Politecnico di Torino

Master of Science in Industrial Engineering and Management

Master of Science Thesis

March 2020

Entropy as a tool to measure characteristic features of financial indexes and of related variables of financial interest

Advisor: Anna Carbone
Candidate: Pietro Murialdo

Abstract

In this work, the information embedded in widely used financial stochastic processes (such as GARCH, ARFIMA) is studied by means of the cluster entropy approach. This study is motivated by a previous analysis conducted on real worlds financial series (such as DAX, Nasdaq) (Carbone, 2013), (Ponta and Carbone, 2018). Specifically, the ability of artificial market series of replicating real-world phenomena is investigated, from the information theory point of view. Data have been generated using different processes such as Autoregressive Fractionally Integrated Moving Average and General Autoregressive Conditional Heteroscedastic. The DMA cluster entropy method consists in partitioning a time series with its moving average to obtain a set of clusters of variable lengths and the corresponding frequencies. Then, a probability distribution is drawn and the Shannon entropy (Shannon, 1948) is used to quantify the embedded amount of information. The results over the artificial series is proved to be consistent with data from financial markets. A power-law correlation for clusters of shorter length than the moving average window and an exponential correlation for clusters of longer lengths are observed. Finally, these information are summarized by the *Market Dynamic Index* presented in Ponta and Carbone (2019) integrating over clusters length and different moving average windows.

Contents

Acknowledgements

I would like to express my deepest gratitude to Professor Anna Carbone for her patient guidance and enthusiastic encouragement during the development of this work.

I wish to thank my parents for their support and encouragement throughout my studies without which it would not have been possible to achieve this goal.

Finally, I would also like to thank my friends in Turin for their support, for the time spent together and all the challenges overcome together.

1 Introduction

In this work time series generated by several financial and mathematical models are investigated by means of a *cluster entropy* approach. The approach to be applied to series requires first a partition of the series itself into disjoint sets. The partition method used here divides a sequence into segments of variable length according to the intersections with its moving average. These segments are called *clusters*. These clusters are ranked according to their length to obtain a probability distribution. The amount of information contained in the cluster length probability distribution is then evaluated with the Shannon entropy. The cluster entropy dynamics is summarized in terms of the *Market Dynamic Index* to provide a comparable figure.

In general, the Shannon entropy is a way to compress a signal in its elementary parts and transmit them to the receiver without loss of relevant information (Shannon, 1948). The Shannon entropy as a measure to quantify the expected information in a message is introduced. Originally used in statistical mechanics, entropy is a measure of the probability that a thermodynamic system would be in such a configuration. Similarly, in information theory entropy is used to study the expected configuration of a message in probabilistic terms. An idea closely linked to Kolmogorov complexity of an object, which defines the minimum length of a program that produces the object as output.

The cluster entropy method used in this work was first applied to study the information content of the 24 chromosome sequences of human genome. Nucleotide composition of each chromosome was the variable under consideration. The nucleotide sequences were mapped to a one-dimensional numerical series. It was found that the cluster entropy of nucleotide sequences could be written as a sum of three components, a constant, a logarithmic term and a linear term. Clusters with logarithmic behavior were found to be power-law correlated and said to be ordered, while clusters with linear behavior were found to be exponentially correlated and said to be disordered. Ordered clusters were able to transmit - on average - the same amount of information about nucleotide composition of the whole series, while disordered ones provided fluctuating information Carbone (2013).

The same approach can then be applied to any one-dimensional correlated series, as is the case of financial time series which are the relevant case in this work.

As stated before, to find a probability distribution for the Shannon functional it is first necessary to partition the continuous phase-space, a continuous set of all the possible configuration that the system can assume, into disjoint sets. The financial series $y(t)$ is partitioned in clusters by the intersection with its moving average $\tilde{y}_n(t)$, with n the moving average window. A cluster is defined as the region bounded by two consecutive crossing points of $y(t)$ and $\tilde{y}_n(t)$. The intersections between the two series yield a set of clusters that can be ranked according to their lengths and frequencies to obtain a probability distribution $P(\tau, n)$ for each n.

A fundamental feature of this method consists in the way used to partition the sequence. The probability distribution allows to distinguish between power-law and exponentially

correlated clusters, retaining determinism out of randomness by varying n and vice versa. It is therefore possible to separate information-bearing clusters from random ones. Last but not least, being the moving average a heavily used tool in technical trading, clusters defined by this partitioning method have a direct connection with the trader perspective.

Also in the case of financial time series the Shannon entropy can be described by a constant, a logarithmic term and a linear term, as anticipated in Carbone (2013) in the case of chromosomes.

To clarify the meaning of these terms reference is made to thermodynamics. In an *isolated system*, entropy increase dS is related to the irreversible processes occurring in the system spontaneously. As the state of maximum entropy is reached, entropy tends to an asymptotic value of the form of the Boltzmann entropy $S = \log \Omega$, with Ω the volume of the isolated system. When the system is free to interact with the external environment, in which case it is said an *open system*, a further increase of entropy dS_{ext} occurs due to the irreversible spontaneous processes arising from the interaction. Therefore, an excess of entropy is generated and the asymptotic limit of the isolated system, expressed by a logarithmic curve, is exceeded.

In the framework of information theory an analogous interpretation of the cluster entropy can be outlined. Clusters whose $\tau < n$ are power-law distributed and their behavior is described by the logarithmic term expressed as a function of τ. These are the ordered clusters that were found able to transmit the whole nucleotide information in the case of chromosomes. Power-law correlated clusters does depend only on the size of the clusters themselves and their entropy tends to an asymptotic value, so their behavior can be compared to the one of entropy for a thermodynamic isolated system. On the other hand, clusters whose $\tau \geq n$ are exponentially correlated and their behavior is described by a linear term expressed as a function of τ and n. These clusters are said to be disordered and represent the excess entropy introduced by the interaction with the external system. The excess entropy component accounts for the additional heterogeneity introduced by the partitioning process and is in fact a function of the moving average window n.

The cluster entropy method was applied to several major European stock indexes to study market dependence of prices and volatilities in Ponta and Carbone (2018). Entropy of price series were found to be practically market invariant while volatility series showed significance market dependence. Cross market homogeneity of price series is the reason behind the choice of further investigating price series of artificial financial models.

In the same work a cumulative information measure was developed to obtain the weights of an efficient portfolio. In Ponta and Carbone (2018) a method based on a cluster entropy approach was proposed to estimate weights of the efficient portfolio. The portfolio weights for each asset were calculated integrating the cluster entropy curves of the volatility series first over the cluster length τ and then over the moving average window n to summarize all the information into a single figure. Values obtained in this fashion for each asset were then normalized and compared with the ones obtained by means of the Markowitz approach. Research has shown that returns are not normally distributed as assumed in the

traditional Markowitz optimal portfolio method. Moreover, it was also shown that financial variables are not only influenced by the immediate past but also by shocks occurred further in time. In other words, financial series present some degrees of persistence. This features compromise Markowitz assumption of normal distributed returns and lead to important implication in portfolio allocation, option and asset pricing and risk management (Franke, Härdle, and Hafner, 2015).

The idea was further developed to study the dynamics of price series in Ponta and Carbone (2019). Here the *Market Dynamic Index* was defined as the integral of the cluster entropy over τ and was used to quantify the dynamics of consecutive time periods of some American assets' prices. The study revealed a systematic dependence of clusters over time.

In this work the analysis by means of the cluster entropy is carried further. Several financial and mathematical models are investigated to understand their ability to embed information and to study their dynamics.

The study developed in this thesis is an application of concepts traditionally developed in statistical physics to economics. After decades where fields of study have focused on different topics and developed their own tools, an interest in interdisciplinary applications has emerged. The introduction of approaches and methods developed in other fields have often proved useful in solving problems and answering new questions (Carbone, Kaniadakis, and Scarfone, 2007).

In recent years mass communications and new technologies, especially the internet, have evolved rapidly, increasing the amount of information exchanged and recorded accurately. A key year for modern finance was 1973, when, as a consequence of the collapse of the Bretton Woods' agreements, currencies value become fluctuating and determined by the market, currencies begun to be traded on dedicated foreign exchange markets 24h a day, and Black and Scholes published their first work about a closed form to evaluate an option.

Since that time, volume of trading increased enormously. Additionally, a second revolution began in the 1980s, when electronic trading and computers become a standard in financial markets. One result is that there is a huge availability of financial data characterized by the properties of being highly reliable and high-frequency, which for liquid markets means that the delay between two records can be of few seconds (Franke, Härdle, and Hafner, 2015).

The immense availability of such data have led scientists to apply methods traditionally belonging to other fields, such as statistical physics and complexity, to investigate social phenomena. In fact, elementary social components interact giving rise to phenomena that cannot be studied linearly but that often require methods able to account for larger number of heterogeneous interaction. Financial markets resemble complex systems, where agents interact between themselves in the presence of feedback from previous interactions.

This thesis is organized as follows. In particular, before presenting the study developed on financial series models, a thermodynamic framework for modern economics is exposed. First, some notions about thermodynamics and power-law distributions will be introduced. Then, a thermodynamic theory of economics is presented. Topics such as in-

come and wealth, elasticity and utility theory, production and consumption, and money are investigated. Ideas are illustrated along with their differences from traditional economics and their limits, to offer an alternative perspective of the economic system (Bryant, 2015). Such thermodynamic perspective of the economic system can be found in Chapter 2. Then, in Chapter 3, the data used in this work are described breaking down the models used to generate artificial time series and the main characteristics of financial series for major US markets used as a benchmark. In Chapter 4 the method used in this analysis and the ideas behind it are outlined, moreover some preliminary results from other works are recalled for the sake of comparison. In Chapter 5 results for *Fractional* and *Geometric Brownian Motion*, *Autoregressive Fractionally Integrated Moving Average* and other processes are presented. Finally, in Chapter 6 an overview and a discussion of the results is presented and a path for future work is suggested. Furthermore, an Appendix containing results for other sets of parameters and MATLAB scripts is available for further insight.

2 A Socioeconomic Perspective of Thermodynamics

In recent years, the continuous growth of human population and economy, the possible advent of resources' shortages, climate and environmental issues, has spawned researchers across the globe to wonder what might occur to the future of the world, to the sustainability of our economic models and the time scales involved in these changes.

Today, economics has to deal with an additional theme: the environment. Additional not in the sense that economics has never considered the environment, but rather in the way economics looks at the environment. Until today human beings has been looking at the surrounding nature as a source of resources to foster their development, or as constraint. Nowadays the climatic changes are forcing us to integrate the environment into the scopes of economics. Societies must find a sustainable way of fostering their development because nature is turning its back against them.

Moreover, after the 2008 financial crisis, economists' reputation was shaken. It is also in this context that *Econophysics* developed. Econophysics is a recent and rapidly evolving field of study that arose because of the methods and the origin/background of its researchers. In reality, the field was born in the '80s, but its attention was primarily concerned with financial markets. It is no surprise that Fisher Black, the driving force behind the Black-Scholes-Merton model, the model used in finance to price option derivatives, was trained as a physicist. In more recent times, econophysics has moved closer to economics, developing models on the probability distributions of money, wealth and income in societies.

Econophysics uses mathematical methods developed in statistical physics to study statical properties of complex systems, such as those where a large number of humans gather in an economic system interacting between each other.

Econophysics emphasizes quantitative analysis of large amount of economic and, especially, financial data. The latter being easier to obtain and more precise since the introduction of computers and internet. Studying mathematical models of large number of interacting agents has to do with agent-based modeling and simulation, which distance itself from the representative-agent approach used in economics, which by definition ignores statistical and heterogeneous aspects of the economy.

Historically, statistical mechanics was developed in the second half of the nineteenth century by James Clerk Maxwell, Ludwig Boltzmann and Josiah Willard Gibbs. These physicist believed in the existence of atoms and developed mathematical methods to describe their properties. Before the 1850s, statistics was considered a tool of political economy, in fact, the term *statistics* origin from its practitioners, *statists*. Then the subject developed into a more general quantitative discipline and encouraged physicist in developing statistical mechanics.

Rudolf Clausius started the development of the kinetic theory of gases, but it was James Clerk Maxwell who, inspired by the popularity of social statistics at the time, made decisive steps in deriving the probability distribution of velocities of molecules in gases.

This approach was further developed by Ludwig Boltzmann, who said: *"The molecules are like individuals, [...] and the properties of gases remain unaltered because the number of these molecules, which on the average have a given state, is constant"*. Then, Boltzmann added: *"This opens a broad perspective, if we do not only think of mechanical objects. Let's consider to apply the method to the statistics of living beings, society, sociology and so forth"*.

However, not only physicist pay attention to the applications of statistical mechanics to gasses or engineering systems, but also to social and economic systems, as they consider it a risky practice. Few remember that the process originated in exactly the opposite way, when it was not absurd to recall human habits to explain the behavior of inanimate particles.

Before entering the core of this work, we want to lay down a theory of human and economic development from an econophysicist point of view. It shall take into account classic thermodynamics to show not only human development through an anthropocentric point of view, but the whole system of earth and universe in a comprehensive manner. The objective of this chapter is to illustrate that the laws of thermodynamics might be used to build an alternative framework to interpret the economic system.

2.1 The Laws of Thermodynamics and Maxwell-Boltzmann Distribution

The concept of entropy, that is the core of this work, was developed during the XIX^{th} century by some of the most important scientist of all time, such as Sadi Carnot, Rudolf Clausius, Willard Gibbs, James Maxwell and Ludwig Boltzmann.

Before introducing the econophysics perspective of economics, some notions of thermodynamics such the ideal equation of gasses, the First and Second Laws of Thermodynamics will be introduced.

In an ideal gas, a number N of molecules are free to move, colliding and exchanging kinetic energy with each other, inside an enclosure of volume V, exerting a pressure P against its walls. Through the application of heat from outside, the molecules are stimulated to move faster, collide more frequently, exchange more energy and so accumulate internal energy; thus the temperature T is raised, and potentially also V and P can increase. The relationship between this factors is the *ideal gas* equation:

$$PV = NkT, \tag{1}$$

where the Boltzmann Constant k is a measure of the average kinetic energy of a gas per degree of temperature, and thus T is a measure of relative kinetic energy of a molecule. The second member of (1) is then a measure of total energy.

For a non-flowing system, N is constant, while V and P vary as a function of T. On the other hand, for a flow system, such as a pipeline, N is a flow of molecules per period of time, with a corresponding flow of gas volume V per unit of time and variable P and T.

The First Law of Thermodynamics says is that when a closed and isolated system is taken through a cycle, the net work delivered to the surroundings is proportional to the net heat taken from the surroundings, and vice versa.

The First Law is about the conservation of energy. Three kinds of energy act on a gas that is closed in a piston. A mechanical force from the piston applied through a *work* G, some *heat* Q that can be transferred across the piston's wall, and the *internal energy* U, measured as a function of temperature T. The First Law states that these energy always have to balance each other out:

$$dQ = dG + dU. \tag{2}$$

According to the Second Law of Thermodynamics in a closed system, one cannot on its own transfer energy from one body at a low temperature to another body having a higher temperature; only the other way. It's the relative temperature difference - or gradient - between the two bodies that determines the direction of heat flow. There is no actual loss of overall energy content of the two bodies combined, but there is a reduction in the higher grade energy potential held by the hotter body which cannot be retrieved. This loss of potential constitutes an increase in entropy.

