II} is in the state j if the subsystem Σ^I is in the state i. If the two subsystems Σ^I and Σ^{II} are independent, the entropy becomes additive with respect to the subsystems, that is

$$S(p) = S(p^I) + S(p^{II}), \tag{4.9}$$

where $S(p^I)$ and $S(p^{II})$ are the entropies of the two subsystems.

The four Khinchin axioms allow us to define, as the lack of information about the measure of an event i that occurs with a probability p_i, the following indicator

$$b_i = -\ln p_i\ . \tag{4.10}$$

We observe that if an event occurs with a probability $p = 1$, the lack of information about that event is zero because that event certainly occurs; b) if an event occurs with a probability $p = 0$, the lack of information about that event is infinity. Finally, if we have a long series of observations where the events $i = 1,\ 2,\ ...,\ R$ occur with

a probability distribution $\{p_i\}$, the four Khinchin axioms uniquely determine the entropy of the system as the average of the lack of information b_i about each events, that is

$$S(p) = -c \sum_{i=1}^{R} p_i \ \ln p_i \ . \tag{4.11}$$

The function $S(p)$ of Eq. (4.11) is the Shannon entropy, its opposite is called the Shannon information. The positive constant c is undetermined and it is chosen to be 1 by convention.

4.2 The Rényi and the Tsallis entropies.

The Shannon entropy is not the only way to measure information of a probability distribution. However, it is the only entropy that fulfills all four Khinchin axioms. If the fourth axiom is replaced by something else, it is possible to define other information measure. Two of them are the Rényi and the Tsallis entropies.

The Rényi entropy [117] is defined by considering the first three Khinchin axioms and replacing the fourth axiom with the simpler condition that the entropy is additive for independent subsystems as shown in Eq. (4.9). Given a probability distribution $\{p_i\}$ of R events, where $i = 1,\ 2,\ ...,\ R$, the Rényi entropy is defined by

$$S_\beta(p) = \frac{1}{1-\beta} \ \ln \sum_{i=1}^{R} (p_i)^\beta \, , \tag{4.12}$$

where β is an arbitrary real number. By using l'Hôpital's theorems, it is easy to prove that for $\beta \to 1$, the Rényi entropy converges to the Shannon entropy (4.11). An important general property of the Rényi entropy is that it is a monotonically increasing function of β for arbitrary probability distributions p.

The Tsallis entropy [118] is defined by considering the first three Khinchin axioms but it does not fulfill the fourth axiom but only a generalization of it [119]. Tsallis entropy is not extensive. Given a probability distribution $\{p_i\}$ of R events, where

$i = 1,\ 2,\ ...,\ R$, the Tsallis entropy is defined by

$$S_q(p) = \frac{1}{1-q}\ \left(1 - \sum_{i=1}^{R} p_i^q\right), \tag{4.13}$$

where q is an arbitrary real number. As for the Rényi entropy, the Shannon entropy can be recovered as the limit of the Tsallis entropy for $q \to 1$. The entropic index q characterizes the degree of nonextensivity according to the following rule

$$S_q(\Sigma) = S_q(\Sigma^I) + S_q(\Sigma^{II}) + (1-q)S_q(\Sigma^I)S_q(\Sigma^{II}), \tag{4.14}$$

where Σ^I and Σ^{II} are two independent subsystems in which the system Σ is divided. The situation of $q > 1$ corresponds to subadditivity or subextensivity. The situation of $q < 1$ corresponds to superadditivity or superextensivity. For $q = 1$, the Tsallis entropy coincides with the Shannon entropy and, therefore, it is additive.

4.3 The Kolmogorov-Sinai entropy.

The Kolmogorov-Sinai entropy was introduced to arrive at the fundamental concept of entropy of a trajectory [116]. It may be used, for example, for analyzing the entire sequences of iterates of a map. Let

$$\{A_R\} = \{A_1,\ A_2,\ ...,\ A_R\} \tag{4.15}$$

be a set of R different symbols. This set may be a partition of the phase space. A trajectory is a symbolic sequence of $\{x_i\}$ symbols chosen from $\{A_R\}$. The index i assumes integer consecutive values, that is, $i = 0,\ 1,\ 2,\ ....$ Given a sample sequence of N symbols chosen from $\{A_R\}$, $\{\omega_N\} = \{\omega_0,\ \omega_2,\ ...,\ \omega_{N-1}\}$, let

$$p(\omega_1,\ \omega_2,\ ...,\ \omega_{N-1}) = \int\limits_{j(\omega_1,\ \omega_2,\ ...,\ \omega_{N-1})} d\sigma(x), \tag{4.16}$$

be the probability of finding the sample set $\{\omega_N\} = \{\omega_1,\ \omega_2,\ ...,\ \omega_{N-1}\}$ in the symbolic sequence $\{x_i\}$. The function $\sigma(x)$ is the probability distribution of the initial values. Let us now measure the Shannon Entropy that, as explained in sec. 4.1, is an appropriate measure of the information that is present in the probability (4.16). By following the prescription of sec. 4.1, the Shannon Entropy is given by

$$S(N) = -\sum_{\omega_0,\ \omega_2,\ ...,\ \omega_{N-1}} p(\omega_0,\ \omega_2,\ ...,\ \omega_{N-1})\ \ln p(\omega_0,\ \omega_2,\ ...,\ \omega_{N-1}), \tag{4.17}$$

where the sum is done upon all possible different sample sequences $\{\omega_N\}$ which may be obtained by using the partition $\{A_R\}$ of R symbols (4.15). The Shannon entropy depends upon the size N, therefore, for making it independent of N, let us measure the Shannon entropy for unit length defined by

$$K_{\{A\}} = \lim_{N\to\infty} \frac{S(N)}{N}\ . \tag{4.18}$$

Finally, in order to construct a fundamental quantity that is independent of the arbitrarily chosen partition $\{A_R\}$, we take the supreme with respect to all possible partitions

$$KS = \sup_{\{A_R\}} h_{\{A_R\}}. \tag{4.19}$$

The value KS is the Kolmogorov-Sinai entropy of the symbolic sequence $\{x_i\}$. The Kolmogorov-Sinai entropy is independent of N and $\{A_R\}$, however, it may depend on the probability distribution of the initial values $\sigma(x)$ that defines the probability (4.16).

As a simple example, let us suppose having a sequence x_i made with only two symbols, $A_1 = +1$ and $A_2 = -1$. If the sequence is perfectly random, for any N, all possible sample sequences $\{\omega_N\}$ have the same probability given by

$$p(\omega_0,\ \omega_2,\ ...,\ \omega_{N-1}) = \frac{1}{2^N}\ . \tag{4.20}$$

By using (4.17) and (4.18), it is easy to prove that the Kolmogorov-Sinai entropy of

this random sequence is given by $KS = \ln 2$.

The Kolmogorov-Sinai entropy may have many applications. It may be used for studying the statistical properties of a map. In this case, the analysis is made by measuring the Kolmogorov-Sinai entropy of entire sequences of iterates of a map. For getting a value independent from the probability distribution of the initial values $\sigma(x)$, the invariant distribution $\mu(x)$ of the map should be used as initial value. The partition $\{A_R\}$ may be done, for example, by dividing the map space in R cells of size $\epsilon = 1/R$. If x_n is the n-th iterate of a map and it belongs to the cell A_i of the partition, x_n is associated to the symbol that characterizes the cell A_i. In this way a map is transformed into a dynamical symbolic sequence whose entropic information is measured by the Kolmogorov-Sinai entropy. For example, in the case of the Bernoulli shift map defined by

$$x_{n+1} = f(x_n) = \begin{cases} 2x_n & for\ x_n \in [0, 0.5] \\ 2x_n - 1 & for\ x_n \in [0.5, 1] \end{cases}, \tag{4.21}$$

it is possible to generate a symbolic sequence of only two symbols $A_1 = +1$ and $A_2 = -1$ according to whether the n-th iterate of the map x_n belongs to the interval $[0, 0.5]$ or to the interval $[0.5, 1]$. The Bernoulli shift map is equivalent to a sequence of symbols $A_1 = +1$ and $A_2 = -1$ chosen randomly. Its Kolmogorov-Sinai entropy is $KS = \ln 2$.

The Kolmogorov-Sinai entropy is also important because it is simply related to the Lyapunov exponent λ of a map. This relation is given by

$$K = \lambda - k, \tag{4.22}$$

where k is the escape rate of a one-dimensional expanding map. If $k = 0$, the Lyapunov exponent coincides with the Kolmogorov-Sinai entropy [116].

The main problem with the Kolmogorov-Sinai entropy is computational. It may be very hard to measure the KS entropic density. The Kolmogorov-Sinai entropy uses an analysis that requires long sample sequences $\{\omega_N\}$. This creates huge problems for a computer because the number of probabilities (4.16) that must be calculated

for each N, increases exponentially with N. Even if only two symbols, $+1$ and -1, are used to create the symbolic sequence, the number of possible sample sequences of N symbols is 2^N. A value of N relatively small is enough to saturate also the most powerful computer! To solve the problem, recourse may be made to some compression algorithm [120] that reduces drastically the need of a computer resources. In this dissertation we focus on the Diffusion entropy Analysis that is a powerful alternative to the Kolmogorov-Sinai entropy analysis for studying large trajectories in the long-time limit.

4.4 Diffusion Entropy Analysis.

In this section, we finally address the main topic of this dissertation, the Diffusion Entropy Analysis (DEA). The purpose of the DEA algorithm is to establish the possible existence of scaling, either normal or anomalous, in the most efficient way as possible without altering the data with any form of detrending. The existence of scaling implies the existence of a pdf $p(x,t)$ that scales according to the equation

$$p(x,t) = \frac{1}{t^\delta} \cdot F\left(\frac{x}{t^\delta}\right), \tag{4.23}$$

where δ is the pdf scaling exponent.

Because we want to study the statistical properties of a time series, first we need an algorithm to obtain a pdf of a time series. Therefore, as we already did in section 3.1, let us consider a time series $\{\xi_i\}$ of N data:

$$\xi_1,\ \xi_2,\ \xi_3,\ \xi_4,\ \xi_5,\ \xi_6,\ \xi_7,\ \xi_8,\ \xi_9,\ \xi_{10},\ \xi_{11},\ \xi_{12},\ ...,\ \xi_{N-1},\ \xi_N\ . \tag{4.24}$$

Let us select first of all an integer number t, fitting the condition $1 \leq t \leq N$. This integer number will be referred to by us as "time". For any given time t, we can find $N - t + 1$ sub-sequences defined by

$$\xi_i^{(s)} \equiv \xi_{i+s}, \qquad s = 0, \ldots, N - t. \tag{4.25}$$

For any of these sub-sequences we build up a diffusion trajectory, labeled with the index s, defined by the position

$$x^{(s)}(t) = \sum_{i=1}^{t} \xi_i^{(s)} = \sum_{i=1}^{t} \xi_{i+s}. \tag{4.26}$$

Let us imagine this position refers to a Brownian like particle that at regular intervals of time jumps been jumping forward or backward according to the prescription of the corresponding sub-sequence of Eq. (4.25).

At each time t, it is possible to estimate a pdf $p(x,t)$ that will be used to evaluate the entropy of this diffusion process. To do that, we have to partition the x-axis into cells of size $\epsilon(t)$. When this partition is made, we have to label the cells. We count how many particles are found in each cell at a given time l. We denote this number by $N_i(t)$. Then, we use this number to determine the probability that a particle can be found in the i-th cell at time l, $p_i(t)$, by means of

$$p_i(l) \equiv \frac{N_i(t)}{(N - t + 1)}. \tag{4.27}$$

At this stage the entropy of the diffusion process at time t is determined and reads

$$S_d(t) = -\sum_i p_i(t) \ \ln[p_i(t)]. \tag{4.28}$$

The easiest way to proceed with the choice of the cell size, $\epsilon(t)$, is to assume it to be independent of t and determined by a suitable fraction of the square root of the variance of the fluctuation ξ_i. In the case in which the numbers ξ_i are +1, 0 and -1, $\epsilon = 1$ is the natural choice.

The method we are adopting is based on the idea of a moving window of size t that makes the $s - th$ trajectory closely correlated to the next, the $(s+1) - th$ trajectory. The two trajectories have $t - 1$ values in common. The motivation for using overlapping windows, with the DEA method, is given by our wish to establish a connection with the Kolmogorov-Sinai entropy described in Sec. 4.3. In fact, the Kolmogorov-Sinai entropy of a symbolic sequence is evaluated by moving a window

of size N along the sequence. Any window position corresponds to a given combination of symbols, and from the frequency of each combination it is possible to derive the Shannon entropy $S(N)$ as shown in Eq. (4.17). Moreover, we use overlapping windows because in this way the number of available trajectories is much higher than that can be obtained by using non-overlapping windows. The non-overlapping window technique is adopted by the DEA discussed in Sec. 3.3. The large number of trajectories generated with overlapping windows is fundamental because to derive the pdf of the diffusion process, we need enough trajectories to calculate a frequency distribution (4.27) that must correspond to the real probability distribution.

Let us consider the simplifying assumption of considering large enough times as to make the continuous assumption valid. Therefore, let us adopt the continuous version, valid for $t \gg 1$, of the Shannon entropy that reads

$$S(t) = -\int_{-\infty}^{\infty} dx\ p(x,t) \ln[p(x,t)]. \tag{4.29}$$

We also assume that

$$p(x,t) = \frac{1}{t^{\delta}}\ F\left(\frac{x}{t^{\delta}}\right) \tag{4.30}$$

and that $F(y)$ maintains its form, namely that the statistics of the process are independent of time. Let us plug Eq.(4.30) into Eq. (4.29). Using a simple algebra, we get the fundamental relation:

$$S(\tau) = A + \delta\ \tau, \tag{4.31}$$

where

$$A \equiv -\int_{-\infty}^{\infty} dy\ F(y)\ \ln[F(y)] \tag{4.32}$$

and

$$\tau \equiv \ln(t/t_0), \tag{4.33}$$

where t_0 is the unit time of diffusion. t_0 may be neglected with the implicit assumption that t is measured in unit of t_0. Eq. (4.31) is the key relation for understanding how the DEA is used for detecting the pdf scaling exponent δ. After a transition regime that may be less or more extended (this depends on the pdf of the data $\{\xi_i\}$) the DEA

shows an asymptotic behavior from which the parameter δ can be detected. However, the fact that the DEA method can be used as a reliable way to detect scaling only in the long-time limit, may have some limitation. In fact, the real data available are finite, thereby there are saturation effects in the long-time regime. However, also the studying of the transition regime may be used for getting important information. Of course, δ does not depend on the value of t_0.

4.5 Non-stationary dynamical transient analysis.

In the previous section, we saw that the DEA method can be used as a reliable way to detect the pdf scaling exponent δ only in the long-time limit. Normally, the stationary thermodynamical condition is reached only after a non-stationary dynamical transient. Steady thermodynamical condition holds when the statistical properties of the diffusion process is well described by the Central Limit Theorem or by the Generalized Central Limit Theorem. For example, in the normal random walk there is the need of waiting a while before the binomial distribution may be decribed by a Gaussian distribution that fulfills the scaling properties. Another example is if the signal has some periodicities. A diffusion process at times shorter than the characteristic periods of the signal is sensitive to the intensity of the periodicities. If the time series $\{\xi_i\}$ of N data is fractional Brownian noise or is distributed according to a Lévy distribution, the stationary thermodynamical condition is reached after the first step of diffusion. In fact, the first step of diffusion gives the pdf of the data $\{\xi_i\}$.

The non-stationary dynamical transient may be simulated by a non stationary pdf of the type

$$p(x,t) = \frac{1}{t^{\delta(t)}}\ F\left(\frac{x}{t^{\delta(t)}}\right), \tag{4.34}$$

where the pdf scaling exponent $\delta(t)$ changes with the diffusion time t. Let us suppose that

$$\delta(t) = \delta_0 + \eta \ln(t). \tag{4.35}$$

Since the scaling parameter δ cannot exceed the ballistic value $\delta = 1$ in the case of a dynamical approach to diffusion with fluctuation of limited intensity, this condition

applies to the time scale defined by

$$\eta \ln(t) < 1 - \delta_0. \tag{4.36}$$

At this stage it easy for us to show the convenience of adopting the non-extensive entropy indicator advocated some years ago by Tsallis [118]. First of all, we notice that in the new non-stationary condition the traditional entropy indicator yields:

$$S(\tau) = A + \delta_0 \tau + \eta \tau^2, \tag{4.37}$$

where, $\tau = \ln t$. According to the Tsallis picture, the entropy undergoes a regime of linear increase in time (the time t rather than the logarithmic time τ) if an entropic index $Q \neq 1$ is adopted. For values of $q < Q$ the entropy $S_q(t)$ undergoes an increase with time faster than the regime of linear increase in time, and for $q > Q$ the entropic increase is slower. On the basis of this result we make the conjecture that in the diffusion regime a linear dependence on τ is recovered if an entropic index $q > 1$ or $q < 1$ is adopted.

Let us see all this in detail. The non-extensive Tsallis indicator [118] reads in the continuous formalism

$$S_q(t) = \frac{1 - \int_{-\infty}^{+\infty} dx \; [p(x,t)]^q}{q - 1}. \tag{4.38}$$

It is straightforward to prove that this entropic indicator coincides with that of Eq.(4.23) when the entropic index q gets the ordinary value $q = 1$. Let us make the assumption that in the diffusion regime the departure from this traditional value is very weak. This can be quantitatively expressed as follows. Let us define first

$$\epsilon \equiv q - 1. \tag{4.39}$$

We make the assumption that

$$\epsilon << 1, \tag{4.40}$$

which, as we shall see, is fulfilled by the process under study in this paper. This

allows us to use the following approximated expression for the non-extensive entropy

$$\begin{aligned} S_{1+\epsilon}(t) &= -\int_{-\infty}^{+\infty} dx\ p(x,t)\ln(p(x,t)) \\ &\quad -\frac{\epsilon}{2}\int_{-\infty}^{+\infty} dx\ p(x,t)\left[\ln(p(x,t))\right]^2 . \end{aligned} \tag{4.41}$$

In the specific case where the non stationary condition of Eq. (4.34) applies, this entropy gets the form

$$\begin{aligned} S_{1+\epsilon} &= A+\delta(t)\ln(t) \\ &\quad -\epsilon\ B-\epsilon\ \delta(t)\ln(t)A-\frac{\epsilon}{2}\ \delta(t)^2\left[\ln(t)\right]^2 , \end{aligned} \tag{4.42}$$

where

$$B \equiv \frac{3}{8}+\frac{1}{4}\ln(2\pi\sigma^2)+\frac{1}{8}\left[\ln(2\pi\sigma^2)\right]^2 . \tag{4.43}$$

It is straightforward to show that the regime of linear increase in time is recovered when ϵ is assigned the value

$$\epsilon = \frac{\eta}{\delta_0^2/2+\eta A}. \tag{4.44}$$

These theoretical remarks demonstrate that this non-extensive approach to the diffusion entropy makes it possible to detect the strength of the deviation from the steady condition. In fact Eq. (4.44) proves that $\epsilon = 0$ implies a steady condition. It is evident that the measure of the departure from the steady condition is given by

$$\eta = \frac{1}{2}\frac{\epsilon\delta_0^2}{1-\epsilon A}. \tag{4.45}$$

This Section shows that the breakdown of the stationary property of Eq. (4.23) can be revealed by the Diffusion Entropy Analysis under the form of an entropic index q slightly departing from the condition of ordinary statistical mechanics, namely $q = 1$.

Fig. 4.1 shows the effect of the entropic index q upon the entropy of the diffusion. The entropic index q has the effect to bend the curve. If $q > 1$ the curve becomes more convex; if $q < 1$ the curve becomes more concave. This effect is illustrated in

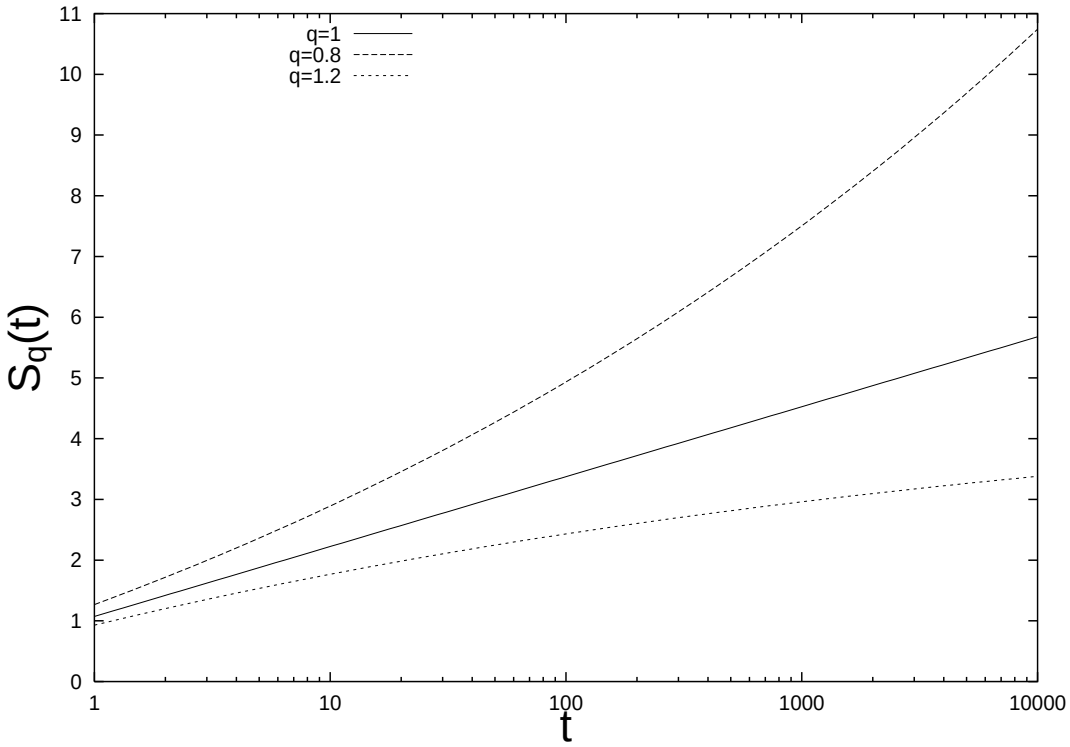

Figure 4.1: Diffusion Entropy by using the non-extensive Tsallis entropy equation (4.38). The dashed line is for $q = 0.8$, the solid line is for $q = 1$, and the dotted line is for $q = 1.2$. The figure shows the bending due to the adoption of $q \neq 1$.

Fig. 4.1 by adopting the following Brownian diffusion equation:

$$p(x,t) = \frac{1}{\sqrt{\pi t}} \exp\left(-\frac{x^2}{t}\right) \ . \tag{4.46}$$

By adopting the non-extensive Tsallis entropy equation (4.38), we get

$$S_q(t) = \begin{cases} (1 - \pi^{0.5(1-q)}\ q^{-0.5}\ t^{0.5(1-q)})/(q-1) & q \neq 1 \\ 0.5 + 0.5\log(\pi t) & q = 1\ . \end{cases} \tag{4.47}$$

In Fig. 4.1, the curves corresponding to $q = 0.8$, $q = 1$ and $q = 1.2$ are plotted.

CHAPTER 5

ARTIFICIAL SEQUENCE ANALYSIS.

In this chapter, we verify the theoretical predictions of Chapter 2 about the pdf scaling exponent δ and the variance scaling exponent, H, by using artificial sequences. we compare the Variance Scaling Analysis with the Diffusion Entropy Analysis.

In Chapter 2, we saw that the Fractional Brownian Motion, Sec. 2.3, is characterized by the fact that the pdf scaling exponent δ and the variance scaling exponent, H, coincide. In the case of Lévy flights, Sec. 2.5, the variance scaling exponent, H, cannot be defined. However, the pdf scaling exponent δ can be measured with the Diffusion Entropy Analysis. In the case of Lévy walks, Sec. 2.6, both H and δ can be measured but they do not coincide and are related one to the other via the following expression

$$\delta = \frac{1}{3-2H} \ . \tag{5.1}$$

We describe the algorithms for generating the artificial sequences. We also analyze an intermittent dynamical model base upon the Manneville map [121] that may be used to produce time series of rare events, that is, a sequence of data $\{\xi_i\}$ distributed according to an inverse power law distribution. These types of sequences may be used to produce Lévy diffusion processes.

All analyses are made by using the following diffusion method. Given a time series $\{\xi_i\}$ of N data:

$$\xi_1,\ \xi_2,\ \xi_3,\ \xi_4,\ \xi_5,\ \xi_6,\ \xi_7,\ \xi_8,\ \xi_9,\ \xi_{10},\ \xi_{11},\ \xi_{12},\ ...,\ \xi_{N-1},\ \xi_N \ . \tag{5.2}$$

we select first of all an integer number t, fitting the condition $1 \leq t \leq N$. This integer number will be referred to by us as "time." For any given time t, we can find $N-t+1$ sub-sequences defined by

$$\xi_i^{(s)} \equiv \xi_{i+s}, \qquad s = 0, \ldots, N-t. \tag{5.3}$$

For any of these sub-sequences, we build up a diffusion trajectory, labelled with the index s, defined by the position

$$x^{(s)}(t) = \sum_{i=1}^{t} \xi_i^{(s)} = \sum_{i=1}^{t} \xi_{i+s}. \tag{5.4}$$

The statistics of the positions $x^{(s)}(t)$ are analyzed by using the basic algorithm for detecting the variance scaling exponent H, Sec. 3.1, and the DEA algorithm for measuring the pdf scaling coefficient, δ, Sec. 4.4.

5.1 Fractional Brownian diffusion.

Fractional Brownian diffusion is produced by Fractional Brownian noise. We generate a time series $\{\xi_{H,t}\}$ of N data by using the algorithm by Mandelbrot that can be found in the book of Feder [70]. Chosen a value of $H \in [0:1]$, let $\{\theta_i\}$ be a set of Gaussian random variables with unit variance and zero mean. The discrete fractional Brownian increment is given by

$$\begin{aligned} \xi_{H,t} = x_H(t) - x_H(t-1) \;\; = \;\; & \frac{n^{-H}}{\Gamma(H+0.5)} \left\{ \sum_{i=1}^{n} i^{H-0.5}\; \theta_{1+n(M+t)-i} + \right. \\ & \left. \sum_{i=1}^{n(M-1)} \left((n+i)^{H-0.5} - i^{H-0.5} \right)\; \theta_{1+n(M-1+t)-i} \right\}, \end{aligned}$$

where M is an integer that should be moderately large, and n indicates the number of the fractional steps for each unit time. Note that the time t is discrete. In the simulation, good results are obtained with $M = 1000$ and $n = 10$. The time series $\{\xi_{H,t}\}$ is then used for generating a diffusion process with the trajectories (5.4).

Fig. (5.1) shows 5000 data generated by the fractional Brownian algorithm with $H = 0.5$. The data $\{\xi_{0.5,t}\}$ are distributed according to a Gaussian distribution with a variance $\sigma^2 = 1$. Fig. (5.2) shows a sample of diffusion trajectories defined by the position $x^{(s)}(t)$ of Eq. (5.4). Ten trajectories are plotted. At $t = 0$, all trajectories depart from the position $x^{(s)}(0) = 0$. As the time t increases, the trajectories give origin to the classical diffusion picture.

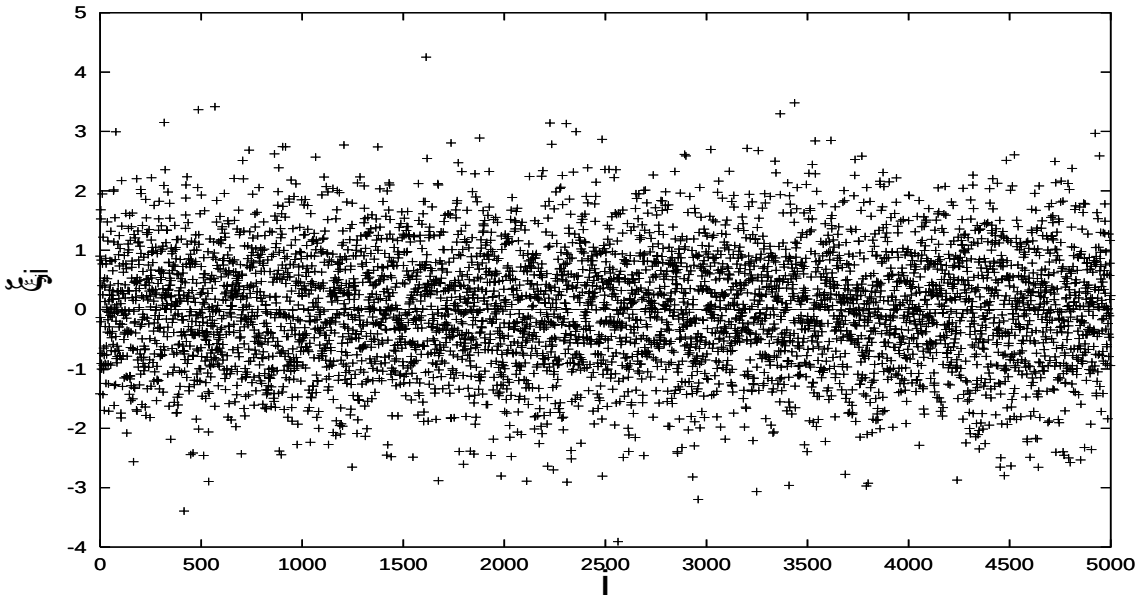

Figure 5.1: Brownian Noise. $H = 0.5$, variance $\sigma^2 = 1$.

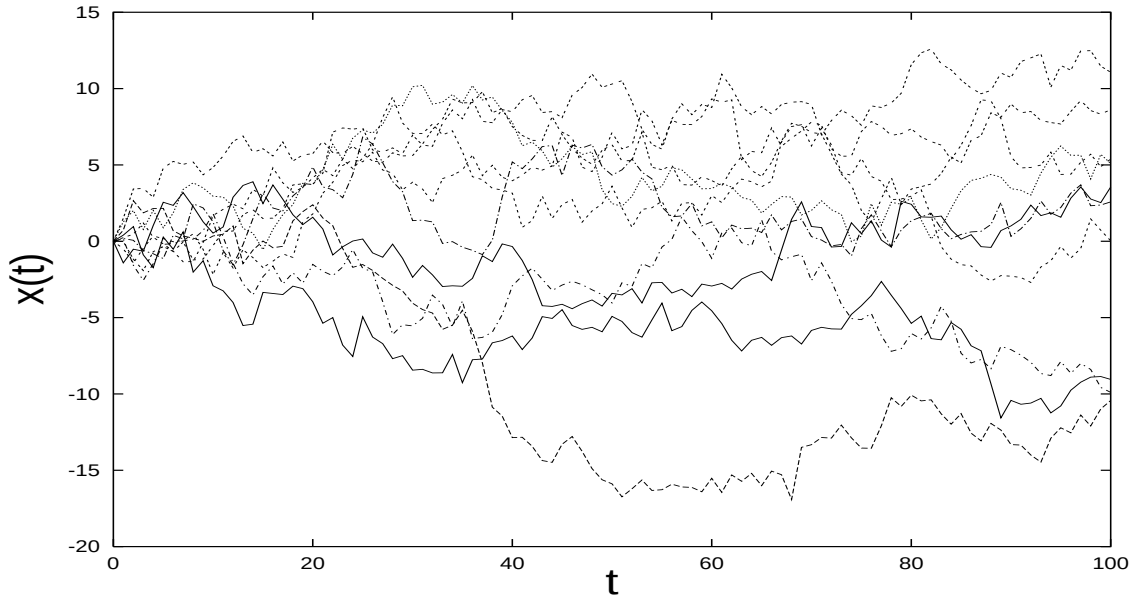

Figure 5.2: Brownian diffusion generated by the trajectories (5.4) and using the data plotted in Fig. (5.1). Only ten trajectories are plotted.

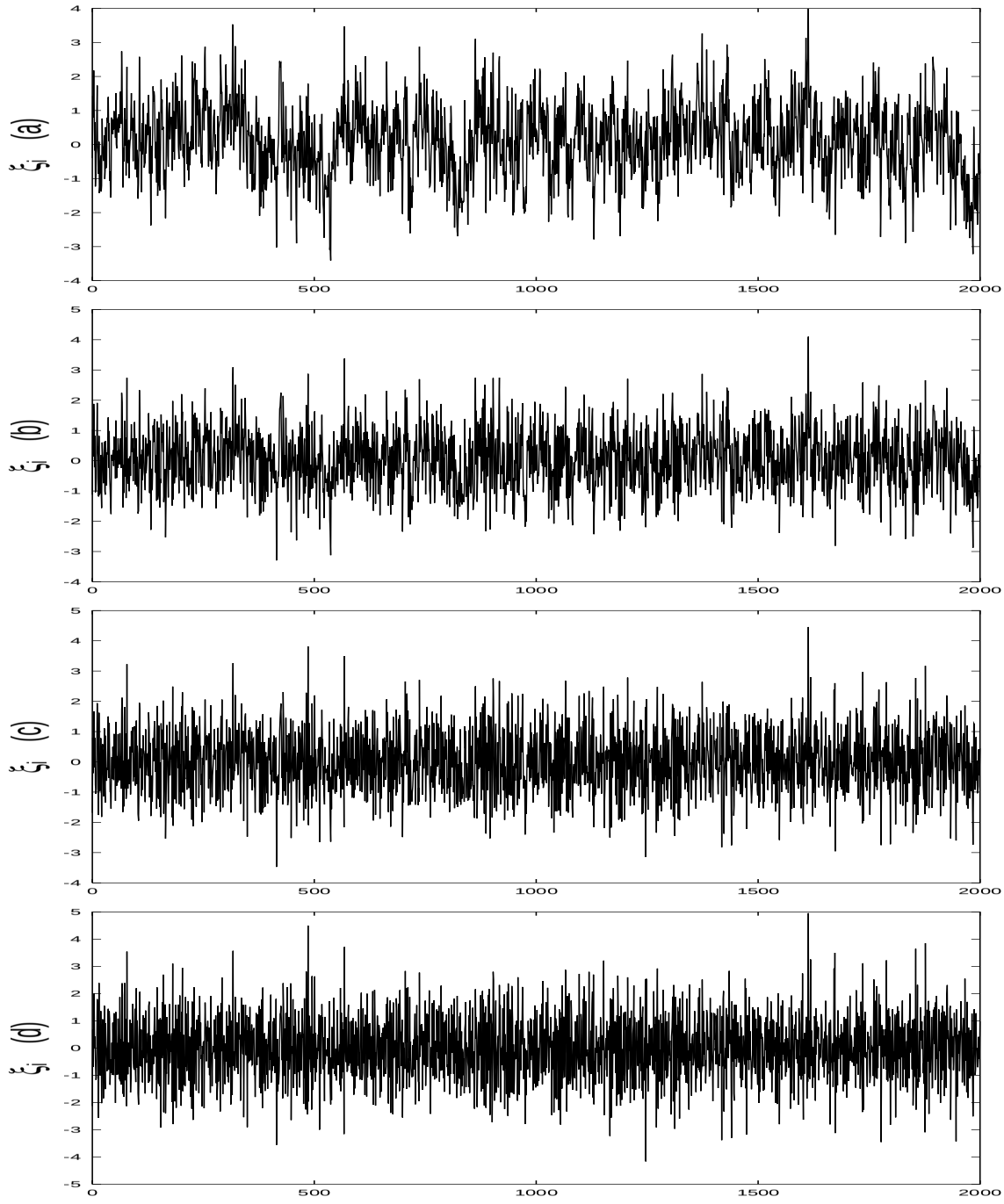

Figure 5.3: Fractional Brownian Noise. (a) $H = 0.8$, (b) $H = 0.6$, (c) $H = 0.4$, (d) $H = 0.2$. Fig. (a) and (b) show persistence, Fig. (c) and (d) show antipersistence.