The Second Law of Thermodynamics is concerned with the direction in which reactions occur. In fact, every process in nature proceeds spontaneously in one direction only, never in the opposite. These processes are said to be *irreversible*. For irreversible processes, happening in a closed system, the entropy of the system always increases.

There are two ways of describing entropy, one is macroscopic and the other microscopic. Starting with the macroscopic description, entropy variation for a *reversible process* is defined as,

$$\Delta S_{rev} = \int_i^f \frac{dQ}{T}, \tag{3}$$

where dQ represents the amount of heat exchanged and T the fixed temperature at which the transformation occurs between an initial and a final state, respectively, i and f. Entropy is measured in $Joule/Kelvin$.

Entropy is a *state function*. Its final value does not depend on the path followed by the transformation from the initial to the final state, but only on the actual state of the system.

Let us imagine a thermal insulator of volume V into which there is an amount of gas n at a temperature T maintained constant during the process by an external thermal source. The gas is initially pressed down on the top by a box of lead balls. The balls are then gradually removed. During this process the gas expands at constant temperature.

The First Law of Thermodynamics states that,

$$dQ + dL = dE_{int}. \tag{4}$$

Since internal energy of a gas depends only on its temperature, in this isothermal trans-

formation $\Delta E_{int} = 0$. Substituting dL with $-pdV$ and p with nRT/V,

$$dQ = -dL = pdV = nRT\frac{dV}{V}, \tag{5}$$

resulting in,

$$Q = \int dQ = nRT \int_{V_i}^{V_f} \frac{dV}{V} = nRT \ln \frac{V_f}{V_i}. \tag{6}$$

Thus the entropy change is equal to,

$$\Delta S = \frac{Q}{T} = nR \ln \frac{V_f}{V_i}. \tag{7}$$

So, the entropy variation depends only on the initial and final state of system. After the expansion, the final volume is larger than the initial one, therefore the entropy change is positive.

However, the entropy postulate applies to *irreversible* and *isolated* transformation only. In this example the transformation is not irreversible and the system is not isolated. to consider the system isolated, the thermal source that maintains the temperature constant must be considered in the system with the enclosure. So, the entropy variation of the gas, $\Delta S_{gas} = -|Q|/T$ is equal but opposite in sign to the entropy variation of the thermal source $\Delta S_{res} = +|Q|/T$, resulting in an overall *null* entropy variation. The entropy postulate must then be slightly changed. In a closed system, the entropy always increases for irreversible processes or remain constant for reversible ones,

$$\Delta S \geq 0. \tag{8}$$

In any case it never decreases. This implies that if somewhere in a closed system entropy decreases, somewhere else it must increase, at least of the same amount.

As it is known, unluckily in nature it is not possible to build a perfectly reversible process, so in reality the equal sign in Equation (8) must be removed.

2.1.1 Statistical Thermodynamics

There is also another method to estimate the entropy of a system. This is a problem of statistical mechanics and deals with the possible states in which atoms and molecules can be found inside a closed box.

Let us distribute 8 molecules, identical to each other, inside a box divided ideally in two halves. When the box is empty, there are 8 different possibilities to distribute the molecules. Once the first is placed, there are 7 left possibilities to distribute the remaining molecules, and so on. The total number of possible arrangements is,

$$8! = 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 40320. \tag{9}$$

However, since the molecules are identical, not all these possibilities are independent. If in one half of the box there are 5 molecules and in the other half there are 3, then the total number of independent possible arrangements is,

$$w = \frac{N!}{N_1!N_2!} = \frac{8!}{5!3!} = 56, \tag{10}$$

Every possible independent arrangement that the molecules can assume inside the box is called a *micro-state*, and the *multiplicity* w the number of micro-states for each configuration. The total number of possible micro-states, that is equal to the sum of multiplicities over every configuration, is $2^8 = 256$.

The fundamental assumption in mechanical statistics is that a system has equal probability to be in any micro-state. So, while moving around, the molecules will spend the same amount of time in each configuration. But each configuration is not equally probable. Configurations with higher multiplicity are more likely to be found.

Calculating the multiplicity for each possible configuration, it is clear that the most likely configuration are those where molecules are evenly distributed. For example, the system will spend 70 times more time in the configuration where 4 molecules occupy each half than the one where all 8 molecules occupy one half and none occupies the other.

In real cases, the number of molecules is much higher. When the number of molecules is of the order of 10^{22}, that is the number of molecules or atoms in a mole of any substance, the major part of micro-states correspond to those of almost equal distribution of the molecules in the two sides of the box. It exist, theoretically, the chance that all the molecules are in one half of the box only, but the probability is so infinitesimally small that is practically impossible.

Therefore, thermodynamic systems on their own tend to assume the most likely configuration, which coincide with the state of maximum entropy. The Austrian physicist Ludwig Boltzmann was the first scientist to merge the multiplicative characteristic of probability, with the additive one of entropy. From here the hypothesis that the relationship between probability and entropy has to include the logarithm, which allows the following operation,

$$\ln(ab) = \ln a + \ln b. \tag{11}$$

The probabilistic entropy relationship found by Boltzmann, which in his honor takes his name, is the *Boltzmann entropy* equation,

$$S = k \ln w, \tag{12}$$

with k the Boltzmann constant and w the multiplicity associated with the configuration under consideration.

Finally, let us imagine to apply the concept of statistical entropy to a cup of coffe that is being mixed by a spoon. In this case the focus will be on molecules *speed* rather than *position*.

The final state, where the molecules are in resting state and moving randomly, has more micro-states than the case where the molecules are moving orientated in the same direction because of the spoon movement. So,

$$w_{rest} > w_{rot}, \tag{13}$$

then, Equation (12) implies that,

$$S_{rest} > S_{rot}. \tag{14}$$

Accordingly to the Second Law of Thermodynamics, coffee, once that is left alone, tends to the state of largest entropy. It never happens the opposite. So coffee tends to move from a rotatory motion to a resting one. No one has ever witnessed a cup of coffee that started rotating by itself.

In conclusion, the concept of entropy is often associated to the one of *disorder*. When coffee is moving in rotatory motion, the movements of the molecules is ordered and entropy is low, while in the final steady state with high entropy, molecules motion is random and therefore more disordered. Therefore, the Second Law states that a closed system tends to the state of maximum entropy, which coincides with the most disordered one. This concept extends when the surrounding are taken into consideration. As it will be discussed further, also in the limit case where the system coincides with the whole Universe, the Second Law is satisfied.

Following the idea that velocity of molecules is directly connected to entropy, Maxwell distribution can be used to calculate the entropy. The Scottish physicist *James Clerk Maxwell 1831 - 1879* first found a relationship that described the velocity of circulation of molecules in a gas containing a high number of molecules.

It is in fact possible to calculate the *root-mean-squared speed* v_{rms} of molecules in a isolated gas relying on the concept of pressure towards the enclosure walls and the density of the gas. However, this does not provide any information on the distribution of the speeds. It is not likely that all molecules will have the same speed as collisions are continuously happening. Nor it is more likely that all particles have speed equal to zero or much greater than v_{rms}. So the *Maxwell-Boltzmann distribution* for a sample of gas at thermodynamic equilibrium, with temperature T and N molecules of mass m is,

$$N(v) = 4\pi N\left(\frac{m}{2\pi kT}\right)^{3/2} v^2 e^{-mv^2/2kT}, \tag{15}$$

with the product $N(v)dv$ equal to the number of molecules having speed in the range v and $v + dv$. Thus,

$$N = \int_0^\infty N(v)dv, \tag{16}$$

is equal to the total number of particles. Choosing appropriately v_1 and v_2 it is possible to find the number of molecules with the velocity in that range as the area under the distribution between the two limits.

Let us imagine some water in an isolated enclosure. The Maxwell-Boltzmann distribution says that only few molecules will have enough speed to escape from water surface and evaporate; the tail is in fact small. When evaporating, the kinetic energy inside the water decreases, with a consequent decrease of molecules' speed and thus a decrease of temperature. This is why evaporation is a cooling process. However, if the enclosure is removed, the energy outflow will be compensated by inflow of heat, and the evaporation process will take place at constant temperature T until there will be no water left.

The Maxwell-Boltzmann distribution can therefore be used to estimate the velocity of particles in a gas. Then this information can be used to evaluate the entropy through relation (12).

2.2 Power-Laws or Pareto distributions

Power-law distributions are used to described phenomena where a large portion of the effects can be linked to a small portion of causes and the remaining portion of effects to a large portion of the causes. Such a behavior is is very common in nature. Examples of phenomena distributed as power-laws are the following: word frequency in natural language, citations of scientific papers, magnitude of earthquakes, diameters of moon craters, wealth of richest people, populations of cities and many more. However, scientists cannot still define if there is a single underlying mechanism that origins such behavior.

Power-law distribution are sometimes called Pareto distributions, in honor of *Vilfredo Pareto 1848 - 1923*, an Italian engineer-economist who first made use of the relationship to explain income distribution, or Zipf's law, an American linguist and philologist who studied statistical behavior of different languages. Two example are now presented to introduce the reader to an illustration of power-law distributions.

The average human being is around 180cm tall, which means that some people are higher and other are shorter. We never see people whose height is very far from the mean, for example we never see people 10cm or 500cm tall. According to the Guinness Book of Records, the shortest and tallest man were respectively 57cm and 272cm, with a ratio of 4.8.

However, as it is well known, many other phenomena are distributed differently. For example sizes of towns and cities vary much more significantly. There are few huge cities with populations above millions, while many more have smaller numbers, in the order of thousands. In the US, the largest city is New York City, with around 8 million people; the smallest, according to the Book of Records, is Duffield in Virginia, with a population of 52 people. In this case, the ratio is 150000.

So, in the former case, it is likely to find the major part of peoples' heights to be centered around the mean with few outliers. In the latter case instead there will be few cities that accounts for a big fraction of the population and many more small cities that accounts for the rest. Heights distribution can be well approximated by a normal distribution. On the other hand, the distribution of cities' dimensions is significantly skewed, so a different

distribution is needed.

Let $p(x)dx$ be the fraction of cities with population between x and $x + dx$. If the histogram on a log-log scale of the data is a straight line with negative slope, then the logarithmic transformation would look like $\ln p(x) = -\alpha \ln x + c$, with α and c two constants. Taking the exponential of both sides:

$$p(x) = Cx^{-\alpha}. \tag{17}$$

Distributions of the form (17) are said to be following a *power-law* distribution, with α the exponent of the power law and $C = e^c$.

Power-laws phenomena can be represented in many ways. The most efficient consists in using a *cumulative distribution function*,

$$P(x) = \int_x^\infty p(x)dx = C\int_x^\infty x^{-\alpha}dx = \frac{C}{\alpha - 1}x^{-\alpha+1}. \tag{18}$$

The cumulative distribution function follows a power-law as well, but with exponent $\alpha - 1$. The plot of $P(x)$ on logarithmic scales is again a straight line, but with a slightly shallower slope.

One should also note that few real-world distributions follow a power-law behavior over their entire range. When $x \to 0$ the function $p(x) = Cx^{-\alpha}$ diverges, so the power-law behavior must deviate to another behavior for some $x < x_{min}$. This is why often it is said that a certain phenomenon has a *power-law tail*.

Now, the principal properties of power-law distributions are derived. Equation 17 says that a continuous real variable x has probability $p(x)$ to assume value between x and $x+dx$ equal to $Cx^{-\alpha}$, with $\alpha > 0$. Thus, the constant C is found imposing the normalization requirement on 18:

$$1 = \int_x^\infty p(x)dx = C\int_x^\infty x^{-\alpha}dx = \frac{C}{1-\alpha}\left[x^{-\alpha+1}\right]_{x_{min}}^\infty. \tag{19}$$

As we have just said, there must be a lower value x_{min} that act as a limit for the power-law to converge.

For the right-hand side of the equation to converge, it must be $\alpha > 1$. Otherwise the equation diverges and cannot be normalized. Solving (19),

$$C = (\alpha - 1)x_{min}^{\alpha-1}, \tag{20}$$

and thus,

$$p(x) = \frac{\alpha - 1}{x_{min}}\left(\frac{x}{x_{min}}\right)^{-\alpha}. \tag{21}$$

Distributions that do not follow a power-law over their entire range might still be normalized independently of the value of α. However, power-law distribution with $\alpha \leq 1$ do not occur normally in nature, if ever.

The first two moments will be derived now and it will be clear that all moments can be derived with a simple generalization. Let X be a power-law distributed random variable $\in (x_{min}, \infty)$. The mean is then given by,

$$E[X] = \int_{x_{min}}^{\infty} xp(x)dx = C\int_{x_{min}}^{\infty} x^{-\alpha+1}dx = \frac{C}{2-\alpha}\left[x^{-\alpha+2}\right]_{x_{min}}^{\infty}. \tag{22}$$

Note that if $\alpha < 2$ the mean does not have a finite value. Indeed, if we calculate the mean of all the observation we will always get a finite value. But in case the sample is infinite, then the mean would diverge. The actual meaning of the divergence of the mean consists in the extreme volatility of the variable distributed as a power-law. The mean is not well defined as the values can vary enormously from one measurement to another. For $\alpha > 2$ the mean is however well defined and equal to,

$$E[X] = \frac{\alpha-1}{\alpha-2}x_{min}. \tag{23}$$

The second moment is,

$$E[X^2] = \frac{C}{3-\alpha}\left[x^{-\alpha+3}\right]_{x_{min}}^{\infty}, \tag{24}$$

and diverges for $\alpha < 3$, meaning that no finite mean square nor variance exist. So, the same argument for the divergence of the mean holds. Otherwise, the second moment is well defined and is equal to,

$$E[X^2] = \frac{\alpha-1}{\alpha-3}x_{min}^2. \tag{25}$$

The result can be generalized as follows,

$$E[X^m] = \frac{\alpha-1}{\alpha-1-m}x_{min}^m. \tag{26}$$

Last, the *Pareto principle* will be presented, also referred as 80/20 rule. The principle states that 80% of the effects are caused by 20% of the causes. The principle was named in Pareto's honor who, during the investigation of wealth distribution in Italy, was the first to state that 80% of the wealth in Italy belonged to 20% of the people. For this reason, in this paragraph, $p(x)$ will be referred to as the wealth distribution.

The principle will now be generalized. First the fraction of wealth belonging to the two halves of the population is found. For any power-law with $\alpha > 1$ the median is well defined. There is then a point that divides the distribution in two equal parts. The point is given by,

$$\int_{x_{1/2}}^{\infty} p(x)dx = \frac{1}{2}\int_{x_{min}}^{\infty} p(x)dx, \tag{27}$$

which leads to,

$$x_{1/2} = 2^{1/(\alpha-1)} x_{min}. \tag{28}$$

The richest half posses a fraction of wealth equal to,

$$\frac{\int_{x_{1/2}}^{\infty} xp(x)dx}{\int_{x_{min}}^{\infty} xp(x)dx} = \left(\frac{x_{1/2}}{x_{min}}\right)^{-\alpha+2} = 2^{-\left(\frac{\alpha-2}{\alpha-1}\right)}, \tag{29}$$

which converges provided that $\alpha > 2$.

Thus, the amount of wealth detained by the richest half depends only on the exponent α that characterizes the power-law distribution. If $\alpha = 2.1$ the richest 50% detains $\simeq 94\%$ of the wealth.

More generally, the fraction of the population whose wealth exceeds x is given by the combining (18) and (20),

$$P(x) = \int_{x}^{\infty} p(x)dx = \frac{C}{\alpha-1} x^{-\alpha+1} = \left(\frac{x}{x_{min}}\right)^{-\alpha+1}, \tag{30}$$

then, the portion of wealth belonging to those people is,

$$W = \frac{\int_{x}^{\infty} p(x)dx}{\int_{x_{min}}^{\infty} p(x)dx} = \left(\frac{x}{x_{min}}\right)^{-\alpha+2}. \tag{31}$$

Assuming $\alpha > 2$ and substituting (30) into (31), the wealth in the richest P of the population is,

$$W = P^{(\alpha-2)/(\alpha-1)}. \tag{32}$$

2.3 Thermodynamic Quantifiers for Socioeconomic Phenomena

We now want to have a look at how humans have built, and keep building, their economic system. That is, how societies have been using the resources from earth, with high potential energy, to build an ordered state for themselves, resulting in a reduction of potential energy and an increase of entropy in the system.

Let us first give a look at how population has been growing and evolving around the world. Data from *Angus Maddison* researches and from the *Maddison Project* are referred to, and are considered as one of the most reliable sources of historic data about economic statistics.

Around 0 A. D. the average life expectancy was about 24 years, and still in 1820 A.D. it did not increase significantly, as it was about 26 years. However, from the nineteenth century until today, life expectancy has grown significantly, and is now an average of 66 years world wide, with a maximum average of 77-81 years in developed countries and a minimum average of 52 years in Africa. Even if in the mid-nineteenth century life had

already extended significantly, half of the population was estimated to die before the 45th year of age.

The increase in life expectancy has not only let humans to live longer, but also allowed populations to grow significantly. From 0 A.D., population on earth has grown from 231 million people to 603 millions at the beginning of the eighteen century. Then, the total population grew significantly to 1040 millions in 1820 A.D., 2527 millions in 1950 A.D., 7162 millions in 2013 A.D., and is forecasted to become 9551 by 2050 A.D.

The first large growth in population happened in Europe, followed by the Americas and Oceania. Then, growth moved to Asia and Latin America. And finally, future growth is predicted to be driven by Africa. These changes are driven by economic features. Populations grow significantly during industrial transformations, then stabilizes when their economies become service oriented and manufacturing is moved in less developed countries that demand lower wages.

During the agricultural and industrial revolution in the eighteen century, man learned to use wind and water for agriculture; animals and boats for longer movements and transportations. Physical work by humans was still heavy and inevitable, and time for non-essential pursuits was scarce.

In the nineteenth century, scientific advances led to innovative agricultural techniques and a mechanization of work. But most importantly, carbon started being employed as a powerful energy source for the production of trains and boats and for their functioning through steam engines.

The eighteen and the nineteen centuries are marked by inventions that change the way humans live. Means of transportation and communication, electricity, vaccines and medicinals are some of discoveries that happened in this period and that contributed to not only the survival of human beings, but to enhance their ability to live their lives. Life expectancy grew and time for non-essential pursuits arose. People became able to accumulate surplus for savings or reinvestments to stimulate the growth process.

The twentieth century, even if it sees two devastating global wars, is characterized by an even further growth and technological advancement. Oil and gas are the new, and higher in potential energy, sources of energy employed on massive scale. Internet, superconductors and artificial hearts are some of the type of inventions that characterize this period.

Now western countries are considered developed and they moved to a service oriented economy, moving large parts of the manufacturing chain in countries that are still in development, where wage cost supported by the industry for the same output is lower. Recently this process has accelerated. Korea and BRICS countries (Brazil, Russia, India, China, South Africa) have acquired the same knowledge of western countries and are moving their manufacturing centers abroad. According to estimates, global GDP has grown by a factor of 73 from 1820 to 2008.

Of course, this growth has been sustained by a continuous growth in energy consumption, almost everything we rely on in our daily life needs energy to work. This energy is mainly originated from fossil fuels' consumption, resulting in emission of CO_2 and other

greenhouse gasses.

An econophysicist might see in this picture of human development a strikingly similar concept with the one of entropy. As humans build a more ordered state for themselves with a negative change of entropy, a more disordered state is delivered to the system, resulting in an overall net positive entropy variation.

We now want to set the roots for the mathematics behind a thermodynamic framework of economics. Economics is about the circulation of goods and money and the flow of value. However, the process through which economics associates value to each object might not be obvious. This is why we present here the properties of value and show an ideal economic cycle.

We first need to define the difference between *productive content* k and *economic value*. The productive content of an abject is given by the characteristics that an object possesses, such as color, weight and shape; it allows the object itself to be defined univocally and to be different from all other objects. The economic value is instead the price agreed by economic agents to make an exchange; for the same object it can be different depending on the circumstances of the exchange, such as geographic region or levels in market demand. But for the same object, the productive content does not vary, unless the object is consumed or used up during an industrial process. In fact, the productive content does vary at each stage of production with a corresponding reduction in potential energy and release of entropy in the system.

The productive content k is similar to the Boltzmann constant: it is just something that exist, it is impossible to define it. Economists argue that in economics it is impossible to assert the value of an object univocally, as it depends on people; it is not deterministic as in the sciences. It is also impossible to associate an economic value to each object with a productive content comparing it to a fixed body of reference: this body of reference does not exist, and even if there was one, it would be impossible to fix values as they vary with respect to each other.

In the past, exchange values were determined by *barter*, that is by equating values between different objects. Then, money was introduced, coupled with a physical standard such as gold for the dollar. Nowadays, following the collapse of the Bretton Woods' agreements, currencies' values are free to fluctuate on the markets and are regulated by institutions that determine interest rates and supply of money. So, given the stability of value granted by these institutions, economists define both the productive content k of money and its exchange value to be 1, of any currency (1 dollar, 1 pound, 1 euro, ...), while the market determines what you can buy with it.

A second problem concern the evaluation of the productive content of human labour. It is determined by many factors: age, skills, motivation. Economists get around the problem defining, also in this case, the productive content of human labour as *one person*. The cost of labour is then associated with the wage rate per unit of time per person. Thus, cost of labour is again defined subjectively by other humans, relatively to purchase power and as variation in productive content, brought to an object, by the work engaged by the worker.

In general, economics associates any object with a productive content k of *1* unit; then the markets determines the exchange value. However the real source of the economic value derives from the productive content itself and the processes through which it is made and consumed, both in terms of inanimate and living resources.

Economic accounting measures economic value in terms of: net value added, type of expenditure it represents, income to humans. Resource consumption is defined by humans as money exchanged. At each stage of production, input is used to produce output, from raw materials to the final product; at each step productive content is constantly refined and waste discarded, but humans are paid and economic value grows. Therefore, we can say that humans make a profit out of building a more ordered state for themselves at the expense, in term of entropy gain, of the universe system which increasingly become more disordered. One of the reasons because this happens is that humans assume ownership over resources and do not have to respond for losses of value or entropy increase to Nature.

An economic-thermodynamic cycle can be represented as follows. Resources are taken from the environment with high potential energy and low entropy and are combined with human labour and capital stock to produce output. Each time some input go through the *economic engine*, in addition to production of output, waste is discarded into the environment. Now, the output produced, can be exchanged for money. This money are under the form of profit and wages on the producer side and expenditures and consumptions on the buyer side. The output produced can then be ready for consumers' consumption or can become new input for further production.

At each stage the input produced is characterized by low entropy and high potential energy, and can be used for consumption or become further input, though for a limited number of cycles before being consumed; on the other hand, waste is characterized by low potential energy and high entropy, so even if discarded back into the environment, it cannot be used as input.

Paul Samuelson noted, during his Nobel lecture, that in the physical world the pressure of a gas can be raised by compressing its volume or by absorbing some energy from a source at a higher temperature, likewise in economics price of a unit of volume output can increase or decrease according to supply and demand.

As between economics and thermodynamics there are many similarities, it is also fundamental to outline the differences. First, volume in science indicates a collection of molecules of a gas in a closed space, flowing or non flowing, such as a balloon or a pipeline. In economics, volume stands for the amount of product per unit of time that flow into or out of a stock.

We can set an economic exchange whose functioning can be described by the same relation (1) used for ideal gasses. N is the number of units or *carriers of value* with an associated and independent productive content k. Given V the volume that goes through the stock per unit of time, two humans can exchange different carriers of value at a chosen price P according to an *Index of trading value* T. However, as already stated, economics

define k to be 1. So the relationship (1) simplifies to:

$$PV = NT, \tag{33}$$

with N being the number of stock units being *turned over* by index T to become input (or output) value flow PV. Note that, even if T is equated to the velocity of circulation of a currency, it can also be referred to any item of exchange.

We should not forget that, the notion of price is associated to a stock of money, which has productive content equal to its price, equal to 1.

If value flow PV varies on one side, then it must be able to vary also on the other side of Equation (33). Especially when the stock number N is fixed, then T must be able to embody both changes in P and V.

Two differences between economics and sciences should be emphasized. First, in sciences, volume flow of gasses is said to act in a tridimensional manner. In economics, volume flow is defined as item flow, without a spatial volume, and is said to act at a point. The second is that, for flow systems, V and N are bonded by the throughput flow per unit of time: both sides of equation must be balanced. For non-flow systems, V refers to the volume which contains the gas that can expand or contract while N is the fixed number of molecules enclosed in the volume.

Economics however has both characteristics of flow and non-flow systems. In this case, V on one side is associated with volume flow; on the other side N does not change, but the concept of flow is transferred to the index of trading value T, which describes the velocity of circulation of the relative item and is also related to the price P.

We can make a comparison between the unit of measure, respectively, in thermodynamics and in economics: P can indicate pressure or price per unit, V volume or units of output per unit of time, N number of molecules or number of carriers of value in a stock, T temperature or index of trading value and k the Boltzmann constant or productive content. Time is balanced out in both cases.

Furthermore, while productive content flowing through a stock does not change, the productive content that goes through a process of production/consumption does change, with an additional undefined efficiency loss.

2.3.1 The First Law of Thermodynamics: An Economic Perspective

Shifting to an economic perspective, we cannot say that energy and value are the same, they are different concepts, but it seems that they obey to similar principles. Recalling the First Law of Thermodynamics exposed in Section 2.2, let us now imagine a stock, with internal energy U, consisting in N units of a good with productive content k equal to 1. At one end, an input work value flow G feeds the stock per unit of time of the good. At the other end, a similar output work value flow G of the good comes out. Additionally, to let the system work, an *entropic value* Q must feed the stock.

Three events can impact on the stock. The first concerns a variation in *Work Value* G per unit of time, defined as a change in *volume flow* V,

$$dG = PdV. \tag{34}$$

The second, occurs when the amount of Q feeding the system changes. Q represents an external factor such as consumer preferences or abundance of a particular good, that can affect the way the work value flow G act on the system.

The third, can occur when there is a change in the economic value U of the stock, which we call *Internal Value*. The internal value is different from the stock productive content. The internal value varies according to the trading index T, that is the amount of units entering and leaving the stock per unit of time. On the other hand, since k is set to 1, the stock productive content is fixed by N.

Moreover, U will also be a function of the relative time spent in stock, compared to the flow of economic activity. This proportion will be called *value capacity or lifetime coefficient* ω. *Industrial stocks have high* ω, as their volume throughput is high compared to their size. Other stocks have lower ω, e.g. money are turned over several times a year while bonds or human capital usually last longer. An alternative measure of the lifetime coefficient is the *rate of return*,

$$r = \frac{1}{\omega}. \tag{35}$$

For example, a stock having 10 units, of which 1 joins every year and 1 leaves every year, has a rate of return of 10% and value capacity of 10, which is also equal equal to the ratio of time spent in stock, i.e. ten years, and the time base of economic activity, i. e. one year. The change in stock internal value is therefore,

$$dU = N\omega dT. \tag{36}$$

The equivalent of the value capacity in thermodynamics is the *specific heat capacity*, that is the amount of energy that one unit of mass of an element needs to raise its temperature by one degree K. The specific heat capacity varies significantly upon the element.

The economic concept of internal energy deals with the changes in volume flow of productive content and entropic valued added or taken away. This concept is explained with the help of the following example. A trader dealing with a stock with productive content $k = 1$, trades with a work value G a volume V at a price P, giving a turnover $G = PV$, which satisfies also the relationship involving the trading index: $PV = NT$. In a good year, an increase of entropic value Q due to an increase in demand might allow the trader to charge higher prices and increase his work value flow G. In such case, the internal energy U is perceived to increase as well, along with T, even if the item has not changed. The opposite for Q decreasing holds. So, the internal value U is a function of the index of trading value T, and T is a measure of both the volume speed and the price at which economic stocks are being turned over.

An economic definition for the First Law of Thermodynamics in (2) might be the following: *"Entropic Value introduced (or taken away) is equal to the change in work value flow rate, plus any change in the internal value of the stock".*

The First Law of Thermodynamics says nothing about the efficiency of a cycle. It just states that work cannot be done without a source of heat, and so that a perpetual machine cannot exist. Similarly in our economic system, it is impossible to have an increase in work output without an increase in supply from additional resources. So it seems possible that an economic system obeys to the First Law, as it works in cycles, recalling the idea of conservation of energy. What the First Law does not mention is the irreversibility of the process, where an input, to be converted in output, always discards some waste in system, which cannot be reused, undermining the reversibility of the cycle.

2.3.2 The Second Law of Thermodynamics: An Economic Perspective

The Second Law of Thermodynamics states that an amount of heat must always be rejected during a thermodynamic cycle: *It is impossible to construct a system that will operate in a cycle, extract heat from a reservoir, and do an equivalent amount of work on the surroundings.*

The easiest example is that of a fossil-fired power station. The temperature of the flue gas is much higher than the temperature released to the surroundings, entailing a loss of energy in the system. Another example is the loss of energy along the power lines used to transport electricity. At each stage of production/transportation, bits of energy Q and work value G are passed along and some heat Q is discarded and lost in the process. We can say that the electricity that we use daily is highly ordered.

As already noted in Section 2.2, in thermodynamics, entropy regards the amount of energy in a cycle that cannot be used to do work. In statistical mechanics, it is a measure of the probability that a system would be in a certain state; which is usually referred to as a measure of disorder.

In economics terms, we can say that *it is impossible to construct an economic system which will operate in a cycle, extract productive content from a resource reservoir and do an equivalent amount of work, in terms of manufacture and productive content.* Economic systems are highly inefficient as they involve large production of waste and a significant level of irreversibility.