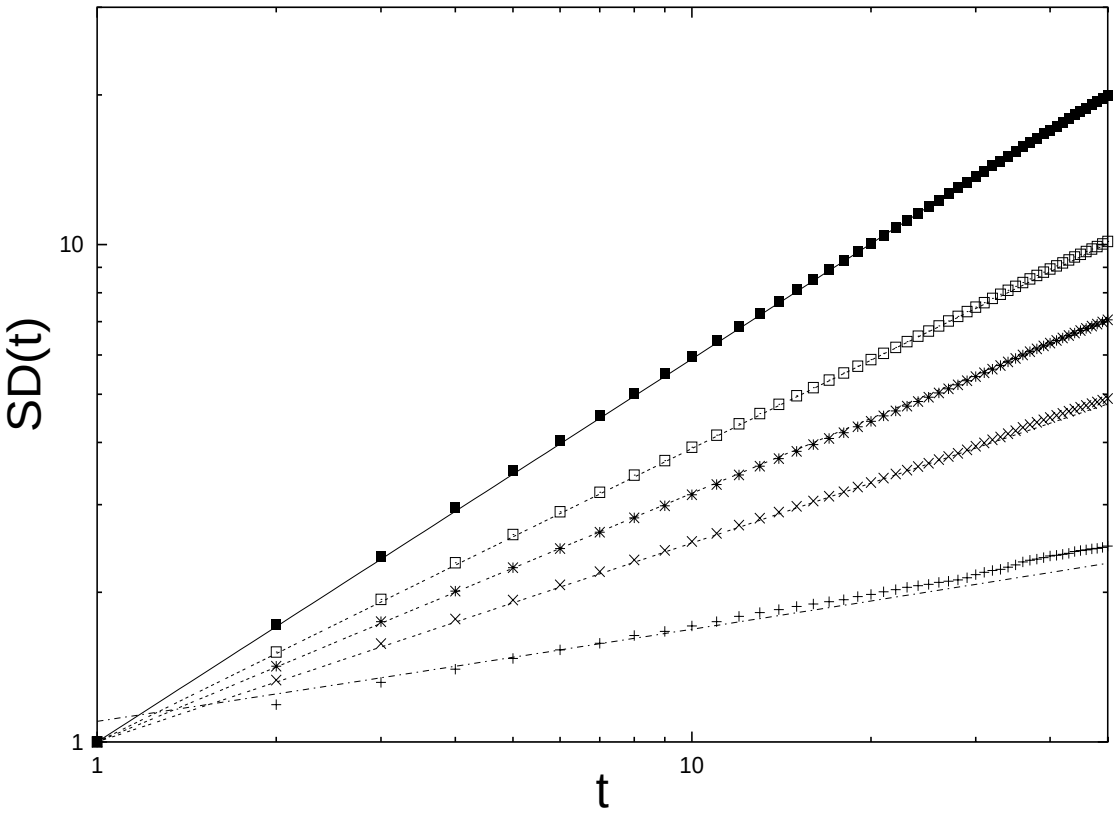

Figure 5.4: Variance scaling analysis. In ordinate it is plotted the standard deviation $SD(t)$ of the positions of the diffusion trajectories in function of the diffusion time t. The data shown in Fig. (5.1) and in Fig. (5.3) are used. The straight lines correspond to the right scaling exponent H for each set of data: from up to down: (1) $H = 0.8$, (2) $H = 0.6$, (3) $H = 0.5$, (4) $H = 0.4$, (5) $H = 0.2$.

Figs. (5.3, a-d) show 2000 data generated by the fractional Brownian algorithm with four different values of H: (a) $H = 0.8$, (b) $H = 0.6$, (c) $H = 0.4$, (d) $H = 0.2$. Fig. (a) and (b) show a noise with a persistence behavior, Fig. (c) and (d) show a noise with antipersistence behavior.

Fig. (5.4) shows the variance scaling analysis of the data shown in Fig. (5.1) and in Fig. (5.3). For convenience, in ordinate we plot the normalized standard deviation $SD(t)$ of the positions of the diffusion trajectories as a function of the diffusion time

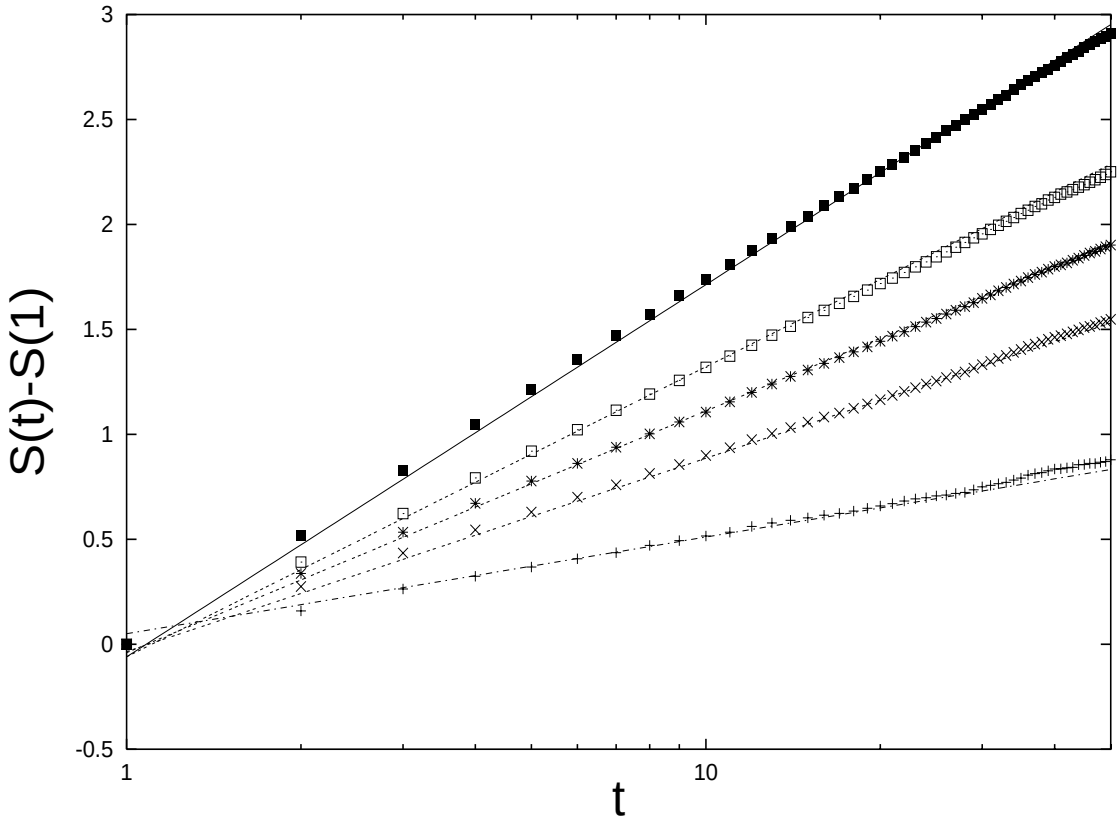

Figure 5.5: Diffusion Entropy analysis. For convenience, in ordinate it is plotted the entropy $S(t) - S(1)$ of the pdf due to the positions of the diffusion trajectories in function of the diffusion time t. At $t = 1$ all curves start from the same ordinate position $= 0$. The data shown in Fig. (5.1) and in Fig. (5.3) are used. The straight lines correspond to the right scaling exponent δ for each set of data: from top to bottom : (1) $\delta = 0.8$, (2) $\delta = 0.6$, (3) $\delta = 0.5$, (4) $\delta = 0.4$, (5) $\delta = 0.2$.

t. The normalized standard deviation $SD(t)$ is related to the variance $var(t)$ by

$$SD(t) = \sqrt{\frac{var(t)}{var(1)}} \ . \tag{5.5}$$

With this choice of $SD(t)$, all curves start from the same ordinate position = 1. The straight lines correspond to the right scaling exponent H for each set of data, that is, from top to bottom: (1) $H = 0.8$, (2) $H = 0.6$, (3) $H = 0.5$, (4) $H = 0.4$, (5) $H = 0.2$. The variance scaling exponent H is defined by

$$SD(t) = t^H \ . \tag{5.6}$$

Fig. (5.5) shows the Diffusion Entropy analysis of the data shown in Fig. (5.1) and in Fig. (5.3). For convenience, in ordinate we plot the entropy

$$S(t) - S(1) \tag{5.7}$$

of the pdf due to the positions of the diffusion trajectories as a function of the diffusion time t. At $t = 1$ all curves start from the same ordinate position = 0. The straight lines correspond to the right scaling exponent δ for each set of data: from top to bottom: (1) $\delta = 0.8$, (2) $\delta = 0.6$, (3) $\delta = 0.5$, (4) $\delta = 0.4$, (5) $\delta = 0.3$. The pdf scaling exponent δ is defined by

$$S(t) - S(1) = \delta \ \ln(t) \ . \tag{5.8}$$

Fig. (5.4) and Fig. (5.5) show clearly that for Fractional Brownian Motion the variance scaling exponent H and the pdf scaling exponent δ coincides. The scaling may not start from the first steps of diffusion because of a transition region due to the artificial discretization introduced by the cell size ϵ used for estimating the diffusion pdf.

5.2 Manneville Map: an intermittent dynamical model for time series of rare events.

In Chapter 2, we analyzed anomalous diffusion processes that cannot be described by fractional Brownian motion that, as we saw in the previous paragraph, are characterized by a variance scaling exponent, H, that coincides with the pdf scaling exponent δ. In Chapter 2, we analyzed Lévy flights (LJM) and walks (SVM), and diffusion models characterized by long rests (SJM and AJM). In this cases, usually, the variance scaling exponent, H, does not coincide with the pdf scaling exponent δ. In the case on Lévy flights it is not possible to define H because the variance diverges. All these diffusion models rest upon the fact that the data are distributed according to an inverse power law pdf of the type

$$\psi(y) = \frac{(\mu - 1)\ T^{\mu-1}}{(T + y)^{\mu}}, \tag{5.9}$$

where T is a positive value and the exponent $\mu \in [2 : 3]$. With this choice of the exponent μ, the second moment of $\psi(y)$ diverges. For the SVM, AJM and SJM, the pdf (5.9) may be considered equivalent to an inverse power-law distribution of waiting times. If we adopt the LJM, the pdf (5.9) may be considered the distribution of the flights. The GCLT assures that a diffusion processes generated by long flights distributed according to Eq. (5.9) converges to a Lévy distribution.

In this paragraph, we introduce an intermittent dynamical model for time series distributed according to Eq. (5.9) [6]. The dynamical model is based on the Manneville Map [121] that has been widely used in the recent past to derive Lévy and describe turbulence processes [122, 123, 124, 144].

The Manneville map reads:

$$x_{n+1} = \Phi(x_n) = x_n + x_n^z \quad (mod.\ 1) \quad (1 \leq z). \tag{5.10}$$

We note that at $z = 1$ the Manneville map becomes equivalent to the Bernoulli shift map used by Zaslavsly [126] to prove that in the case of fully chaotic systems the distribution of the Poincaré recurrence times is an exponential whose decay rate is

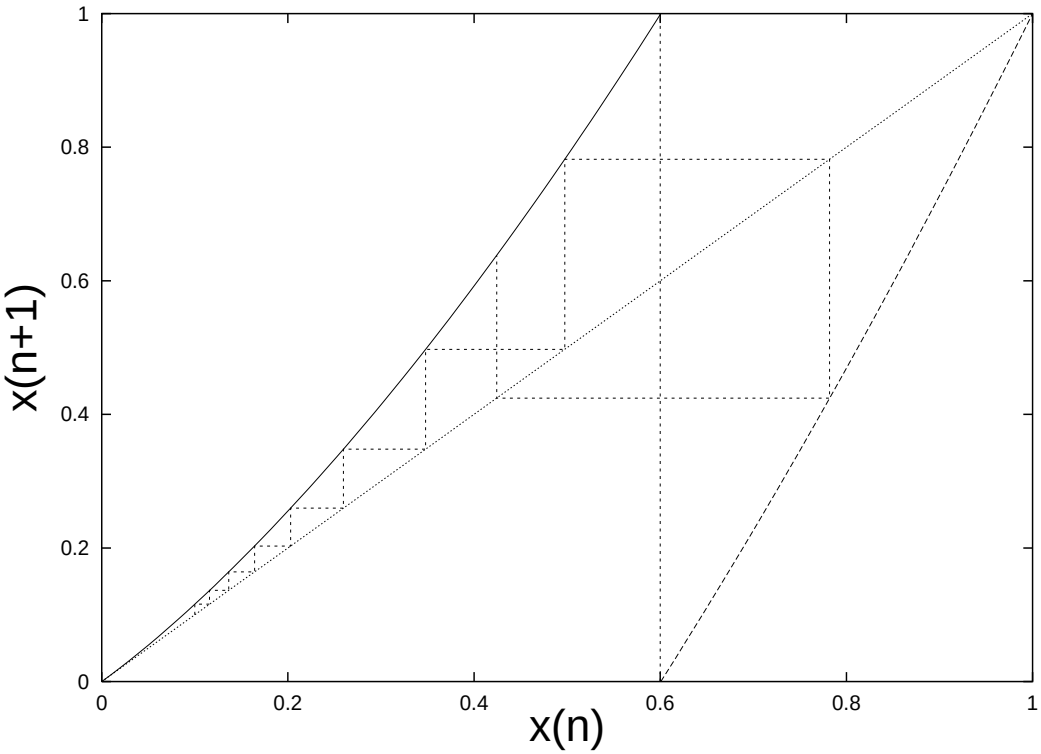

Figure 5.6: The Manneville Map. $z = 1.8$, $d(z) = 0.6$. The laminar region, $[0, d(z)]$, and the chaotic region, $[d(z), 1]$.

the Kolmogorov-Sinai (KS) according to the following prescription:

$$P_R(t) \propto exp(-h_{KS}t) \ . \tag{5.11}$$

Fig. (5.6) shows the Manneville Map. For $z > 1$ the interval $[0, 1]$ is divided in two regions, the laminar region, $[0, d(z)]$, and the chaotic region, $[d(z), 1]$, with $d(z)$ defined by

$$d(z) + d(z)^z = 1 \tag{5.12}$$

We review here the arguments used by Geisel and Thomae [127] to derive an analytical expression for the distribution of the times of sojourn in the laminar region. First of all we assume that the injection point x_0 is so close to $x = 0$ as to replace eq.(5.10)

with:

$$\frac{dx}{dt} = \lambda x^z. \tag{5.13}$$

The coefficient λ can be fixed to 1. Thus we obtain the following time evolution:

$$x(t) = \left[x_0^{1-z} + (1-z)t\right]^{1/(1-z)}. \tag{5.14}$$

Hence the time necessary for the trajectory to get the border $x = d(z)$ is given by

$$t = T(x_0) \equiv \left(\frac{1}{x_0^{z-1}} - \frac{1}{d^{z-1}}\right) \frac{1}{z-1}. \tag{5.15}$$

Note that in the special case where the initial conditon is so close to $x = 0$ as to fulfill the condition $x_0 << d$, the exit time $T(x_0)$ can be satisfactorily approximated by

$$T(x_0) \approx \frac{1}{x_0^{z-1}} \; \frac{1}{z-1}, \tag{5.16}$$

which, as we shall see in Section III, can be used to define the time at which we lose control of the trajectories departing from a region of the map very close to $x = 0$.

Note that the distribution function $\psi(t)$ is related to the injection probability $p(x_0)$ by

$$\psi(t)\, dt = p(x_0)\, dx_0. \tag{5.17}$$

Assuming equiprobability for the injection process we have:

$$p(x_0)\, dx_0 = \frac{1}{d(z)} \left|\frac{dx_0}{dT}\right| \, dT. \tag{5.18}$$

Thus, we obtain:

$$\psi(t) = \frac{1}{d(z)} \left|\frac{dx_0}{dt}\right|. \tag{5.19}$$

By differentiating t with respect to x_0 we finally arrive at

$$\psi(t) = d^{z-1} \left[1 + d^{z-1}(z-1)t\right]^{-z/(z-1)}. \tag{5.20}$$

We observe that the exponent μ of the basic inverse power-law Eq. (5.9) is related to the exponent z of the Manneville map (5.10) via the equation

$$\mu = \frac{z}{z-1} \ . \tag{5.21}$$

It is important to observe that the mean waiting time T_{av} is given by

$$T_{av} = \frac{1}{d^{z-1}} \ \frac{1}{2-z}, \tag{5.22}$$

with $T_{av} = \infty$ for $2 \leq z$.

We see that Eq. (5.20) implies that the region corresponding to $z > 2$ or $\mu < 2$ is characterized by a diverging first moment. This region is in conflict with the Kac theorem [128] and for this reason we do not take it into account. The region pertaining to the interval $1.5 < z < 2$ or $2 < \mu < 3$ is characterized by a finite second moment and a diverging second moment. This is the region of interest for us, since it corresponds to that generating Lévy diffusion according to the recent work of Ref. [129]. The region $1 < z < 1.5$ or $\mu > 3$ is characterized by a finite second moment. Of course, the ideal condition $z = 1$ implies all the moments to be finite. This is a region which would correspond, in the perspective of Ref.[129], to ordinary Gaussian diffusion. We note that in the region $1 < z < 1.5$ the waiting function distribution must make a transition from the inverse power law behavior of Eq.(5.20) to the exponential regime, where the arguments of Zaslavsky leading to Eq.(5.11) apply. In fact, at $z = 1$, the Manneville map becomes identical to the Bernoulli shift map. In that case, the theoretical remarks of Zaslavsky yield:

$$\psi(t) \propto exp(-t \ln 2). \tag{5.23}$$

For computational purposes we found to be more convenient to evaluate the population of the laminar region, $M(t)$, rather than the waiting time distribution $\psi(t)$. The two functions are related the one to the other by

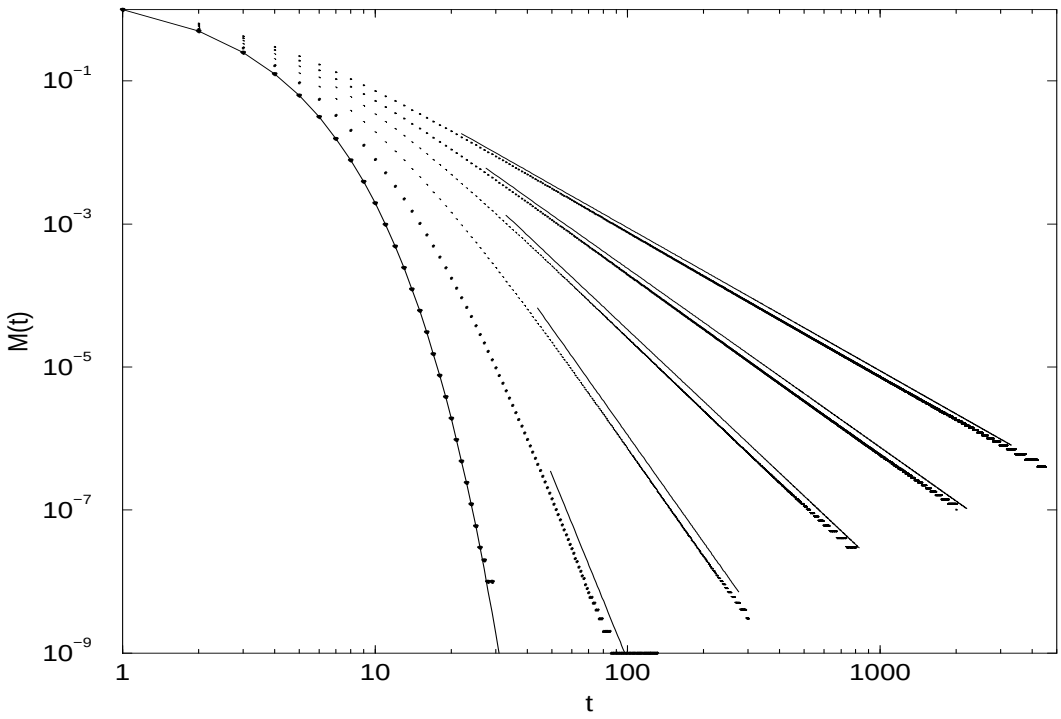

Figure 5.7: M(t) as a function of time. The meaning of the six solid lines is as follows. The lowest solid line is the function $M(t) = exp(-tln2)$. All the other solid lines denote the long-time inverse power law $M(t) = 1/t^{\frac{1}{z-1}}$. The dotted lines are the numerical result. All the full lines but the lowest have been shifted to the right to make them distinguishable from the numerical result. The value of the parameter z, from the bottom to the top is: $z = 1, 1.1, 1.2, 1.3, 1.4, 1.5$.

$$M(t) = 1 - \int_0^t \psi(t')dt'. \tag{5.24}$$

Thus, the analytical expression of Eq.(5.20) yields

$$M(t) = \left[1 + d^{z-1}(z-1)t\right]^{-1/(z-1)}. \tag{5.25}$$

Figs. (5.7) and (5.8) illustrate the results of our numerical treatment. In Fig. (5.7) we illustrate the result of the numerical calculation with the parameter z in the interval $[1, 1.5]$. We see that the long-time limit fits for all values of z considered but $z = 1$, the theoretical prescription of the inverse power law $t^{-\frac{1}{z-1}}$. However,

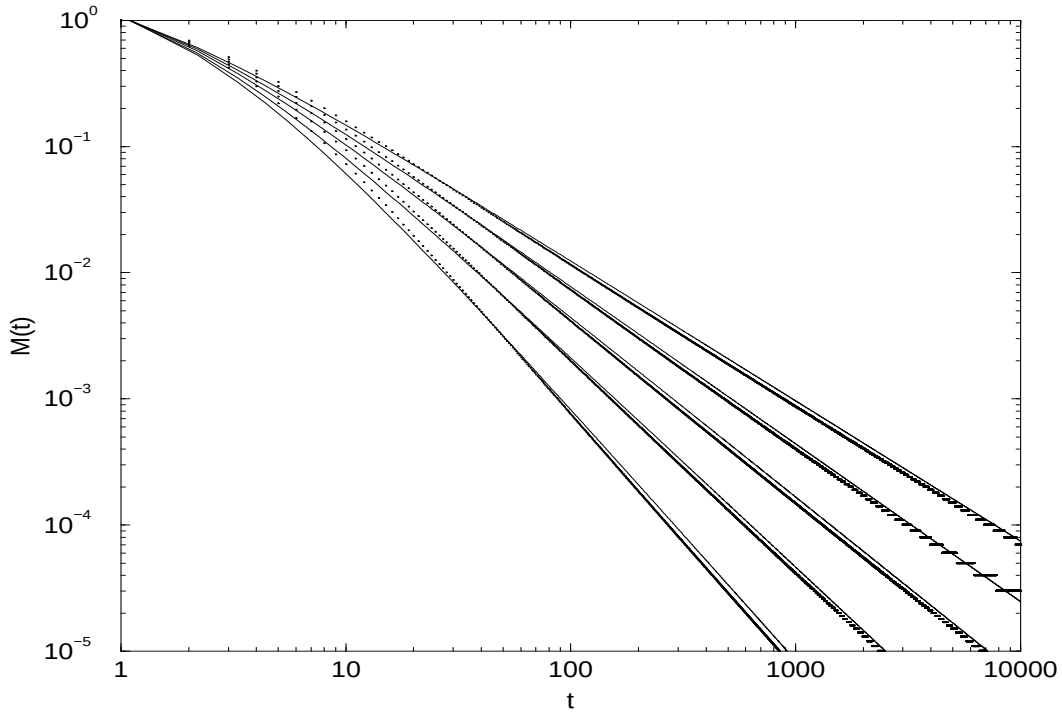

Figure 5.8: M(t) as a function of time. The meaning of the four pairs of lines is as follows. The solid lines denote the function $M(t$ of Eq.(5.25) and the dotted lines denote the numerical results. To make the solid lines distinguishable from the dotted lines we shifted them to the right by the quantity $\epsilon = 0.1$. In the logarithmic representation adopted, this is equivalent to replacing t of $M(t)$ with $texp(-\epsilon)$. The value of the parameter z from the bottom to the top changes as follows: $z = 1.5, 1.6.1.7, 1.8, 1.9$.

this inverse power law regime is reached after an extended transition regime, which seems to be exponential-like. The duration of this transition regime becomes more and more extended with z coming closer and closer to $z = 1$. At $z = 1$ this transition regime becomes infinitely extended and coincident with the theoretical prediction of Eq.(5.23). In Fig. (5.8), devoted to studying the relaxation of the Manneville map with z in the interval $[1.5, 2]$, we see that the prediction of Eq.(5.25) is very accurate and tends to become exact with increasing the values of z.

5.3 Symmetric velocity model simulations.

Section 2.6 was dedicated to the Lévy walk diffusion. In Ref. [111] it is proved that a Lévy walk diffusion may be generated by the so-called Symmetric Velocity Model (SVM). This type of diffusion is characterized by the fact that the walker travels with a constant velocity throughout a whole time interval t chosen from a waiting time pdf of the kind

$$\psi(t) = (\mu - 1)\frac{T^{\mu-1}}{(T+t)^{\mu}} \ . \tag{5.26}$$

At the end of the time interval, the walker may or may not change direction and travel through the next time interval with the same or the opposite constant velocity. We focus our attention upon the interval $2 < \mu < 3$ that is characterized by a finite characteristic waiting time T and by an divergent second moment as required for a Lévy diffusion statistics. Lévy walk diffusion model is more realistic than the Long Jump Model or Lévy flights because it implies that, for making a jump, a walker needs a time proportional to the length of the jump itself. In Section 2.6, we proved that SVM implies a second moment scaling exponent H different from the pdf scaling exponent δ. The two exponents are related to the power exponent μ of Eq. (5.26) according to the following equations

$$H = \frac{4-\mu}{2} \tag{5.27}$$

and

$$\delta = \frac{1}{1-\mu} \ . \tag{5.28}$$

Fig. 5.9 shows δ and H against μ. Eqs. (5.27) and (5.28) implies that Lévy walk diffusion is characterized by an easy relation between the second moment and the pdf scaling exponents:

$$\delta = \frac{1}{3-2H} \ . \tag{5.29}$$

Let us describe how an artificial sequence ξ_i, generating a Lévy walk diffusion process, is generated. To construct such a sequence first of all we need to generate a series of N integer numbers $\{L(i)\}$ according to a probability distribution $p(L)$:

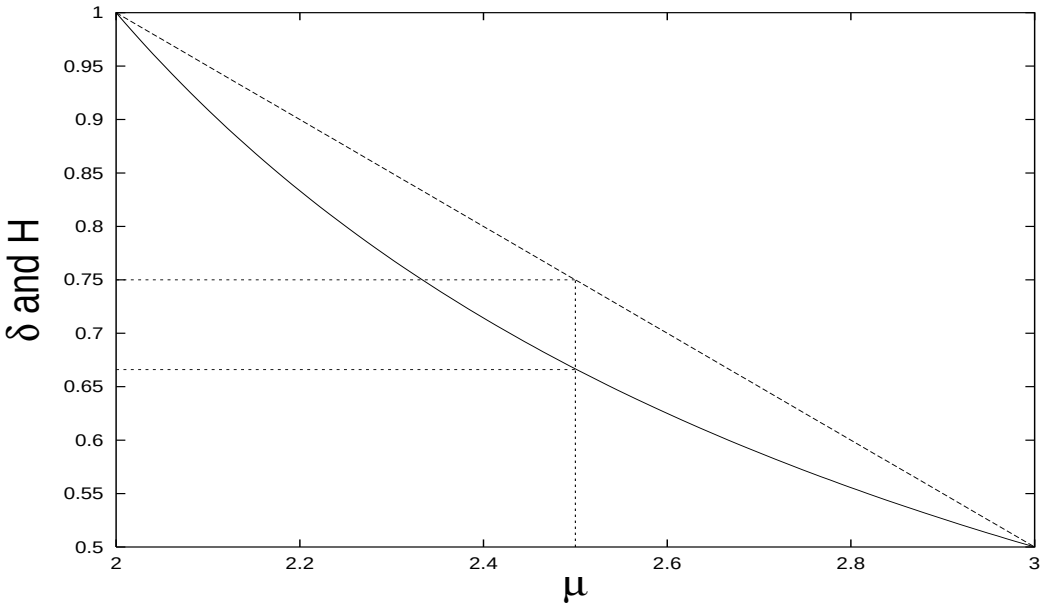

Figure 5.9: δ (solid line) and H (dashed line) against μ. For $\mu = 2.5$, $H = 0.75$ and $\delta = 0.666$.

these numbers can be interpreted as the lengths of strings of the sequence to build up. Then, for any string, we toss a coin and we decide to fill it with $+1$'s or -1's, according to whether we get head or tail. The integer numbers $L(i)$ are distributed according to the following inverse power law pdf:

$$p(L) = \frac{C}{(T+L)^{\mu}}, \tag{5.30}$$

where T and $C = \left(\sum_{L=1}^{\infty} \frac{1}{(T+L)^{\mu}}\right)^{-1}$ are two constants. We focus our attention on the condition $2 < \mu < 3$. This condition is known [130, 131, 132] to yield a Lévy diffusion. A technical method to create a distribution as in equation (5.30) is the following. We divide the interval [0,1] of real numbers into infinite sectors. The L-*th* sector covers the space

$$sector(L) \equiv \left[X(L), X(L) + \frac{C}{(T+L)^{\mu}}\right), \tag{5.31}$$

where

$$X(L) = \begin{cases} 0 & \text{if} \quad L = 1, \\ C \sum_{n=1}^{L-1} 1/(T+n)^{\mu} & \text{if} \quad L > 1. \end{cases} \tag{5.32}$$

The length of the L-*th* sector is equal to the probability $p(L)$ given by the Eq.(5.30). Then, by using a computer, we generate a sequence of rational random numbers $\Upsilon(i)$ uniformly distributed between 0 and 1: if the rational number $\Upsilon(i)$ belongs to the L-*th* sector, the value L will be assigned to the element $L(i)$ of the sequence of integer numbers. The described algorithm and the uniformity of the sequence of rational random numbers $\Upsilon(i)$ assure that the sequence of integer numbers $L(i)$ is distributed exactly according to the power law given by the equation (5.30). An alternative method can be found in Ref. [133]. The illustrated algorithm is based upon the discrete assumption, it is possible to use an algorithm based upon the continuous assumption that the element $L(i)$ is related to the random number $\Upsilon(i)$ by

$$L(i) = Int\left[T\,(y+1)^{-\frac{1}{\mu-1}} - 1\right], \tag{5.33}$$

where Int[y] is the integer part of y. However, in our simulations we use the discrete algorithm because it seems to give more precise results than the continuous one. The only disadvantage in using the discrete algorithm is that it is slower than the continuous one. The discrete algorithm has the advantage of producing a series of $L(i)$ distributed exactly according to the inverse power-law distribution (5.31). After generating the artificial sequence ξ_i, it is possible to apply the procedures exposed in the Chapters 3 and 4 to determine the variance scaling coefficient H via several forms of methods that detect variance scaling, as well as, the pdf scaling exponent δ via the diffusion entropy.

The simulations are made by using 5 million data ξ_i. The sequence $L(i)$ is distributed according to Eq. (5.30) where it is chosen $T = 0$. In fact, T is only a time scaling, and large T have the effect of increasing the transition region before reaching the scaling region. Because we are interested in studying the scaling region, T is fixed as small as possible. We generate five sets of data ξ_i by using five different exponents μ: $\mu = 2.8$, $\mu = 2.6$, $\mu = 2.5$, $\mu = 2.4$, $\mu = 2.2$. Table 5.1 shows the theoretical

μ	2.200	2.400	2.500	2.600	2.800
H	0.900	0.800	0.750	0.700	0.600
δ	0.833	0.714	0.667	0.625	0.556

Table 5.1: Theoretical relation between the waiting time distribution power exponent μ and the variance scaling exponent H and the pdf scaling exponent δ.

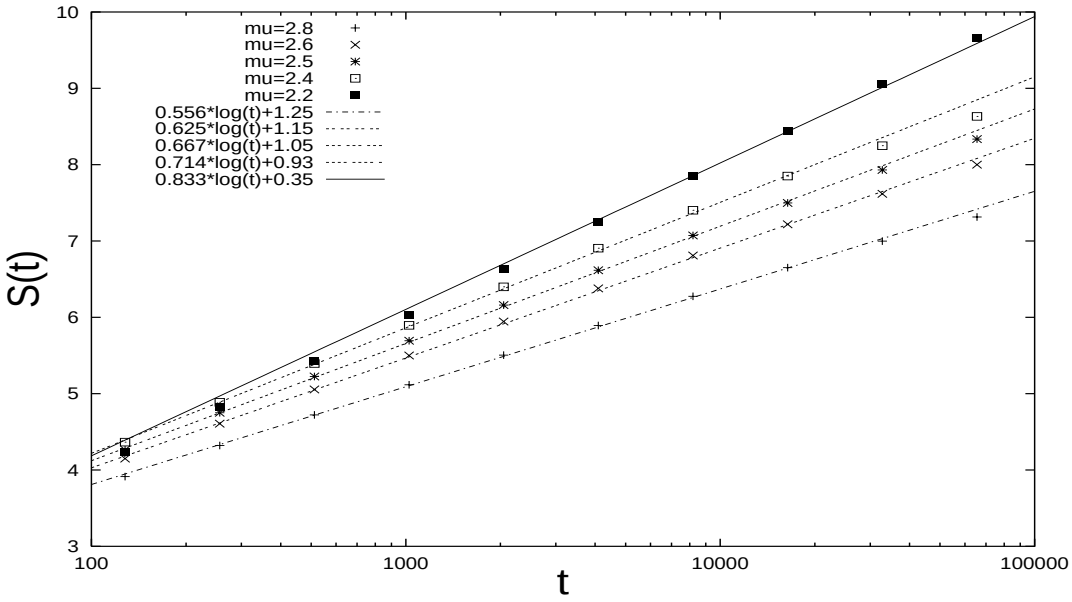

Figure 5.10: Diffusion Entropy Analysis of Lévy walk. Five sets of data corresponding to $\mu = 2.8$, $\mu = 2.6$, $\mu = 2.5$, $\mu = 2.4$, $\mu = 2.2$. The values of the pdf scaling exponent δ are in Table 5.1.

relation between μ and the scaling exponents δ and H.

Fig. 5.10 shows the results obtained by using the diffusion entropy method. The five sets of data scale according to the theoretical values of δ shown in Table 5.1. For large t we note a slight saturation effects due to the fact that the number of data is limited.

Figs. 5.11, 5.12, 5.13, 5.14 show the Hurst Analysis, the Detrended Fluctuation Analysis, Variance Scaling Analysis and the Wavelet Variance Analysis respectively applied to the five sets of data corresponding to $\mu = 2.8$, $\mu = 2.6$, $\mu = 2.5$, $\mu = 2.4$, $\mu = 2.2$. The four pictures show that the four techniques are practically equivalent.

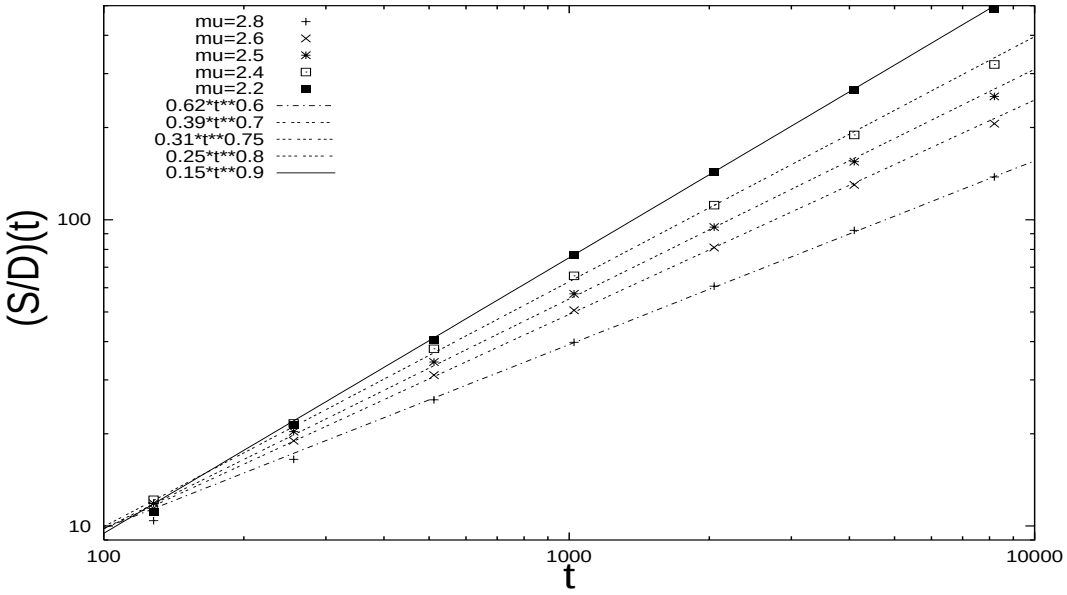

Figure 5.11: Hurst Analysis of Lévy walk. Five sets of data corresponding to $\mu = 2.8$, $\mu = 2.6$, $\mu = 2.5$, $\mu = 2.4$, $\mu = 2.2$. The values of the scaling exponent H are in Table 5.1.

They all give the same scaling exponents. These exponents coincide with the theoretical values of H of Table 5.1. To graphically show the same scaling exponent H, the Variance Scaling Analysis is made by plotting the standard deviation, that is, the square root of the variance. The Wavelet Variance Analysis is made upon the integral of the data series ξ_i. Finally, we plot the square root of the spectral density according to the prescriptions of Section 3.6. The wavelet spectral density is calculated by using the Maximum Overlap Discrete Wavelet Transform [115]. For the calculation the Daubechies H4 discrete wavelet is used .

The five first figures of this section clearly show the efficiency of the Diffusion Entropy Analysis in detecting the real statistical properties of a data set. In fact , all techniques related to the variance –Hurst Analysis, Detrended Fluctuation Analysis, Standard Deviation Analysis, Wavelet Spectral Analysis– are practically equivalent and are able to detect only the second moment scaling exponent H. The problem is that if a data set is analyzed only with one of these techniques based upon the variance and a scaling exponent H is found, the data set may be mistaken for Fractional

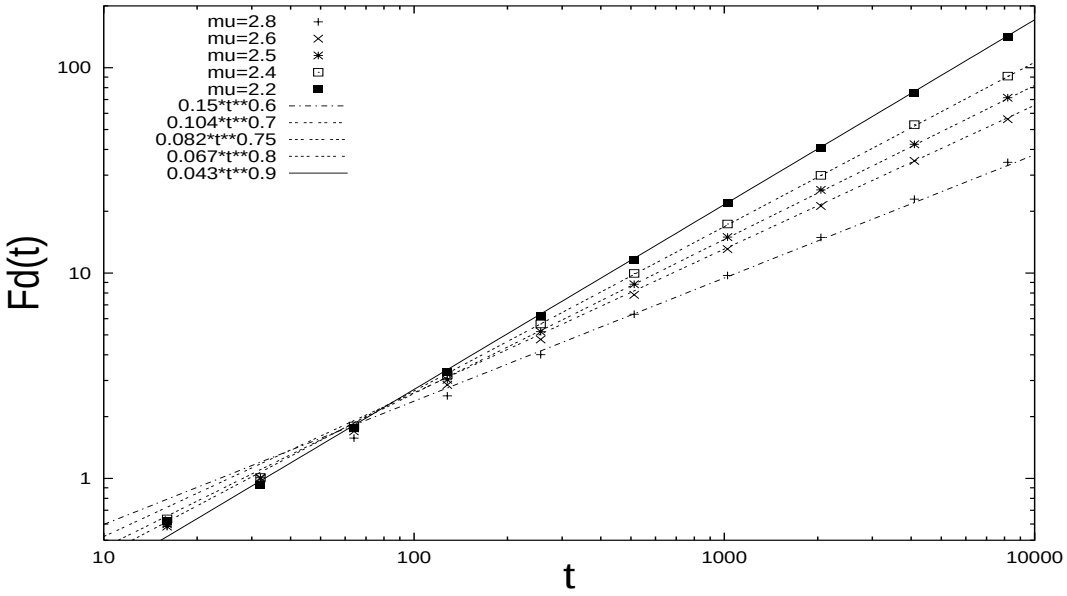

Figure 5.12: Detrended Fluctuation Analysis of Lévy walk. Five sets of data corresponding to $\mu = 2.8$, $\mu = 2.6$, $\mu = 2.5$, $\mu = 2.4$, $\mu = 2.2$. The values of the scaling exponent H are in Table 5.1.

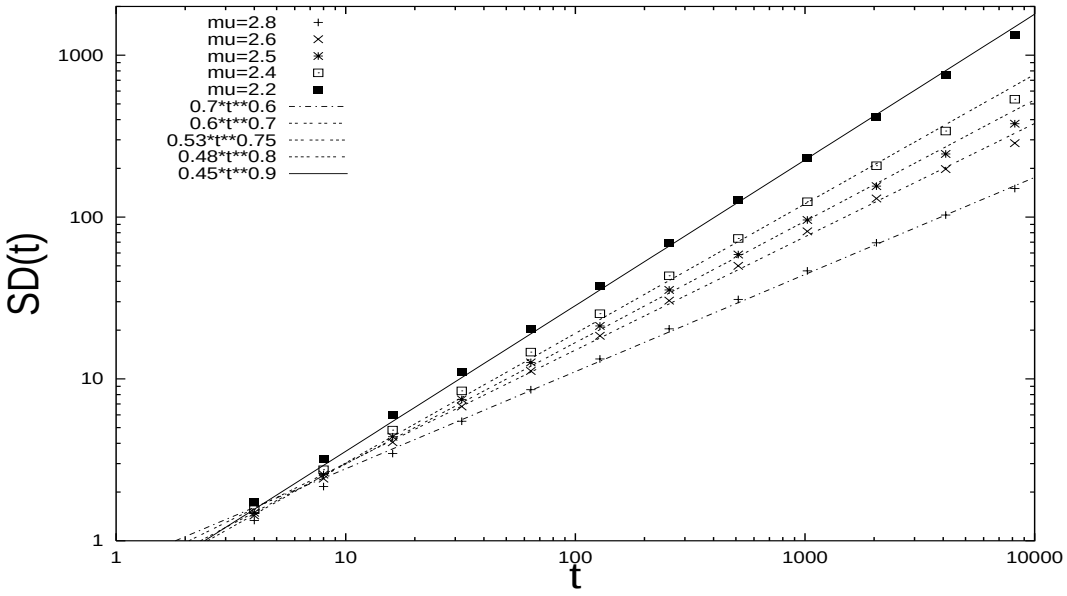

Figure 5.13: Standard Deviation of Lévy walk. Five sets of data corresponding to $\mu = 2.8$, $\mu = 2.6$, $\mu = 2.5$, $\mu = 2.4$, $\mu = 2.2$. The values of the scaling exponent H are in Table 5.1.

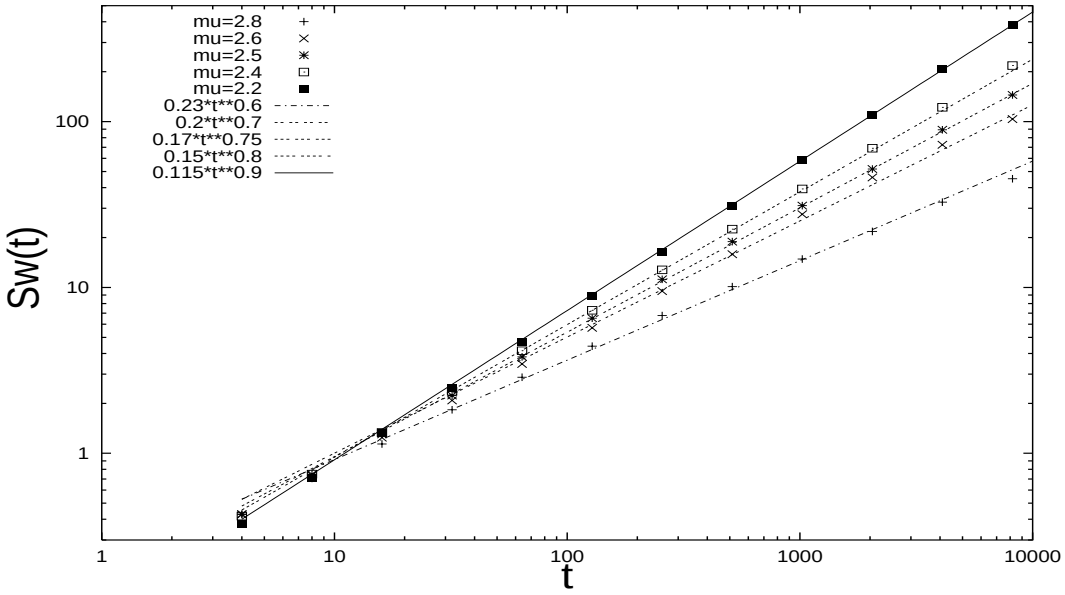

Figure 5.14: Wevelet Variance Analysis of Lévy walk. Five sets of data corresponding to $\mu = 2.8$, $\mu = 2.6$, $\mu = 2.5$, $\mu = 2.4$, $\mu = 2.2$. The values of the scaling exponent H are in Table 5.1.

Brownian noise that is based upon the Gaussian assumption while the truth may be that Lévy and not Gauss statistics characterize the data set. The Diffusion Entropy Analysis can distinguish between the two types of noise. If $\delta = H$ the noise may be Fractional Brownian Noise, whereas if $\delta \neq H$ the noise cannot be Fractional Brownian. If δ and H are related by Eq. (5.29), the data show clear Lévy properties.

The reason is because Diffusion Entropy Analysis detects a scaling different from that revealed by the variance scaling analysis. DEA studies the scaling of the diffusion distribution itself whereas the variance is related to only one of the moments of the distribution. The Gaussian diffusion is characterized by the fact that all moments, $M_\eta(t) =< |x - \overline{x}|^\eta >$, scale with the same exponent of the distribution. A Lévy diffusion has divergent moments. Therefore, it is not true that all moments scale as the distribution does. Fig. 5.15 shows the first moment of the distribution $M_1(t)$ against the diffusion time t. Whereas the second moment scaling H does not coincide with the distribution of scaling exponent δ, the first moment scaling exponent H_1 is equal to δ. This is due to the fact that the first moment of a Lévy distribution is

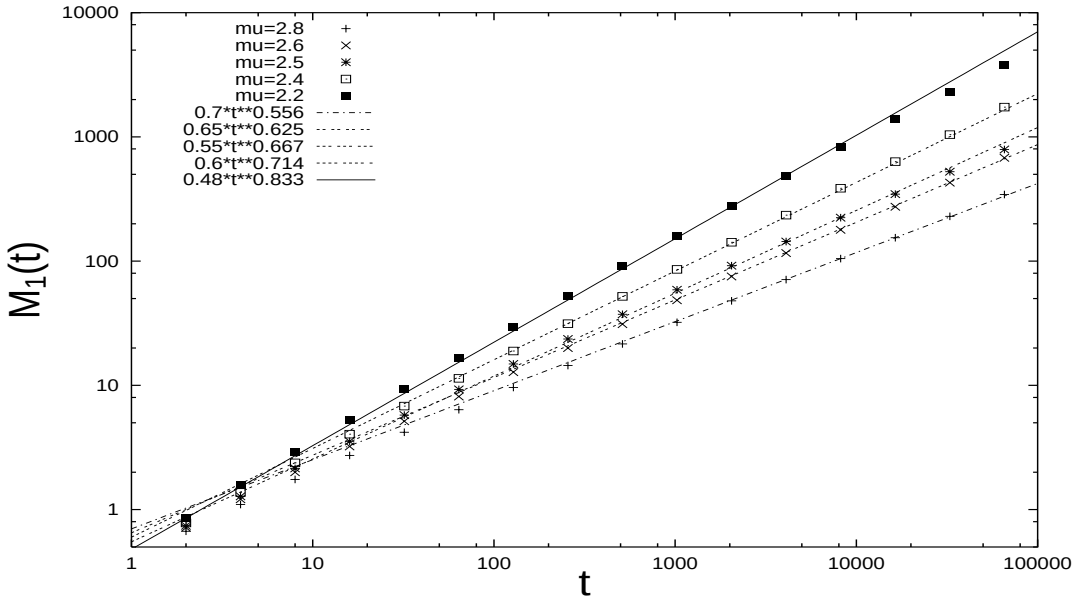

Figure 5.15: First Moment Analysis of Lévy walk. Five sets of data corresponding to $\mu = 2.8$, $\mu = 2.6$, $\mu = 2.5$, $\mu = 2.4$, $\mu = 2.2$. The values of the scaling exponent H_1 are in Table 5.1. We have that $H_1 = \delta$.

convergent. In the next two sections we analyze the scaling properties of the long jump model, Section 2.5. For long jumps distributed according to an inverse power law distribution (5.30) with $2 < \mu < 3$, the first moment analysis is still able to detect the correct scaling exponent as the DEA does. The methods based upon the variance fail completely in detecting a true scaling. If $1 < \mu < 2$, even the first moment fails to detect the real scaling. Only the Diffusion Entropy Analysis is able to measure the real scaling in all situations.

5.4 Long jump model: case $2 < \mu < 3$.

In this section and in the next one we simulate the long jump model theoretically discussed in Sec. 2.5 and analyze its scaling properties. We generate a sequence of five million jumps, $\{\xi_i\}$, distributed according to the following inverse power law

$$\psi(\xi) = (\mu - 1)\frac{T^{\mu-1}}{(T+\xi)^{\mu}} \ . \tag{5.34}$$

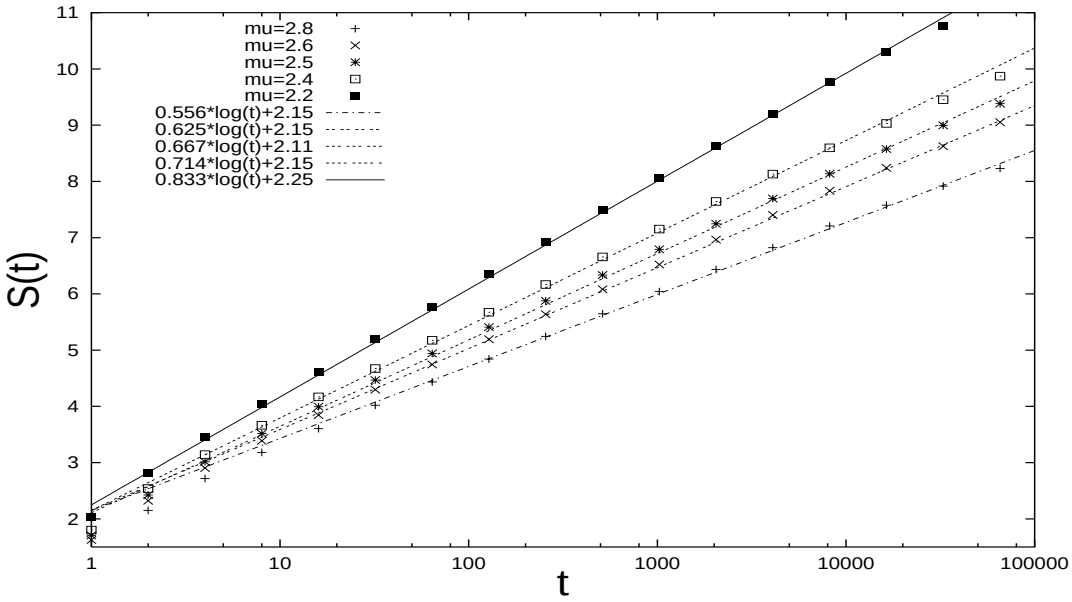

Figure 5.16: Diffusion Entropy Analysis of Long Jump Model with $2 < \mu < 3$. Five sets of data corresponding to $\mu = 2.8$, $\mu = 2.6$, $\mu = 2.5$, $\mu = 2.4$, $\mu = 2.2$. The values of the pdf scaling exponent δ are in Table 5.1.

The diffusion process is generated by a walker that at each unit time makes a jump forward or backward equal to the jump length ξ. The generalized central limit theorem assures that a diffusion process generated in this way is a Lévy diffusion whose pdf scales according to the equation

$$\delta = \frac{1}{\mu - 1}, \tag{5.35}$$

as it is proved in Sec. 2.5. If $2 < \mu < 3$, the variance diverges; Therefore, all scaling detectors based upon the variance are expected to fail in detecting the real scaling of the diffusion distribution given by Eq. (5.37) because there is not an equation like Eq. (5.29) that relates δ to H as is possible for the Symmetric Velocity Model. On the contrary, the modulous of the first moment, $M_1(t) =< |x - \overline{x}| >$ of the distribution does not diverge. It should be able to detect the scaling of the distribution. The Diffusion Entropy Analysis directly studies the scaling properties of the distribution; Therefore, DEA detects real scaling.

Fig. 5.16 shows the Diffusion Entropy Analysis of five sequences of data $\{\xi_i\}$

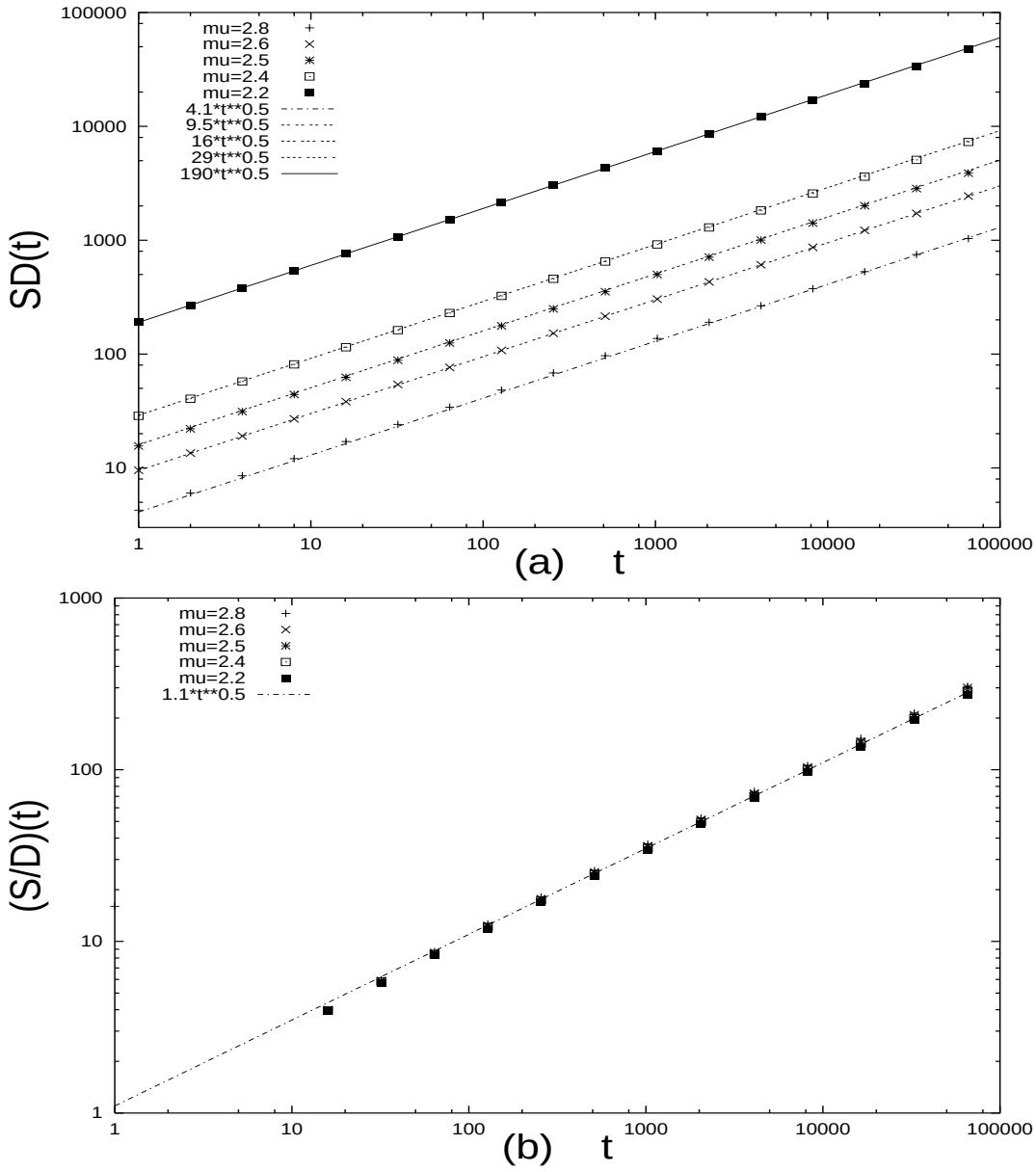

Figure 5.17: (a) Standard Deviation Scaling Analysis (SDSA) and (b) Hurst R/S analysis of Long Jump Model with $2 < \mu < 3$. Five sets of data corresponding to $\mu = 2.8$, $\mu = 2.6$, $\mu = 2.5$, $\mu = 2.4$, $\mu = 2.2$. The figures show that the scaling analysis method based upon the study of the variance are unable to detect the real scaling of the distribution.

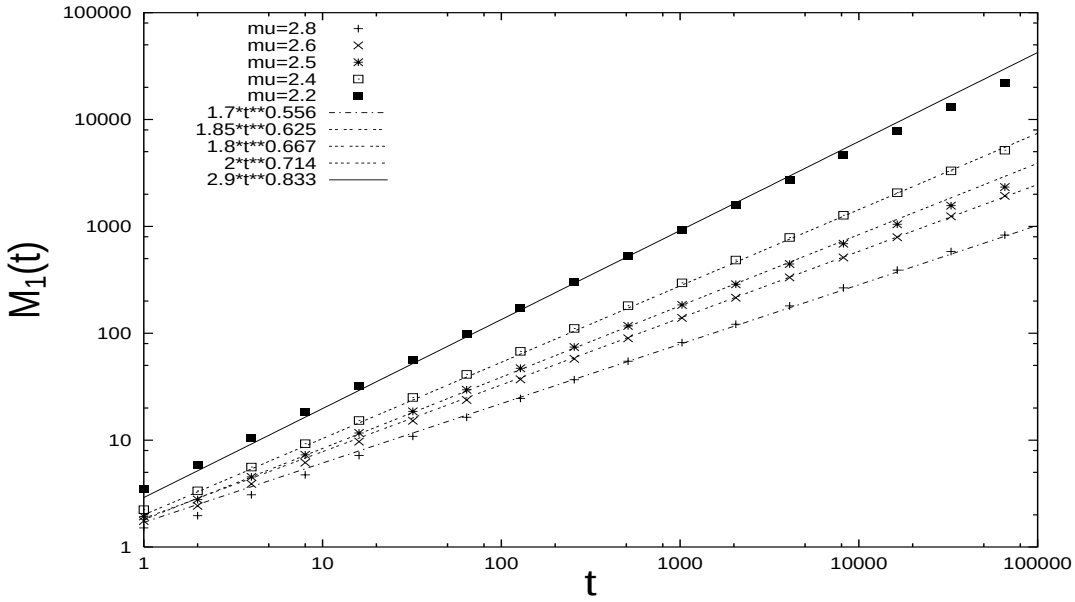

Figure 5.18: First Moment Scaling Analysis of Long Jump Model with $2 < \mu < 3$. Five set of data corresponding to $\mu = 2.8$, $\mu = 2.6$, $\mu = 2.5$, $\mu = 2.4$, $\mu = 2.2$. The values H_1 coincide with the pdf scaling exponent δ present in Table 5.1.

generated by using the discrete algorithm illustrated in the previous section with the constant $T = 0$, and $\mu = 2.8$, $\mu = 2.6$, $\mu = 2.5$, $\mu = 2.4$ and $\mu = 2.2$. The figure shows clearly that DEA detects the real scaling exponent δ whose values are in Table 5.1.

Figs. 5.17 show the Standard Deviation scaling analysis and the Hurst analysis for the same five sequences analyzed before. The figures show that the scaling coefficients H do not change with μ. $H = 0.5$ in all cases. This proves that the variance scaling analysis methods are unable to detect the real pdf scaling exponent δ. Because the Variance scaling methods give $H = 0.5$, by using them it is possible to reach the wrong conclusion that these diffusion pdfs are Gaussian, whereas they are Lévy. Fig. 5.17a shows a finite standard deviation because the number N of data is finite. By increasing N, the variance increases. Fig. 5.17b shows that all five curves coincide. This is due to the fact that the Hurst Analysis normalizes the data by dividing by the standard deviation.

Finally, Fig. 5.18 shows the First Moment Scaling Analysis (FMSA) of Long Jump Model with $2 < \mu < 3$. The values H_1 coincide with the pdf scaling exponent δ present in Table 5.1. This is due to the fact that for $2 < \mu < 3$ the first moment of the diffusion pdf is finite and, therefore, it detects the right scaling. In the next section, we show that for $1 < \mu < 2$ while the FMSA is unable to detect the right scaling, the DEA is still able to detect the right scaling. This makes the Diffusion Entropy the best method of detection of the pdf scaling.

5.5 Long jump model: case $1 < \mu < 2$.

As done in the previous section, we generate a sequence of five million jumps, $\{\xi_i\}$, distributed according to the following inverse power law

$$\psi(\xi) = (\mu - 1)\frac{T^{\mu-1}}{(T + \xi)^{\mu}} \ . \tag{5.36}$$

The diffusion process is generated by a walker that for each unit time makes a jump forward or backward equal to the jump length ξ. The generalized central limit theorem assures that a diffusion process generated a diffusion whose pdf scales according to the equation

$$\delta = \frac{1}{\mu - 1}, \tag{5.37}$$

as it is proved in Sec. 2.5. If $1 < \mu < 2$, not only the second moment but also the first moment diverges. This means that even the first moment scaling analysis is expected to fail to detect the real scaling. The Diffusion Entropy Analysis, instead, directly detects the scaling of the distribution. Therefore, it is expected to succeed in detecting the right scaling exponent related to the power exponent μ via Eq. (5.37). Figs. 5.19 and 5.20 show the DEA and the FMSA applied to the same set of data respectively. Table 5.2 shows the relation between μ and δ for the four sets of data used in the simulations. Fig. 5.19 shows clearly that DEA is able to detect the real scaling of the distribution, whereas Fig. 5.20 shows that the first moment scaling analysis cannot be used for detecting the real scaling. The first moment diverges for $1 < \mu < 2$. The finite values shown in Fig. 5.20 are due to finite set of data.

μ	1.900	1.800	1.700	1.600
δ	1.111	1.250	1.429	1.667

Table 5.2: Theoretical relation between the waiting time distribution power exponent μ and the pdf scaling exponent δ.

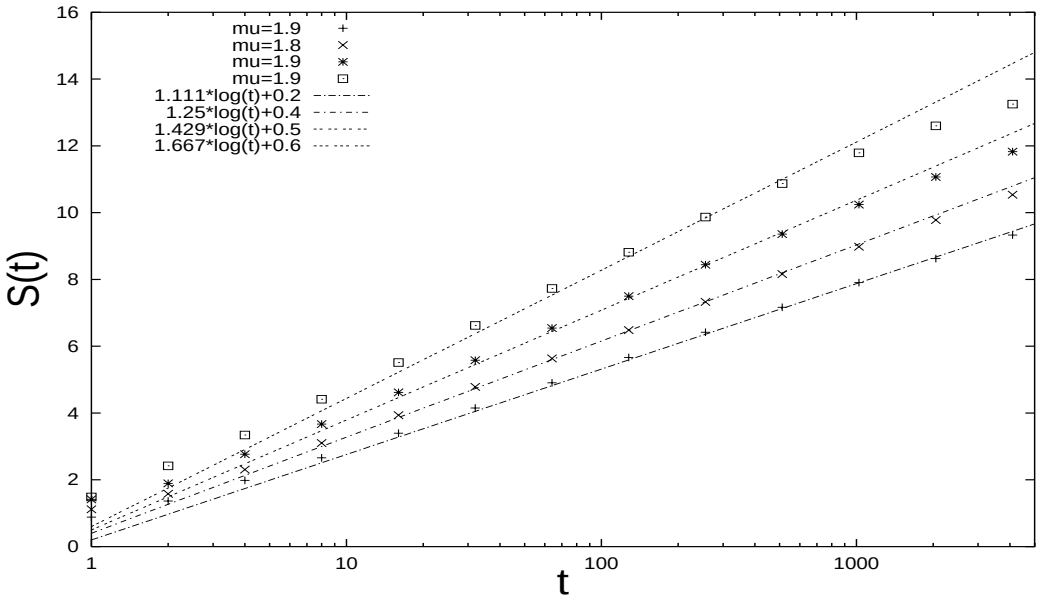

Figure 5.19: Diffusion Entropy Analysis of Long Jump Model with $1 < \mu < 2$. Four sets of data corresponding to $\mu = 1.9$, $\mu = 1.8$, $\mu = 1.7$, $\mu = 1.6$. DEA detects the right pdf scaling exponents δ; Table 5.2.

$M_1(t)$ increases with the number of data analyzed, therefore it has not an universal meaning.

5.6 Diffusion Entropy Analysis is the best scaling detector.

The results of the previous sections show clearly the superiority of the Diffusion Entropy Analysis over all standard scaling detector methods like the variance scaling analysis, Hurst's analysis, detrended fluctuation analysis, relative dispersion analysis, spectral analysis, wavelet spectral analysis and first moment scaling analysis. The moment scaling analysis methods are based upon the assumption that the scaling

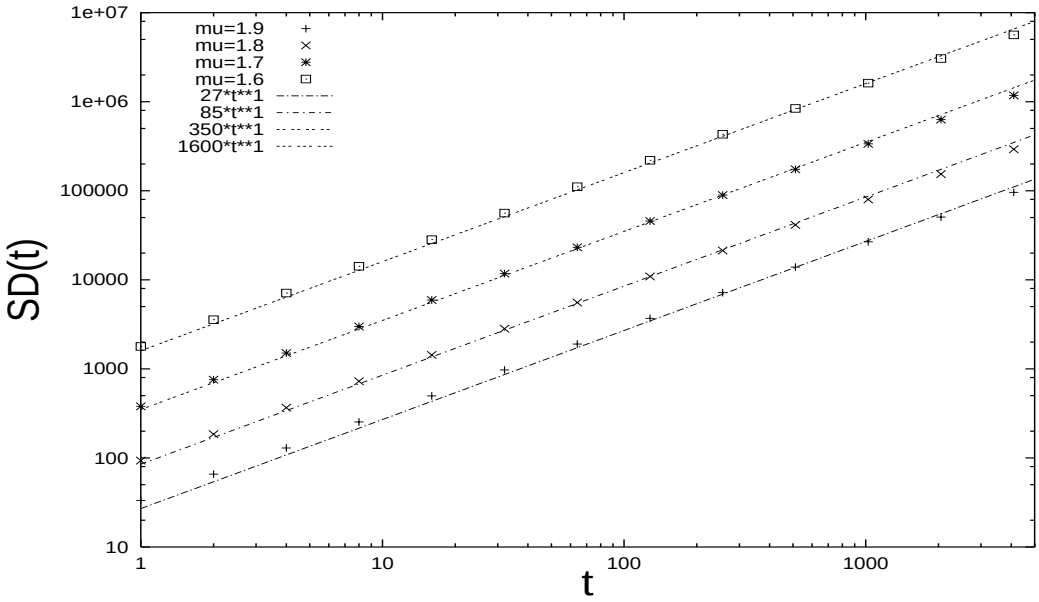

Figure 5.20: First Moment Scaling Analysis of Long Jump Model with $1 < \mu < 2$. Four sets of data corresponding to $\mu = 1.9$, $\mu = 1.8$, $\mu = 1.7$, $\mu = 1.6$. All four sets of data show $H_1 = 1$. This proves that FMSA is unable to detect the real scaling.

detected by one of the moment, M_η of the distribution coincides with the scaling of the distribution. This is true in some cases like in the Fractional Brownian diffusion, but it is not true in general. Diffusion processes characterized by, for example, a Lévy statistics as the symmetric velocity model or the long jump model, are characterized by the fact that the exponent H is different from the true scaling exponent δ. Only the Diffusion Entropy Analysis is able to detect the true scaling exponent δ in all situations because DEA measures directly the scaling of the distribution. However, to know whether a time sequence is characterized by Gaussian or non-Gaussian statistics, DEA is not enough. There is the need of a simultaneous use of the DEA with one of the variance scaling analysis methods. If the two exponents H and δ are equal, the Gaussian hypothesis may be realistic whereas it cannot be granted if $H \neq \delta$. If by adopting the symmetric velocity model $\delta = 1/(3 - 2H)$, the theory assures us that the time sequence is characterized by Lévy properties.

5.7 Non-stationary condition induced by weak and persistent memory.

In this section we show that a sequence of data generated by fast fluctuations around a weak and slowly fluctuating drift, produces the same effects as those illustrated in Sec. 4.5. This supports our conviction that the breakdown of the stationary condition discussed earlier is a manifestation of weak but persistent memory. In fact, the connection between η and ϵ is proven to be the same as that of Eq. (4.44) and Eq. (4.45). This means that for short periods of the logarithmic time τ the effect of persistent memory becomes indistinguishable from the breakdown of the stationary condition. We have in mind the notable effect, illustrated in Sec. 4.5, of the structure of of Eq. (4.35) yielding the entropic property of Eq. (4.37).