However, it seems like economics does not obey to the Second Law since no subtraction in financial accounts is made to take into account losses in efficiency and productive content. Any individual pays for what he or she consumes, with no concern about the energy lost in the process, that is, for the increase of entropy. We can then say that in economics an idea of reversibility is assumed and Equation (3) can be written,

$$dS = dQ/T_{rev} \tag{37}$$

By combining Equation (37) with the relationships for the First Law for a polytropic

system with a single unit of stock, we obtain

$$dS = (\omega - \omega n + 1)\frac{dV}{V_{rev}} \tag{38}$$

which describes the entropy variation as change in volume throughput per unit of time and as a function of the lifetime coefficient and of the elastic index.

When integrated Equation (38) looks similar to Equation (12):

$$S = (\omega - \omega n + 1)\ln V + C, \tag{39}$$

with C a constant of integration, and the factor $(\omega - \omega n + 1)$ is called *Marginal Entropic Index*.

recalling (35), we can substitute and obtain,

$$S = \left(1 + \frac{1-n}{r}\right)\ln V + C \tag{40}$$

Two things can be noted. When the Marginal Entropic Index is zero, then the entropy generation is zero, which means that the equilibrium is maintained. Then, even if the volume flow rate V does not change, an entropy variation can still appear due to a change in prices P or in the index of trading value T.

Finally we can note that entropy variation in (38) is logarithmic in nature, which means that changes dS can be viewed as per cent changes in the volume flow rate, price or index of trading value - corrected by the Marginal Entropic Index - and with k equal to 1.

2.3.3 Income and Wealth

Global productive content contribution is assumed to be attributed to humankind and measured by GDP per capita. Relationship (33) can be adjusted substituting the index of trading value T with income rate w and allowing N to represent the total population, to obtain:

$$PV = Nw. \tag{41}$$

This relationship is a mean value among all people considered in N, there is then a individual distribution according to individual contribution. Equation (41) can be regarded as the kinetic value that measures the amount of value circulating between people.

It is well known in economics that income has a skewed distribution. As mean income rises, the curve broadens as income distribution becomes more uniform and moves right. To describe this relationship, economists use different distributions, with the log-normal the most common, but others include *Gamma* and *Maxwell-Boltzmann* distributions.

Similar skewed distribution is found in analyzing kinetic energies of gasses' particles, which, as discussed above, are distributed following a Maxwell-Boltzmann distribution about a mean energy value kT.

2.3.4 Elasticity

In a thermodynamic system, pressure and volume are not always inversely proportional, but are related through an index n of compression or expansion. This gives rise to the following *polytropic* relationship:

$$PV^n = C, \tag{42}$$

where C is a constant. Characteristic of polytropic equations is that they can be adapted to fit any situation. Any relationship can be obtained with (42). If $n = 0$, a constant price situation is found. If $n = \infty$ a constant volume flow position is obtained.

Similar relationship fits also for economics, where price and volume are not always simply inversely proportional. Furthermore, combining (42) with (33) and setting $N = 1$, we can let Equation 42 cover the index of trading value T too,

$$T = CV^{1-n}. \tag{43}$$

2.3.5 Utility

In utility theory, each consumer associates an amount of utility to each good, which depends on personal interests, and given a budget constraint, decides how to allocate its budget in order to maximize its utility. A basic rule of utility theory, is that as a consumer demands more of the same good, the utility gained by each incremental unit diminishes. This behavior is referred to as diminishing marginal utility and is formalized as following:

$$\frac{(\delta Y/\delta V)_a}{P_a} = \frac{(\delta Y/\delta V)_b}{P_b} = ... = \frac{(\delta Y/\delta V)_n}{P_n} \tag{44}$$

where $(\delta Y/\delta V)_i$ is the ratio of the marginal utility per volume variation of item i and P_i is the price of item i. This relationship states that the variation in marginal utility of a good, with respect to its variation in volume and its price, is equal to that of another good. Consumer equilibrium in Equation (44), if reached, is never maintained.

An example of diminishing marginal utility is the following. A human or an animal, might feel hungry and desperate if they have not eaten for long time. But once they find some food, their benefit from eating it decreases as they sate their hunger; they will not keep eating forever. In modern societies, there is a large difference between utility preferences. Advanced societies demand almost entirely unnecessary goods while poor one still starve for daily survival.

We can say that a good we are interested in, which might have gone through many production stages, has a low level of entropy and a high level of order of productive content. But as the good is consumed, a part of productive content is assimilated by the consumer, satisfying a portion of his utility; the rest is discarded contributing to entropy rise. Thus, we now try to find a link between entropy and utility.

A consumer who buys a good, passes from a moment where his level of entropy was high before the purchase, to one where the entropy level has decreased at the moment of

the purchase. In the act of purchasing, the consumer has gained some *negentropy* because he has increased order for himself. However, to purchase the good, an amount of money must have left his stock, resulting in an overall entropy variation very small; but large enough for the purchase to make sense. After the purchase however, as the good is being consumed, new entropy would gradually be created. Also in the case the consumer did not purchase the item, entropy creation would have happen through other consumptions or loss of value of the money stock.

Utility can then be defined as *potential economic entropy*. Utility resides at the beginning of ownership and entropy is generated as the purchased item goes through time.

A thermodynamic process must be able to describe variations in price and utility independently from the underlying productive content. The entropic value Q which incorporates changes in demand, new money and other features, can be used to describe how individual utility is influenced.

For a single consumer, we can set his stock to be made by one unit $N = 1$. Thus, Equation (38) links the marginal entropy change for one unit of product, to the marginal entropy change of the consumer for that product.

Further, his wage w can be related to the total value flow PV that he can purchase:

$$w = PV \tag{45}$$

Then, $w = 1$ can be set and one can imagine that its budget can be spent on one product only. Thus, inverting for V Equation (45) and substituting into Equation (38), we obtain,

$$\left(\frac{dS/dV}{P}\right) = (\omega - \omega n + 1). \tag{46}$$

It is straightforward to see the similarity of the left terms of (46) with Equation (44) for diminishing marginal utility. Additionally, the relationship could further be simplified if we assume the elastic constant n to be 1, resulting in the price P being inversely proportional to volume V, independently on the the value of the lifetime coefficient ω.

We can then say that entropy and utility are related to each other, with the entropy proceeding in the opposite direction, thus called *negentropy*. Moreover, entropy is additive, so summing all contributions from different consumers together can bring to a measure of social good.

This is not an equilibrium process, but one forever seeking it. A thermodynamic representation of the economic system is therefore of non-equilibrium and obeys to *Le Châtelier* principle, which states that, *if a change occurs in one of the factors under which a system is in equilibrium, then the system will tend to adjust itself so as to annul the effects of that change*. It does not mean the system will reach it, but that it will continually try to reach the equilibrium. Therefore, differently from economics, where the constraint is associated with the budget or wage, a thermodynamic representation can take into account a multiple of relevant factors, allowing the constraint to be variable as well.

2.4 Production and Consumption

Human needs are met by a combination of production and consumption, measured in monetary terms by GDP. However, this value does not describe the dynamics of the economic system.

The traditional economic framework that describes the production dynamics is the theory of the firm which relies on production and marginal productivity functions and relates inputs and outputs of a process. An important production function is the *Cobb-Douglas*'s that follows,

$$V_O = Ae^{\lambda t}(N_K)^{\alpha}(N_L)^{\beta}, \tag{47}$$

where, V_O is the output volume flow, A represents usage of stocks per unit of time, N_K and N_L capital and labour stocks, α and β the elastic coefficients, and the exponential function is the *Hicks-neutral* technical progress function and represents the impact of technical progress over time.

In Equation (47) does not appear any reference to possible short-time shortages or constraints. Economics, gives for granted the supply of energy and consequently α and β are a measure of the marginal productivities of capital and labour, measured as shares of capital and labour costs, while all other factors are measured through the technical progress. however, the $Cobb - Douglas$ production function assumes supply of resources and forgets about productive content consumption and entropy generation.

In the short-term, resources can be divided in renewable with an extended lifetime, or non-renewable with short lifetime, i.e. that needs continuous substitution. For example, human contribution lasts for long time, and is renewed with the birth and education process. On the other hand, fuels usually have a very short lifetime. Resources operates also at different activity levels during their lifetime. Ultimately, in the long-term all resources must be substituted.

In this section an alternative framework, typical of chemical reactions, is proposed to describe the production system. A system works according to a *fixed* process, that is a process that needs an exact combination of resources in input to yield a final product as output. Input volume flows of energy and resources arise from stocks or from external sources. Thus, to produce one unit of output O, the production function needs α units of capital stock K, β units of labour L and δ units of resources R. The production process works then similarly to a chemical reaction and can be written:

$$\alpha K + \beta L + \delta R \rightleftharpoons 1O. \tag{48}$$

There are however some key differences. In economics, reactions only proceeds *forward*. Then, reactants always need some treatment to be used for producing output. Consequently, an amount of waste with low productive content and high entropy value must be discarded with a resulting increase in entropy in the system. The consumer will therefore pay according to the process occurred and for the efficiency losses in the chain.

To account for efficiency losses, production waste will be indicated by D. Then Equation (48) can be corrected into,

$$\alpha K + \beta L + \delta R \rightleftharpoons 1O + \rho D. \tag{49}$$

We can then represent the creation of an ordered product with low entropy and the generation of high entropy waste through the combination/consumption of resources. Ultimately, when the final output is consumed as well, a further increase in entropy is generated.

The thermodynamic approach is based on *Le Châtelier* principle. The system does not reach an equilibrium, but continually seek to proceed to such state. We can see that, from the thermodynamic perspective, we operate in a non-equilibrium system, while from the macroeconomic accounts perspective, that is, in terms of money, we operate in an equilibrium system. In fact, the equality symbol is used in Equation (47), while the reversible chemical reaction symbol is used in (48) and (49). We then need to identify the source of disequilibrium.

Let us redefine the *volume flows*, separating them into *active* component referred as a and an *inactive* one referred as $(1 - a)$, respectively indicating the part of the stock that are actively contributing to the output and the part that is not being used. In chemistry, the concept of *activity* is common; it indicates the effective concentration of a species in a mixture. In an economic context, a is variable. It depends on the portion of resources used to fuel productivity, but refers also to stock constraints due to stock renewability or eventual shortages.

We could represent the proportion of active stock as $N_a = a \cdot N$ and the portion of inactive stock as $N_i = (1 - a) \cdot N$. As a consequence, the volume flow can be indicated as,

$$V = v \cdot N_a = v \cdot a \cdot N, \tag{50}$$

with v the *volume flow rate per unit of time*.

For example, the output flow on the right side of (49) contributed by active labour, is offset by potential output that could be generated by inactive labour, which therefore does not contribute to generate value for the system but rather absorbs it. Likewise, it be would preferable to have a production system that is always fully active rather than one interrupted by down-times. Another example is the case of debt used to finance production flows. If interest rates increase a portion of the active stock of money must be transferred to interest payments, decreasing activity levels.

Additionally, although it does not appear in national accounts, resource stocks exhibit property of *utilization*. Flow in production and consumption are variable and subject to levels of demand, depletion, human intervention, and could also face saturation or inactivity of the stock itself, with consequential obsolescence and depreciation.

We now look at volume input flow on the left side of (49) as reactants of capital, such as labour and resources, which are fed in the production process to deliver *steady* output flows defined through *activity levels*, that indicates the proportion of each input stock that actively contributes to the production process and output flow.

In addition, an amount of *motive force of anticipated benefits* can be injected to further increase the activity levels of the production flow.

However, the rise in activity levels is not consequential. The maximum level that can be reached depends upon inactive availabilities and ultimate output demand. For example, an input factor that was already fully active cannot contribute to further production flow and eventually become a *constraint*.

To explain the impact of motive force on production flows the concept of *Free Energy* is recalled. The concept was pioneered by *Josiah Willard Gibbs* and *Herman von Helmholtz*. It expresses the available amount or free energy that can be used up in a reaction to equilibrium. In economics we apply the variant, that is used in chemistry, to describe the *chemical potential* μ,

$$\mu = \mu^{std} - NkT \ln V, \tag{51}$$

expressed as a logarithmic function of the volume and a constant, called *standard state value* μ^{std}, measured empirically.

In the economic representation, the free value of an economic component can be associated to both volume flow or activity rate. Since in economics the factor k becomes 1, we can write,

$$\mu = \mu^{std} - NT \ln V \tag{52}$$

$$\mu = \mu^{std} - NT \ln a \tag{53}$$

with N the stock number and T the index of trading value.

Instead of using energies, we refer to the *free values* of each reactant, both in input and output. Considering the Euler exponents set out in the Cobb-Douglas production function, to recognize that a fixed amount of input factors are needed to make a product, the *net free value available* is then equal to

$$1[\mu]_O + \rho[\mu]_D - \alpha[\mu]_K - \beta[\mu]_L - \delta[\mu]_R \tag{54}$$

The equation describes the parts of *inactive* inputs that become active when they combine together to form product.

Two features should be noted. First, in economics the components melt together through a common denominator: money. Second, change in free value $\Delta\mu$ can also be expressed as negative change in entropy $-T * \Delta S$; for example, in a forward reaction or in a growing economy, free value is consumed to fuel growth, with a resulting increase in entropy.

Another expression that can be written is the following,

$$\left(\frac{V_2}{V_1}\right) = e^{S_2 - S_1} \left(\frac{V_2}{V_1}\right)_K^{\alpha} \left(\frac{V_2}{V_1}\right)_L^{\beta} \left(\frac{V_2}{V_1}\right)_R^{\delta} \left(\frac{V_2}{V_1}\right)_D^{-\rho}, \tag{55}$$

$$\left(\frac{V_2}{V_1}\right) = e^{S_2 - S_1} \left(\frac{a_2}{a_1}\right)_K^{\alpha} \left(\frac{a_2}{a_1}\right)_L^{\beta} \left(\frac{a_2}{a_1}\right)_R^{\delta} \left(\frac{a_2}{a_1}\right)_D^{-\rho}. \tag{56}$$

On the right side of Equations (55) and (56) changes in output volume flows are expressed on the left as a function of an exponential term, accounting for entropy change, and a series of factors accounting for changes in production flows. In Equation (55) changes are expressed in term of volume flows; while in Equation (56) changes are expressed in term of activity levels.

So, change in demand to consume more is compensated by changes in production factors corrected by an entropy factor related to the whole system. A perfectly efficient production function would allow a change in entropy equal to zero; in other words, an input combination that is fully transformed in output without production of waste would be possible. But we know that this would violate the Second Law of Thermodynamics.

Even if Equations (55), (56) have similarities with the Cobb-Douglas function, there are some key differences. If a rise in the system entropy is met by a change in input and output volumes, such that the entropy generated in the system is null, then the system entropy has been transferred to the system relating input and output consumption and waste production. In Equation (49), a technical progress function is not required, as this is incorporated in changes in the mixture of inputs and production flows. Finally, no assumption is made on forward growth or decline; any position is a continual dynamic feedback situation and equilibrium is just a temporary moment.