We create a sequence of data, $\xi_{\Delta}(t)$, where Δ denotes the memory intensity, in the following way. Firstly, using a random noise generator we create a sequence of fluctuations. More specifically, we generate a set of 100,000 random rational numbers, $F(n)$, belonging to the interval $[0, 1]$. The variable n is an integer number running from 1 to 100,000. Secondly, we create an artificial memory through a square periodic function, $f_{\Delta}(t)$, with period equal to 2000, average equal to 0.5, and amplitude equal to the parameter Δ, which is, as mentioned earlier, the memory intensity. In the first period, for t from 0 to 2000, the function $f_{\Delta}(t)$ is:

$$f_{\Delta}(t) = \begin{cases} 0.5 + \Delta & if \quad 0 < t \leq 1000 \\ 0.5 - \Delta & if \quad 1000 < t \leq 2000 \end{cases} . \tag{5.38}$$

In the third and final stage we convert the sequence $F(n)$ into the dichotomous sequence, $\xi_{\Delta}(t)$, of numbers "+1" and "-1" using the following prescription:

$$\xi_{\Delta}(t) = \begin{cases} +1 & if \quad F(t) > f_{\Delta}(t) \\ -1 & if \quad F(t) < f_{\Delta}(t) \end{cases} . \tag{5.39}$$

The numbers "-1" and "1" can be interpreted as the discrete jumps of a random walker. This means that with the data $\xi_{\Delta}(t)$, interpreted as the steps made, forward or backward, by a random walker moving on the x-axis, we can build a trajectory, denoted by Ξ_{Δ}. This trajectory specifies the random walker's position on the same

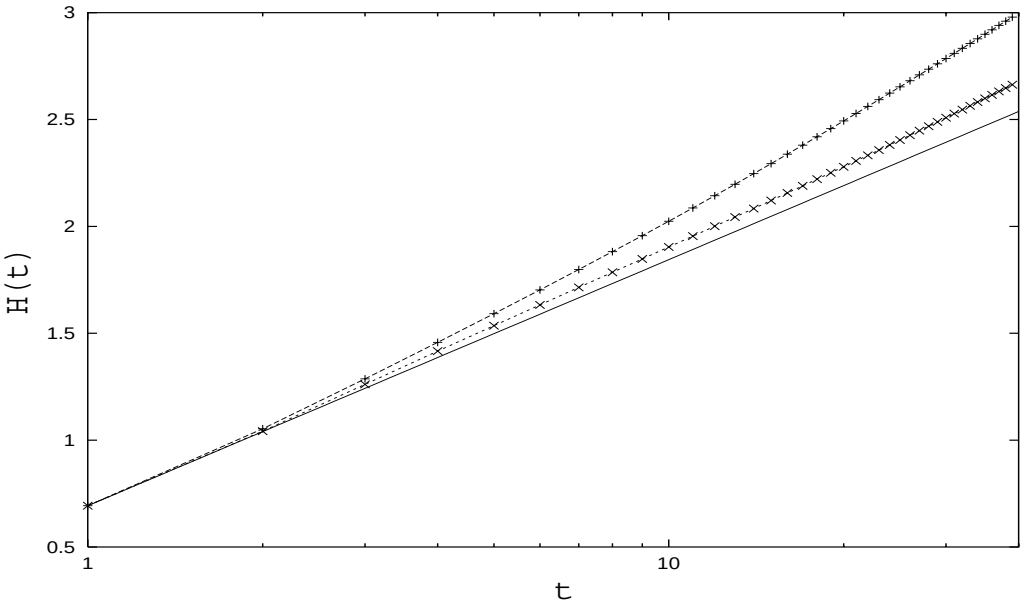

Figure 5.21: The diffusion entropy as a function of time. The three curves refer to the three sets of data with $\Delta = 0$ (solid line), $\Delta = 0.04$ (the curve denoted by $\times$) and $\Delta = 0.10$ (the curve denoted by $+$), respectively. The case $\Delta = 0$ results in the diffusion entropy of a stationary process, the ordinary random walk, in this case. The corresponding curve, as expected, is a linear function of the logarithmic time $\tau \equiv \ln(t)$, see Eq.(5.40). The other two curves, corresponding to non-vanishing memory strength, result in an evident departure from the linear dependence on logarithmic time, larger for the case of larger memory (larger Δ). This is a clear illustration of the breakdown of the stationary condition caused by a memory of weak but non-vanishing intensity.

x-axis, at any given time N. With $\Delta = 0$ there is no memory: The data $\xi_{\Delta=0}(t)$ are statistically equivalent to those that one would obtain by tossing a fair coin.

At this point, it is possible to calculate the diffusion entropy associated with the weak memory controlled by the parameter Δ.

Fig. 5.21 shows the results of this numerical analysis. We plot three curves corresponding to $\Delta = 0$, $\Delta = 0.04$, and $\Delta = 0.10$. For $\Delta = 0$ the diffusion entropy refers to a stationary diffusion process. According to the theory of Sec. 4.5, this condition is expected to produce an entropy increase linear with respect to the logarithmic time $\tau = \ln(t)$. In the case under study, the stationary condition is that of an ordinary

Brownian diffusion, which yields for the scaling coefficient δ_0 the value 0.5. Thus, the diffusion entropy is given by

$$H(t) = 0.5 \ln(t) + \ln(2), \tag{5.40}$$

which fits very well the numerical result of Fig. 5.21. For the other two curves, corresponding to $\Delta = 0.04$ and $\Delta = 0.10$, respectively, there is a significant deviation from the linear dependence on the logarithmic time τ, which is larger with the larger memory strength. In conclusion, Fig. 5.21 shows that the numerical evaluation of the diffusion entropy detects the breakdown of the condition of stationary diffusion, even if this is caused by a very weak memory. In the case under study, the memory strength, given by the parameter Δ, is equivalent respectively to the 4% and 10% of the signal.

Let us discuss now how to measure the intensity of the breakdown of the stationary condition. In accordance with the prescriptions of Sec. 4.5. This can be done in two ways. The first method is direct one. It is based on fitting the curves with the quadratic approximation of Eq.(4.37). This allows us to determine the coefficients A, δ_0, and, η. On the basis of the theoretical remarks of Sec. II we realize that the latter parameter is the property of interest to measure. We can establish a connection with the second method by determining the entropic index $q = 1 + \epsilon$ by means of Eq.(4.44). The second method rests on the determination of the entropic index q, as the "magic" value of q making the non-extensive Tsallis entropy linear with respect to logarithmic time τ. In practice, this means that we have to look for the value of q that results in the maximum the coefficient of linear correlation. The form of the non-extensive Tsallis entropy used in this case is:

$$H_{q,\Delta}(t) = \left(1 - \sum_{x_\Delta} p_\Delta(x_\Delta, t)^q\right) / (q-1), \tag{5.41}$$

where, as in the earlier case, the sum is understood on all available positions x_Δ. For $q = 1$ Eq.(5.41) is identical to the Shannon Entropy.

We illustrate the results of the numerical calculations based on the adoption of

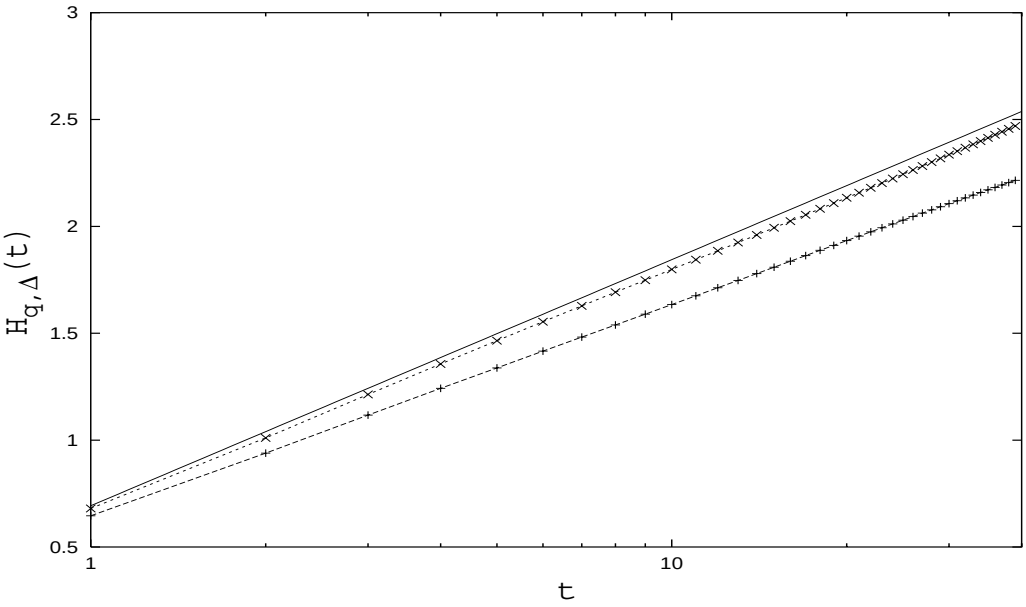

Figure 5.22: The non-extensive Tsallis entropy as a function of time t. The three curves are the numerical realization of Eq.(5.41) with $q = 1$ (solid line), $q = 1.054$ (symbol $\times$) and $q = 1.205$ (symbols $+$) and correspond to different values of the memory strength Δ, which are $\Delta = 0$, $\Delta = 0.04$ and $\Delta = 0.10$, respectively. The choice of different entropic indices q for the different values of Δ has been done with the criterion of selecting the value of q resulting in the most extended linear regime with respect to the logarithmic time $\tau \equiv \ln(t)$.

Eq.(5.41) in Fig. 5.22. This figure refers to the same physical conditions as those of Fig. 5.21 and proves that for any of those conditions an entropic index q can be found so that the non-extensive Tsallis entropy becomes a linear function of τ. Thus, within the context of the problems under discussion in this paper, Tsallis entropy takes on the following "thermodynamic" meaning. This new type of entropic indicator allows us to imagine the processes departing from the stationary condition as also being stationary. The departure of q from the ordinary value $q = 1$ increases with increasing δ. We find that $\Delta = 0$, $\Delta = 0.04$ and $\Delta = 0.10$ correspond to $q = 1$, $q = 1.054$ and $q = 1.205$., respectively. All this is in complete accordance with the theoretical remarks of Sec. 4.5.

We make a judgment of the results obtained by means of the numerical treatment

of the simple model of this section with the help of Table 5.3. In the first three columns of this table we report, for different values of the parameter Δ, the values of the coefficients A, δ_0, η and $q = 1 + \epsilon$ calculated using the Eq.(4.44), and thus using the former of the two methods illustrated earlier. In the last column of this table we report the values of q obtained using the latter method. The fitting procedure adopted to generate the values of this table have been limited to time windows whose size does not exceed the value of 30. We see that, as expected, q is an increasing function of Δ and that within the statistical accuracy the two methods yield the same value for q. This means that Eqs.(4.44) and (4.45) are correct. Of course, this implies that the memory strength is weak enough. From the values reported in Table I we see that the accuracy of the theoretical prediction is satisfactory for values $\epsilon \leq 0.207$, a fact which implies the maximum value $\Delta = 0.09$ for the memory strength. Beyond this value the validity of the quadratic approximation necessary to evaluate the parameter η is broken. For higher values of the memory strength it is probably convenient to use the latter method. However, in this case the theoretical connection with the breakdown of stationary diffusion is still missing and further research work is required.

Δ	η	δ_0	A	$q_1 = 1 + \epsilon$	q_2
0.03	0.0018	0.517	0.690	1.014	1.017
	$\pm$.0005	$\pm$.002	$\pm$.002	$\pm$.004	$\pm$.008
0.04	0.0063	0.512	0.692	1.047	1.054
	$\pm$.0006	$\pm$.002	$\pm$.002	$\pm$.005	$\pm$.005
0.05	0.0113	0.506	0.693	1.083	1.092
	$\pm$.0007	$\pm$.003	$\pm$.003	$\pm$.006	$\pm$.004
0.06	0.0165	0.501	0.695	1.120	1.127
	$\pm$.0007	$\pm$.003	$\pm$.003	$\pm$.007	$\pm$.004
0.07	0.0214	0.498	0.696	1.154	1.157
	$\pm$.0008	$\pm$.003	$\pm$.003	$\pm$.008	$\pm$.004
0.08	0.0261	0.497	0.696	1.184	1.181
	$\pm$.0007	$\pm$.003	$\pm$.00	$\pm$.008	$\pm$.004
0.09	0.0300	0.498	0.695	1.207	1.197
	$\pm$.0006	$\pm$.003	$\pm$.002	$\pm$.007	$\pm$.004
0.10	0.0329	0.503	0.693	1.221	1.205
	$\pm$.0005	$\pm$.002	$\pm$.002	$\pm$.005	$\pm$.004

Table 5.3: Entropic index q resulting from two distinct fitting procedures. The coefficient q_1 is calculated by using the first method via the measure of the coefficients η, δ_0, A and ϵ. The coefficient q_2 is calculated by using the Tsallis entropy.

Part II

APPLICATIONS

CHAPTER 6

LÉVY STATISTICS IN CODING AND NON-CODING NUCLEOTIDE SEQUENCES

In this chapter we apply the Diffusion Entropy, DEA, to the statistical analysis of nucleotide sequences. The recent progress in experimental techniques of molecular genetics has made available a wealth of genome data (see for example the NCBI's Gen-Bank data base of Ref. [134]), and raised the interest for the statistical analysis of DNA sequences [135, 136, 137, 138]. These pioneer papers mainly focused on the controversial issue of whether long-range correlations are a property shared by both coding and non-coding sequences or are only present in non-coding sequences. The results of more recent papers [139, 140] yield the convincing conclusion that the former condition applies. However, some statistical aspects of the DNA sequences are still obscure, and it is not yet known to what extent the dynamic approach to DNA sequences proposed by the authors of Ref. [141] is a reliable picture for both coding and non-coding sequences. The later work of Refs. [142] and [143] established a close connection between long-range correlations and the emergence of non-Gaussian statistics, confirmed by Mohanti and Narayana Rao [139]. According to the dynamic approach of Refs. [141, 144] this non-Gaussian statistics should be Lévy, but this property has not yet been assessed with compelling evidence. The reason for this failure is that there exists no reliable method of scaling detection. In this chapter, we aim at filling this gap and we show that the Diffusion Entropy Analysis (DEA) realizes the ambitious goal of affording the genuine scaling value. Furthermore, we prove that the joint use of the DEA method and of the Variance Scaling Analysis (VSA) methods like the Standard Deviation Scaling Analysis, the Detrended Fluctuation Analysis (DFA) [145], and the Wavelet Spectral Analysis (WSA) [140, 146], allow us to:

1) establish the presence of long-range correlations in coding as well as in non-coding sequences;

2) assess the Lévy nature of the resulting non-Gaussian statistics.

This is possible because, as shown in the previous chapters, DEA detects the true scaling of a time series. In fact, DEA directly detects the pdf scaling exponent δ, whereas the other techniques based upon the Variance Scaling Analysis are able to detect the true scaling only in the reductive case of Fractional Brownian noise.

As shown in the previous chapters, scaling is the property of diffusion processes relating the space variable x to the time variable t via the key relation $x \propto t^H$. The symbol H stands for Hurst, as a recognition by Mandelbrot of the earlier work of Hurst [70], and is interpreted as a scaling parameter. Mandelbrot's arguments are based on the so called fractional Brownian motion (FBM), an extension of ordinary Brownian motion to anomalous diffusion. If the FBM condition applies, H is the real scaling of the time series as shown in Section 5.1. In this case the departure from ordinary diffusion is given by $H \neq 1/2$, with no departure, though, from Gaussian statistics. When the FBM condition and, therefore, the Mandelbrot's argument do not apply, only the DEA can establish the true scaling. It is worth stressing that *our definition of scaling is given by the asymptotic time evolution of the probability distribution of x, obeying the property*

$$p(x,t) = \frac{1}{t^\delta} F\left(\frac{x}{t^\delta}\right), \tag{6.1}$$

where the symbol δ denotes the true scaling, which exists also when the second moment of $F(y)$ is divergent. The Lévy nature of the resulting non-Gaussian statistics of the nucleotide sequences is assessed on the basis of the theoretical results of Section 2.6 and on the numerical results of Section 5.3 that are about the Lévy walk and the Symmetric Velocity Model.

6.1 DNA genome data and its numerical representation.

In the last few years, thanks to the recent progress in experimental technics in molecular genetics, a wealth of genome data has become available (see for example Ref.[134]). This has triggered a large interest both in the study of mechanics of folding [147], and

on the statistical properties of DNA sequences. In particular, genomes can be considered as long messages written in a four-letter alphabet, in which we have to search for information (signal). Recently, there have been many papers pointing out that DNA sequences are characterized by long-range correlation, this being more clearly displayed by non-coding than by coding sequences [144, 135, 138, 145].

In this chapter we consider various DNA sequences: The human T-cell receptor alpha/delta locus (Gen Bank name HUMTCRADCV) [145], a *non-coding* chromosomal fragment (it contains less than 10% coding regions); The Escherichia Coli K12 (Gen Bank name ECO110K) [145], and the Escherichia Coli (Gen Bank ECOTSF) [143], two genomic fragments containing mostly *coding regions* (more than 80% for ECO110K). The three sequences have comparable lengths, respectively $M = 97634$ basis for HUMTCRADCV, $M = 111401$ basis for ECO110K and $M = 91430$ basis for ECOTSF.

The first two sequences have been analyzed in Ref. [145] by means of the Detrended Fluctuation (DFA). The fundamental difference between them is that the non-coding sequence, namely HUMTCRADCV, shows the presence of long-range correlation at all scales, while the sequence ECO110K, a coding sequence, shows the presence of long-range correlation only at the large-length scale. The third sequence, ECOTS, has been studied in Ref. [143] with the interesting conclusion that the large-length scale shows non-Gaussian statistics. The authors of Ref. [145] using the illuminating example of the lambda phage genome, pointed out that the DFA does not mistake the presence of patches of different strand bias for correlation. This is an important property, shared by the DEA method, which is widely independent of the presence of biases, since the entropy increases mainly as a consequence of the trajectories departing from one another. We want to prove that the DEA method makes it possible to relate the non-Gaussian statistics and the anomalous scaling of the large-length scale to the same cause: the onset of Lévy statistics.

The usual way to study the statistical properties of DNA is to consider a sequence of four bases: adenine, cytosine, guanine, and thymine (respectively A, C, G, and T), at the simplified level of a dichotomous sequence of two symbols, purine (for A and G) and pyrimidine (for C and T). A trajectory, the so-called DNA walk, can

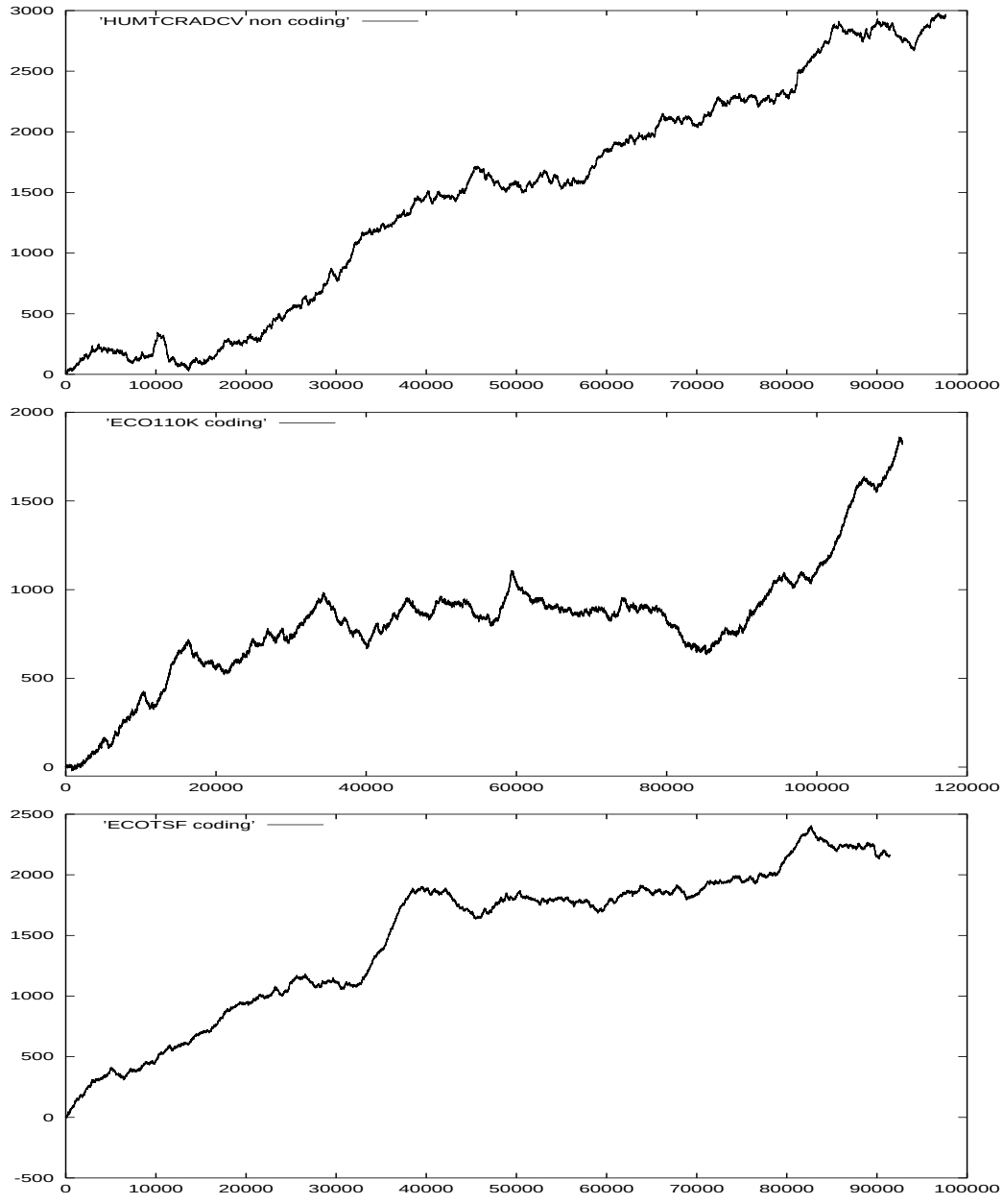

Figure 6.1: The DNA walk. Fig. 6.1a shows the DNA walk relative to HUMTCRADVC, a non-coding chromosomal fragment. Figs. 6.1b and 6.1c show the DNA walk relative to ECO110K and ECOTSF, two coding genomic fragments.

be extracted by considering a one-dimensional walker associated to the nucleotide sequence in the following way: the walker takes one step up when there is a pyrimidine in the nucleotide and a step down when there is a purine. The DNA sequence is therefore transformed in a sequence $\xi_i, \quad i = 1, ..., M$, of numbers $+1$ or -1. Here i can be conceived as a discrete value of "time", and the walker makes a step ahead or backward, according to whether at "time" i the random walker sees $+1$ or -1, namely if the i-th site of the DNA sequence hosts a pyrimidine or a purine. The displacement of the walker after l steps is $x(l) = \sum_{i=1}^{l} \xi_i$ and is reported in Fig. 6.1 for the three sequences under consideration.

6.2 Detrended Fluctuation Analysis and Wavelet Spectral Analysis.

The first thing we notice is that all the three series present "patches", i.e. excess of one type of nucleotide. In the DFA method of ref. [145] Stanley and collaborators define a detrended walk by subtracting the local trend from the original DNA walk and then they study the variances $F(l)$ of the detrended walk. If the walk is totally random, as in the ordinary Brownian motion, no correlations exist and $F(l) \sim l^{1/2}$. On the contrary, the detection of $F(l) \sim l^H$ with either $H > 1/2$ or $H < 1/2$ is expected to imply the presence of extended correlation, which, in turn, is interpreted as a signature of the complex nature of the observed process. Stanley et al. found a scaling exponent $H = 0.61$ for the non-coding intron sequence HUMTCRADCV, and $H' = 0.51$ for the intronless sequence of ECO110K in the short-time region. We indicate the scaling at short-time H' and that at long-time H. They claim their method is able to avoid the spurious detection of apparent long-range correlations which are the artifacts of the patchiness by detrending. The detection of the true scaling, as we have seen, often involves the adoption of detrending procedures, since a steady bias hidden in the data produces effects which might be mistaken for a striking departure from Brownian diffusion, while the interesting form of scalings must be of totally statistical nature.

Fig. 6.2 shows the Detrended Fluctuation Analysis compared with the Standard

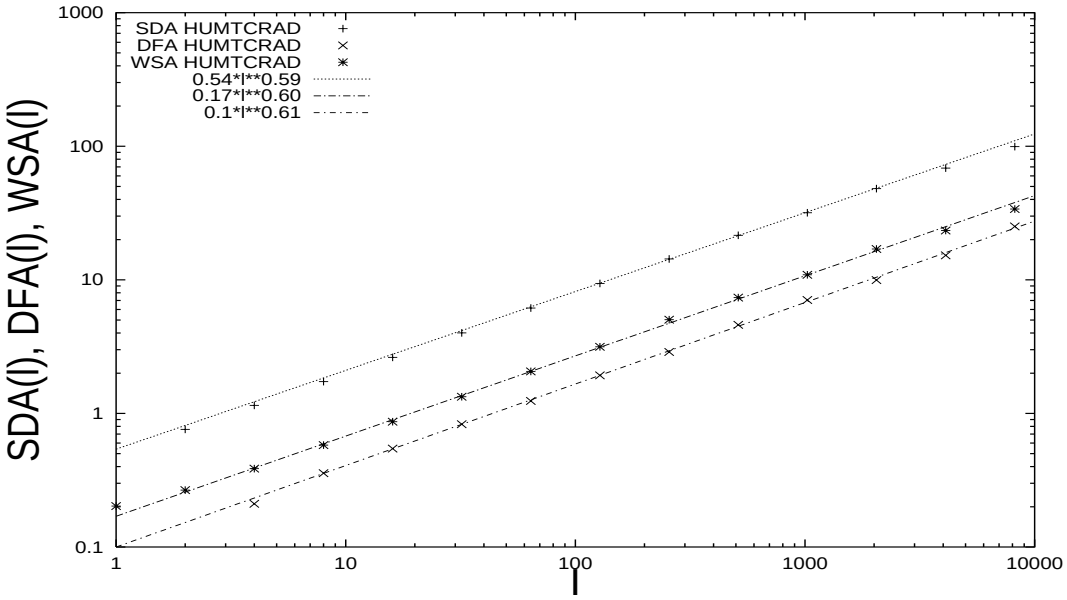

Figure 6.2: SDSA, DFA and WSA of the HUMTCRADVC, a non-coding chromosomal fragment. The scaling exponent H is 0.59±0.01 (SDSA), 0.60±0.01 (DFA), 0.61±0.01 (WSA). H is the same both at short-time and long-time regions.

Deviation Analysis and the Wavelet Spectral Analysis of the HUMTCRADVC, a non-coding chromosomal fragment. Wavelet Spectral Analysis is conducted on the DNA walk and the square root of the wavelet variance is plotted in the ordinate axes. In this way we obtain a scaling directly compatible with the one obtained by the SDSA and the DFA. The scaling exponent H is 0.59 ± 0.01 (SDSA), 0.60 ± 0.01 (DFA), 0.61 ± 0.01 (WSA). H is the same both at short and long time region. The advantages of using the DFA are not evident here. Figs. 6.3 show the Detrended Fluctuation Analysis compared with the Standard Deviation Analysis and the Wavelet Spectral Analysis of the ECO110K and ECOTSF, two coding genomic fragments. The scaling exponent H' is 0.53 ± 0.01 (SDSA), 0.52 ± 0.01 (DFA), 0.52 ± 0.01 (WSA) in the short-time region. H is 0.73 ± 0.01 (SDSA), 0.75 ± 0.01 (DFA), 0.74 ± 0.01 (WSA) at long-time region. Both figures show the three adopted methods detect the same scaling exponents H' and H both in short and long time region. However, the advantages/disadvantages of using the DFA are evident. DFA detects the scaling in the long-time region later

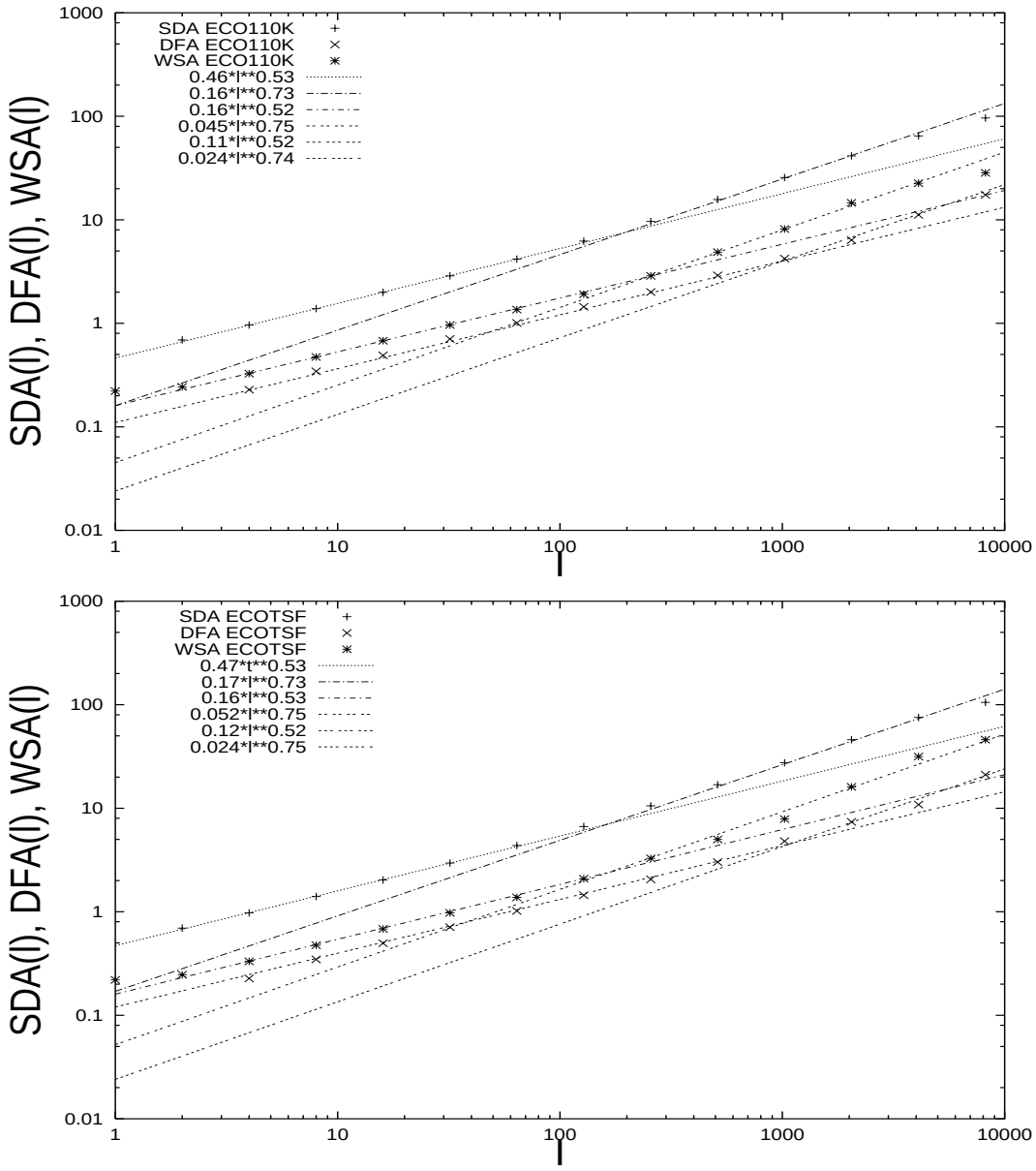

Figure 6.3: The DNA walk. SDSA, DFA and WSA of ECO110K and ECOTSF, two coding genomic fragments. The scaling exponent H' is 0.53 ± 0.01 (SDSA), 0.52 ± 0.01 (DFA), 0.52 ± 0.01 (WSA) at short-time region. H is 0.73 ± 0.01 (SDSA), 0.75 ± 0.01 (DFA), 0.74 ± 0.01 (WSA) at long-time region.

because of the detrending that cuts off long local trend. In Ref. [145], Stanley and collaborators were interested in studying the scaling in the short-time region in order to distinguish the non-coding from the coding DNA sequences. The DFA works for extending such a regime. If the interest is in studying the signal as it is, that is, studying the statistical properties of what appears to be local trend as well, it is better to adopt a non-detrending method of analysis. The Wavelet Spectral Analysis is adopted by Arneodo and collaborators in Ref. [140, 146]. Figs. 6.2 and 6.3 show that there is no difference between SDA and WSA. This is because the Wavelet Transform behaves, Sec. 3.3, like the Fourier Transform that studies the variance of the signal. Therefore, WSA, as Fourier Spectral Analysis, can detect the true scaling only in the Gaussian case. In all other cases, WSA can detect only the variance scaling that may not coincide with the true scaling, see Sec. 5.3.

By using the DEA algorithm we can detect the existence of scaling, either normal or anomalous, Gaussian or Lévy, in a very efficient way, and without altering the data with any form of detrending. We analyze the data of both the coding and non-coding sequences. Starting from the sequence $\xi_i,\ i = 1, ..., N$ we create the diffusion trajectories and we compute the diffusion entropy $S(l)$ according to equation (4.28). The results are reported in Figs. 6.4-6.6. We extract δ which identifies the scaling with the asymptotic tangent to the curve $S(l)$ vs. $\log(l)$.

6.3 The Copying Mistake Map: a model for DNA sequences.

According to the dynamical model of Ref. [141] a *non-coding* DNA sequence corresponds to an artificial sequence with long-range inverse power law correlation as those studied in Section 2.6. On the other side, a *coding sequence* can be obtained by adopting a model called Copying Mistake Map (CMM) [141]. According to CMM, Nature has, at her disposal, two independent sequences. The first one is of the same kind as the inverse power law correlated sequence discussed in Section 2.6 (and corresponding, as already said, to non-coding sequences). The second sequence is a random walk. Then Nature decides to build up the DNA sequence using a random criterion. To any site of the sequence to build, Nature assigns a symbol drawn, with a probability

p_R, from the corresponding site of the latter sequence (the random sequence), and drawn, with probability p_L, from the corresponding site of the former sequence (the correlated sequence). In the case of coding sequences usually the condition

$$p_R \gg p_L \tag{6.2}$$

applies. The authors of Ref. [143] pointed out that the CMM model is equivalent to an earlier model [148, 138] called Generalized Lévy Walk (GLW). The CMM (and the GLW, as well, of course) yields, for short times, a diffusion process indistinguishable from ordinary Brownian motion. At large times, however, the long-range correlation predominates. In Ref. [143] the CMM was adopted to account for the properties of prokaryotes, for which a significant departure from Gaussian statistics occurs. One of the coding sequences studied here, namely ECOTSF, is the same as one discussed in Ref. [143]. It produces strong deviations from Gaussian statistics. On the basis of that, and of the conjectures made in Section 3 D, we expect also for coding sequences at large "times" a scaling parameter δ corresponding to the Lévy statistics

$$\delta = \frac{1}{\mu - 1}, \tag{6.3}$$

This is different from the result provided by using the DFA,

$$H = \frac{4 - \mu}{2}. \tag{6.4}$$

The proof of Eq. (6.4) can be given using the dynamic approach to diffusion discussed in Refs. [144, 141]. The discussion made in Section 2.6, with artificial sequences, and consequently, with a known value of μ, proves that the DE method detects in the long-time limit the correct scaling of Eq. (6.4). In the case of the real DNA sequences here under discussion, the value of μ is not known. Both the DFA and the DEA method can be used as reliable methods of determining μ, if we give credit to the dynamic approach to DNA sequences of Ref. [141], adopted also in the ensuing papers of Refs. [143, 142]. By determining μ by means of the DFA and

plugging the corresponding expression in Eq. (6.3), we get

$$\delta = \frac{1}{(3-2H)}. \tag{6.5}$$

We denote this relation as Lévy Condition (LC). If the Gaussian case, which is the only case where a method based on measuring variances yields a reliable scaling prediction, we would get

$$\delta = H. \tag{6.6}$$

We call this Gaussian Condition (GC). We prove that the DEA makes it possible to assess which of the two, LC or GC, applies.