We can say that in an economic system, change in active output volume demanded is represented by a change in the system entropy function and a change of flow of input factors available to combine together. A restriction on any input factor influences the whole system and might become a constraint.

The production process outlined is dynamic in the sense that we have built an economic system that links multiple parts in an economic chain, emphasizing the interacting nature of the system, where the free energy used by a component of the system is replaced by that of another, resulting in an overall increase of entropy. Moreover, if a particular constraint has an effect on a component of the economic chain, it might happen to cause a reaction on other parts of the chain.

The thermodynamic nature of the economic perspective outlines that output is a complex non-equilibrium function, influenced by many dynamically interacting stocks and flows, with an equilibrium position that is itself continuously varying to seek entropy maximization states, according to eventual constraints.

In general terms, we can write an entropy function, recalling the Boltzmann entropy, which describes entropy as a logarithmic function of economic activity V and the level of constraints X present in the system:

$$S = \ln \frac{V}{X}. \tag{57}$$

We will now use an example to show the growth of entropy in our economic system. Imagine a man living outside civilized society. He wants to build an house from local material to improve his well-being and protect from natural events. Using resources and

his physical effort, the man builds a more ordered state for himself. However, without effort to maintain the house, it will fall into despair and eventually fall down and lose its value, resulting in an entropy increase.

During the production process, assembling resources with his effort, the man has created *order* and caused an entropy reduction. On the other hand, unusable waste has bee created, resulting in an increase of entropy. Furthermore, through the life of the house, other entropy is generated as the house is being *consumed* by the man and the environment. If the man eventually take care of the house or builds another one, the process would restart from beginning, with a consequent perpetual increase in entropy.

In modern societies, humans have assembled together and made technological advancements that allowed to transfer efforts to other resources at his own benefit, becoming an overseer of the process. The discovery of new sources of energy has allowed humans to accelerate the rate of production and consumption, even resulting in what we call extreme consumerism and planned obsolescence. At this stage, entropy production rate has increased enormously, but is still dependent on supply of energy and resources.

We then see that both economic and thermodynamic systems favors conditions of maximum entropy production. This principle causes the pursue of position where activity rates of inactive inputs can be easily and with minimum cost increased, e.g. investments in economic structures that allow higher output yields or bypass eventual constraints and pursuit of short-term project rather than sustainable long-term ones.

2.5 Money

We have set until now an economic system in terms of flows and stocks, giving an idea of how large can the interconnections be and thus the complexity degree of the whole system. Money is central for this system to work. It can be seen as a common stock used as a mean to exchange value between different flows. Money is therefore considered as a unique large stock accounting for all production flows, but in the opposite direction.

A monetary equivalent of value, equal to price times volume flow $P \cdot V$, flows in and out of the stock, in the opposite direction of value itself. The stock of money is made of N units, with productive content k equal to 1 of any currency, and it is turned around by the index of trading value T, also called *Velocity of Circulation*, according to (33).

In the classic theory of money, the quantity N of money is taken as identity measure. Here, since the framework we are building is an attempt at demonstrating that, like anything else in Nature, also economics follows the rule of thermodynamics, it is reasonable to think that money too behave accordingly to these rules, and then a general entropy function can be given.

A problem that arises in economics, is that while k in thermodynamics is fixed, and we fixed its value also in our economic system, the dilutive effect of inflation must change all other parameters. In other words, reflection of change in the real value k of money appear as a change in the factors of (33), where N represents the stock of money.

Moreover, if also the amount of money N is fixed, then, again, the index of trading value T must be able to reflect changes in both the volume flow V and prices P.

Another difference between the thermodynamic and economic system resides in the number of unknowns. In a *non-flow* thermodynamic system, k represents the Boltzmann constant, N the number of molecules, P the pressure, V the volume and T the temperature; k and N are fixed, while P, V and T are variable, resulting in three unknowns. In a *flow* system, V and N become the volume flow and the molecule flow respectively. N is now variable as well, resulting in four unknowns. However, volume flow V and molecular flow N can be summarized with the specific volume $v = V/N$, reducing the number of variables for flow systems to three variables, as in non-flow systems.

In money systems however all factors are variable. Even the productive content k, which is actually a nominal value, can vary under extensive inflation. Further, it is not possible to substitute V and N with a ratio of the two. Dividing a volume output by a depreciating money stock number might compromise the definition of volume at constant prices. It is also preferable to preserve the volume flow measure as it is, since it is used to describe the size of an economy.

The risk is that change in money stock N can find a way of influencing all other variables. One, imperfect, solution consists in dividing the output price level P by the number of units of monetary stock N, obtaining a *specific price* $P_N = P/N$. In this way we leave undisturbed output volume flow V and velocity of circulation T, although the nature of price level P and the variable size of stock N is changed. The new relation is the following,

$$P_N V = T, \tag{58}$$

where both k and N are set to 1. By taking logs and differentiating, we obtain,

$$\frac{dP_N}{P_N} + \frac{dV}{V} = \frac{dT}{T}. \tag{59}$$

The velocity of circulation of money can then be influenced by both, a change in the specific price P_N or in the output volume flow V. However, if the two factors on the left side balance each other, the velocity of circulation can be unchanged.

The reasons for this variations might be found looking at the elastic index n as a function of time. By taking the logarithm, differentiating and rearranging a polytropic expression of the kind $P_N V^n = C$, we find that,

$$n = -\left(\frac{dP_N}{P_N}\right) \Big/ \left(\frac{dV}{V}\right) \tag{60}$$

It is now possible to lay out a general entropy generation function in the case of money. Recalling Equation (38) and the case where we substitute the lifetime coefficient ω with the rate of return $1/r$, the entropy generation is related to three factors: (i) the elastic index n, (ii) the growth rate of volume flow $\frac{dV}{V}$, and (iii) the lifetime coefficient ω, or

its inverse, the rate of return r. For money, the rate of return r can be considered the long-term average level of the velocity of circulation.

The presence of constraints for volume flow was discussed in the previous sections. Now the focus is shifted on constraints acting on money. Interest rates and central banks overall monetary base are some of these. In the thermodynamic perspective, changes in monetary base N impact on Specific Price P_N. If the monetary base increases, specific price decreases, since the two are inversely related. However, this is likely to lead to a rise in prices P, which corresponds to an inflation phenomenon due to printing of money, balancing back the specific price.

In the previous section, concept of active and inactive stocks led to a relation between change in entropy generation, output volume flow and constraints acting on output flows. Similarly, we study the forces acting on money, which continually flows into and out of the economy. Here, a thermodynamic system for economics seems to operate with a polytropic relationship between specific price, output volume and velocity of circulation, with variations in the elastic index affecting the relationship. Additionally, the link between entropy and utility is extended to include money utility as well.

A general rule, in macroeconomics, is that money demand is positively related to income and output, and it is negatively related to interest rates. A rise in interest rates usually constitutes a restraining force on an economy, while a decrease, an expansionary force. We say that interest rates are negatively related to money entropy generation; when entropy increases, and so does money demand, price inflation must be balanced by interest rates. Interests represent then a form of constraint on money flow and a *negative entropy change*.

Money balance exists primarily to facilitate flow of value G in an economy. However, money are often deposited on bank accounts, where they earn interests. A rise in interest rates, would potentially decrease the stock of active money in the economy: when interest rates increase, who borrows money, must pay larger interests, resulting in a decrease in investments. Therefore, it is useful to find a *cumulative interest index*, indicated by I, which accounts for interest compound cumulations along time:

$$I_t = I_0 \cdot (1 + i_1) \cdot (1 + i_2) \cdot ... \cdot (1 + i_t), \tag{61}$$

with I_0 an initial rate. Since money balance exists to help value flow G, and an increase in interest rates diminishes the amount of money circulating, the index of interest rate subtracts from the potential value flow it could have generated in an economy. It can be stated that, in the *long-term*, value flow G and interest index I proceeds in tandem as in,

$$G = I^{\theta} \tag{62}$$

with θ close to 1.

Proceeding similarly to section 5, where we stated that entropy changes are related to differences in rates changes of volume flows, we can state for money that entropy changes

depend on the difference between the rates of change of output value flow G and money interest index I,

$$dS = \left(\frac{dG}{G} - \frac{dI}{I} \right). \tag{63}$$

3 Data

In this chapter, the different types of artificially generated series used throughout the work are illustrated. The analysis has been performed over different mathematical and financial processes. Among the mathematical processes, the following are studied: Fractional and Geometric Brownian Motion (FBM and GBM); Cox-Ingersoll-Ross (CIR); Hull-White-Vasicek (HWV). Then, the following financial processes are investigated: Autoregressive Fractionally Integrated Moving Average (ARFIMA); Generalized Autoregressive Conditional Hetereskedasticity (GARCH). Furhermore, the main features of some financial assets that are referred to in this work are presented. Last, the sampling method used to arrange data into comparable sequences is outlined. However, first of all the most important properties and definition of elementary statistics and time series are recalled.

3.1 Stochastic Properties of Random Data

3.1.1 Expectation and Variance

Given a random variable X, the first moment, or *expectation* is equal to,

$$E[X] = \int_{-\infty}^{\infty} x f_x(x) dx. \tag{64}$$

The second moment, or *variance* is equal to,

$$E[X^2] = Var[X] = E[X - \mu]^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f_x(x) dx. \tag{65}$$

As we will see later, the variance can be indicated as γ_0.

3.1.2 Covariance, Autocovariance and Autocorrelation

In statistics, the covariance refers to a measure of dependability of two variables with each other; a measure of how two variables vary together. Covariance is equal to the expectation of the product of the distances between the variables and their respective mean,

$$E[X, Y] = Cov[X, Y] = E[(X - E[X])(Y - E[Y])] \tag{66}$$

If we substitute the random variable Y with X we obtain the variance seen above. Moreover, for a time series, we can define the *autocovariance*, which is a measure of how a variable and its lagged value vary together. In other words, is a measure of time dependability of a variable with it self.

If the same stochastic process, in t and $t+j$ is equal to, respectively, $E[X_t] = \mu + \epsilon_t$ and $E[X_{t+1}] = \mu + \epsilon_{t+j}$. The autocovariance is then given by

$$\begin{aligned} E[X_t, X_{t+j}] &= E[(X_t - E[X_t])(X_{t+j} - [E_{t+j}])] \\ &= E[(\mu + \epsilon_t - \mu)(\mu + \epsilon_{t+j} - \mu)] \\ &= E[\epsilon_t \epsilon_{t+j}] \\ &= 0. \end{aligned} \tag{67}$$

and is indicated as γ_j.

Then, the *j-th autocorrelation* ρ_j is defined as the j-th autocovariance divided by the variance:

$$\rho_j = Corr(Y_t, Y_{t-j}) = \frac{Cov(Y_t, Y_{t-j})}{\sqrt{Var(Y_t)}\sqrt{Var(Y_{t-j})}} = \frac{\gamma_j}{\sqrt{\gamma_0}\sqrt{\gamma_0}} = \frac{\gamma_j}{\gamma_0} \tag{68}$$

with $|\rho_j| < 1$ for all j.

3.1.3 Stationarity

For a stochastic process X_t, if the mean and the autocovariances do not depend on time, the process is said to be *covariance-stationary* or simply *stationary*:

$$\begin{aligned} E(X_t) &= \mu, && \text{for any t} \\ E(X_t - \mu)(X_{t-j} - \mu) &= \gamma_j, && \text{for any t and j} \end{aligned}$$

Notice that if a process is covariance stationary, the covariance between X_t and X_{t-j} depends only on the lag j and not on the date t.

3.1.4 Ergodicity

A covariance-stationary process is said to be *ergodic for the mean* when the following conditions are satisfied. The time average of the process converges in probability to its mean,

$$\lim_{t\to\infty} \frac{1}{T} \sum_{t=1}^{T} x_t = \lim_{t\to\infty} \bar{x} = E(X_t), \tag{69}$$

and provided that the autocovariance γ_j are *absolutely sumable*,

$$\sum_{j=0}^{\infty} |\gamma_j| < \infty. \tag{70}$$

3.1.5 Long memory

Finally, models (i) and (ii) have in common the long range dependence, which means that observations widely separated in time present significant dependence. One definition of long range dependence or *long memory* is defined by specifying an hyperbolic autocorrelation function ρ_j,

$$\rho_j \sim j^{2d-1} L(j) \text{ as } \mathrm{j} \to \infty,\ 0 < \mathrm{d} < 0.5. \tag{71}$$

with d the memory parameter and $L(j)$ a slowly varying function.

3.2 Artificial Data Series

3.2.1 Fractional Brownian Motion

The *Fractional Brownian Motion* is a long memory processes first introduced by Mandelbrot and Van Ness (1968) referring to it as a *self-similar* process. A stochastic process X_t, with $t \in \mathbb{R}$, is said to be self-similar if there exist $H > 0$ such that for any *scaling factor* $c > 0$,

$$X_{ct} \overset{\mathcal{L}}{=} c^H X_t, \tag{72}$$

with H the Hurst exponent and ($\overset{\mathcal{L}}{=}$) equivalence in distribution. Self-similar processes are stochastic models where a scaling in time is equivalent, *in term of distribution*, to an appropriate scaling in space. Moreover, for any k, the distribution of $(X_{t_1+c} - X_{t_1+c-1}, ..., X_{t_k+c} - X_{t_k+c-1})$ does not depend on c, the X_t is said to be self-similar with *stationary increments*.

So, a gaussian process B_t^H is called a *fractional brownian motion*, if it satisfies,

1. B_t^H is self-similar with $0 < H < 1$,
2. B_t^H has stationary increments.

When $H = 0.5$ we obtain a simple Brownian Motion with independent increments. When $0 < H < 0.5$ the *Fractional Brownian Motion* is said to be anti-persistent, which means that increments tend to be opposite signed. Conversely, when $0.5 < H < 1$ the it is said to be persistent, which means that increments tend to be equally signed.

3.2.2 Geometric Brownian Motion

The Geometric Brownian Motion is used as a basis for the *Black-Scholes-Merton* model used to price options,

$$dX_t = \mu(t) X_t dt + D(t, X_t)\sigma(t) dW_t, \tag{73}$$

$\mu(t)$ indicates the level of return, $D(t, X_t)\sigma(t)$ the volatility and dW_t is a simple Wiener process. Volatility is deterministic and constant and there are no jumps. Increments are independent on previous states.

3.2.3 Cox-Ingersoll-Ross

The Cox-Ingersoll-Ross is a model with ergodic and stationary distribution, used to simulate interest rates:

$$dX_t = s(t)[\mu(t) - X_t]dt + D(t, X_t^{1/2})\sigma(t)dW_t. \tag{74}$$

The drift factor $s(t)[\mu(t) - X_t]$ assures *mean reversion* of the process, $s(t)$ represents the speed of convergence to the mean level of return, $\mu(t)$ indicates the level of return, $\sigma(t)$ the volatility, and dW_t is a simple Wiener process. The process is mean reverting, that means the process tends to its mean when the it goes to infinite. Another feature of the Cox-Ingersoll-Ross model is that it never touches the zero level. When the process comes closer to zero, the drift factor becomes dominant as the standard deviation, represented by the term $D(t, X_t^{1/2})\sigma(t)$, becomes very small by definition. As a consequence, the effect of further shocks in interests is limited.

3.2.4 Hull-White-Vasicek

Hull-White and Vasicek models are two different specification of processes used to model interest rates. In the former, parameters vary with time deterministically, in the latter they are fixed:

$$dX_t = s(t)[\mu(t) - X_t]dt + \sigma(t)dW_t. \tag{75}$$

The drift factor of the process is $s(t)[\mu(t) - X_t]$, $s(t)$ represents the speed of convergence to the mean level of return, $\mu(t)$ indicates the level of return, $\sigma(t)$ the volatility, and dW_t is a simple Wiener process.

3.2.5 White Noise

The following definition of *white noise* is used in the financial models presented below and therefore it is here introduced. A sequence $\{\epsilon_t\}_{t=-\infty}^{\infty}$ whose elements have mean zero, variance σ^2 and for which the elements of ϵ are uncorrelated with each other across time,

$$E(\epsilon_t) = 0 \tag{76}$$

$$E(\epsilon_t^2) = \sigma^2 \tag{77}$$

$$E(\epsilon_t \epsilon_\tau) = 0 \ \text{ for } t \neq \tau \tag{78}$$

Furthermore, if we substitute (78) with a stronger condition that ϵ's are independent with each other and if we say that ϵ follows a normal distribution:

$$\epsilon_t, \ \epsilon_\tau \ \text{ independent for any } t, \tau \tag{79}$$

$$\epsilon_t \sim N(0, \sigma^2), \tag{80}$$

then we have a *Gaussian white noise process.*

3.2.6 Autoregressive Fractionally Integrated Moving Average

The model of an *autoregressive fractionally integrated moving average* process (*ARFIMA*) of a time series of order (p,d,q) with mean μ, may be written, using the lag operator L, as:

$$\Phi(L)(1-L)^d(y_t - \mu) = \Theta(L)\epsilon_t, \tag{81}$$

with ϵ_t *i.i.d.* and $\sim (0, \sigma_\epsilon^2)$

The autoregressive part of the process is represented by the factor

$$\Phi(L) = 1 - \phi_1 L - ... - \phi_p L^p,$$

where the lag operator of order p shifts the value of y_t back to p observations, so that one obtains:

$$\Phi(L)y_t = (1 - \phi_1 L - ... - \phi_p L^p)y_t = y_t - \phi_1 y_{t-1} - ... - \phi_p y_{t-p}.$$

Analogously, the moving average part of the process is represented by the factor

$$\Theta(L)\epsilon_t = (1 + \theta_1 L + ... + \theta_q L^q)\epsilon_t = \epsilon_t + \theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q}.$$

Finally, $(1-L)^d$ is the fractionally differencing operator defined by:

$$(1-L)^d = \sum_{n=0}^{\infty} \frac{\Gamma(k-d)L^k}{\Gamma(-d)\Gamma(k+1)}.$$

For $d < 1/2$ the ARFIMA process is said to exhibit long memory. So, when $d > 1/2$ one can represent the process in an alternative form that still maintain the invertibility property as:

$$(1-L)^{d-1}(1-L)\Phi(L)(y - \mu) = \Theta(L)\epsilon_t.$$

3.2.7 Generalized Autoregressive Conditional Heteroskedasticity

The third model is the *autoregressive conditional heteroskedastic. ARCH(q)* models are used to generate processes with stochastic volatility. The reason why this model was introduced is the following.

We say that ϵ_t follows an ARCH(q) when its conditional variance is defined as,

$$\sigma_t^2 = \omega + \alpha_1 \epsilon_{t-1}^2 + ... + \alpha_q \epsilon_{t-q}^2, \tag{82}$$

with $\omega > 0$, $\alpha_i \geq 0, i = 1, ..., p$, $\sigma_t = \epsilon_t / z_t$ and $z_t \sim N(0, 1)$.

As we said before, ϵ_t is white noise but it is not an independent process. The unconditional variance of the process is defined as $\sigma^2 = \omega/(1 - \sum_i \alpha_i)$ with $\sum_i \alpha_i < 1$ to hold for covariance-stationarity of the process.

When ϵ_t is an ARCH(q) process, the conditional variance $\epsilon^2 = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \eta_t$, with $\eta_i = \sigma_t^2(z_t^2 - 1)$, is a AR(q) process.

The model used in this work is the *Generalized Autoregressive Conditional Heteroskedasticity* or *GARCH(p,q)*, that takes into account also p lags of conditional variances:

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2, \tag{83}$$

with $\omega > 0$, $\alpha_i \geq 0$ for $i = 1, ..., q$, $\beta_i \geq 0, i = 1, ..., p$, $\sigma_t = \epsilon_t / z_t$ and $z_t \sim N(0, 1)$.

In this model the unconditional variance is given by $\sigma^2 = \omega/(1 - \sum_i^q \alpha_i - \sum_j^p \beta_j)$. If ϵ_t follows a GARCH process, then ϵ_t^2 follows an ARMA process with conditional heteroskedastic terms η_t.

3.3 Market Data

The objective of this thesis is to compare financial and mathematical models to financial markets by means of a cluster entropy approach. Results for financial markets where developed in Ponta and Carbone (2018) and Ponta and Carbone (2019). This section recalls major features of the data used those works.

The cluster entropy approach was applied to data of some European assets, namely BUND, BOBL, DAX, Euro Currency, Euro Stoxx and FIB30 to obtain a quantification of cross-market heterogeneity and the weights of the optimal portfolio, and to some of the most important USA's indexes, namely the S&P500, NASDAQ and Dow Jones Industrial Average, to study the sensitivity of the method to prices and provide a measure of dynamics. Results are reported only for NASDAQ and Dow Jones Industrial Average, while results for S&P500 can be found in Appendix A.7. Data were downloaded from `www.bloomberg.com/professional`.

Note that while NASDAQ is an index resulting from all the public firms quoted on the market, the DJIA and S&P500 are indexes representative of a selected number of public firms. For each index, data include tick-by-tick prices from January 2018 to December 2018. Main information about the data series are reported in Table 1.

3.4 Sampling

To study the dynamics of financial series different time horizons are compared, but so that the comparison is possible, series must be of the same lengths. It is therefore necessary to propose a sampling method that makes the series comparable but that do not discards useful information contained in the series.

Raw data are provided under the form of *monthly series*, i.e. tick-by-tick series from the first tick in a month to the last one. Note that lengths of the monthly series might vary: months have different number of trading days and the number of transactions per unit of time is not regular.

The following method is then applied. The series are cumulatively summed to obtain twelve *cumulative series*. The first cumulative series ranges from the first transaction of January 2018 to the last one of January 2018; the second ranges from the first transaction in January 2018 to the last of February 2018; continuing in this fashion, the twelfth ranges from the first transaction in January 2018 to the last of December 2018, which is equivalent to the whole year. Because each cumulative series ranges from the first tick of 2018 to the last tick of the relative month, the twelve series have very different lengths.

To sample such cumulative series, twelve *sampling frequencies*, i.e. twelve integers indicating for each series the number of samples to skip for every number reported, must be found. Sampling frequencies are obtained dividing the length of each cumulative series by the length of the shortest and then rounding to the closest integers. Thence, each cumulative series is sampled with the relative sampling frequency to yield a *sampled series*: for each sample in the sampled series, a number of samples equal to the sampling frequency has been discarded in the *cumulative series*.

The sampled series obtained are *approximately* of equal lengths. To obtain twelve series of *exactly* equal length the observations exceeding the length of the shortest series are cut off.

The result consists in twelve sampled series that are equal in length but that represent different time horizons.

To generate comparable artificial series a similar method is used. We took as reference the lengths of NASDAQ cumulative series reported in Table 1. Then, a series of length equal to that of NASDAQ in 2018 (i.e. the length of the twelfth cumulative series) was generated. Thence, we divided the series in the respective *cumulative series* according to the lengths obtained for NASDAQ cumulative series reported in Table 1. Then the sampling method proceeds analogously from the calculation of the sampling frequency.

Such sampling method was applied to series generated by artificial financial models to make sure that the information content would be comparable to that of real-world financial series.

Month	S&P500	NASDAQ	DJIA
Jan	516635	586866	516644
Feb	984046	1117840	984101
Mar	1500662	1704706	1500764
Apr	2017282	2291572	1623779
May	2558504	2906384	2165044
Jun	3075125	3493250	2681708
Jul	3580946	4069315	3187571
Aug	4146769	4712062	3753440
Sep	4614188	5243031	4220776
Oct	5180010	5885785	4786628
Nov	5685831	6461851	5292492
Dec	6142450	6982025	5749152

Table 1: Cumulative indexes' lengths

4 Methods

In this section the methods used throughout the work are recalled. The cluster entropy approach developed in Carbone (2013) and in Ponta and Carbone (2018) is now recalled along with some of the results for the sake of clarity.

4.1 Moving Average Cluster Entropy

As noted before Shannon wanted to measure the amount of information embedded in a message to separate the corresponding sequence into the the portion that is actually carrying the relevant information and the portion that is not necessary to reproduce the initial message.

The Shannon functional is written as,

$$S(\tau, n) = \sum P(\tau, n) \log P(\tau, n), \tag{84}$$

where $P(\tau, n)$ is a probability distribution associated with the time series $y(t)$.

To obtain such a probability distribution it is necessary to partition the continuous phase space into disjoints sets. The method traditionally adopted divides the sequence into segments of equal lengths. In this work, however, a method developed recently in Carbone (2013) is used.

A time series $y(t)$, such as those in Figures 1.a and 1.b, is partitioned in *clusters* by the intersection with its moving average $\tilde{y}_n(t)$, with n the size of the moving average. The simplest type of moving average is defined at each t as the average of the n past observation from t to $t - n + 1$,

$$\tilde{y}_n(t) = \frac{1}{n} \sum_{k=0}^{n-1} y(t-k). \tag{85}$$

Note that while the original series is defined from 1 to m, the moving average series is defined from 1 to $m - n + 1$ because n samples are necessary to initialize the series. The original series and the moving average series are indicated as $\{y(t)\}_{t=1}^{m}$ and $\{\tilde{y}_n(t)\}_{t=1}^{m-n+1}$ respectively. Consecutive intersections of the time series and of the moving average series yield a partition of the phase space into a series of *clusters*. Each cluster is defined as the portion of the time series $y(t)$ between two consecutive intersection of $y(t)$ itself and its moving average $\tilde{y}_n(t)$ and has length (or duration) equal to:

$$\tau_j \equiv ||t_j - t_{j-1}||, \tag{86}$$

where t_{j-1} and t_j refers to two subsequent intersections of $y(t)$ and $\tilde{y}_n(t)$.

In this work a different type of moving average called *Detrending Moving Average (DMA)* algorithm is used. For each time series $y(t)$ and moving average window n (assuming n to be odd) three different types of moving averages are computed: (a) *backward*, (b)

centered and (c) *forward*,

$$\tilde{y}_{n,backward}(t) = \frac{1}{n}\sum_{k=0}^{n-1} y(t-k) \tag{87}$$

$$\tilde{y}_{n,centered}(t) = \frac{1}{n}\sum_{k=-(n-1)/2}^{(n-1)/2} y(t-k) \tag{88}$$

$$\tilde{y}_{n,forward}(t) = \frac{1}{n}\sum_{k=-(n-1)}^{0} y(t-k). \tag{89}$$

For each moving average window n the probability distribution function $P(\tau, n)$ of the lengths τ can be obtained by counting the number of clusters $\mathcal{N}_j(\tau_j, n)$ with length τ_j, $j \in \{1, m\}$. The result is,

$$P(\tau, n) \sim \tau^{-D}\mathcal{F}(\tau, n), \tag{90}$$

where $D = 2-H$ and H indicate respectively the fractal dimension and the Hurst exponent of the sequence. The Hurst exponent is widely used to indicate the degree of persistence of long-range correlated clusters. Long-range correlation means that the clusters are organized in a similar way along the time series, even for clusters far away in time from each other, a fact that is defined as *self-similarity*. Note also that, for a given moving average window n, the frequency of a cluster of length τ_j is averaged over the number of moving average types that generated it, i.e. if a cluster of length τ_k is generated by both the backward and centered DMA then the two frequencies are summed and averaged by two.

In Equation (90) the term $\mathcal{F}(\tau, n)$ takes the form,

$$\mathcal{F}(\tau, n) \equiv e^{-\tau/n}, \tag{91}$$

to account for the drop-off of the power-law behavior for $\tau < n$ and the onset of the exponential decay when $\tau \geq n$ due to the finiteness of n. When $n \to 1$ the lengths τ of clusters tend to be centered around a single value. When $n \to m$, that is when n tends to the length of the whole sequence, only one cluster with $\tau = m$ is generated. For middle values of n however a broader range of lengths is obtained and therefore the probability distribution spreads all values.

When the probability distribution in (90) is fed into the Shannon functional in (84) the result is the following,

$$S(\tau, n) = S_0 + \log \tau^D - \log \mathcal{F}(\tau, n), \tag{92}$$

which, after substituting (91), becomes,

$$S(\tau, n) = S_0 + \log \tau^D + \frac{\tau}{n}, \tag{93}$$

where S_0 is a constant, $\log \tau^D$ accounts for power-law correlated clusters related to τ^{-D} and τ/n accounts for exponentially correlated clusters related to the term $\mathcal{F}(\tau, n)$.

The term S_0 can be evaluated in the limit $\tau \sim n \to 1$, which results in $S_0 \to -1$ and $S(\tau, n) \to 0$, that corresponds to the fully deterministic case, where each cluster has size equal to 1. On the other hand, when $\tau \sim n \to N$, the maximum value for the entropy is obtained with $S(\tau, n) = \log N^D$, which corresponds to the case of maximum randomness, where there is one cluster coinciding with the whole series.

Equation (93) shows that power-law correlated clusters, characterized by having length $\tau < n$, are described by a logarithmic term as $\log \tau^D$, and their entropy do not depend on the moving average window n. However, for values of $\tau \geq n$, which represent exponentially correlated clusters, the term τ/n becomes predominant. Cluster entropy increases linearly as τ/n, with slope decreasing as $1/n$. Hence, due to the finite size effects introduced by the partitioning method, in $\tau = n$ the behavior of entropy changes and its values exceeds the curve $\log \tau^D$. In other words, clusters that are power-law correlated does not depend on n, are said to be *ordered* and represent deterministic information. Clusters that are exponentially correlated does depend on n, are said to be *disordered* and represent random clusters.

As stated in the introduction, the meaning of entropy in information theory can be compared with the meaning of entropy in thermodynamics. In an *isolated system*, the entropy increase dS refers to the irreversible processes occurring spontaneously within the system. In an *open system* however a further increase in entropy dS_{ext} occurs due to the irreversible processes spontaneously occurring with the external environment.

The term $\log \tau^D$ should be interpreted as the entropy of the isolated system. It is independent on n, that is it is independent on the partitioning method. It is also of the form of the Boltzmann entropy, that can be written as $S = \log \Omega$, with Ω the volume of the isolated system. Therefore the quantity τ^D corresponds to the volume occupied by the fractional random walker.