6.4 DEA of non-coding and coding DNA Sequences.

We are finally ready to discuss the results of the application of the DEA. First of all we focus on the *non-coding sequence* HUMTCRADCV. Fig. 6.4a shows that the DEA results in a scaling changing with time. This is pointed out by means of two straight lines of different slopes, $\delta' = 0.615 \pm 0.01$ and $\delta = 0.565 \pm 0.01$, corresponding respectively to the short-time and long-time region. Anomalous diffusion shows up at both the short-time and the long-time scale, and this seems to be a common characteristic of non-coding sequences, supported also by the application of our technique to other non-coding DNA sequences. Moreover, we notice that the scaling in the short-time regime $\delta' = 0.615 \pm 0.01$ coincides exactly with the value found by means of the DFA analysis [145], $H' = 0.61 \pm 0.01$. The authors of Ref. [145] assign this scaling value to both the short and the long-time regime, while here it appears clearly that the true long-time regime scaling is different. We note that this change of scaling corresponds to a transition from the the GC of Eq. (6.6), valid in the short-length scale, to the LC of eq. (6.5), valid in the large-length scale.

Fig. 6.4b shows the result of the DEA applied to an artificial sequence built up according to the CMM prescription so as to mimic the sequence HUMTCRADCV. In this case, the intensity of the random component is not predominant as in the case of the coding sequences, which are known [143] to require the condition of Eq. (6.2). In

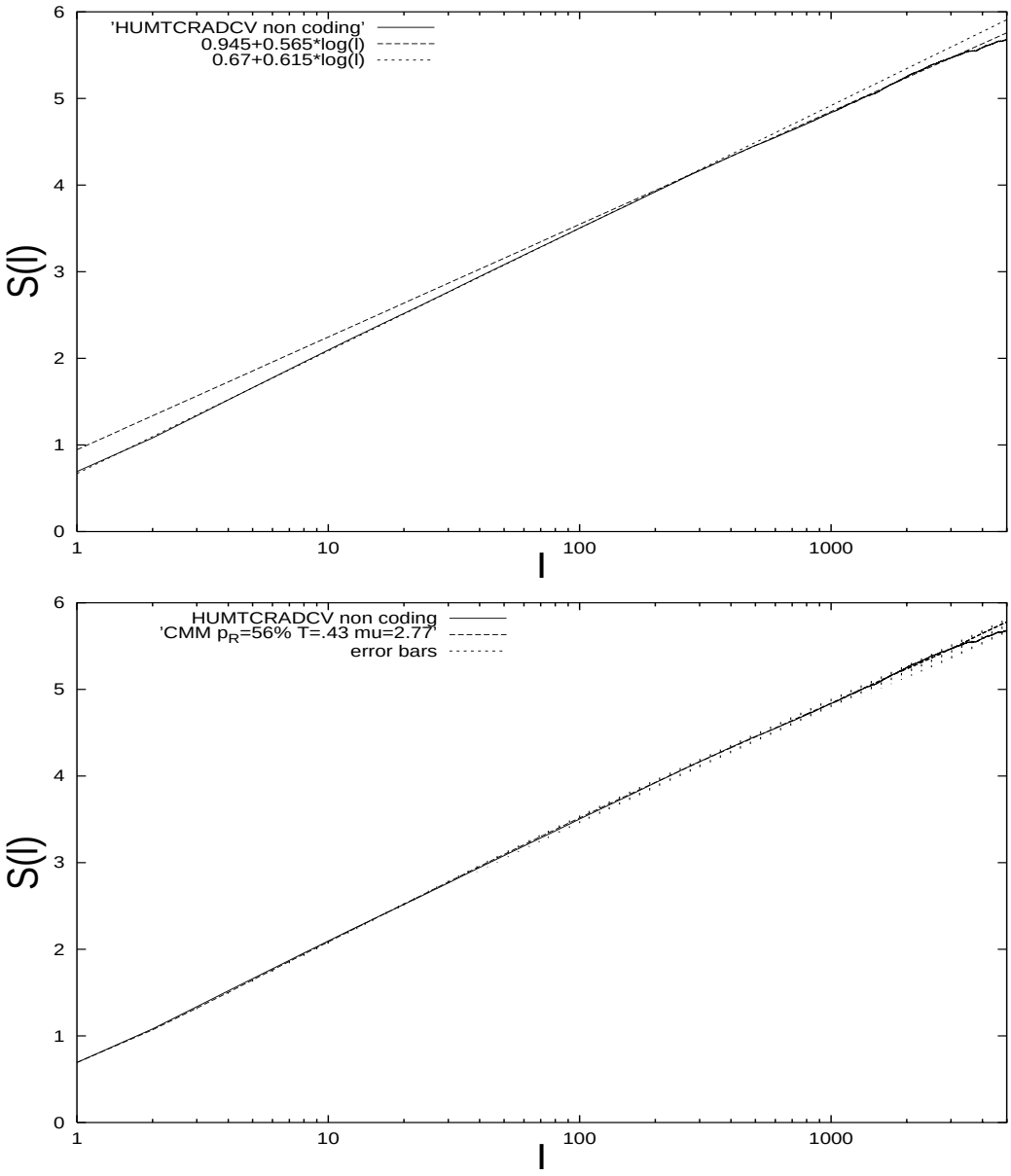

Figure 6.4: Diffusion Entropy and CMM simulation for the HUMTCRADCV, non-coding chromosomal fragment. Fig. 6.4a shows that the DE analysis results in a scaling changing with time. The slope of the two straight lines is $\delta' = 0.615 \pm 0.01$ at short-time regime, and $\delta = 0.565 \pm 0.01$ at long-time regime. Fig. 6.4b shows the comparison between the DEA of the real non-coding sequence and an artificial sequence corresponding to the CMM model: $p_R = 0.56 \pm 0.02$, $T = 0.43$, $\mu = 2.77 \pm 0.02$.

fact, in this case the best fit between the real and the CMM sequence is obtained by setting $p_R = 0.56 \pm 0.02$. Note that we set $\mu = 2.77 \pm 0.02$, which corresponds to the prescription of Eq. (6.3) in the short-time limit. In fact, if we plug $\mu = 2.77 \pm 0.02$ into Eq. (6.4), we get $H = 0.615 \pm 0.01$. This is the slope of the DE curve in the short-time limit. If, on the contrary, we plug $\mu = 2.77 \pm 0.02$ into Eq. (6.3) we obtain $\delta = 0.565 \pm 0.01$, which is the slope of the DEA curve in the long-time regime. This means that the random component has only the effect of improving the accuracy of the fitting, the scaling of whole time regime before the time l of the order of about 100 being essentially determined by the anomalous index $\mu = 2.77 \pm 0.02$. The important property confirmed by the CMM sequence is the crossover from short-time GC of Eq. (6.6) to the long-time LC of Eq. (6.5).

In Figs. 6.5 and 6.6 we turn to the more delicate problem of the *coding sequence*. The first sequence (ECO110K) has already been studied by means the DFA analysis in Ref. [145]. The DFA finds $H' = 0.52 \pm 0.01$ at the short-length scale and $H = 0.75 \pm 0.01$ in the large-length scale. The second sequence (ECOTSF) has been analyzed in Ref. [141] by using four different methods. The first was the second moment analysis of the diffusion process. This is a method of analysis less sophisticated than the DFA, since does not imply any local detrending. The second and third methods were the DFA and the Hurst analysis [70], respectively. The fourth method used was the Onsager regression analysis, a method that, in that context, provides information on the correlation function of the fluctuation ξ, which has an inverse power dependence on time l with the power index $\beta = \mu - 2$. The authors of Ref. [141], by using essentially the first method and the Onsager regression analysis, reached the conclusion that the most plausible value of the scaling parameter in the long-time region is $H = 0.75 \pm 0.01$ that is equivalent to the exponent $H = 0.74 \pm 0.01$ found in Figs. 6.3. It is interesting to remark that the coincidence among the different predictions about scaling, and especially that between the second moment technique and the Hurst analysis implies the adoption of the Gaussian assumption [149]. On the other hand, when that condition does not apply and the two scaling predictions are different, to the best of our knowledge, it does not seem to be known what is the meaning of any of them. Furthermore, the authors of Ref. [141] pointed out that the

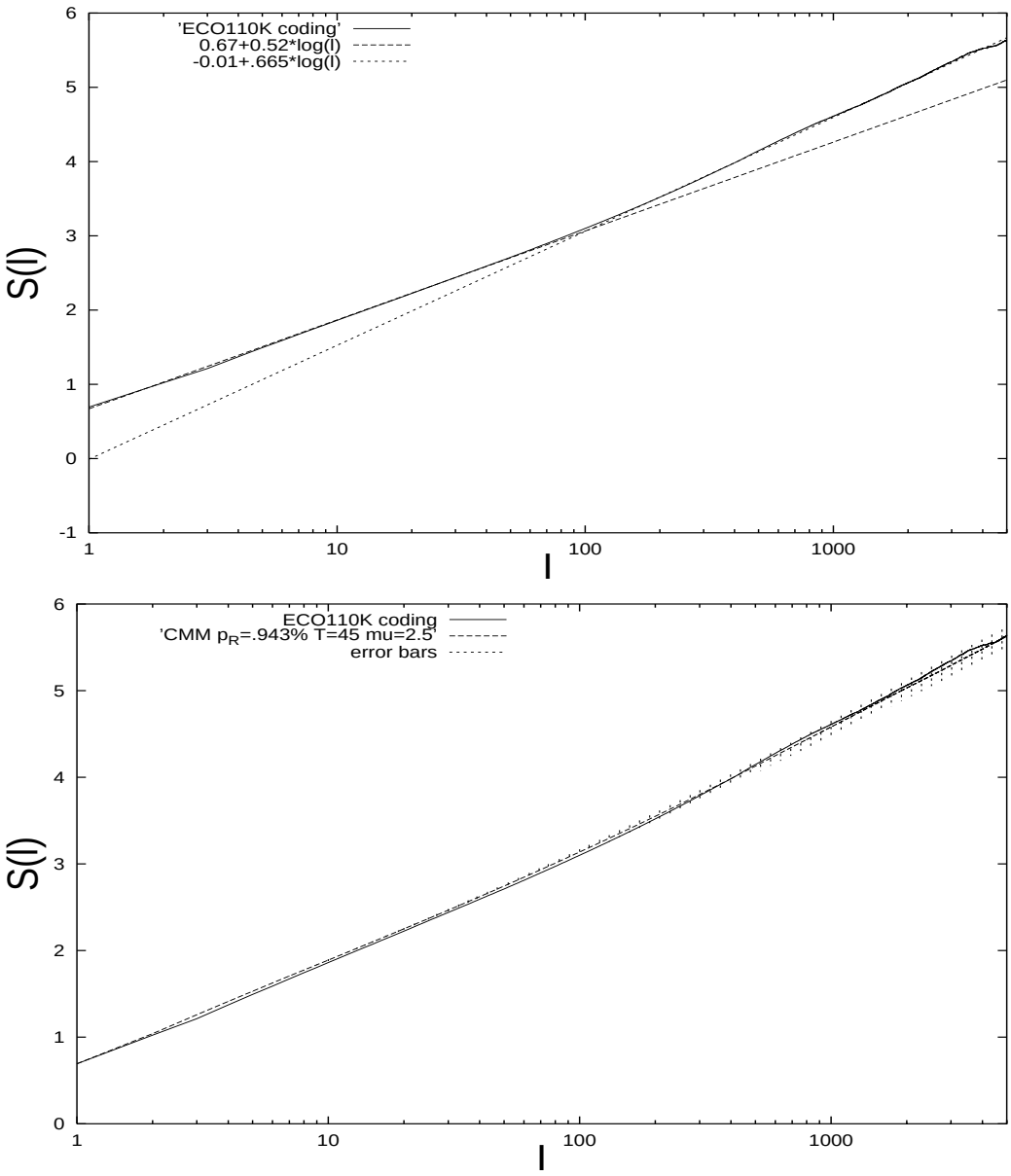

Figure 6.5: Diffusion Entropy and CMM simulation for the ECO110K, coding genomic fragment. Fig. 6.5a shows that the DEA results in a scaling changing with time. The slope of the two straight lines is $\delta' = 0.52 \pm 0.01$ at short-time regime, and $\delta = 0.665 \pm 0.01$ at long-time regime. Fig. 6.5b shows the comparison between the DE analysis of the real coding sequence and an artificial sequence corresponding to the CMM model: $p_R = 0.943 \pm 0.01$, $T = 45$, $\mu = 2.5 \pm 0.02$.

statistics of the long-time regime is too poor to support any claim on the departure from the Gaussian condition.

Finally the DEA illustrated in Figs. 6.5 and 6.6 affords a convincing settlement of the problems left open by the analysis of Ref. [141]. Figs. 6.5a and 6.6a clearly show the difference between the slope at short time, which, in this case, is very close to that of ordinary random walk, and the slope at long time that corresponds to $\delta = 0.665 \pm 0.01$. Since we know that in both cases the long-time slope provided by the DFA is $H = 0.75 \pm 0.01$, we conclude that in both cases the LC of Eq. (6.5) applies. Figs. 6.5b and 6.6b are instead devoted to a comparison with the CMM. A very good agreement is obtained by setting $p_R = 0.943 \pm 0.01$ for ECO110K (Fig. 5b) and $p_R = 0.937 \pm 0.01$ for ECOTSF (Fig. 6b). The results shown in the Figs. 6.5 and 6.6 are very close to each other. and are in accordance with the physical reasons that led the authors of Ref. [141] to propose the CMM model for coding sequences. In fact, with such a high weight, assigned to the random component, the scaling $\delta' = 0.52 \pm 0.01$ and $\delta' = 0.53 \pm 0.01$, very close to the conventional scaling $\delta = H = 0.5$, lasts for an extended period of time. At larger times a transition to a larger scaling takes place.

We note that the authors of Ref. [146] find anomalous diffusion in a statistical condition that they claim to be Gaussian. According to the result of Ref. [142] the Gaussian condition is incompatible with a stationary diffusion process generated by a dichotomous fluctuation yielding a non integrable correlation function with an inverse power law character. However, the authors of Ref. [142], with the help of a fractal model for the DNA folding, proved that the fractional Brownian motion advocated by the paper of Ref. [146] is possible as a form of non-stationary process. Thus, in principle we cannot rule out the possibility that the change of slope with time illustrated by Figs. 6.5a and 6.6a is a manifestation of that condition, implying Gauss statistics throughout the whole time range explored by the statistical analysis.

We see, on the contrary that Figs. 6.5 and 6.6, and Table 6.1 prove that the LC of Eq. (6.5) applies to both sequences. This means that in both cases the long-time limit is characterized by Lévy statistics and that this is the form of non-Gaussian statistics revealed by the analysis of Ref. [142].

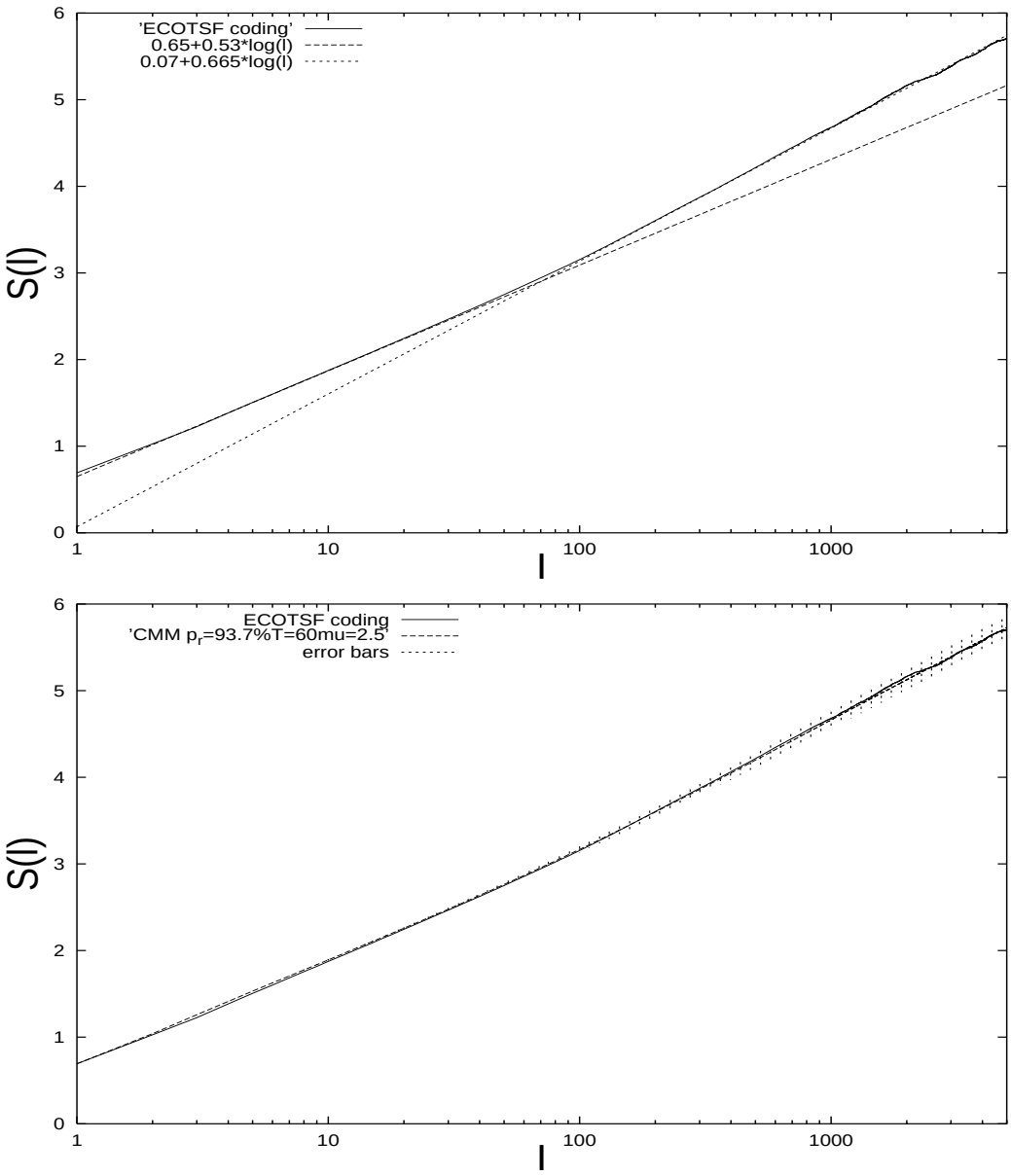

Figure 6.6: Diffusion Entropy and CMM simulation for the ECOTSF, coding genomic fragment. Fig. 6.6a shows that the DEA results in a scaling changing with time. The slope of the two straight lines is $\delta' = 0.53 \pm 0.01$ at short-time regime, and $\delta = 0.665 \pm 0.01$ at long-time regime. Fig. 6.6b shows the comparison between the DEA of the real coding sequence and an artificial sequence corresponding to the CMM model: $p_R = 0.937 \pm 0.01$, $T = 60$, $\mu = 2.5 \pm 0.02$.

Non-Coding	N	H	δ_H	δ
HUMTCRADCV	97630	0.61	0.56	0.56
CELMYUNC	9000	0.71	0.63	0.635
CHKMYHE	31109	0.78	0.69	0.70
DROMHC	22663	0.72	0.64	0.65
HUMBMYHZ	28437	0.58	0.54	0.54
Coding				
ECO110K	111401	0.74	0.66	0.66
ECOTSF	91430	0.74	0.66	0.66
LAMCG	48502	0.85	0.77	0.76
CHKMYHN	7003	0.74	0.66	0.66
DDIMYHC	6680	0.68	0.61	0.61
DROMYONMA	6338	0.69	0.62	0.64
HUMBMYH7CD	6008	0.63	0.57	0.58
HUMDYS	13957	0.69	0.62	0.62

Table 6.1: Values of the scaling exponents H and δ for coding and non-coding genomes. In the first column there is the GenBank name [134]. In the second column there is the length N of the genome. For all measures the error is ± 0.01. δ_H of the fourth column is the theoretical value for δ if the Lévy Condition applies, Eq. (6.5). If the length of the genome is larger than 20,000 the fitted region is $100 < l < 2000$. If the length of the genome is shorter than 20,000, the statistics are not very good for large l. In this case, the fitted region is $20 < l < 200$.

.

6.5 Significance of the results obtained.

To properly appreciate the significance of the results of this chapter, it is necessary to say a few words about the two different scaling prescriptions of Eqs.(6.3) and (6.4). The scaling prescription of Eq.(6.4) is determined by the adoption of the variance method, as clearly illustrated by the dynamical approach to the DNA sequences of Ref. [141]. This prescription is not ambiguous if the condition of Gaussian statistics applies. In fact, a Gaussian distribution drops quickly to zero, and the existence of a finite propagation front does not produce any significant effect. It has to be pointed out, in fact, that the adoption of the Brownian landscape proposed in the pioneer papers of Refs. [135, 138, 145] implies the existence of a propagation front moving with ballistic scaling ($\delta = 1$). In other words, if we find a window of length l filled with only 1 's or with only -1's, this means a trajectory travelling with uniform velocity, and the x-space at distances from the origin larger that l is empty. The existence of a propagation front does not have big consequences in the case of Gaussian statistics, since the population at the propagation front is essentially zero in that case. It is not so in the case of Lévy statistics, though, due to the existence of very long tails in that case. Therefore the Lévy processes resulting from these sequences are essentially characterized by the presence of two distinct scaling prescriptions, the Lévy prescription of Eq.(6.3), concerning the portion of distribution enclosed between the two propagation fronts, and the scaling $\delta = 1$, of the propagation front itself. The scaling of the variance of Eq.(6.4) does not reflect correctly either of these two different scaling prescriptions, being a kind of compromise between the two. The scaling of the distribution enclosed by the two propagation fronts is, on the contrary, a genuine property that corresponds to the prediction of the generalized central limit theorem [150]. It is very satisfactory indeed that the DEA method makes this genuine form of scaling emerge. Furthermore, the DEA is a very accurate method of scaling detection, as proved by the fact that it reveals the existence of Lévy statistics in the case of the coding sequence. In this case, as pointed out by the authors of Ref. [141], the ordinary methods become inaccurate due to the poor statistics available in the long-time limit.

Another important result of this chapter is that it confirms the validity of the

CMM model. This model is expected to generate Lévy statistics not only in the case of non-coding sequences, where it is easier to reveal this property. It predicts Lévy statistics also in the case of coding sequences as the one here analyzed. In Ref. [141] the emergence of Lévy statistics was conjectured but not proved, due to the fact that in that paper the observation was made monitoring the probability distribution $p(x,t)$. The proof of the emergence of Lévy statistics would imply the detection of distribution tails with an inverse power law, a property difficult to assess, due to the poor statistics of the long-time limit. In Ref. [143] a clear deviation from the Gaussian condition was detected in the long-time limit, but, again, no direct evidence was found that this deviation from Gaussian statistics is due to the emergence of Lévy statistics. The results of this chapter prove, with the help of the artificial sequences of Sections 2.6 and 6.3, that the DEA is method of analysis is so accurate as to consider the detection of the property LC of Eq. (6.5) as a compelling evidence that the CMM is a correct way of modelling DNA sequences.

CHAPTER 7

HARD X-RAY SOLAR FLARES

The study of solar flares is becoming popular among the researchers working at the frontier of statistical mechanics, due to the widely shared conviction that they are a signature of a significant departure from the condition of ordinary Brownian motion [151, 152, 153, 154]. As pointed out by Wheatland, [155], the distribution of times between flares, gives information on how to model flare statistics. Although the agreement on the fact that flare statistics depart from ordinary statistical mechanics is general, there seems to be the still unsettled issue of what is the proper model which will account for this form of anomalous statistics. Does this form of statistics reflect self-organized criticality or turbulence [153]? We think that the settlement of this delicate issue is made difficult by the fact that, although many authors claim that $\psi(\tau)$ is an inverse power law with power index μ, the actual value of μ still seems to be uncertain. In fact, the authors of Ref. [151] propose $\mu = 1.7$ and those of Ref. [152] claim that $\mu = 2$ is the proper power law index. Boffetta et al. [153] propose $\mu = 2.4$. Finally, Wheatland explains the origin of the power law behavior with a model yielding $\mu = 3.0$, [155].

This chapter is devoted to analyzing the distribution of time distances τ between two nearest-neighbor flares. This analysis is based on the joint use of two distinct techniques. The first is the direct evaluation of the waiting time distribution function $\psi(\tau)$, or of the probability, $\Psi(tau)$, that no time distance smaller than a given τ is found. We adopt the paradigm of the inverse power law behavior, and we focus on the determination of the inverse power index μ, without ruling out different asymptotic properties that might be revealed, at larger scales, with the help of richer statistics. The second technique is the Diffusion Entropy Analysis that rests on the evaluation of the entropy of the diffusion process generated by the time series. The details of the diffusion process depend on walking rules. We adopt the SJM, AJM, and LJM that are discussed in Chapter 2. This three rules determine the form and the time duration of the transition to the scaling regime, as well as the scaling parameter δ.

With the first two rules the information contained in the time series is transmitted, to a great extent, to the transition, as well as to the scaling regime. The same information is essentially conveyed, by using the third rule, into the scaling regime, which, in fact, emerges very quickly after a fast transition process. We show that the significant information hidden within the time series concerns memory induced by the solar cycle, as well as the power index μ. The scaling parameter δ becomes a simple function of μ, when memory is annihilated by shuffling the data. Thus, the three walking rules yield a unique and precise value of μ if the memory is carefully taken under control, or cancelled by shuffling the data. All this leads to a compelling conclusion that $\mu = 2.138 \pm 0.01$ and, at the same time, proves that the hard x-ray solar flares statistics are more complex than expected on the basis of the waiting time distribution alone.

7.1 Statistical analysis of the real data: $\psi(\tau)$ and $\Psi(\tau)$.

In this section we plan to derive the waiting time distribution $\psi(\tau)$ directly from the statistical analysis of the real data, the x-rays emitted by solar flares in the case here under study. At first sight, one might think that a direct determination of $\psi(\tau)$ is more convenient than any indirect approach. Actually, it is not so. As mentioned in Section I, we find that the evaluation of the probability of finding no time distance larger than a given τ, denoted by $\Psi(\tau)$, defined by

$$\Psi(\tau) \equiv \int_{\tau}^{\infty} \psi(t)dt, \tag{7.1}$$

is more appropriate than the direct evaluation of $\psi(\tau)$. In later sections we shall prove a striking property: the evaluation of μ through the DEA, an approach less direct than the evaluation of $\Psi(\tau)$, is still more efficient.

The data are a set of 7212 hard x-ray peak flaring event times obtained from the BATSE/CGRO (Burst and Transient Source Experiment aboard the Compton Gamma Ray observatory satellite) solar flare catalog list. The data is a nine-year series of events from 1991 to 2000. If the time Δt between two consecutive solar flares

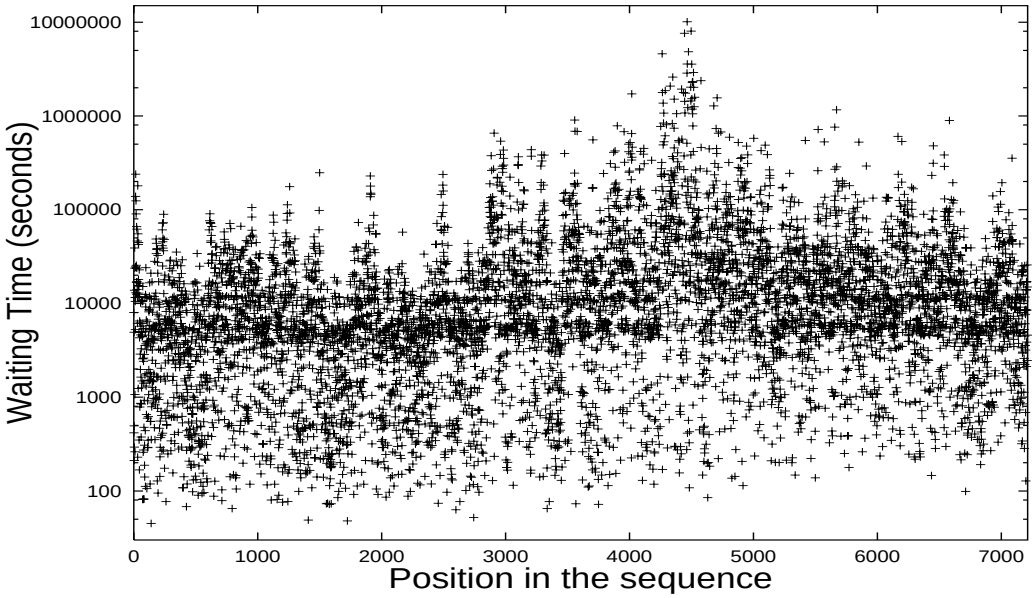

Figure 7.1: The original sequence of the solar flares waiting times. Note the logarithmic scale of ordinates.

is expressed in seconds, the range goes from 45 to 10,000,000 seconds, as shown in Fig. 7.1. Fig. 7.2 shows the rate of solar flares per month from April 1991 to May 2000. The set of data studied here concerns a time period of 9 years, and, consequently, a large part of the whole 11-year solar cycle. Fig. 7.2 shows that during a large portion of this 11-year cycle the flare rate undergoes big changes, thereby significantly departing from the uniform distribution. Furthermore, it is worth remarking that, as shown by Fig. 7.3, the 11-year solar cycle is not a mere harmonic oscillation with the period of 11 years, but a complex dynamic process with many components.

The direct evaluation of the waiting time distribution, $\psi(\tau)$, needs the data to be distributed over many bins with the same size. When only a few data are available, the bin size cannot be too small, and, in turn, the adoption of bins of large size can produce incorrect power law indices. In proceeding with the direct evaluation of the key parameter μ, first of all, we have to adopt a proper criterion to determine the size Δ_i of the i-th bin. We note that the waiting time distribution is expected to be an inverse power law. If we adopted bins of equal size, those corresponding to

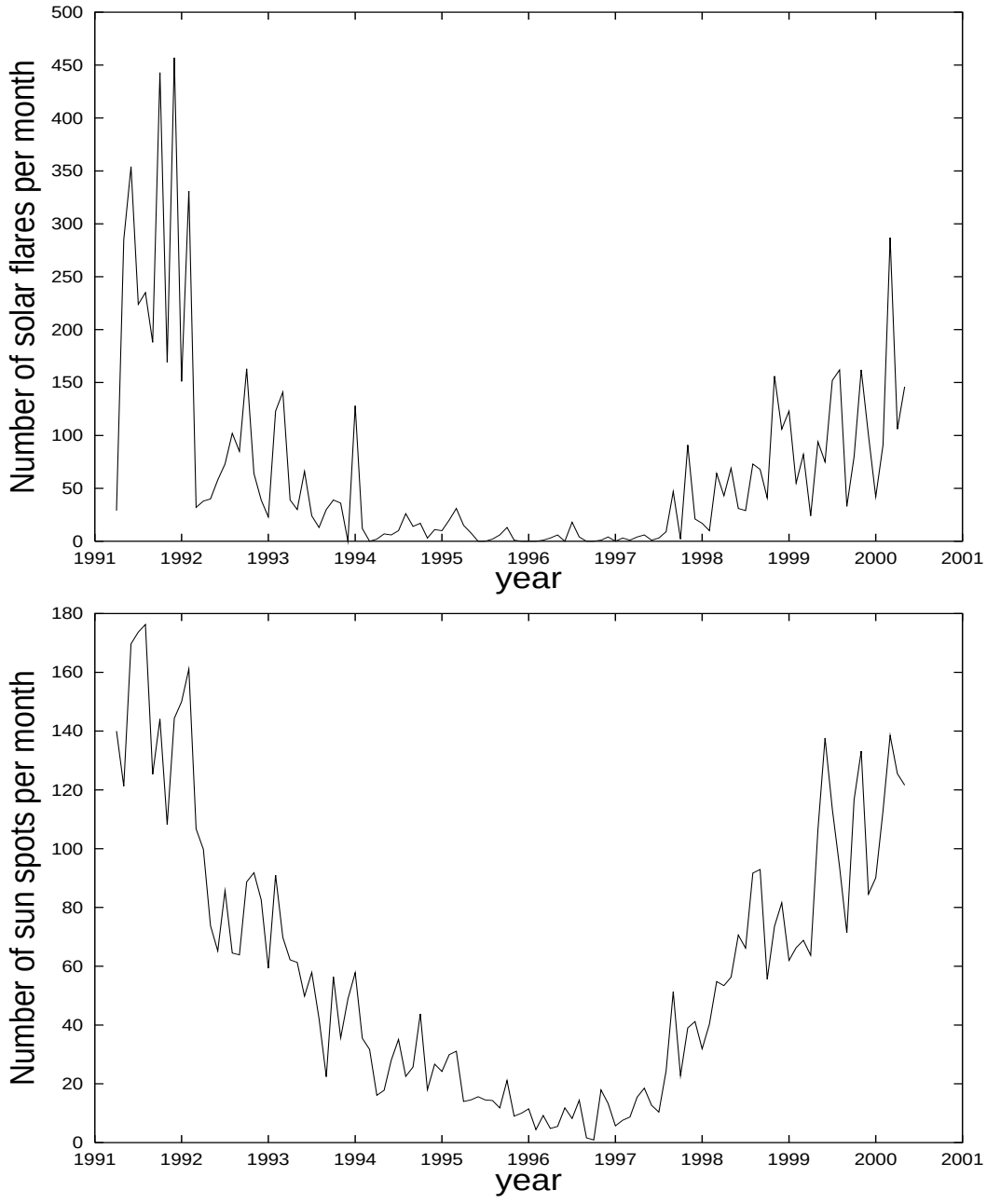

Figure 7.2: Number of solar flares and sun spots per month from April 1991 to May 2000. The two phenomena follow the same solar cycle.

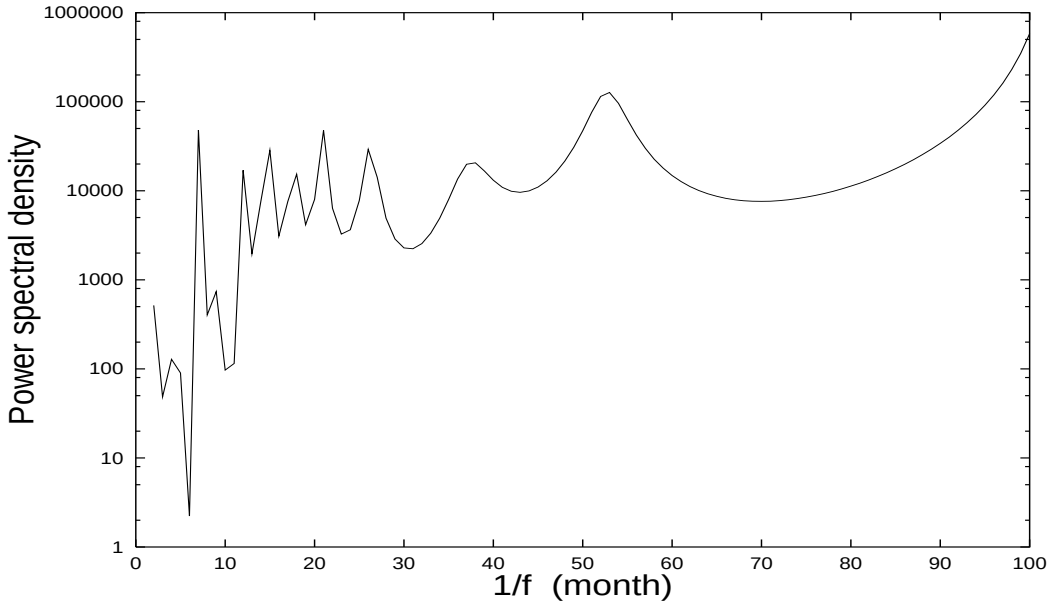

Figure 7.3: The solid curve was obtained by using the maximum entropy method [133].

large times would collect a very limited amount of data, thereby resulting in a non reliable evaluation of the frequencies. To bypass this difficulty we adopt bin sizes that are constant in the logarithmic scale. This means that $ln(\tau_i) - ln(\tau_{i-1})$ is constant, where τ_i and τ_{i-1} are the middle times of two consecutive bins. We define the width of the i-th bin as $\Delta = \tau_i - \tau_{i-1}$, thereby making it become an exponentially increasing function of the sequence position, so as to widely compensate for the density decrease. In this representation the probability density $\psi(\tau_i)$ is expressed by

$$\psi(\tau_i) = \frac{N_i}{N\Delta_i} \ , \tag{7.2}$$

where N is the total number of data points, N_i is number of points located within the i-th bin, and Δ_i, as earlier said, is the width of the i-th bin.