On the other hand, the term τ/n represents the excess entropy introduced by the partitioning method and in fact depends on the moving average window n. If same size boxes were chosen, the excess entropy term τ/n would vanish and entropy would reduce to the logarithmic term. When a moving average partition is used, the term τ/n emerges to account for the additional heterogeneity introduced by the randomness of the process. Thence, for exponentially correlated clusters entropy exceeds the limit of the logarithmic term.

Cluster entropy results obtained for price series of (a) NASDAQ and (b) Dow Jones Industrial Average markets are reported in Figures 1.c and 1.d from Ponta and Carbone (2019). Plots of cluster entropy curves represent 18 curves, one for each moving average window n used to partition the series, with $n = 30, 50, 100, 150, 200, 300, ..., 1500$. The time horizon analyzed is indicated as M, i.e. $M4$ indicates that the cumulative series ranging from January 2018 to April 2018 is studied.

4.2 Cumulative Entropy Indexes

One important step is to quantify the property of the entropy series. In order to improve the accuracy of the method one can consider consider the integral of the entropy function over the clusters length τ, a cumulative measure able to embed all information in a single figure:

$$I(n) = \int S(\tau, n) d\tau, \tag{94}$$

which for discrete sets reduces to,

$$I(n) = \sum_{\tau} S(\tau, n). \tag{95}$$

4.2.1 Market Heterogeneity Index

In Ponta and Carbone (2018), the analysis is performed over prices and volatilities series for different European assets, namely BUND, BOBL, DAX, Euro Currency, Euro Stoxx and FIB30 to obtain a quantification of cross-market heterogeneity and the weights of the optimal portfolio.

The authors define (95) as the *Market Heterogeneity Index* and use it to summarize the information for each single market over the clusters lengths. A first integration over τ yields a quantification of cross-market heterogeneity. Over different assets, the authors found that while (95) for the series of the prices is almost invariant, it varies significantly for the series of the volatilities. Furthermore, the authors integrate one more time the *volatility series* over the moving average window n to obtain the weights of the optimal portfolio, in comparison to the Sharp approach.

4.2.2 Market Dynamic Index

On the other hand, in Ponta and Carbone (2019), the index is calculated with the purpose of obtaining the horizon dependence and quantify the market dynamics. It was defined as *Market Dynamic Index*. In Figures 1.e and 1.f results of the Market Dynamic Index are shown only for NASDAQ and Dow Jones Industrial Average series, results on S&P 500 are reported in Appendix A.7. For small values of n, horizon dependence is small, but it increases as n increases too. Horizon dependence is larger for NASDAQ than for Dow Jones Industrial Average.

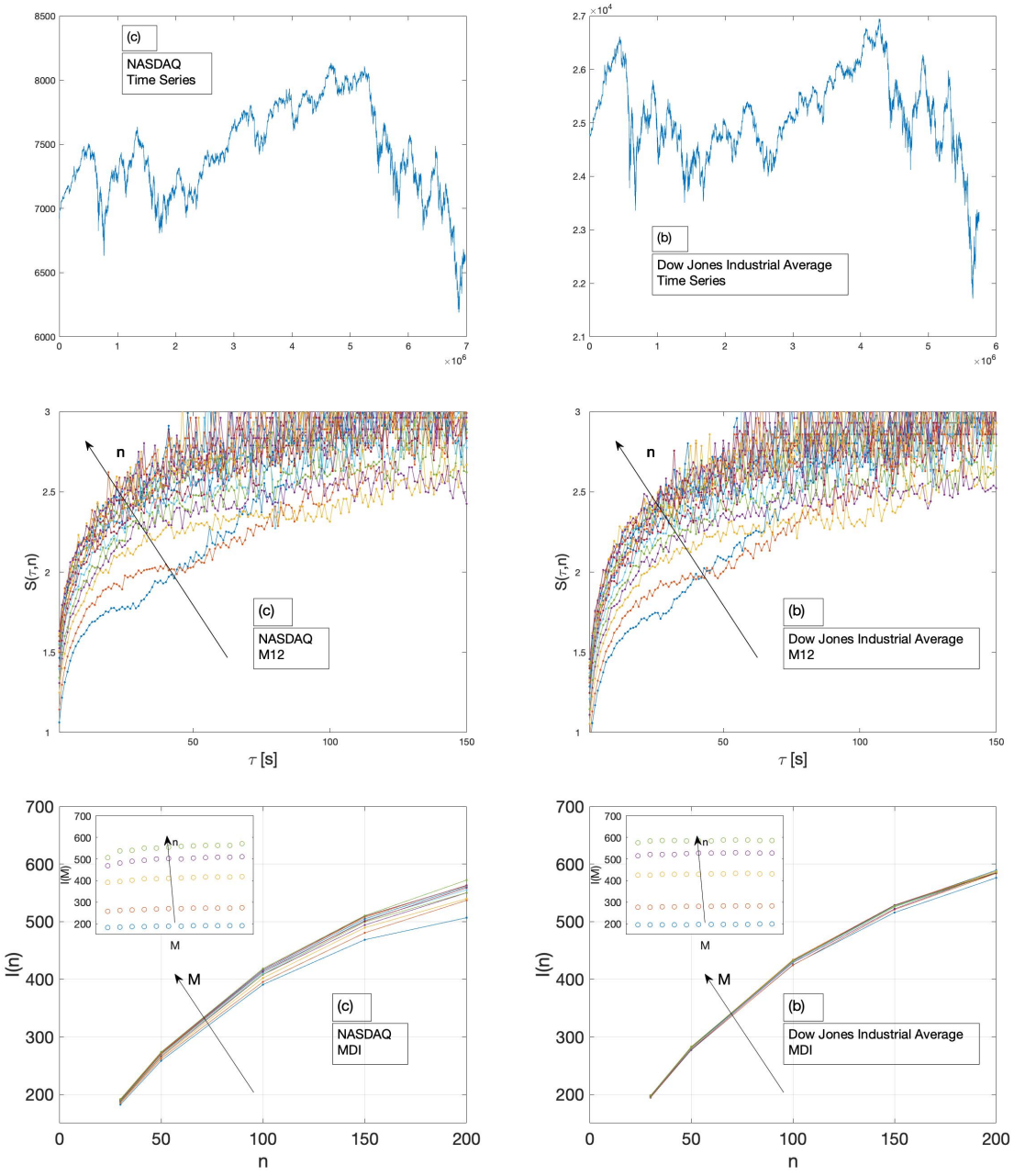


Figure 1: Time series, cluster entropy curves (M12) and Market Dynamic Index for (a) NASDAQ and (b) Dow Jones Industrial Average. Results are reported from Ponta and Carbone (2019)

5 Results

In this section, Cluster Entropy and Market Dynamic Index have been calculated over series obtained by a variety of financial and mathematical models. The artificial models considered are: Geometric and Fractional Brownian Motion (GBM and FBM); Cox-Ingersoll-Ross (CIR); Hull-White-Vasicek (HWV); Autoregressive Fractionally Integrated Moving Average (ARFIMA); Generalized Autoregressive Conditional Heteroskedasticity (GARCH).

The investigation was motivated by the results obtained in real-world financial markets (NASDAQ, DJIA and S&P500). As shown in Chapter 4, cluster entropy in financial markets was proved to be significantly *market* and *horizon* dependent (Ponta and Carbone, 2019).

5.1 Fractional Brownian Motion

According to the classical financial theory, all the information are reflected by market prices and it is not possible to take advantage of past observations to predict future outcomes. Therefore, subsequent price deviations are said to be identically and independently distributed (*iid*). If that were true, correlation would be null and prices would be modelled correctly by a *fractional brownian motion* with Hurst exponent $H = 0.5$, which is equal to a simple *brownian motion*. However, several studies have shown that real world markets do not adhere faithfully to the standard theory of perfectly informed and rational investors.

The fractional brownian motion is first investigated in this thesis. Results of the cluster entropy analysis over fractional brownian motion series with Hurst exponent $H = 0.5$ (see Chapter 3 for a short overview on FBM). The fractional brownian motion were generated by means of the FRACLAB MATLAB tool available at `https://project.inria.fr/fraclab/`. The results are shown in the top panel of Figure 2. In the middle panel of the figure, plots of the entropy cluster at horizon M12 are reported. Cluster Entropy calculated at other time horizon presents a similar behavior. In general, one can expect that power-law correlated clusters follow a smooth logarithmic behavior, then for $\tau \geq n$ the exponential decay sets on and entropy increases linearly with the term τ/n dominating. In the bottom plot, values of the market dynamic index are shown. One can note that they overlap at any moving average window n and time horizon M. At small n, values of $I(n, M)$ are small too. Then, at larger n, when a broader range of clusters' lengths τ is spanned in the power-law distribution, values of $I(n, M)$ increase too.

A number of other fractional brownian motion series generated with Hurst exponent varying in the range $0.1 \leq H \leq 0.9$ were analysed. The results are reported in the Appendix A.1.

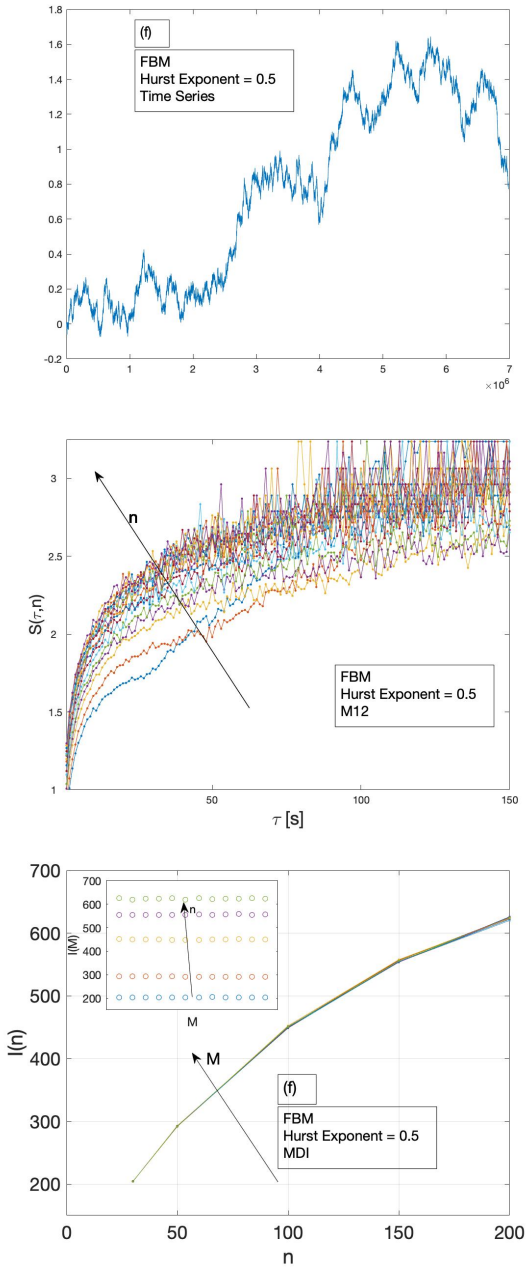


Figure 2: *Time Series* (top), *Cluster Entropy* (middle) and *Market Dynamic Index* (bottom) for series generated by Fractional Brownian Motion with Hurst exponent $H = 0.5$ obtained by means of the FRACLAB tool available at `https://project.inria.fr/fraclab/`.

5.2 Geometric Brownian Motion

The Geometric Brownian Motion (*GBM*) is frequently used to model asset prices (see Chapter 3 for a short overview). Geometric Brownian Motion series are generated by means of the MATLAB tool available at `https://it.mathworks.com/help/finance/gbm.html`.

Cluster entropy results for series generated by means of Geometric Brownian Motion processes, with $r = 1 \cdot 10^{-7}$ and $\sigma = 5 \cdot 10^{-4}$, are shown in Figures 3. Time horizons shown are $M3$, $M6$, $M9$ and $M12$. As one can see cluster entropy curves are quite similar over different time horizons.

In Figure 4.a results for *Market Dynamic Index* are shown. Again, at every moving average n and time horizon M, Market Dynamic Index values overlap.

Another Geometric Brownian Motion series was generated. Details on the analyzed parameters are reported in Table 2 and results obtained for other set of parameters are reported in the Appendix A.2.

GBM		μ	σ
1		0	$5 \cdot 10^{-6}$
2		$1 \cdot 10^{-7}$	$5 \cdot 10^{-4}$
CIR	s	μ	σ
1	$2 \cdot 10^{-7}$	$1 \cdot 10^{-7}$	$5 \cdot 10^{-4}$
2	$1 \cdot 10^{-7}$	$9 \cdot 10^{-7}$	$9 \cdot 10^{-4}$
HWV	s	μ	σ
1	$2 \cdot 10^{-7}$	$6 \cdot 10^{-7}$	$6 \cdot 10^{-4}$
2	$2 \cdot 10^{-6}$	$9 \cdot 10^{-7}$	$6 \cdot 10^{-4}$

Table 2: Parameters choice for Stochastic Differential Equation (SDE) processes. Models reported are the following: Geometric Brownian Motion *GBM*, Cox-Ingersoll-Ross *CIR* and Hull-White-Vasicek *HWV*. Parameter μ indicates the level of return, σ the standard deviation and s the speed of convergence to the mean of the process.

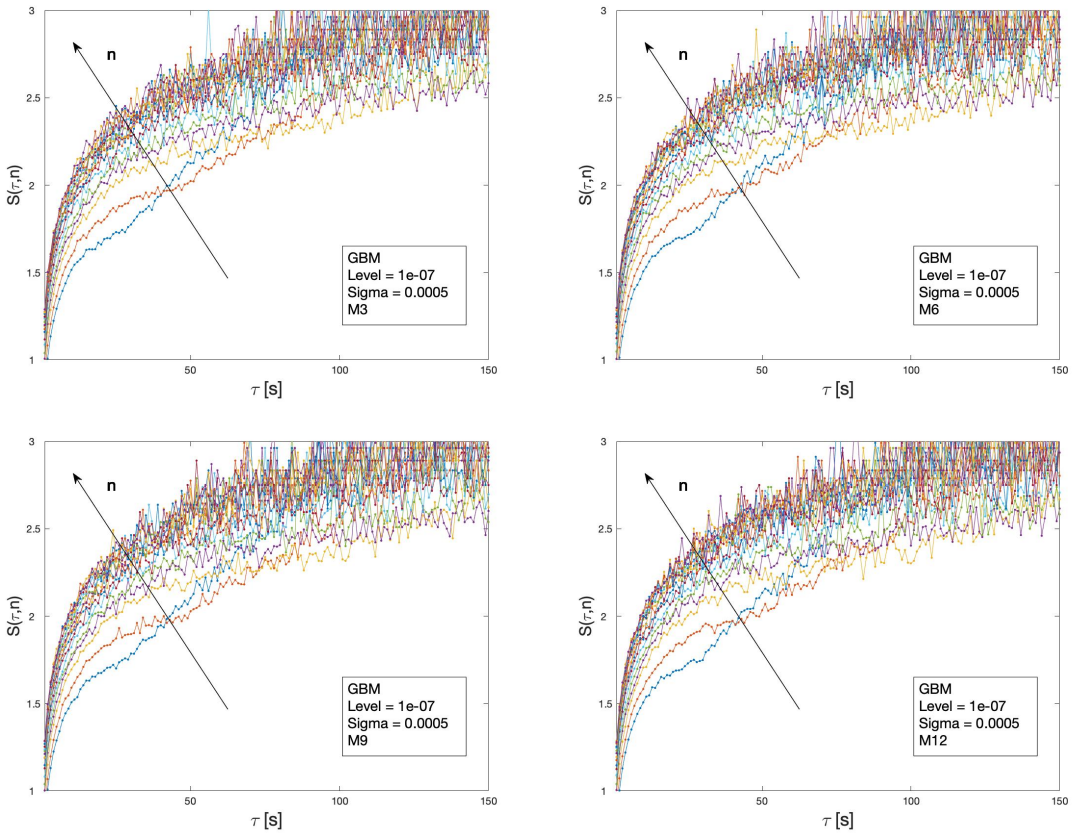


Figure 3: Cluster entropy results for series generated by means of Geometric Brownian Motion process. To generate the series the following parameters are used: $\mu = 1 \cdot 10^{-7}$ and $\sigma = 5 \cdot 10^{-4}$. Labels indicate time horizons M3, M6, M9 and M12.