The fitting is done by using the prescription of a inverse power law of the type

$$\psi(\tau) = \frac{A_1}{(T+\tau)^{\mu}} \ , \tag{7.3}$$

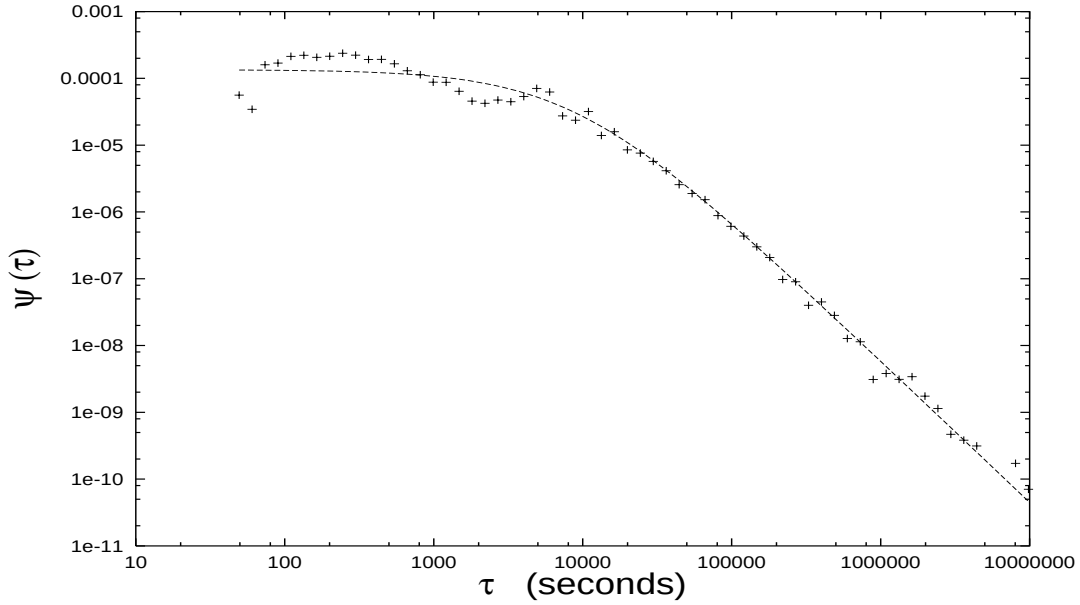

Figure 7.4: The waiting time distribution $\psi(\tau)$ as a function of τ. The crosses refer to real data. The dashed line is the fitting function of Eq. (7.3) with $A_1 = 31006$, $T = 8787$ and $\mu = 2.12$.

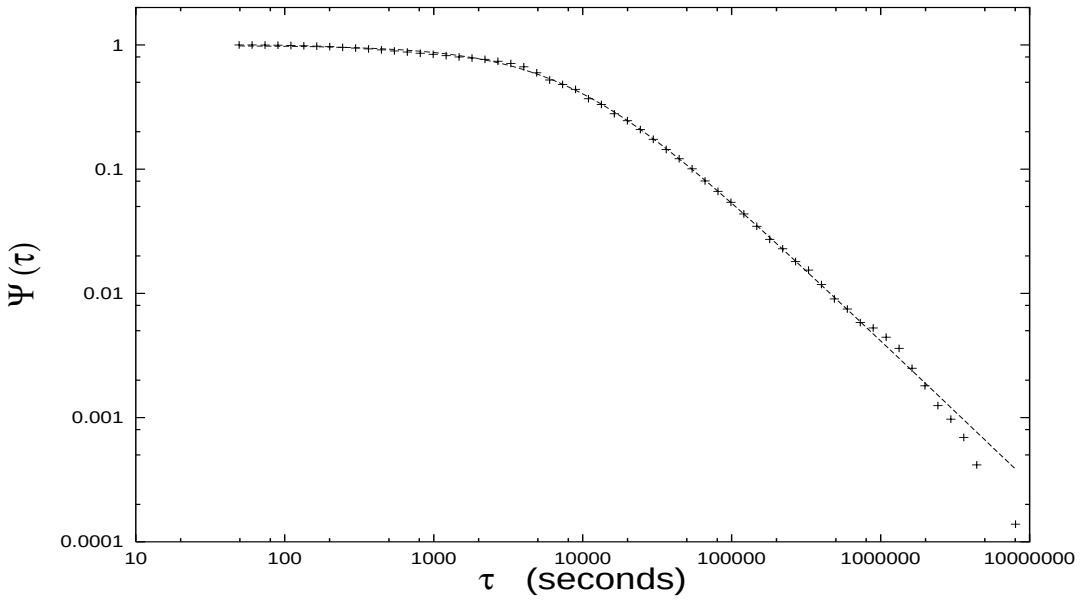

Figure 7.5: $\Psi(\tau)$ as a function of τ. The crosses refer to real data, and the dashed line denotes the fitting function of Eq. (7.4) with $A_2 = 30567$, $T = 8422$ and $\mu = 2.144$.

with A_1, T and μ being three independent fitting parameters. It is worth noting that the normalization condition reduces the three independent parameters to two which are a function of only T and μ. We find it necessary to adopt three independent fitting parameters, with the understood proviso that the departure of A_1 from the value $(\mu - 1)T^{\mu-1}$ can be interpreted as a way to estimate the inaccuracy of the adopted fitting procedure.

The fitting is done by using an implementation of the nonlinear least-squares (NLLS) Marquardt-Levenberg algorithm [156]. The NLLS algorithm may not give unique values for the fitting parameters. It needs initial guesses for the free parameters and the final results may change or be affected by huge errors. This fitting procedure yields: $T = 8787$, $\mu = 2.12 \pm 0.32$ and $A_1 = 31006$. The evaluated value of A_1 is not far from the value 29236 that would be required by the normalization condition. However, there are very large errors of the order of 100%, with an error on the parameter μ of the order of 15%, thereby implying $1.80 < \mu < 2.44$. This means that the result of this fitting procedure would prevent us from assessing the important question on whether the process is stationary ($\mu > 2$) or non stationary ($\mu < 2$). The large error of this procedure depends upon the initial values assigned to the three fitting parameters, T, μ and A_1, whose choice requires a more efficient criterion. It also depends on the fact that there are oscillations around the fitting curve, as clearly illustrated by Fig. 7.4.

As earlier mentioned several times, a more accurate fitting is obtained using the function $\Psi(\tau)$. Again we do not pay attention to the normalization constraints and we adopt the following fitting function

$$\Psi(\tau) = A_2 \left(\frac{1}{T+\tau}\right)^{\mu-1} . \tag{7.4}$$

As shown by Fig. 7.5, the fitting of the real data is now much more accurate than that of Fig. 7.4. The fitting parameters used are: $A_2 = 30657 \pm 16590$, $T = 8422 \pm 500$, $\mu = 2.144 \pm 0.05$. This sets on the key parameter μ the constraint $2.094 < \mu < 2.194$, which has the very attractive property of establishing the stationary nature of the dynamic model behind the solar flares fluctuations. The results of this search for μ,

$\psi(\tau)$	$\Psi(\tau)$
$1.80 < \mu < 2.44$	$2.094 < \mu < 2.194$

Table 7.1: μ evaluated by using $\psi(\tau)$ and by using $\Psi(\tau)$.

based on the direct evaluation of $\psi(\tau)$ and on the use of $\Psi(\tau)$, are summarized in Table 7.1. We note that the uncertainty interval associated with the use of $\Psi(\tau)$ is contained within the wider uncertainty interval produced by the use of $\psi(\tau)$. This means that we are coming closer to the real value of μ. The width of the uncertainty interval will be further reduced by using the DEA.

7.2 Diffusion Entropy of solar flares.

This section is devoted to the analysis of the solar flares data by means of the DEA. The final result will be given by $\mu = 2.138 \pm 0.01$, namely a value for μ even more accurate than that obtained in Section 6.1 by using $\Psi(\tau)$. We shall prove also that the DEA allows us to establish some aspects of the dynamics behind solar flares that would be overlooked by an analysis based only on the use of the waiting time distribution.

The first issue that we have to solve is how to process the data so as to apply the three walking rules: SJM, AJM, and LJM that are discussed in Chapter 2. For commodity, let us report here the scaling properties of the three rules. The prescriptions are the following:

$$\delta = \begin{cases} \mu - 1, & 1 < \mu < 2 \\ 1/(\mu - 1), & 2 < \mu < 3 \\ 0.5, & \mu > 3, \end{cases} \tag{7.5}$$

$$\delta = \begin{cases} 0.5(\mu - 1), & 1 < \mu < 2 \\ 0.5, & \mu > 2 \end{cases} \tag{7.6}$$

and

$$\delta = 1/(\mu - 1), \qquad \mu > 1, \tag{7.7}$$

for rules No. 1 or AJM, No. 2 or SJM and No. 3 or LJM, respectively.

The data accessible to us are the times $\tau_i = t_i - t_{i-1}$, with t_i and t_{i-1} denoting the time of occurrence of the i-th and the (i-1)-th solar flare, respectively. However, the direct adoption of these numbers would result in technical difficulties that are bypassed by referring ourselves to the new sequence of numbers

$$\beta_j = Int\left[\frac{\Delta t_j}{\Lambda}\right] + 1, \tag{7.8}$$

where $Int[x]$ denotes the integer part of x. The adoption of $\Lambda = 1$ would be virtually equivalent to referring ourselves to the original sequence of numbers. However, preliminary trials with changing values of Λ led us to conclude that there are problems with the adoption of both excessively small and excessively large values of Λ. The adoption of excessively small values of Λ would make the computer analysis too slow and would require an excessively large amount of computer memory. This is the reason why we cannot use the original sequence of numbers. The adoption of excessively large values of Λ, on the other hand, would produce statistical saturation, and a consequent sub-diffusion process that would not accurately reflect the dynamics behind the data. We adopted the criterion of using the largest value of Λ compatible with negligible saturation effect. Preliminary attempts made it possible for us to assess that this convenient value is given by $\Lambda = 3600$.

After processing the data, we have to realize the three walking rules. We note that diffusion is generated by the random walker jumping at any time step. The random walker makes jumps of intensity $|\xi_i|$, ahead or backward, according to whether $\xi_i > 0$ or $\xi_i < 0$. Thus, we create a new sequence ξ_i, of 0's and 1's, with the following prescription. We consider a sequence of infinite empty sites, labelled by the integer index i, considered as a discrete time, running from $i = 1$ to $i = \infty$. We divide this sequence into patches of width β_j. The first patch consists of the sites $i = 1$, $i = 2$,.., $i = \beta_1$, the second patch consists of the sites $i = \beta_1 + 1$, $\beta_1 + 2$, .., $\beta_1 + \beta_2$, and so on. We assign the value 0 to all the sites of the same patch but the last site. This means that the random walker walks only at the end of the patch, namely, at the occurrence time of an event. To apply rule No. 1, AJM, with the random walker

always moving in the same direction, we always assign to the last site of a given patch the value of 1. To apply rule No. 2, SJM, we assign to the last site of any patch either the value 1 or the value -1, according to the coin tossing rule. The coin tossing prescription is realized by using a random number generator. To reduce the risk of artificial periodicity we create 10 different sequences, each corresponding to a different random distribution of 1's and -1's. For any sequence, we run the DE method and then we make the average over the 10 resulting DE curves. To apply the rule No. 3, LJM, which will be shown in action in Section 6.3.3, we have to identify ξ_i with β_i.

The DEA results obtained applying rule No. 1 are illustrated in Fig. 7.6. This figure shows one of the benefits of the DEA. According to rule No. 1, we have to use the prescription of Eq. (7.5). The most accurate of the values of μ, discussed in Section V, is $\mu = 2.144$. This value, being smaller than 3 and larger than 2, makes us adopt the formula $\delta = 1/(\mu - 1)$, and yields the scaling parameter $\delta = 0.874$, which is the slope of the straight line of Fig. 7.6.

This theoretical prediction implies that the times τ_i of the sequence $\{\tau_i\}$ are not correlated with each other. In the specific case of seasonal periodicity described by harmonic oscillations, the numerical results of Ref. [7] prove that the scaling detected by the DE, as well as by other methods to detect scaling, is higher than the Brownian motion scaling $\delta = 0.5$. This is so even when there is no correlation in addition to seasonal periodicity. We eliminate this effect, by shuffling the data. The DEA can be applied to both the original sequence of β_i and to the shuffled sequence. If the DEA yields two different curves, this is a proof of the fact that there is memory in the original sequence. This is an important property that cannot be revealed by the analysis of the waiting time distribution, $\psi(\tau)$. Fig. 7.6 shows that this is the case. In fact we see that the DEA curve corresponding to the shuffled data, after the transition region at short time and before saturation, has a slope distinctly smaller than the curve referring to the non shuffled data. Furthermore, this slope is closer to the slope of the straight line corresponding to the finding of Section 7.2, which yields $\mu = 2.144$, and, consequently, according to Eq. (7.5), $\delta = 0.874$. However, both shuffled and non-shuffled data yield saturation effects at a time scale of the order of $t_{sat} = 1,500$

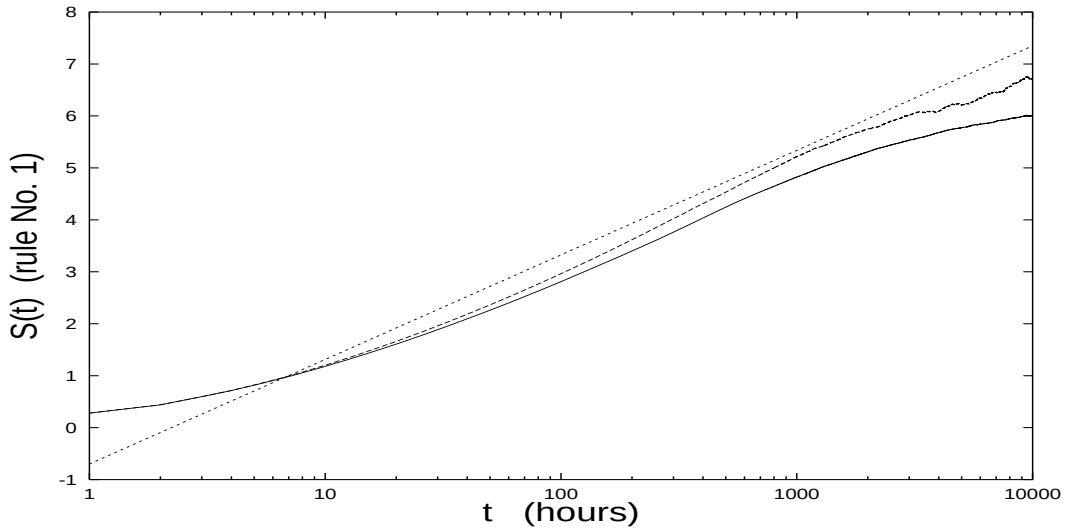

Figure 7.6: DE as a function of time according to rule No. 1. The dotted straight line illustrates the slope of entropy increase corresponding to $\mu = 2.144$, and $\delta = 0.874$, which is the best value of μ afforded by the analysis of Section V. The dashed line is the DEA curve generated by the non-shuffled real data. The solid line is the DEA curve generated by the shuffled real data.

hours. These saturation effects set limits to the accuracy of the determination of the value of μ by means of the DEA.

In Fig. 7.7 we illustrate the results obtained by using rule No. 2. It is remarkable that in this case the shuffled data yield, with the DEA, an entropy increase faster (rather than slower) than the non-shuffled data. This is a consequence of the fact that in this case the deviation from ordinary diffusion, produced by time periodicity, would generate sub-diffusion rather than super-diffusion. We notice that the difference between the shuffled and non-shuffled curves is smaller than that in the case of Fig. 7.6 (rule No. 1) and that the saturation effects show up at later times. We thus conclude that rule No. 2 is much less sensitive to periodicities and to saturation effects than rule No. 1.

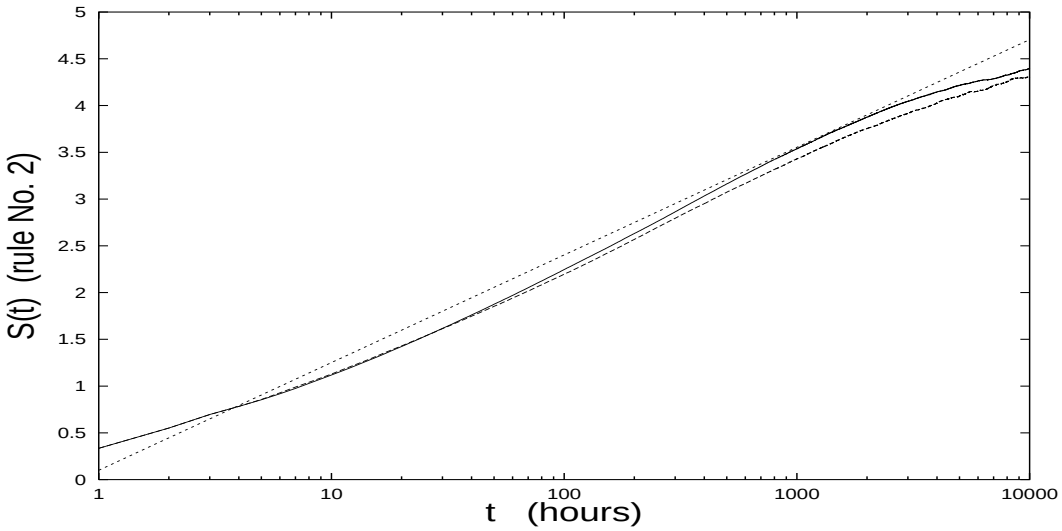

Figure 7.7: DE as a function of time according to rule No. 2. The dotted straight line illustrates the slope of entropy increase corresponding to $\mu = 2.144$, $\delta = 0.5$, which is the best value of μ afforded by the analysis of Section V. The dashed line is the DEA curve generated by the non-shuffled real data. The solid line is the DEA curve generated by the shuffled real data.

7.3 A further improvement: use of artificial sequences.

We have seen that the DEA reveals the existence of memory effects that are overlooked by the direct evaluation of the waiting time distribution. However, as illustrated by the numerical results of Section 7.2, the time region where the DE method might be fruitfully used to detect scaling, is reduced to an intermediate time region, after the transition from dynamics to thermodynamics, and before the saturation effects. This has the unwanted effect of setting limitations to the accuracy of the DE method. To bypass this difficulty we generate artificial sequences with the same statistical limitations of the real data, and then we search for the parameter μ that establishes the most accurate fitting with the DEA curves derived from real data.

To make this procedure as reliable as possible we proceed as follows. We assume that $\psi(\tau)$ has the form

$$\psi(\tau) = \frac{A}{(T+\tau)^{\mu}} \ , \tag{7.9}$$

where T and μ are our fitting parameters. The constant A is determined by the normalization condition through

$$\frac{1}{A} \equiv \int_{45}^{\infty} \frac{1}{(T+\tau)^{\mu}} d\tau \ . \tag{7.10}$$

The fitting parameters are made to change around the mean values established by the results of Section V which yield $\mu = 2.144 \pm 0.05$ and $T = 8422 \pm 500$. Note that in the real data no time exists with a value smaller than $\tau = 45$ sec. This is the reason why the integration in Eq. (7.10) is done from 45 to ∞ rather than from 0 to ∞. The number of data available to us are 7211. Thus we produce 7211 values of τ_i, according to the prescription

$$\tau_i = \left[\frac{1}{(T+45)^{\mu-1}} - \frac{(\mu-1)\ y_i}{A} \right] - T \ , \tag{7.11}$$

with the number y_i randomly selected in the interval $[0, 1]$. It is straightforward to prove that the resulting distribution of τ_i is the same as that of Eq. (7.9) and fits the condition of Eq. (7.10). At this stage we are ready to compare the DEA curves generated by the artificial data to the DEA curves generated by the real data, using both rule No. 1 and rule No. 2. The comparison is made with the DE curves corresponding to shuffled data, since the artificial sequences are generated without correlation among the numbers τ_i.

Let us discuss first the results concerning rule No. 1. These results are illustrated in Figs. 7.8. In Fig. 7.8a we show the effect of changing μ in the interval $[2.094, 2.194]$, with $T = 8422$ and in Fig. 7.8b we show the effect of changing T in the interval $[7922, 8922]$, with $\mu = 2.144$. We see that the DE curves of the artificial sequences fluctuate within an error strip containing the DEA curve of the real data. The size of this error strip increases upon change of time and we see that the spreading caused by the change of T is much smaller than that caused by the change of μ. From a qualitative point of view, the results concerning rule No. 2, shown in Figs. 7.9a and 7.9b, are very similar.

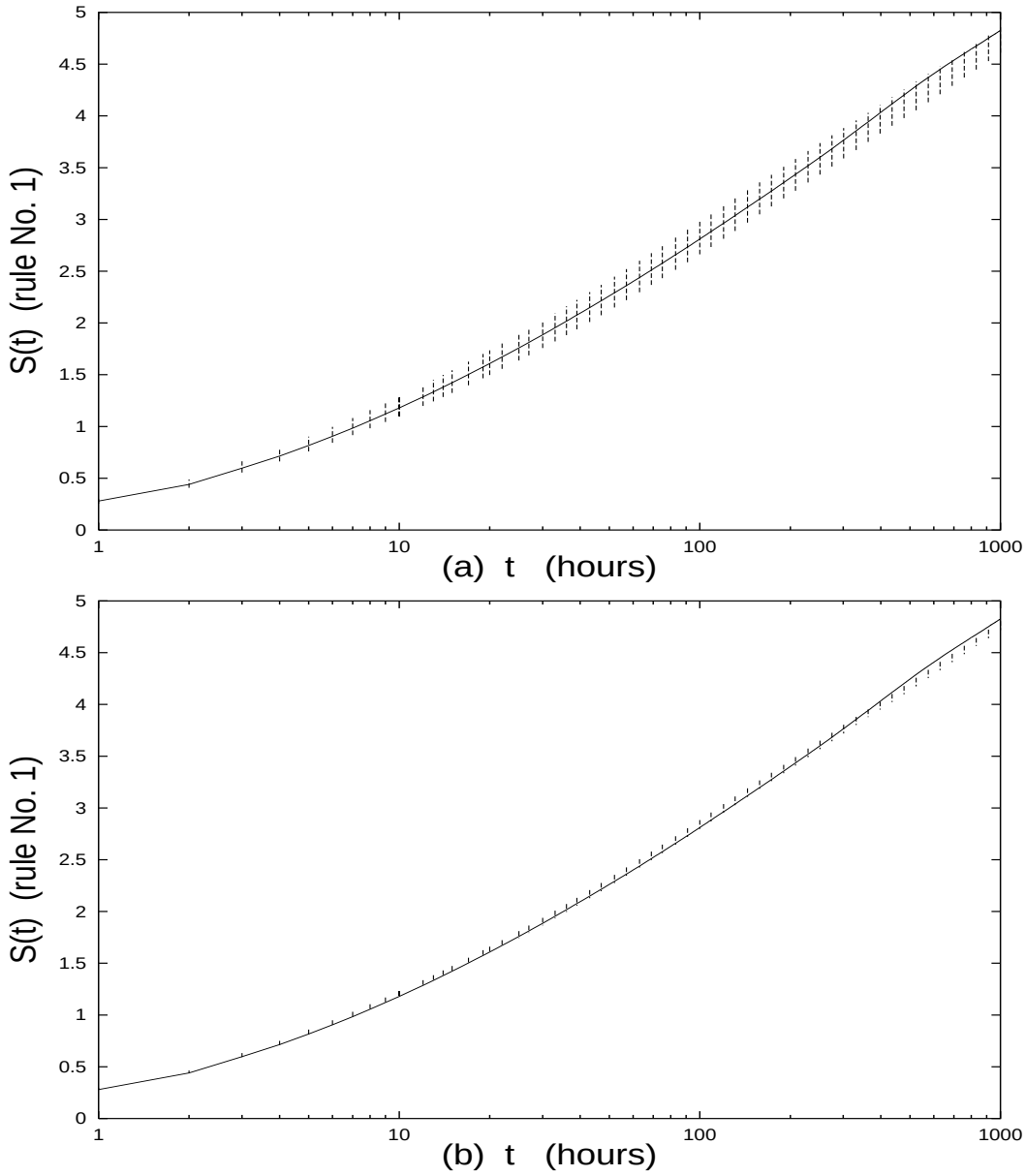

Figure 7.8: DE as a function of time according to the rule No. 1. The two solid curves denote the DEA curve corresponding to the shuffled real data. (a) The vertical bars indicate the changes of the DE curves resulting from the artificial sequences described in the text with $T = 8422$ and μ moving in the interval [2.094, 2.194]. (b) The vertical bars indicate the changes of the DE curves resulting from artificial sequences described in the text with $\mu = 2.144$, and T moving in the interval [7922, 8922].

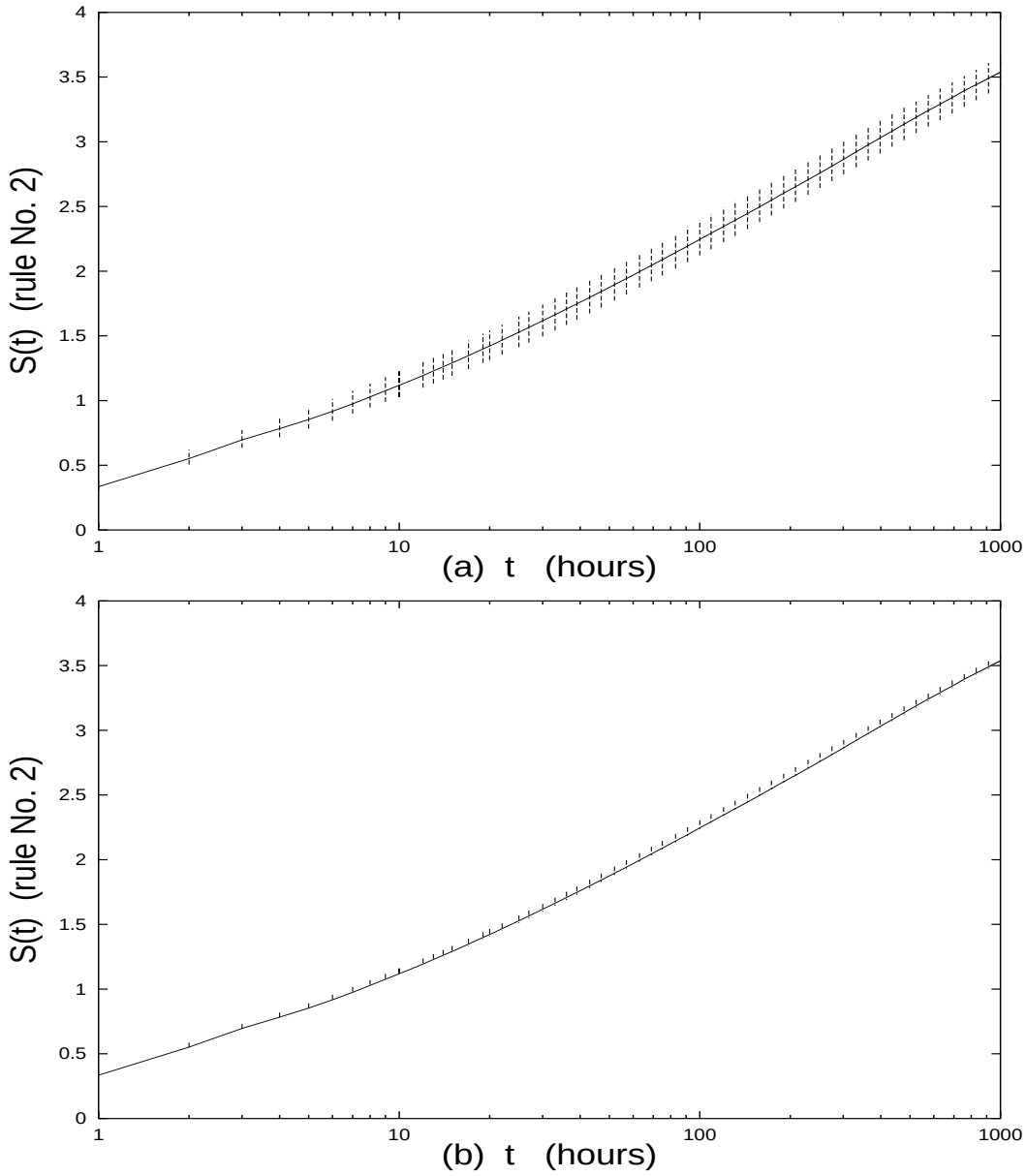

Figure 7.9: DE as a function of time according to the rule No. 2. The two solid curves denote the DEA curve corresponding to the shuffled real data. (a) The vertical bars indicate the changes of the DE curves resulting from the artificial sequences described in the text with $T = 8422$ and μ moving in the interval [2.094, 2.294]. (b) The vertical bars indicate the changes of the DE curves resulting from artificial sequences described in the text with $\mu = 2.144$, and T moving in the interval [7922, 9922].

7.3.1 A more accurate measurement of μ.

We have seen that the area of the T-error strip is significantly smaller than that of the μ-error strip, at least five times smaller. Therefore, we can improve the accuracy of μ by assigning to T a fixed value and looking for the value of μ ensuring the best fitting of the real data. We assign to T the value of 8422, and we proceed with the search for the best fitting. The results are illustrated in Figs. 7.10a and 7.10b. The result concerning rule No. 1 is good, as seen in Fig. 7.10a. As expected, Fig. 7.10b shows that the result concerning rule No. 2 is even better, and we think that it can be judged to be excellent. This extremely accurate result is due to the DEA curve of the artificial sequence coinciding with the DEA curve of real data over the wide range of 1000 hours of diffusion. On the basis of this excellent fitting, we conclude that

$$\mu = 2.138 \pm 0.01 \ . \tag{7.12}$$

7.3.2 Non shuffled data and an artificial sequence with suitable memory.

In Section 7.2, we have noticed that the result of the DEA depends on whether the real data are shuffled or not. We think that in the original data there are signs of the 11-year solar cycle and other subcycles. This makes it harder to establish a connection between the scaling δ and the power index μ. However, if our conclusion that $\mu = 2.138 \pm 0.01$ is correct, it should be possible to fit the DEA curve of the non-shuffled original data with no further change of the fitting parameters T and μ, provided that we sort the artificial sequence in such a way as to mimic the solar periodicity. Rather than doing that with a model we proceed in a more direct way, according to the following procedure. Let us call R_i and A_i the $i-th$ numbers of the real and artificial sequence used in subsection A, respectively. The $i-th$ number of the sorted artificial sequence is denoted by S_i. The subscript i ranges from 1 to N. The number S_1 is fixed by selecting from the set of A_i's the number that is closest to R_1, this being, let us say, $A_{j(1)}$. We thus set $S_1 = A_{j(1)}$. The number $A_{j(1)}$ is eliminated from the artificial sequence. Then, we move to R_2 and from the set of the

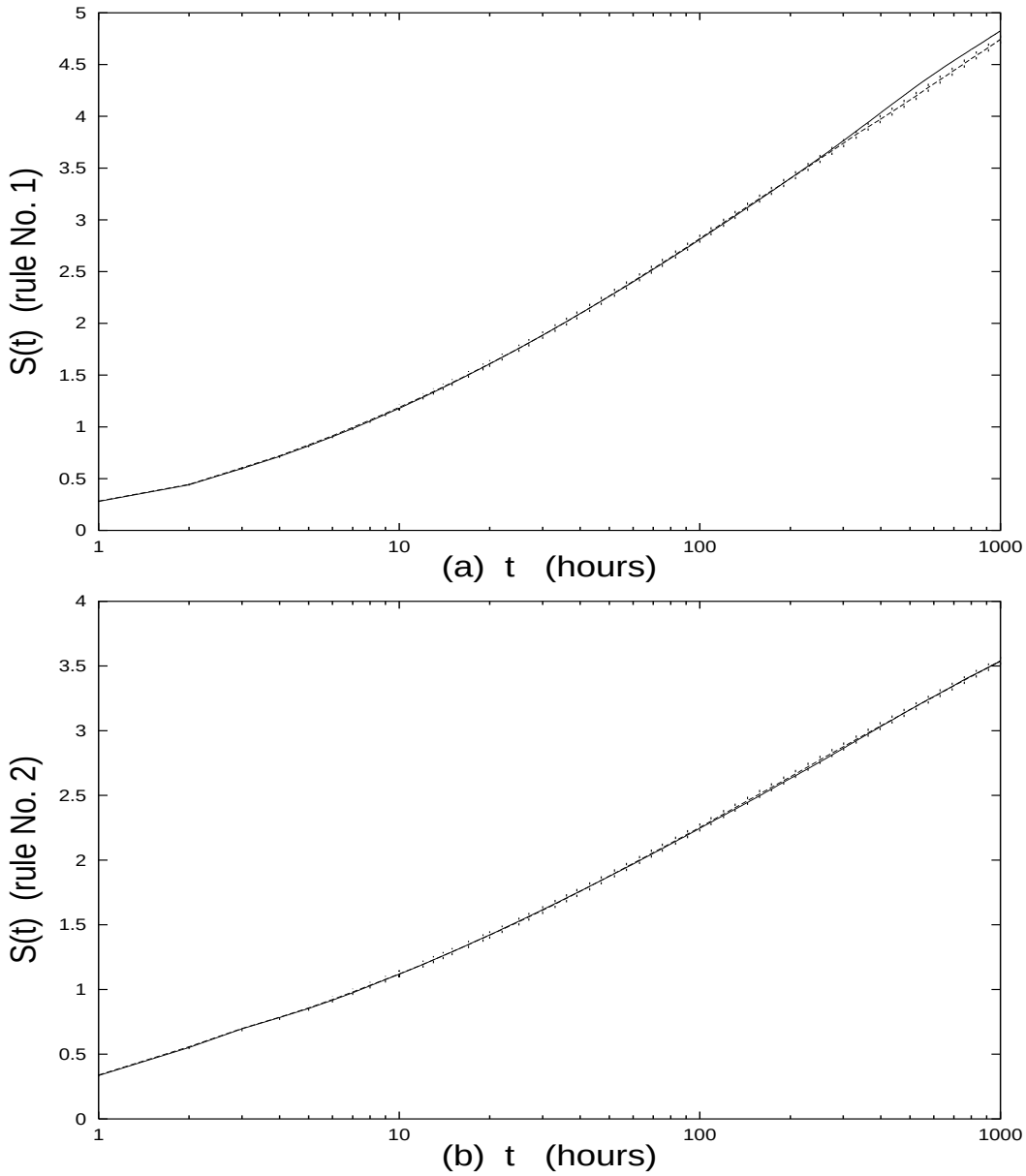

Figure 7.10: DE as a function of time. The solid lines denote the DEA curve generated by the shuffled real data, and the dashed lines, which almost coincide with the solid lines, denote the DEA curves resulting from the artificial sequence with $\mu = 2.138$ and $T = 8422$. (a) Rule No. 1. (b) Rule No. 2.

remaining N-1 numbers of the artificial sequence we select the closest one to it, this being, let us say, $A_{j(2)}$. We proceed with the same criterion until we exhaust all the numbers of the artificial sequence. It is evident that the adoption of this procedure assigns to the artificial data a time order reflecting the complex dynamics illustrated by Figs. 7.1 and 7.2.

At this stage, we evaluate the corresponding DEA curve and we compare it to the DEA curve generated by the non-shuffled real data. As earlier mentioned, the sorted artificial data are the same as those used to produce the excellent fitting of the DE curves derived from the shuffled original data. Thus, the fitting parameters are the same as those used for Figs. 11. We illustrate the new result in Figs. 7.11, which show that the fitting accuracy is as good as (and for rule No. 1 even slightly better than) the fitting of Figs. 7.10. This is a very remarkable result since Figs. 7.6 and 7.7 show that shuffling the data produces a significant effect. Thus, Figs. 7.10 and 7.11 prove that the memory of the data is totally under our control.

7.3.3 Third rule in action.

According to Lepreti, Carbone and Veltri [157] the waiting time distribution $\psi(\tau)$ is already Lévy. This would imply that the adoption of the third rule yields an infinitely fast transition from dynamics to thermodynamics. This is so because Lévy distribution is stable and the convolution between two distinct Lévy distributions is a Lévy distribution [158]. According to our analysis, $\psi(\tau)$ is a shifted inverse power law. It is plausible that the difference between the shifted power law distribution of Fig. 7.4 and the Lévy distribution of Ref. [157] is small. Consequently, the transition to thermodynamics is expected to be very fast. This expectation is confirmed by the numerical results illustrated in Fig. 7.12. The transition to the scaling regime is so fast that it is possible to detect a wide regime of linear dependence of the entropy on $log(l)$, which allows us to derive for μ the value $\mu = 2.138$, in total agreement with the conclusion of the earlier analysis done by means of rules No. 1 and No. 2. We see that in this case the memory of the non-shuffled data yields a δ slightly larger than the scaling parameter of the shuffled data. The adoption of rule No. 3 implies

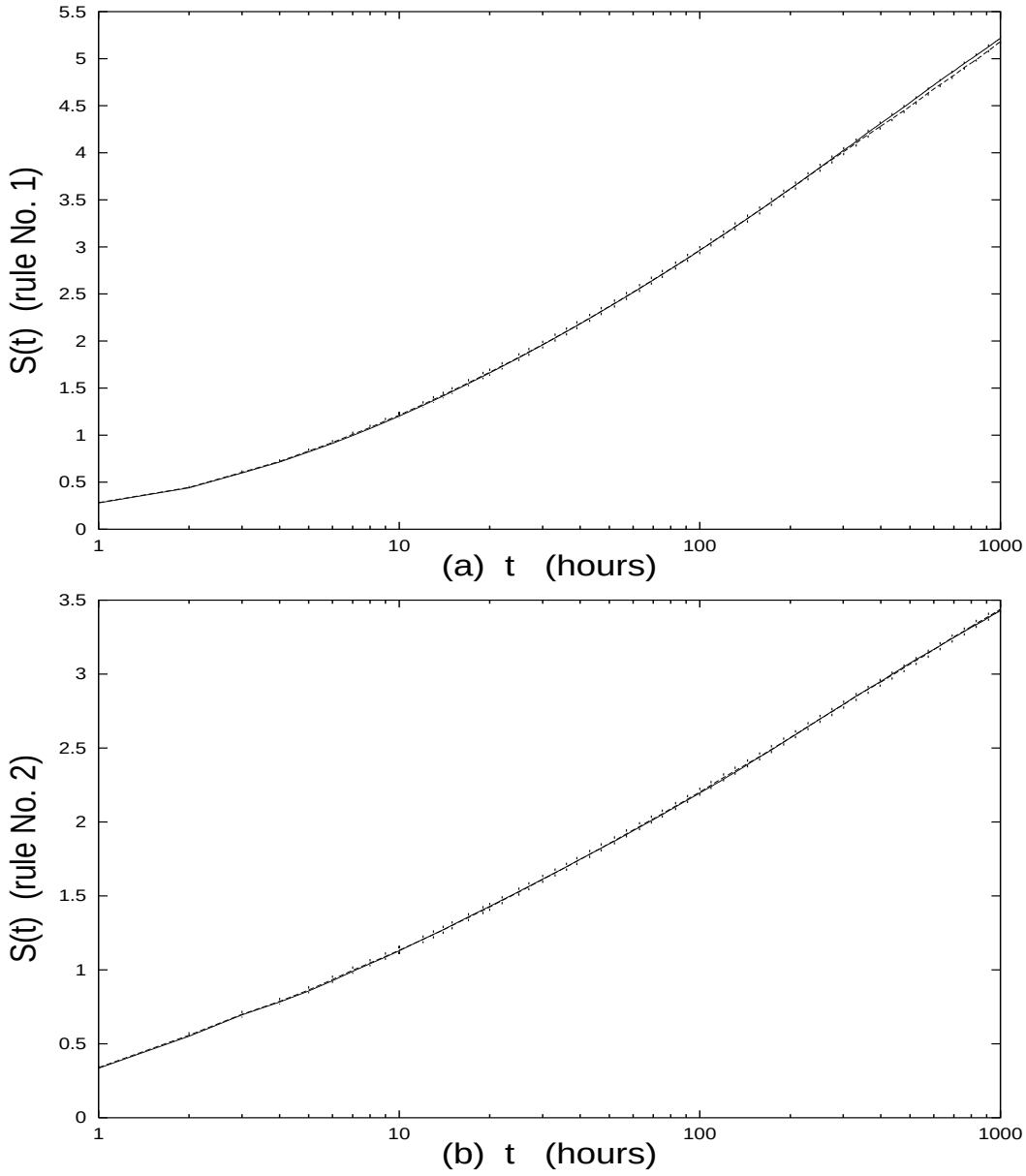

Figure 7.11: DE as a function of time. The solid lines denote the DEA curve generated by the unshuffled real data, and the dashed lines, which almost coincide with the solid lines, denote the DEA curves resulting from the artificial sequence with $\mu = 2.138$ and $T = 8422$ with a modulation mimicking the influence of the 11-years solar cycle. (a) Rule No. 1. (b) Rule No. 2.

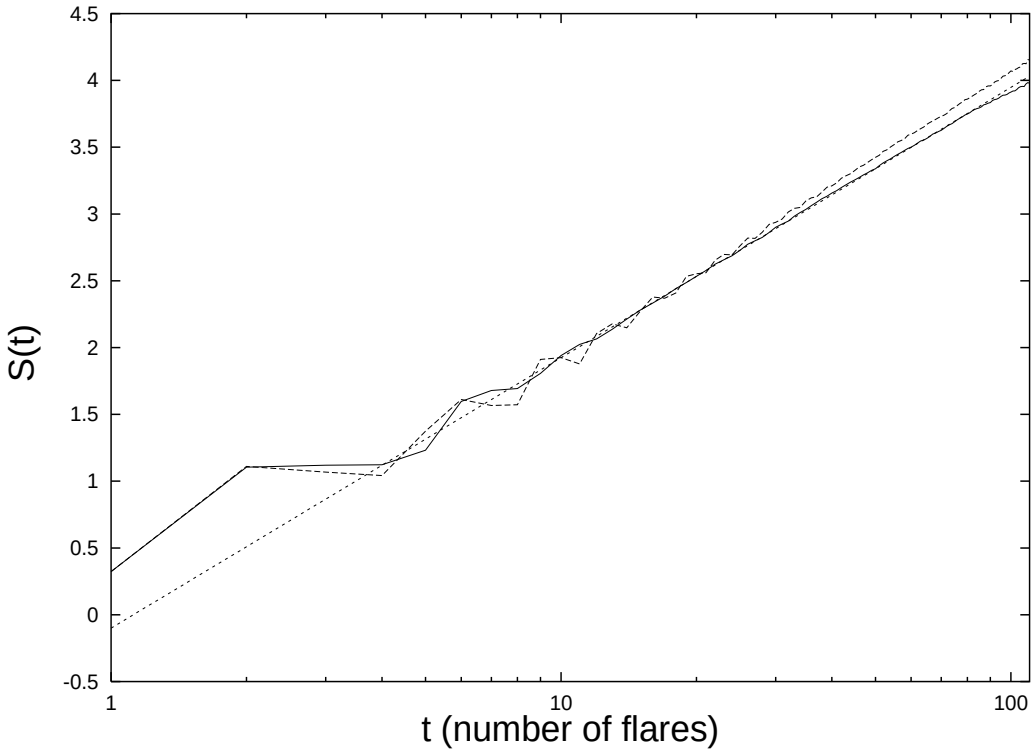

Figure 7.12: DE as a function of time, according to rule No. 3. The solid lines denote the DEA curve generated by the shuffled real data. The dotted straight line illustrates the slope of entropy increase, $\delta = 0.879$, which corresponds to $\mu = 2.138$. The dashed line denotes the DEA curve resulting from the unshuffled real data. Note the superdiffusion of the unshuffled real data DE due to the memory in the original signal.

a statistical accuracy smaller than that of the other two rules, due to fact there is no limitation to the jumps intensities, thereby decreasing the number of particles located in the same cell. This has the effect of making the evaluation of p_i and consequently that of the entropy less accurate. However, this disadvantage is widely compensated by the emergence of a much more extended scaling region that yields as a result a value of μ fully confirming that of the other two rules.

7.4 Concluding remarks.

We see that the uncertainty on the value of μ for solar flares has been significantly reduced. The current literature, if we give the same credit to all the authors, yields values of μ ranging from 3 to 1.7. We provide the compelling conclusion that $\mu = 2.138 \pm 0.01$. However, this is not the main result of his paper. We think that this paper shows that the DEA is a remarkably accurate technique of analysis that goes much beyond the direct evaluation of the waiting time distribution $\psi(\tau)$. This is so because complex processes are characterized by two different kinds of memory. The memory of first kind is the main object of the research work done in the field of the Science of Complexity. To make clear the nature of this kind of memory, let us recall [159] that a Markov master equation, namely a stochastic process without memory, is characterized by a waiting time distribution $\psi(\tau)$ with an exponential form, thereby implying memory for a marked deviation from the exponential condition. This is why the search for an inverse power law distribution with a finite value of μ (the exponential distribution means $\mu = \infty$) can be interpreted as a search for memory. This is the memory of the first kind, to which the prescriptions of Ref. [160] are referred to. For real data, in addition to this form of memory, another type of memory might be present, denoted by us as memory of the second type, under the form of correlation among the values τ_i. In this paper we have seen that this second form of memory is given, in this case, by the 11-year solar periodicity. It is possible that this form of additional memory is present in many other complex processes for different reasons. It is also evident that it is difficult, or perhaps impossible to reveal this form of additional memory by means of the direct evaluation of $\psi(\tau)$. This paper proves that joint use of the direct evaluation of $\psi(\tau)$ (or of $\Psi(\tau)$) and of the DE method is a very useful supplement to the ordinary technique, and that it can be profitably used to shed light on the dynamics behind the time series generated by complex processes.

This paper yields a convincing conclusion concerning the distinction between two possible forms of non-stationary behavior. As pointed out in Section 5.2, the claim that the waiting time distribution $\psi(\tau)$ has the form of Eq. (5.9) is equivalent to assuming that the dynamics of the flaring process is driven by the model of Eq. (5.13)

with the assumption that the trajectories are injected back randomly. This is a stationary model that in the case where $z > 2$, ($\mu > 2$), would be incompatible with the existence of an invariant distribution [161] and consequently with "thermodynamic equilibrium". The inaccuracy of the analyses done by the earlier work in this field would prevent us from distinguishing this form of non-stationary behavior from a genuinely form of non-stationary behavior. By genuinely non-stationary behavior, we mean the existence of rules changing with time. This form of genuinely non-stationary behavior might be modelled, for instance, by assuming that the parameter λ of Eq. (5.13) is time dependent. If we make the assumption that the time dependence of λ has a period of 11 years, and we make our analysis over a period of time that is not much larger than this time period, as we have done, then the process must be perceived as being genuinely non stationary. Our analysis is so accurate as to rule out the former form of non-stationary behavior and to detect significant effects stemming from the latter, or, equivalently, from the existence of the memory of the second type.

In this paper we do not take side with either the proponents of self-organized criticality [162] or with those of turbulence [153, 154]. The dynamical model of Section 5.2 is inspired by the models of turbulence, but we mainly use it to generate artificial sequences mimicking the real ones with no claim that it is an exhaustive picture of the dynamics behind solar flares. The fitting of Fig. 7.5 seems as good as the fitting of Fig. 1 of Ref. [157]. However, our analysis does not rest only on the waiting time distribution. In a very recent paper Wheatland [163] criticized the work of Ref. [157] as being based on the assumption that rate of solar flares is constant. This is not so, as shown by Fig. 7.2. On the other hand, modelling the time dependence of this rate is not easy, since it does not correspond only to a 11-year periodic motion but to a much more complex condition, as illustrated in Fig. 7.3. In fact, this figure shows that there are many other components in action. This is the reason why we decided to mimic the time dependence of the solar flare rate sorting the artificial sequence in the way described in Section 7.3.2. We found that this yields a fitting with the real data as good as the fitting between the DEA curve produced by the artificial sequence, with no sorting induced memory, and the DEA curve produced by the shuffled real data. This is, in our opinion, a strong indication

that the value of $\mu = 2.138$ is a genuine property of real data. On the other hand, the dynamical model of Section 5.2 can also be adapted to reproducing the modulated Poisson process advocated by Wheatland [163]. This is left as a subject of future investigation. Even in this case, the role of the DEA is expected to be crucial, and the result is expected to strongly depend on wether the modulation process involves randomness or only quasi-periodicity.

In our notation, the power index found by the authors of Ref. [157] is $\mu = 2.38$, a value that turns out to be compatible with the uncertainty interval associated to the determination of μ by means of the direct evaluation of $\psi(\tau)$. Our analysis establishes a connection with Lévy statistics, in accordance again with the conclusions of Ref. [157]. However, we adopt a perspective that is different from that of the authors of Ref. [157]. Our diffusion process reaches the Lévy regime after the process of transition from dynamics to thermodynamics that has been discussed in detail in the earlier sections. This process is very fast if the rule No. 3 is adopted, but it is not infinitely fast as in the perspective of the authors of Ref. [157] who assume the waiting time distribution $\psi(\tau)$ to obey already the Lévy statistics. It is worth pointing out that the perspective adopted in this paper makes it possible to take into account the time dependence of the solar flare rate. We do not rule out the possibility that $\psi(\tau)$ is a stretched exponential [164]. In fact, a stretched exponential would not conflict with the attainment of Lévy statistics in the long-time limit of the diffusion process. Although a truncation of $\psi(\tau)$ at large values of τ generates a finite second moment, and consequently Gaussian statistics in the long-time limit, the transition to the conventional thermodynamic regime is ultra slow [165]. It is known [166] that a much earlier transition to Lévy statistics occurs and that the Lévy regime lasts for a very extended period of time. The transition to the Gaussian regime probably takes place at times much larger than the saturation time, and might be made visible only in the ideal case of infinitely large sequences.

CHAPTER 8

THE THERMODYNAMICS OF SOCIAL PROCESSES: THE TEEN BIRTH PHENOMENON

In this Chapter we apply the Diffusion Entropy Analysis to the study of the teen birth phenomenon and we find that the unmarried teen births are a manifestation of a social process with a memory more intense than that of the married teens [7]. To arrive to this conclusion, we argue that a process of social interest is a balance of order and randomness, thereby producing a departure from a stationary diffusion process. The strength of this effect vanishes if the order to randomness intensity ratio vanishes, and this property allows us to reveal, although in an indirect way, the existence of a finite order to randomness intensity ratio. We aim at detecting this effect by proving that Diffusion Entropy Analysis makes it possible for us to build up a memory detector, which signals the presence of even very weak memory, provided that this is persistent over large time intervals. In Sec. 4.5 it is explained how to handle a non-stationary dynamical transient analysis, and in Sec. 5.7 it is proved that a non-stationary condition can be induced by weak and persistent memory. In this Chapter we use the results of Sec 4.5 and 5.7 and we prove that, after having removed the seasonal periodic properties from the two sets of data regarding the married and unmarried teen births, the two groups are different in regard to residual memory. This allows us to classify the two groups as dynamically different. Our analysis is about the teen birth phenomenon in Texas.

8.1 The teen birth phenomenon.

In this section we give a brief introduction to the teen birth phenomenon in Texas and place the results of the diffusion entropy method of analysis within the context of earlier research.

Texas is second only to California in the number of births to teens in the United States. Rates of births to teens of all ages and racial/ethnic groups have been dropping

since 1990, Ref.[167]. However, the size of the problem in Texas is increasing.

In 1996, in Texas there were 80,490 pregnancies and 52,273 births to girls 15-19 years old, Ref.[168]. The U.S. rate of pregnancy among young women 15 to 19 years old was 97 per 1000 girls of that age, the rate in Texas was 113 per 1000 in 1992. The mean age of teens giving birth was 17.62 years in Texas. Approximately 66% of teen births in Texas were out of wedlock and 24% of births to teens were to girls who had given birth at least once previously.

Data used in West et al. (1999) [169] to study the nonlinear dynamics in teen birth data included daily counts of all births to teens in Texas from 1980 through 1998. Findings demonstrated the teen birth data obeyed a scaling law. The authors[169] concluded the scaling relation tied together what happened at the shortest time scales with what happened at the longest time scales, thus, resulting in long-term memory. When found, such long term correlation and complexity in time series suggests there is strong feedback across time scales in the process. These authors[169] suggest the phenomenon is dominated by a self-induced stability which may increase with population density, mobility and interaction among persons, between persons and social institutions, and among social institutions. To date these conjectures have not been tested.

The authors of Ref.[169] did not detrend or smooth the data prior to their analysis, nor did they investigate the effects of marital status on the scaling process. It was their intent to study the gross behavior of the time series of all births. The approach used here provides a deeper look into the endurance of memory and its possible source for this data.

Data for the study reported here were abstracted from birth certificates obtained from the Texas Department of Health. The original time series was constructed from the daily count of births from January 1, 1994 through December 31, 1998. Every recorded birth to a woman under the age of 20 was included. Data on the marital status of the mother allowed us to analyze married and unmarried births separately. Reliable and valid birth certificate information regarding marital status did not become available in Texas until January 1, 1994. See Table 8.1 for further description of the data.

Data Set	All Teens	Married Teens	Unmarried Teens
Mean # Daily Births	149.14	50.39	98.52
Range	143	68	97
Minimum	81	22	55
Maximum	224	90	152
Standard Dev	23.52	10.34	16.68
Variance	553.03	106.89	278.15
Total Births	272,328*	92,006	179,893

Table 8.1: Marital Status Data Set Used N=1994-1998=1826 days. *Marital status was missing on 429 teen birth certificates.

Data Set	All Teens	Married Teens	Unmarried Teens
Lags of 7 Days	.665	.466	.602
Lags of 364 Days	.536	.370	.464

Table 8.2: Autocorrelation in Married and Unmarried Teens.

The reason marital status is relevant to the analysis that follows is based on the observation that amount of linear memory differs between time series of births to married and to unmarried teens. See Table 8.2 for autocorrelations for lags of one weak and approximately one year in total teen births and married and unmarried teen births separately. The fewest births to both married and unmarried teens occurred in the second quarter of each year (April, May and June). The third quarter of every year (July, August and September) had significantly more births to teens than any other quarter.

Large third quarter peaks in births and autocorrelation = 0.536 at lags of 364 days were found for all teen births. Atmospheric factors such as light and temperature have

been suggested as reasons for annual cycles in human births, Ref.[170]. However, it is not clear why such factors would have a stronger effect on unmarried than on married teens as indicated by differences in autocorrelation at lags of 364 days.

Strong weekly periodicities (autocorrelation = 0.665 at lags of 7 days for all teens) were found. These may have been imposed on the process through provider preference. Scheduled inductions of labor and cesarean sections generally occur early in the week in order to assure patients are out of the hospital by the weekend when staffing is a greater problem. However, further investigation is needed to determine the veracity of the assumption that these preferences are related to the weekly periodicity observed in the data. In addition, investigation is needed to determine the reason for the disparity in autocorrelation for lags of 7 days between married and unmarried births.

8.2 Data and Preliminary detrending.

The goal of this chapter is to discover weak memory remaining after the removal of linear correlations from the two original sets of data regarding the married and unmarried teen births. We are aware that any form of detrending carries with it the risk of loss of information and obliteration of long-range memory. However, in this case we believe detrending provided us an opportunity to test our new method for sensitivity to long-lasting memory of extremely weak intensity. We believe earlier findings [169] may be the result of long-lasting memory of large intensity, perhaps the result of annual periodicities not removed by detrending. We make the conjecture that, although weak, there is a residual memory left after detrending annual periodicity, steady drift and removing births occurring on weekends and holidays. In Sec. 4 we shall try to explain the implications of the results. With these warnings in mind, we proceed with our detrending prescription as follows.

First of all, we notice that the number of births corresponding to Saturdays and Sundays are much smaller than those of the week days. In addition to Saturdays and Sundays, there are other days with very small number of births that are identified with holidays. All data identified as holidays are erased and replaced by empty sites. The data of Figs. 8.1 report the daily number of births with empty sites that are

not visible in the scale of that figure. The data refer to the births to married (a) and unmarried (b) teens, in the state of Texas from 1994 to 1998. In both cases the data show ostensible signs of seasonal periodicities. We remind the reader that the purpose of this paper is to prove that the Diffusion Entropy Analysis is an efficient way of detecting residual memory after detrending trivial periodic properties, thus revealing the balance between order and randomness. The seasonal recurrences shown in Figs. 8.1 seem to be a significant example of an obvious periodic property.

Thus, we detrend it proceeding as follows. We assume that the periodic property is described by

$$\Xi(t) = A + Bt + Ccos(\omega t) + Dsin(\omega t). \tag{8.1}$$

This is a deterministic process whose form is determined by seasonal periodicity, the harmonic part, and by the fact that the number of births is proportional to the increase (or decrease) of the number of births to teens within the two categories with time, the term Bt. In the case of the unmarried teens the fit gives A = 97.5, B = 0.00893, C = 1.29, D = -6.30; we set $\omega = 2\pi/365.25$. In the case of the married the fit gives A = 57.8, B = -0.00353, C = - 0.277, D = -4.14; we set $\omega = 2\pi/365.25$. The resulting curves are illustrated by the solid lines of Fig. 8.1a and Fig. 8.1b. In Fig. 8.2a and 8.2b we show the time series after application of this detrending procedure. It is evident that the data of Figs. 8.2 look more erratic than the original data of Figs. 8.1. A mere sight inspection of Fig. 8.2 seems to indicate that the data concerning unmarried teens are more correlated than those of the married. The diffusion entropy method of this paper aims at a quantitative assessment of this property. To proceed to this quantitative assessment that will be illustrated in Sec. IV C, we have to prepare the data in a suitable way. We call these new data $\zeta_a(n)$, if they refer to unmarried teens, and ζ_b, if they refer to unmarried teens. The symbol n denotes, as usual, the discrete "time" n = 1,2,3.... It is important to point out that this time does not correspond exactly to the days of the original data, because the empty sites are not counted, and if two given days are separated by holidays, the latter day is labelled as being immediately subsequent to the former.

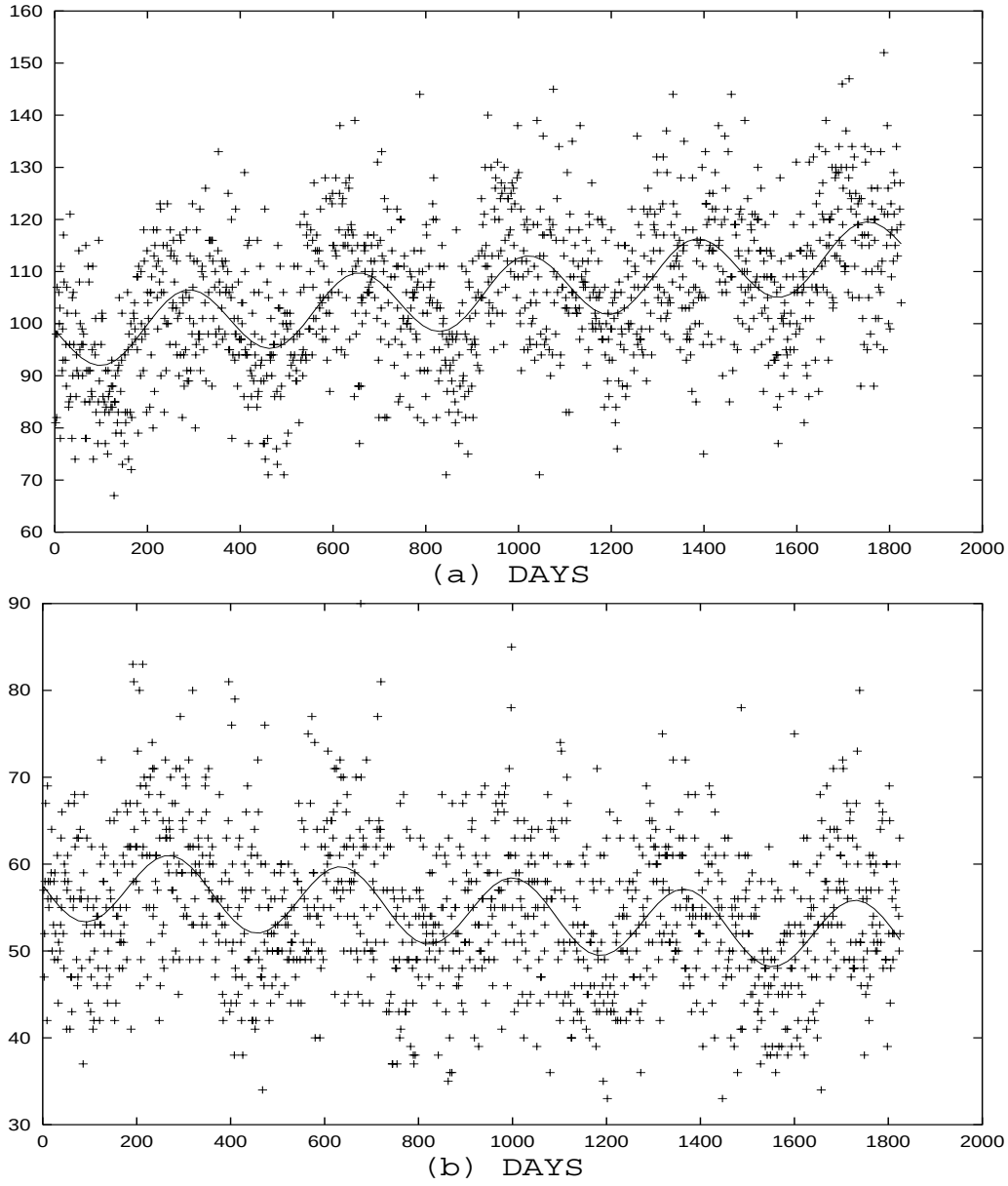

Figure 8.1: (a) The day by day births from unmarried teens in Texas from 1994 to 1998. (b) The day by day births from married teens in Texas from 1994 to 1998. The data has been obtained cancelling all the holidays (see the text). The solid lines illustrate the choice made to detrend seasonal periodicities. This means the analytical expression of Eq. (30) with (a) A = 97.5, B = 0.00893, C = 1.29 , D = -6.30 and $\omega = 2\pi/365.25$; (b) A = 57.8, B = -0.00353, C = -0.227 , D = -4.14 and $\omega = 2\pi/365.25$.

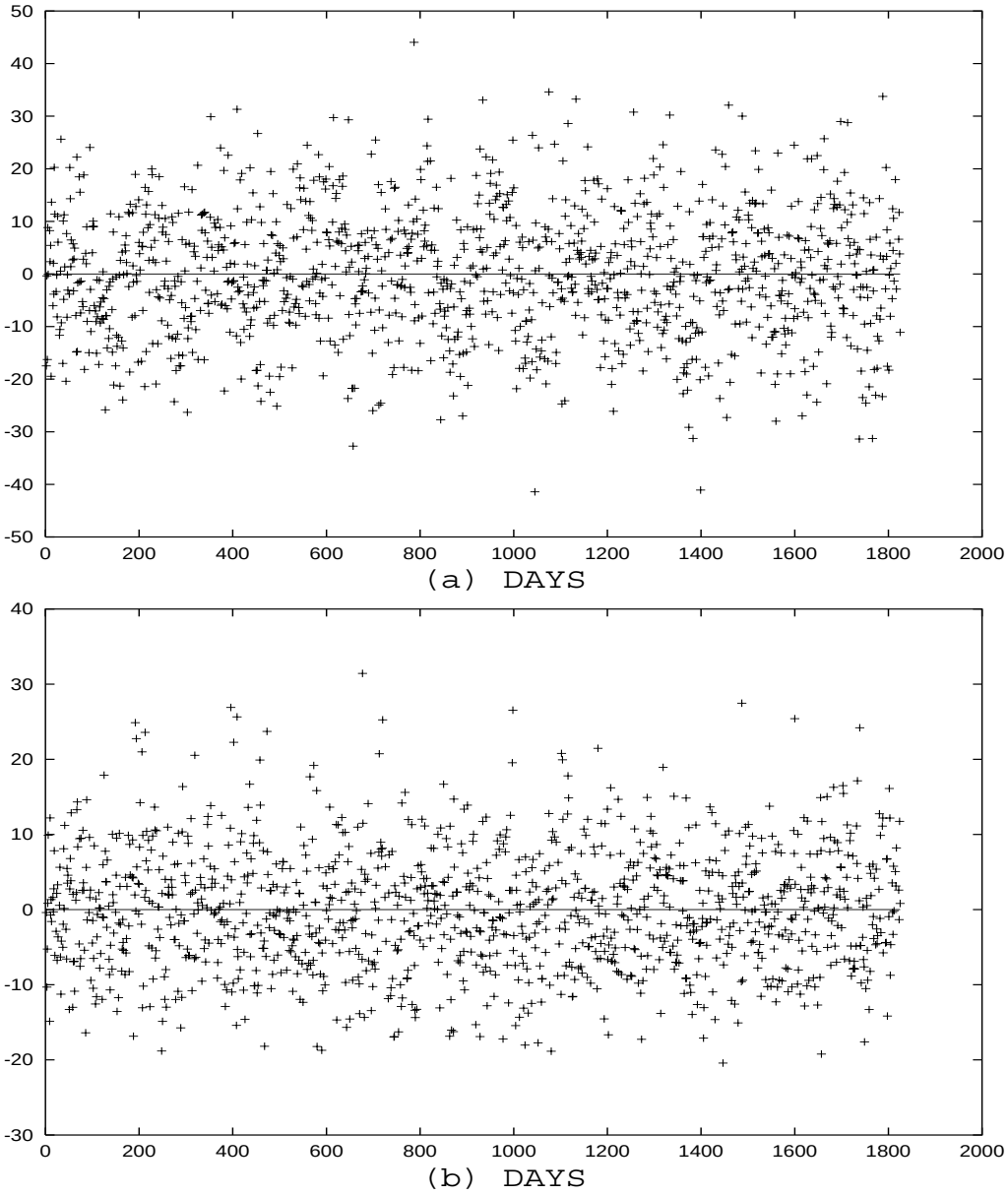

Figure 8.2: (a) and (b) show the data after the detrending of seasonal periodicity, and of all Saturdays, Sundays and holidays respectively for births to unmarried and married teens in Texas from 1994 to 1998 . These are the data that we analyze with the diffusion entropy method.

8.3 Diffusion entropy used to detecting memory.

We apply the Diffusion Entropy Analysis to the new data $\zeta_a(n)$ and $\zeta_b(n)$ of Fig. 8.2a and Fig. 8.2b using the recipe illustrated in Sec. 8.2. To make easier for us to adopt this recipe, we will transform these two sequences into two dichotomous sequences. To do that we adopt the following prescription

$$\xi_{a/b}(n) = \begin{cases} +1 & if \quad \zeta_{a/b}(n) > 0 \\ -1 & if \quad \zeta_{a/b}(n) < 0 \quad . \end{cases} \tag{8.2}$$

As done in Sec. 5.7, we create a large number of trajectories described mathematically by

$$x_{a/b}(r,t) \equiv \sum_{i=0}^{t} \xi_{a/b}(i+r), \tag{8.3}$$

with $r = 1, 2, 3, \ldots$. Then, following again Sec. 5.7, we calculate the probability $p_{a/b}(x_{a/b}, t)$ of finding the random walker in the position $x_{a/b}$ after t jumps. Finally, we calculate the diffusion entropy of the two sets of data, $H_{a/b}(t)$, with the following formula:

$$H_{a/b}(t) = -\sum_{x_{a/b}} p_{a/b}(x_{a/b}, t) \log(p_{a/b}(x_{a/b}, t)) \ . \tag{8.4}$$

We show the results of this statistical analysis in Fig. 8.3. This figure shows that the diffusion entropy of both sets of data exhibits deviation from the stationary condition, indicated by the dotted line. The departure from ordinary statistical mechanics of the unmarried teens is much larger than that of the married teens. The curves of Fig. 8.3 are remarkably similar to those of Fig. 5.21 and this fact, by itself, suggests that births to unmarried teens contain more memory than is the case for married teens.

Following the prescriptions of Sec. 5.7, we also use the Tsallis entropy, which, in

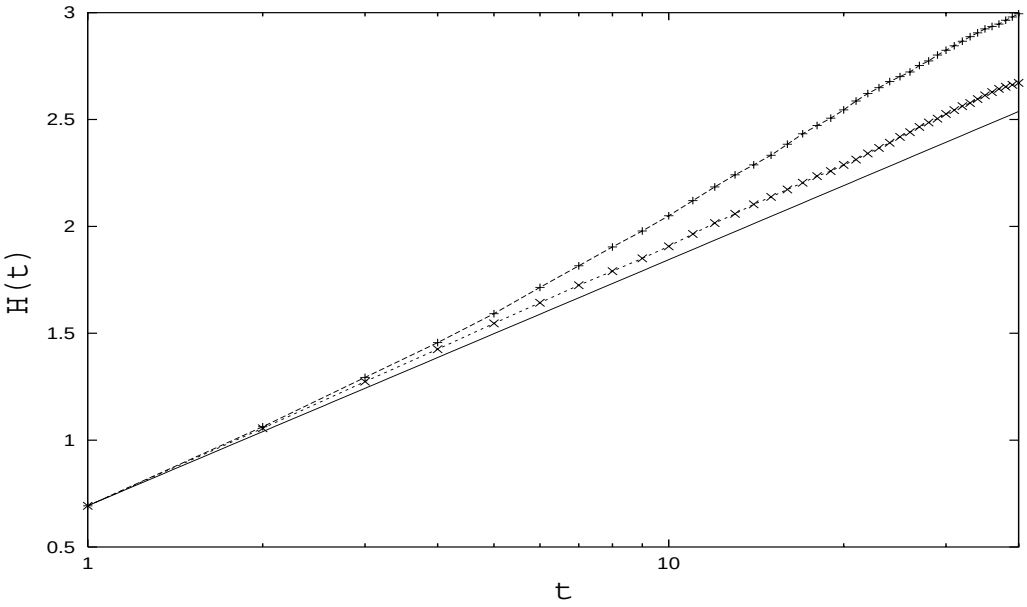

Figure 8.3: The teen births diffusion entropy as a function of time. The solid line corresponds to the prediction of Eq.(5.40) and serves the main purpose of indicating to the reader how the entropy time evolution of a stationary process of diffusion would look in the scale of this figure. The case of unmarried teens is denoted by the symbols $+$ and the case of married teens is denoted by symbols $\times$. The deviation from the straight line of the stationary diffusion process of the unmarried teens is stronger than that of the married teens.

this case, reads:

$$H_{q,a/b}(t) = \left(1 - \sum_{x_{a/b}} p_{a/b}(x_{a/b}, t)^q\right) / (q-1)\,. \tag{8.5}$$

We illustrate the results of this analysis in Fig. 8.4. This figure shows that both the diffusion entropy of the unmarried teen and the diffusion entropy of the married teens have a linear dependence on the logarithmic time τ if we use for the former case $q = 1.204$ and for the latter $q = 1.050$.

Finally, as done in Sec. 5.7, in Table 8.3 we compare the two distinct methods adopted for the detection of the entropic index q. We see that the results of the latter method are very close to those of the former, thereby ensuring the validity of the

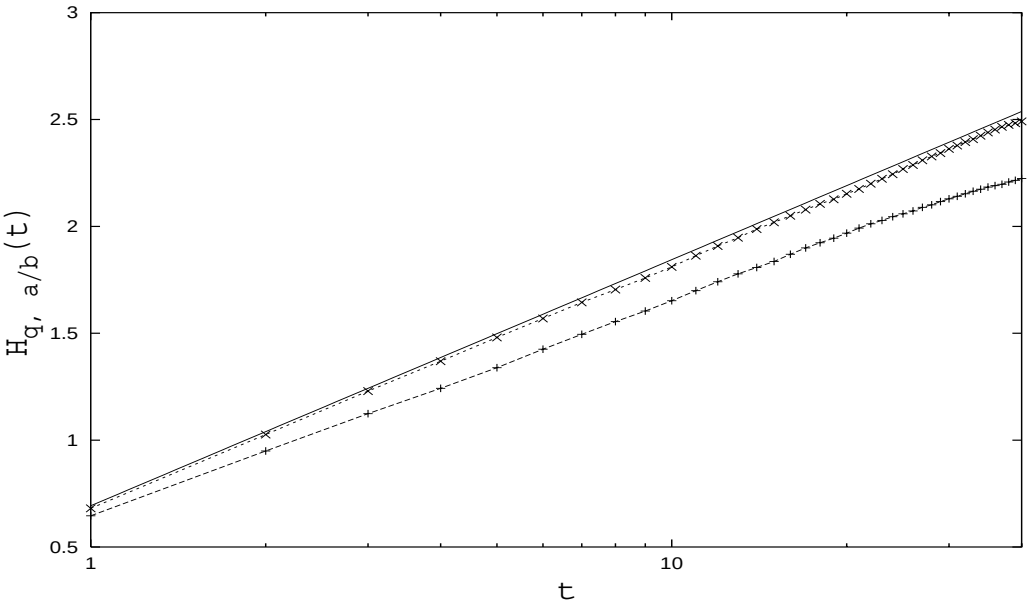

Figure 8.4: The non-extensive diffusion entropy of the teen birth phenomenon as a function of time. The solid line refers to the prescription of Eq.(5.40). The case of unmarried and married teens are denoted by the symbols + and $\times$, respectively. The entropic indices resulting in an entropy increase linear with respect to the logarithmic time τ are q= 1.204 and q= 1.050, for unmarried and married teens, respectively.

parabolic fitting of the former method. Actually, the agreement in the case of the married teens is better than in the case of the unmarried teens, which are characterized by memory strength larger than that of married teens. This is in agreement with the remarks of Sec. 5.7 suggesting that the strong memory case can produce a conflict between the two methods. However, we think that the hidden memory, detected by the new method of this paper, is weak enough as to ensure the validity of our main conclusion that the married and unmarried cases show a remarkable similarity with the memory strengths $\Delta = 0.04$ and $\Delta = 0.10$ of Fig. 5.22, respectively. More precisely, according to Table 8.3, the married and unmarried case correspond to $\eta = 0.006$ and $\eta = 0.035$, respectively. We conclude that the method of diffusion entropy affords a reliable proof that births to unmarried teens have stronger 'weak' memory than those to married teens.

One might wonder if the procedure of making the signal dichotomous might have

	η	δ_0	A	$q = 1 + \epsilon$	q
unmarried teens	0.035 $\pm$ 0.002	0.513 $\pm$ 0.007	0.687 $\pm$ 0.007	1.23 $\pm$ 0.02	1.204
married teens	0.006 $\pm$ 0.001	0.514 $\pm$ 0.004	0.697 $\pm$ 0.004	1.046 $\pm$ 0.009	1.050

Table 8.3: Entropic index q as resulting from two distinct fitting procedures.

produced statistical effects distinct from those resulting from the actual signal. Numerical calculations, not reported here, prove that it is not so, and that the actual data produce similar results. For the quantitative purpose of measuring the memory strength the adoption of a dichotomous signal yields the benefit of resting only on integer numbers for the production of the histograms necessary for the entropy calculations. Furthermore, the adoption of dichotomous signals provides a deeper connection with the model described in Sec. 5.7, and it is in fact the reason for the surprising similarities between Figs. 5.21 Fig. 8.3 and between Figs. 5.22 and 8.4.

8.4 CONCLUSION

The first interesting result we can conclude is that there is an intimate relation between memory and the breakdown of the stationary condition expressed by the scaling property of Eq. (4.1). It has to be pointed out that the departure of the entropic index q from the ordinary value $q = 1$ does not signal a departure from ordinary statistics, as claimed in Ref.[171]. Rather, it signals the breakdown of the scaling property of Eq. (4.1), which, in turn, is due to the fact that the signal under study consists of fast fluctuations with a time dependent bias that are a much slower function of time. If the bias is weak, the deviation from the ordinary entropic index $q = 1$ can be related to the memory strength η, which can be independently derived from the adoption of the ordinary Shannon entropic indicator. In other words, the parameter η is more significant than the parameter $\epsilon \equiv q - 1$, since it correctly suggests that the effect revealed by the Diffusion Entropy Analysis has to do with the time dependence of the scaling parameter δ. We recover from a different perspective the conclusions of

earlier work [130]. The Lévy processes are an interesting example of non-Gaussian statistics. Yet, their diffusion entropy would yield a linear increase with respect to the logarithmic time $\tau \equiv ln(t)$, described by Eq. (4.31), even if this were the case δ_0 would be a scaling parameter larger than 0.5. This is so because the dynamical derivation of the Lévy processes, after a process of memory erasure [130], yields the scaling property of Eq. (4.1).

The second important result is that Diffusion Entropy Analysis is very sensitive to weak but persistent memory. The similarities between Fig. 5.21 and Fig. 8.3 and between Fig. 5.22 and Fig. 8.4 are impressive. These figures mean that it is plausible to conjecture that the detrending process used to analyze the birth data does not eliminate all forms of memory. The form of memory adopted in Sec. 5.7 to create the artificial sequence is a deterministic process with the time period of 2000 days. This cannot be considered as a proof that this is the order of magnitude of the hidden periodicity. We can only conjecture that the periodicity responsible for the deviation of entropy from the linear increase with respect to the logarithmic time τ, is larger than the maximum observation time, which is of the order of 30 days. Beyond 30 days the number of trajectories available is too scarce to ensure statistical stability.

Here we make two conjectures regarding sources of weak memory detected in the detrended data. First, it is likely that weekly and annual periodicities removed by detrending were not the only periodicities present in the data. For example, there are weak cycles with periods of approximately one half year remaining in the data for unmarried teens after detrending. Similar cycles do not appear in the data for married teens. There are sociological factors such as school schedules and holiday breaks that may account for the difference in weak sub-annual cycles between married and unmarried teens. While these effects have not yet been investigated, further study could reveal their association to the memory strengths of the data. However, this conjecture seems simplistic in light of our second conjecture regarding fractal scaling in the data. Traditional time series analysis methods used in the social sciences are based on the assumption that complicated time series data represent the superposition of numerous frequencies of varying periods. A frequency for which cause is understood or considered trivial is modelled and/or removed. Frequencies are, thus, explained

and removed iteratively until all useful information in the data is accounted for and the data remaining represent white noise. However, the time series data used to test the methods described here are known to have fractal scaling properties [169]. These properties are the result of feedback across all time scales within the time series (days, weeks, months, years). Fractal scaling processes cannot be fully characterized by the superposition of few independent, additive frequencies. Instead, the data are nonlinear in that all frequencies are folded together in a complex pattern. In such a case, detrending, as we prescribed, would be insufficient to obliterate the effects of weekly and annual periodicities folded into periodicities on other time scales. Thus, the weak memory remaining may be the result of fractal scaling of the data that resists simple detrending techniques common in the social sciences. The difference in scaling properties of married and unmarried teen births has not been thoroughly investigated but suggests unmarried teens are affected by a stronger feedback mechanism than are married teens.

All these conjectures suggest that the logarithmic oscillations detected in Ref. [169] are of such large intensity because of the fact that the yearly periodicities are not detrended. Their detrending, as shown in this paper, makes the resulting signal much closer to the ordinary random walk. The analysis of this paper reveals, however, that this is not an ordinary random walk, and that a fractal cascade of frequencies of smaller and smaller intensity might exist. This is a challenge for future application of the entropic method of analysis developed in this paper. We plan to shed light on this and other intriguing issues raised by our method by means of the joint analysis of real data and of artificial sequences like that of Sec. 5.7, with fluctuating periodicities driven by an inverse power law prescription.

Part III

CONCLUSION

CHAPTER 9

CONCLUSION

In this dissertation we introduced a new method of statistical analysis of time series, the Diffusion Entropy Analysis (DEA). We showed that this methods out performs all the methods currently used in the field of the Science of Complexity, due to one essential property: it is the only method that establishes the correct scaling of a time series, if this exists. This was proved theoretically, using artificial sequences with assigned properties, designed to prove our assertion. We compared the results from DEA with those derived from different methods currently adopted to detect scaling. These methods are popular and are frequently used: Variance Scaling Analysis, Hurst R/S Analysis, Detrended Fluctuation Analysis, Relative Dispersion Analysis, Spectral Analysis, Spectral Wavelet Analysis. These traditional methods are based on the assumption that the variance scaling coincides with the true scaling, and, so, on the assumption of Gaussian statistics. The scaling detected by the variance methods, denoted by the symbol H, may not exist or may not coincide with the correct scaling, δ. If the time series is a realization of what Mandelbrot called Fractional Brownian Motion, namely a Gaussian diffusion, with anomalous as well as normal scaling, we have $H = \delta$. Consequently, the scaling can be correctly detected by using the variance methods. If, on the contrary, the time series generates, for example, a diffusion of the Lévy type [4, 5], $H \neq \delta$ and the variance methods fail to detect the true scaling.

The scaling of a time series is determined as follows. By summing the terms of a time series, thought of as fluctuations, we get a trajectory and the trajectory can be used to generate the diffusion of the variable x, collecting all these fluctuations. There is scaling if the probability density function (pdf) at time t, $p(x,t)$, of this diffusion process fits the property:

$$p(x,t) = \frac{1}{t^{\delta}} \, F(\frac{x}{t^{\delta}}). \tag{9.1}$$

The coefficient δ is the scaling exponent of the time series under study. The DEA

detects the correct scaling exponent, δ, of a time series because it is based upon the evaluation of the Shannon entropy of the pdf of the diffusion process that reads

$$S(t) = -\int_{-\infty}^{+\infty} dx\ p(x,t)\ \ln\left[p(x,t)\right]. \tag{9.2}$$

If the scaling condition (9.3) applies, the time evolution of the entropy, S(t), is linear with respect to the logarithmic time, $\tau \equiv ln(t/t_0)$, which makes Eq. (9.3) read

$$S(\tau) = A + \delta\ \tau. \tag{9.3}$$

Eq. (9.3) states that the scaling exponent δ is determined by the asymptotic slope of the entropy $S(\tau)$.

The DEA out performs other methods of analysis due to the fact that the information extracted from the pdf, expressed under the form of entropy, is larger than the information extracted from the pdf variance. Note that the entropy indicator needs not to be the Shannon indicator. To detect scaling the Renyi entropy is as effective as the Shannon entropy; From this dissertation is a research project emerging, with a form of DEA based on the Renyi entropy, as a way to shed light on multi-fractal statistics. To study the long-lasting regime of transition in teen births, we have seen that the non-additive form of entropy advocated by Tsallis can also afford some benefits. The methods of analysis resting on variance, afford much more limited information, and, as we have seen, can be trusted only in the Gaussian case. On the other hand, when the true scaling of the DEA departs from that detected by the variance methods, this can be taken as a clear indication that the statistics are not Gaussian. Only the joint use of the two scaling analysis methods, the variance methods and the DEA, can assess the real nature, Gauss or Lévy, of a time series.

Furthermore, the benefits stemming from the entropic method of analysis of a diffusion process (the DEA) are not limited to the detection of the true asymptotic scaling δ. We can explore the still unknown regime of transition from dynamics to thermodynamics, and we can also address the ambitious issue of studying the time series produced by non-stationary processes. Finally, we can bypass the limitation

posed by a scarce number of data that would produce saturation before reaching the scaling regime. As we have seen, the DEA is useful even when the scaling regime is not reached and only the transition or the non-stationary regime may be studied.

This dissertation focuses on theory and on the discussion of artificial sequences, as a way to check the efficiency of the new method of analysis. However, the efficiency of the DEA was illustrated also by showing this analysis in action on real data, referring to three different kinds of processes. The DNA sequences are long enough as to show the DEA in action to detect the true scaling. The direct detection of scaling and the auxiliary study of an artificial sequence generated by the CMM prove that DNA sequences are characterized by Lévy statistics. The efficiency of the DEA to study the transition regime is made clear by its application to Hard x-ray solar flare waiting times. Of special relevance to settle the intriguing problem of the correct power index μ has be the adoption of different walking rules. The comparison between the real data in the natural order and the real data in shuffled order made it possible for us to prove the existence of complex dynamics, and of memory effect, which would be overlooked completely by an analysis only based on the waiting time distribution. This may turn out to be useful to shed light on the dynamics of a turbulent phenomenon. The births data from teenagers in Texas affords another example of memory properties, this time of social interest, emerging from the adoption of the DEA. We could assess that the births to unmarried teenagers reveal memory effects more intense than for births to married teenagers.

On the basis of the results of this dissertation, we reach the conclusion that the DEA is of fundamental importance to detect the real statistical and dynamical properties of the time series. In general, DEA does not substitute or make the standard methods of analysis of a time series obsolete. In fact, as we have seen in the case of DNA sequences, the genuine statistics of the process emerge from the joint use of DEA and ordinary variance scaling methods, these latter being based on variance. Moreover, its ductile and sensitive entropic nature makes the DEA ideal for the detection of even little changes occurring over time in the statistics of a time series. Innumerable applications in any field of science are expected to follow.

BIBLIOGRAPHY

[1] B.B. Mandelbrot, *The Fractal Geometry of Nature*, (Freeman, New York, 1983).

[2] M. Schroeder, *Fractals, Chaos, Power Law*, (Freeman,New York, 1991).

[3] H. E. Hurst, R. P. Black, Y. M. Simaika, *LongTerm Storage: An Experimental Study* (Constable, London).

[4] P. Lévy, Calcul des Probabilits, (Gauthier-Villars, Paris, 1925).

[5] P. Lévy, Théorie de l'addition des variables Aléatoires, (Gauthier-Villars, Paris, 1954).

[6] G. Aquino, P. Grigolini, N. Scafetta, Chaos, Soliton and Fractals 12 (2001) 2023.

[7] N. Scafetta, P. Hamilton, P. Grigolini, Fractals 9, 193 (2001).

[8] N.Scafetta, V. Latora, P. Grigolini, cond-mat 0105041.

[9] P. Grigolini, D. Leddon, N. Scafetta, cond-mat 0108229.

[10] A. Montagnini, P. Allegrini, S. Chillemi, A. Di Garbo, P. Grigolini, Physics Letters A 244 (1998) 237.

[11] R. Metzler, J. Klafter, Physics Reports 339 (2000) 1-77.

[12] R. Brown, Phil. Mag. 4 (1828) 161; Ann. Phys. Chem. 14 (1828) 294.

[13] A. Fick, Ann. Phys. (Leipzig) 170 (1955) 50.

[14] A. Einstein, Ann. Phys. (Leipzig) 17 (1905) 549.

[15] C. W. Gardiner, *Handbook of Stochastic Methods*, Springer-Verlag Berlin Heidelberg New York (1997).

[16] B.D. Hughes, Random Walks and Random Environments, Vol. 1: Random Walks, (Oxford University Press, Oxford, 1995).

[17] J.-P. Bouchaud, A. Georges, Phys. Rep. 195 (1990) 12.

[18] B.J. West, W. Deering, Phys. Rep. 246 (1994) 1.

[19] B.V. Gnedenko, A.N. Kolmogorov, Limit Distributions for Sums of Random Variables, (Addison-Wesley, Reading, MA, 1954).

[20] L. F. Richardson, Proc. Roy. Soc. 110 (1926) 709.

[21] H. Scher, E.W. Montroll, Phys. Rev. B 12 (1975) 2455.

[22] G. Pfister, H. Scher, Phys. Rev. B 15 (1977) 2062.

[23] G. Pfister, H. Scher, Adv. Phys. 27 (1978) 747.

[24] G. Zumofen, A. Blumen, J. Klafter, Phys. Rev. A 41 (1990) 4558.

[25] P.W.M. Blom, M.C.J.M. Vissenberg, Phys. Rev. Lett. 80 (1998) 3819.

[26] H. Scher, M.F. Shlesinger, J.T. Bendler, Phys. Today 44 (1991) 26.

[27] B. Rinn, W. Dieterich, P. Maass, Phil. Mag. B 77 (1998) 1283.

[28] A. Klemm, H.-P. Müller, R. Kimmich, Phys. Rev. E 55 (1997) 4413.

[29] A. Klemm, H.-P. Müller, R. Kimmich, Physica 266A (1999) 242.

[30] F. Klammler, R. Kimmich, Croat. Chem. Acta 65 (1992) 455.

[31] R. Kimmich, NMR: Tomography, Diffusometry, Relaxometry, (Springer, Berlin, 1997).

[32] H.W. Weber, R. Kimmich, Macromol. 26 (1993) 2597.

[33] E. Fischer, R. Kimmich, N. Fatkullin, J. Chem. Phys. 104 (1996) 9174.

[34] E. Fischer, R. Kimmich, U. Beginn, M. Moeller, N. Fatkullin, Phys. Rev. E 59 (1999) 4079.

[35] R. Kimmich, R.-O. Seitter, U. Beginn, M. Moeller, N. Fatkullin, Chem. Phys. Lett. 307 (1999) 147.

[36] K. Binder, W. Paul, J. Polym. Sci. B Pol. Phys. 35 (1997) 1.

[37] P.-G. de Gennes, Scaling Concepts in Polymer Physics, (Cornell University Press, Ithaca, 1979).

[38] M. Doi, S.F. Edwards, The Theory of Polymer Dynamics, (Clarendon Press, Oxford, 1986).

[39] S. Havlin, D. Movshovitz, B. Trus, G.H. Weiss, J. Phys. A 18 (1985) L719.

[40] M. Porto, A. Bunde, S. Havlin, H.E. Roman, Phys. Rev. E 56 (1997) 1667.

[41] W. Young, A. Pumir, Y. Pomeau, Phys. Fluids A 1 (1989) 462.

[42] O. Cardoso, P. Tabeling, Europhys. Lett. 7 (1988) 225.

[43] F. Amblard, A.C. Maggs, B. Yurke, A.N. Pargellis, S. Leibler, Phys. Rev. Lett. 77 (1996) 4470.

[44] E. Barkai, J. Klafter, Phys. Rev. Lett. 81 (1998) 1134.

[45] E.R. Weeks, H.L. Swinney, Phys. Rev. E 57 (1998) 4915.

[46] W.D. Luedtke, U. Landmann, Phys. Rev. Lett. 82 (1999) 3835.

[47] G. Matheron, G. de Marsily, Water Res. Res. 16 (1980) 901.

[48] G. Zumofen, J. Klafter, A. Blumen, J. Stat. Phys. 65 (1991) 991.

[49] G.K. Batchelor, Quart. J. Roy. Meteor. Soc. 76 (1950) 133.

[50] M.F. Shlesinger, B.J. West, J. Klafter, Phys. Rev. Lett. 58 (1987) 1100.

[51] P. Tabeling, A.E. Hansen, J. Paret, in [55].

[52] I. Sokolov, A. Blumen, J. Klafter, Europhys. Lett. 47 (1999) 152.

[53] O.V. Bychuk, B. O'Shaugnessy, Phys. Rev. Lett. 74 (1994) 1795.

[54] S. Stapf, R. Kimmich, R.-O. Seitter, Phys. Rev. Lett. 75 (1995) 2855.

[55] J. Bodurka, R.-O.Seitter, R. Kimmich, A. Gutsze, J. Chem. Phys. 107 (1997) 5621.

[56] J. Klafter, A. Blumen, G. Zumofen, M.F. Shlesinger, Physica 168A (1990) 637.

[57] A. Ott, J.-P. Bouchaud, D. Langevin, W. Urbach, Phys. Rev. Lett. 65 (1990) 2201.

[58] M. Sahimi, Phys. Rep. 306 (1998) 214.

[59] S. Schaufler, W.P. Schleich, V.P. Yakovlev, Europhys. Lett. 39 (1997) 383.

[60] S. Schaufler, W.P. Schleich, V.P. Yakovlev, Phys. Rev. Lett. 83 (1999) 3162.

[61] G. Zumofen, J. Klafter, Chem. Phys. Lett. 219 (1994) 303.

[62] E. Barkai, R. Silbey, Chem. Phys. Lett. 310 (1999) 287.

[63] R. Balescu, Phys. Rev. E 51 (1995) 4807.

[64] J. Klafter, B.S. White, M. Levandowsky, Biological Motion, Lecture Notes in Biomathematics, Vol. 89, W. Alt and G. Hoffmann, eds., (Springer, Berlin, 1990).

[65] M. Levandowsky, B.S. White, F.L. Schuster, Acta Protozool. 36 (1997) 237.

[66] R. Nossal, J. Stat. Phys. 30 (1983) 391.

[67] H. Linder, Biologie, J.B. Metzler, Stuttgart, 1984.

[68] C.K. Matthews, K.E. van Holde, Biochemistry, 2nd Edition edition, (Benjamin/Cummings, Menlo Park, CA, 1996).

[69] G.M. Viswanathan, V. Afanasyev, S.V. Buldyrev, E.J. Murphy, P.A. Prince, H.E. Stanley, Nature 381 (1996) 413.

[70] J. Feder, Fractals, (Plenum, New York, 1988).

[71] H. Takayasu, Fractals in the Physical Sciences, (Manchester University Press, Manchester, 1990).

[72] J.-F. Gouyet, Physique et Structures Fractales, (Masson, Paris, 1992).

[73] K. Falconer, Fractal Geometry: Mathematical Foundations and Applications, (Wiley, Chichester, UK, 1990).

[74] B.B. Mandelbrot, J.W. van Ness, SIAM Rev. 10 (1968) 422.

[75] B. O'Shaugnessy, I. Procaccia, Phys. Rev. Lett. 54 (1985) 455

[76] J. Klafter, M.F. Shlesinger, G. Zumofen, Phys. Today 49 (2) (1996) 33.

[77] B. Berkowitz, H. Scher, Phys. Rev. E 57 (1998) 5858.

[78] H. Scher, M. Lax, Phys. Rev. B 7 (1973) 4491, 4502.

[79] J. Klafter, R. Silbey, Phys. Rev. Lett. 44 (1980) 55.

[80] J. Klafter, A. Blumen, M.F. Shlesinger, Phys. Rev. A 35 (1987) 3081.

[81] G.M. Zaslavsky, S. Benkadda, Chaos, Kinetics and Nonlinear Dynamics in Fluids and Plasmas, (Springer, Berlin, 1998).

[82] E. Barkai, V. Fleurov, Phys. Rev. E 56, 6355.

[83] R. Kutner, P. Maass, J. Phys. A 31 (1998) 2603.

[84] J. Bernasconi, W.R. Schneider, W. Wyss, Z. Phys. B 37 (1980) 175.

[85] V. Seshadri, B.J. West, Proc. Natl. Acad. Sci. USA 79 (1982) 4501.

[86] B.J. West, V. Seshadri, Phsyica A 113 (1982) 203.

[87] F. Peseckis, Phys. Rev. A 36 (1987) 892.

[88] H.C. Fogedby, Phys. Rev. Lett. 73 (1994) 2517.

[89] H.C. Fogedby, Phys. Rev. E 50 (1994) 1657.

[90] R. Kubo, M. Toda, N. Hashitsume, Statistical Physics II, Solid State Sciences, Vol. 31, (Springer, Berlin, 1985).

[91] R. Muralidhar, D. Ramkrishna, H. Nakanishi, D. Jacobs, Physica A 167 (1990) 539.

[92] K.G. Wang, L.K. Dong, X.F. Wu, F.W. Zhu, T. Ko, Physica A 203 (1994) 53.

[93] K.G. Wang, M. Tokuyama, Physica A 265 (1999) 341.

[94] V.M. Kenkre, E.W. Montroll, M.F. Shlesinger, J. Stat. Phys. 9 (1973) 45.

[95] V.M. Kenkre, Phys. Lett. A 65 (1978) 391.

[96] D. Bedeaux, K. Lakatos, K. Shuler, J. Math. Phys. 12 (1971) 2116.

[97] I. Oppenheim, K.E. Shuler, G.H. Weiss (Eds.), Stochastic Processes in Chemical Physics: The Master Equation, (MIT Press, Cambridge, Massachusetts, 1977).

[98] C. Tsallis, S.V.F. Levy, A.M.C. Souza, R. Maynard, Phys. Rev. Lett. 75 (1995) 3589.

[99] C. Tsallis, D.J. Bukman, Phys. Rev. E 54 (1996) R2197.

[100] L. Borland, Phys. Rev. E 57 (1998) 6634.

[101] A. Compte, D. Jou, J. Phys. A 29 (1996) 4321.

[102] D.H. Zanette, P.A. Alemany, Phys. Rev. Lett. 75 (1995) 366.

[103] L. E. Reichl, *Statistical Physics*, J. Wiley. New York (1998).

[104] M. F. Shlesinger, J. Stat. Phys. 10, 421 (1974).

[105] P. Grigolini, L. Palatella, G. Raffaelli, in press on Fractals, cond-mat/0104166.

[106] E. Barkai, R. Metzler, J. Klafter, Phys. Rev. E 61, 132 (2000).

[107] A. Compte, Phys. Rev E 53 (1996) 4191.

[108] R. Metzler, J. Klafter, I. Sokolov, Phys. Rev E 58 (1998) 1621.

[109] B.J. West, P. Grigolini, R. Metzler, T.F. Nonnenmacher, Phys. Rev. E 55 (1997) 99.

[110] S. Jesperrsen, R. Metzler, H. C. Fogedby, Phys. Rev. E 59 (1999) 2736.

[111] P. Allegrini, P. Grigolini, B. J. West, Phys. Rev. E. 54, 4760 (1996).

[112] C.-K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, A. L. Goldberger, Phys. Rev E 49, 1685 (1994).

[113] K.Hu, P. Ch. Ivanov, Z. Chen, P. Carpena, H.E. Stanley, Phys. Rev. E 64, 011114 (2001).

[114] J. B. Bassingthwaighte, L. S. Liebovitch and B. J. West, *Fractal Physiology* (Oxford University Press, Oxford, 1994).

[115] D. B. Percival and A. T. Walden, *Wavelet Methods for Time Series Analysis*, Cambrige University Press, Cambrige (2000).

[116] C. Beck, F. Schlögl, *Thermodynamics of chaotic systems*, Cambridge University Press, Cambridge (1993).

[117] A. Rényi, *Probability Theory*, North-Holland, Amsterdam (1970).

[118] C. Tsallis, J. Stat. Phys. 52, 479 (1988).

[119] R. J. V. dos Santos J. Math. Phys. 38 (8) 4104, August 1997.

[120] F. Argenti, V Benci, P.Cerrai, A. Cordelli, S. Galatolo and G. Menconi, Chaos, Soliton and Fractals 13 (2002) 461-469.

[121] Manneville P. J Phys (Paris) 1980; 41:1235.

[122] G. Zumofen and J. Klafter, Phys. Rev. E 47, 851 (1993).

[123] G. Zumofen and J. Klafter, Physica (Anmsterdam) 69D, 436 (1993).

[124] G. Trefan, E. Floriani, B.J. West, and P. Grigolini, Phys. Rev. E bf54, 4760 (1994).

[125] P. Allegrini, P. Grigolini, and B.J. West, Phys. Rev. E 54, 4760 (1996).

[126] G. M. Zaslavsky, *Physics of Chaos in Hamiltonian Systems*,Imperial College Press, London (1988).

[127] T. Geisel and S. Thomae, Phys. Rev. Lett. 52, 1936 (1984).

[128] M. Kac, *Probability and Related Topics in Physical Sciences* (Interscience, New York,1958).

[129] M. Buiatti, P. Grigolini, A. Montagnini, Phys. Rev. Lett. 82, 3383 (1999).

[130]] M. Bologna, P. Grigolini, and J. Riccardi, Phys.Rev. E 60, 6435 (1999).

[131] M. Annunziato, P. Grigolini, Phys. Letters A 269, 31 (2000).

[132] G. Zumofen, J. Klafter, Phys. Rev. E 47, 851 (1993).

[133] M. Buiatti, P. Grigolini, L. Palatella, Physica A 268, 214 (1999).

[134] National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/

[135] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons and H.E. Stanley, Nature 356, 168 (1992).

[136] W. Li, Int. J. Bifurcation Chaos Appl. Sci. Eng. 2, 137 (1992); W. Li and K. Kaneko, Europhys. Lett. 17, 655 (1992); W.Li, T. Marr, and K. Kaneko, Physica (Amsterdam) D 75, 392 (1994).

[137] R. Voss, Phys. Rev. Lett. 68, 3805 (1992).

[138] S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H.E. Stanley, Phys. Rev. E 47, 4514 (1993).

[139] A.K. Mohanti and A.V.S.S. Narayana Rao, Phys. Rev. Lett. 84, 1832 (2000).

[140] B. Audit, C. Thermes, C. Vaillant, Y. d'Aubenton-Carafa, J.F. Muzy, and A. Arneodo, Phys. Rev. Lett. 86, 2471 (2001).

[141] P. Allegrini, M. Barbi, P. Grigolini and B.J. West, Phys. Rev. E 52, 5281 (1995).

[142] P. Allegrini, M. Buiatti, P. Grigolini and B.J. West, Phys. Rev. E 57, 4588 (1998).

[143] P. Allegrini, M. Buiatti, P. Grigolini and B.J. West, Phys. Rev. E 58, 3640 (1998).

[144] P. Allegrini, P. Grigolini, B.J. West, Phys. Rev. E 54, 4760 (1996).

[145] C. -K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley and A.L. Goldberger. Phys. Rev. E 49, 1685 (1994).

[146] A. Arneodo, E. Bacry, P.V. Graves, and J.F. Muzy, Phys. Rev. Lett. 74, 3293 (1995).

[147] A. Torcini, R. Livi, A. Politi, A dynamical approach to protein folding, cond-mat/0103270.

[148] M. Araujio S. Havlin, G.H. Weiss, and H.E. Stanley, Phys. Rev. A 43, 5240 (1991).

[149] R. Mannella, P. Grigolini, and B.J. West, Fractals 2, 81 (1994).

[150] B.V. Gnedenko and A.N. Kolmogorov, Limit Distributions for Sums of Independent Random Variables, Addison-Wesley Publishing Company, Inc. Cambridge (1954).

[151] E.N. Parker, Astrophys. J. 330, 474 (1988); E.N. Parker, Sol. Phys. 121, 271 (1989).

[152] R.P. Lin, R.A. Schwarz, S.R. Kane, R.M. Pelling, and K.C. Hurley, Astrophys. J. 283, 421 (1984); S. Sturrock, P. Kaufmann, P.L. Moore, and D.F. Smith, Sol. Phys. 94, 341 (1984); N.B. Crosby, M. J. Aschwanden, and B.R. Dennis, Sol. Phys. 143, 275 (1993).

[153] G. Boffetta, V. Carbone, P. Giuliani, P. Veltri, and A. Vulpiani, Phys. Rev. Lett. 83, 4662 (1999).

[154] P. Giuliani, V. Carbone, P. Veltri, G. Boffetta, A. Vulpiani, Physica A 280, 75 (2000).

[155] M. S. Wheatland, Astrophys. J. 536 L 109 (2000).

[156] W. H. Press and others, *Numerical Recipes in C*, Cambridge University Press, Cambridge (1992).

[157] F. Lepreti, V. Carbone, and P. Veltri, Astrophys. J. 536, L133 (2001).

[158] B.V. Gnedenko, A.N. Kolmogorov, *Limit Distributions for Sum of Independent Random Variables*, Addison Wesley, Reading (1954).

[159] D. Bedeaux, K. Lakatos Lindenber and K.E. Shuler, J. Math. Phys. 12, 2166 (1971).

[160] P. Grigolini, L. Palatella, G. Raffaelli, in press on Fractals, cond-mat/0104166.

[161] M. Ignaccolo, P. Grigolini, A. Rosa, Rev. E 64, 026210 (2001).

[162] M.S. Wheatland, P.A.Sturrock, and J.M. McTiernan, Astrophys. J. 509, 448 (1998).

[163] M. S. Wheatland, arXiv:astro-ph/0107147.

[164] J. Laherrére and D. Sornette, Eur. Phys. J. B 2, 525 (1998).

[165] R. N. Mantegna and H. E. Stanley Phys. Rev. Lett. 73, 2946 (1994)

[166] E. Floriani, R. Mannella, P. Grigolini, Phys Rev E 52, 5910 (1995).

[167] Ventura, S. J, Curtin, S. C., Matthews, T.J., *National Vital statistics Reports,* 47 (26), 1-14 1999).

[168] *The National Campaign to Prevent Teen Pregnancy,* 1-5 (1999).

[169] B.J. West, P. Hamilton, D.J. West, *Fractals* 7,113 (1999).

[170] Lam, David A.; Miron, Jeffrey A. Silver Spring, Maryland. *Demography,* Vol. 33, No. 3, 291-305, Aug 1996).

[171] C. Tsallis, Physica A 221, 277 (1995).