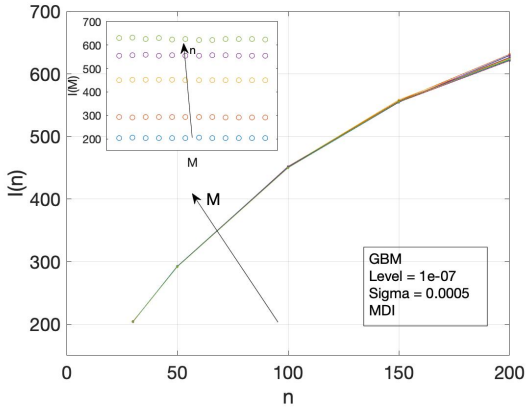


Figure 4: Market Dynamic Index for series generated by means of Geometric Brownian Motion process. To generate the series the following parameters are used: $\mu = 1 \cdot 10^{-7}$ and $\sigma = 5 \cdot 10^{-4}$.

5.3 Cox-Ingersoll-Ross

The Cox-Ingersoll-Ross is a mathematical model used to simulate interest rates (see Chapter 3 for a short overview). Cox-Ingersoll-Ross series are generated by means of the MATLAB tool available at `https://it.mathworks.com/help/finance/cir.html`.

Cluster entropy results for series generated by means of Cox-Ingersoll-Ross processes, with $s = 2 \cdot 10^{-7}$, $\mu = 1 \cdot 10^{-7}$ and $\sigma = 5 \cdot 10^{-4}$, for time horizons $M3$, $M6$, $M9$ and $M12$ are shown in Figure 5. Cluster entropy values do not present significant variation at different time horizons. The logarithmic term of the form $\log \tau^D$ describes power-law correlated clusters, while the linear term τ/n describes exponentially correlated clusters.

Market Dynamic Index results are shown in Figure 6. Curves overlap also in this case at every moving average window n and time horizon M.

Another Cox-Ingersoll-Ross series was generated. Details on the specification analyzed are reported in Table 2 and results obtained for other set of parameters are reported in the Appendix A.3.

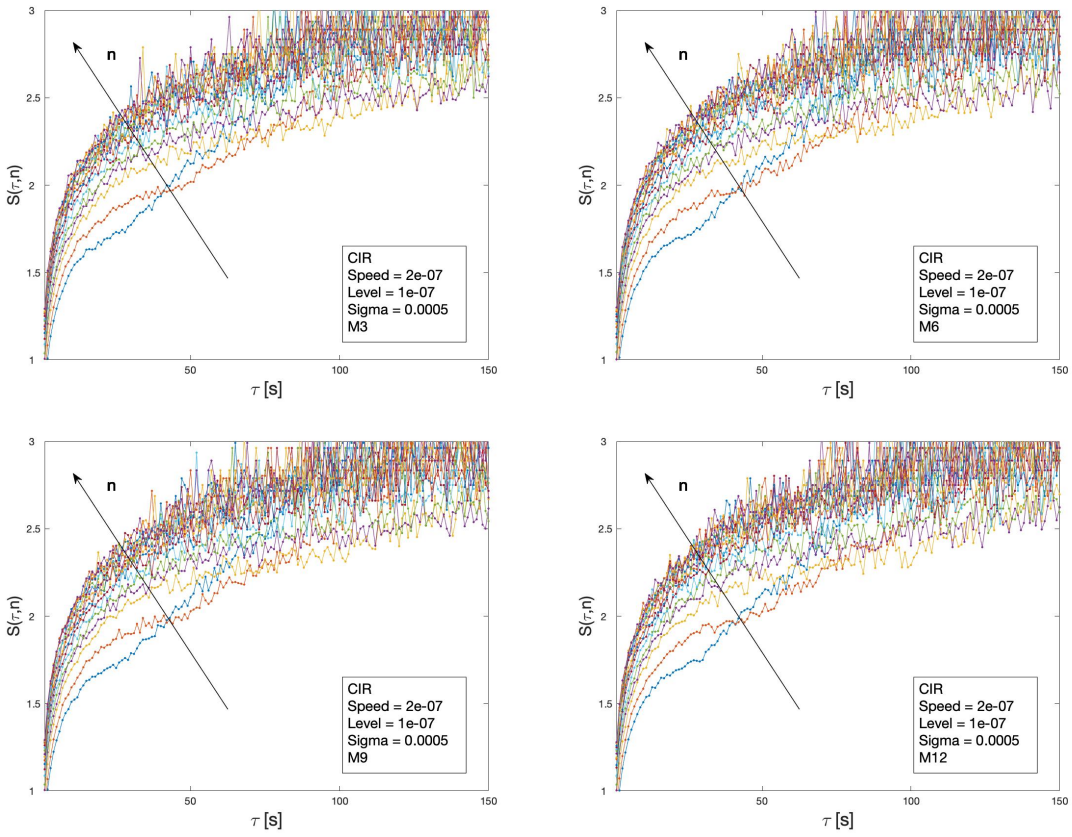


Figure 5: Cluster entropy results for series generated by means of Cox-Ingersoll-Ross process. To generate the series the following parameters are used: *Speed of convergence to the mean* $s = 2 \cdot 10^{-7}$, $\mu = 1 \cdot 10^{-7}$ and $\sigma = 5 \cdot 10^{-4}$. Labels indicate time horizons M3, M6, M9 and M12.

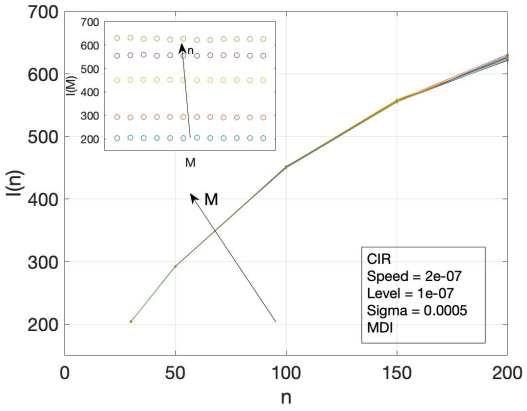


Figure 6: Market Dynamic Index for series generated by means of Cox-Ingersoll-Ross process. To generate the series the following parameters are used: $s = 2 \cdot 10^{-7}$, $\mu = 1 \cdot 10^{-7}$ and $\sigma = 5 \cdot 10^{-4}$.

5.4 Hull-White-Vasicek

The Hull-White-Vasicek is another mathematical model used to simulate interest rates (see Chapter 3 for a short overview). Hull-White-Vasicek series are generated by means of the MATLAB tool available at `https://it.mathworks.com/help/finance/hwv.html`.

Results for the cluster entropy analysis on series generated by means of Hull-White-Vasicek processes with $s = 2 \cdot 10^{-6}$, $\mu = 9 \cdot 10^{-7}$ and $\sigma = 6 \cdot 10^{-4}$, for time horizons $M3$, $M6$, $M9$ and $M12$ are shown in Figure 7. No significant difference can be found between cluster entropy results at different time horizons. Again, a smooth logarithmic behavior and a linear behavior describe respectively power-law and exponentially correlated clusters.

Then, in Figure 8 results for the MDI are shown. Values of $I(n, M)$ are identical for every moving average window n and time horizon M.

Another Hull-White-Vasicek series was generated. Details on the specification analyzed are reported in Table 2 and results obtained for other set of parameters are reported in the Appendix A.4.

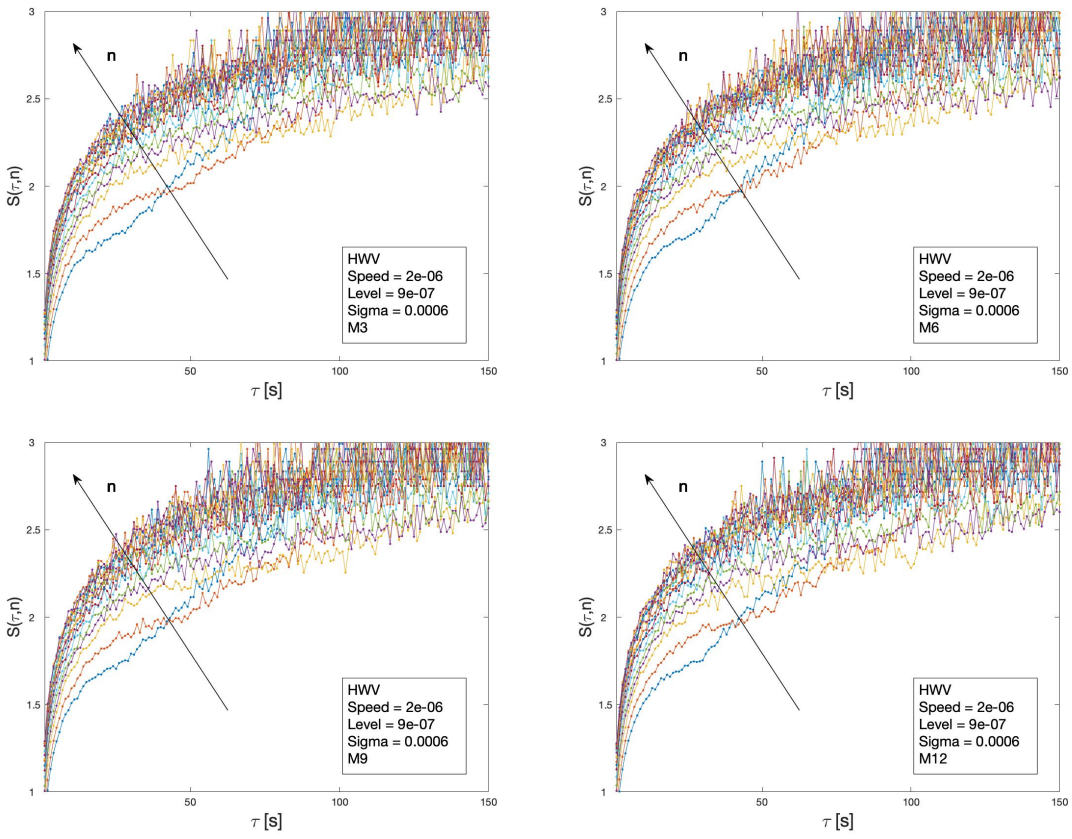


Figure 7: Cluster entropy results for series generated by means of Hull-White/Vasicek process. To generate the series the following parameters are used: $s = 2 \cdot 10^{-6}$, $\mu = 9 \cdot 10^{-7}$ and $\sigma = 6 \cdot 10^{-4}$. Labels indicate time horizons M3, M6, M9 and M12.

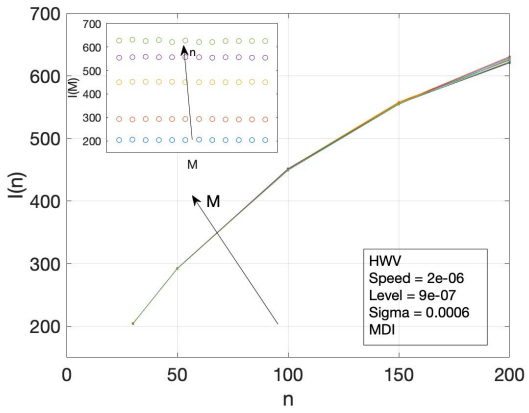


Figure 8: Market Dynamic Index for series generated by means of Hull-White/Vasicek process. To generate the series the following parameters are used: *Speed of convergence to the mean* $s = 2 \cdot 10^{-6}$, $\mu = 9 \cdot 10^{-7}$ and $\sigma = 6 \cdot 10^{-4}$.

5.5 Autoregressive Fractionally Integrated Moving Average

Among financial models, the *autoregressive fractionally integrated moving average* is one of the most common processes used to model prices of long-range correlated assets (see Chapter 3 for a short overview). ARFIMA series are generated by means of the MATLAB tool available at
`https://www.mathworks.com/matlabcentral/fileexchange/25611-arfima-simulations`.

Figures 9 and 10 illustrate cluster entropy curves for series obtained by means of ARFIMA models at different time horizons. In Figure 9 the parameters are $\phi = 0.3$, $d = 0.25$, $\theta = 0.4$. In Figure 10 the parameters are $\phi_1 = 0.4$, $\phi_2 = 0.16$ and $\theta_1 = 0.9$, $\theta_2 = 0.81$, $\theta_3 = 0.729$ and $d = 0.11$. Each plot refers to a different time horizon, namely M3, M6, M9 and M12. In both figures, cluster entropy varies over the time horizons considered. One can note a significant difference in the degree of choppiness of power-law correlated clusters: Figure 9, with a higher $d = 0.25$ is choppier than Figure 10, with $d = 0.11$. Then, exponentially correlated clusters are described by the linear term τ/n.

The Market Dynamic Index is shown in Figure 11. *MDI* curves vary significantly for each time horizon M. Plots 11.a, 11.b refer to the models in Figures 9, 10 respectively; in Figure 11.a, $I(n, M)$ assumes different value for each n and M. On the other hand, in Figure 11.b $I(n, M)$ is almost M-invariant for small values of n, but a dependence on M is found for larger n. The other plots represent results for different combination of parameters. Also in these cases $I(n, M)$ behaves similarly.

Several other combinations of parameters were studied (see Table 3), results obtained for other set of parameters are reported in Appendix A.5.

Model Number	ϕ	θ	d
1	0.3	0.4	0.25
2	0.3	0.85	0.25
3	0.9	0.85	0.25
4	0.3	0.4	0.48
5	0.9	0.4	0.48
6	0.3	0.85	0.48
7	0.9	0.85	0.48
8	0.4, 0.16	0.9, 0.81, 0.729	0.11
9	0.4, 0.16	0.9, 0.81, 0.729	0.3

Table 3: Parameters choice for ARFIMA processes

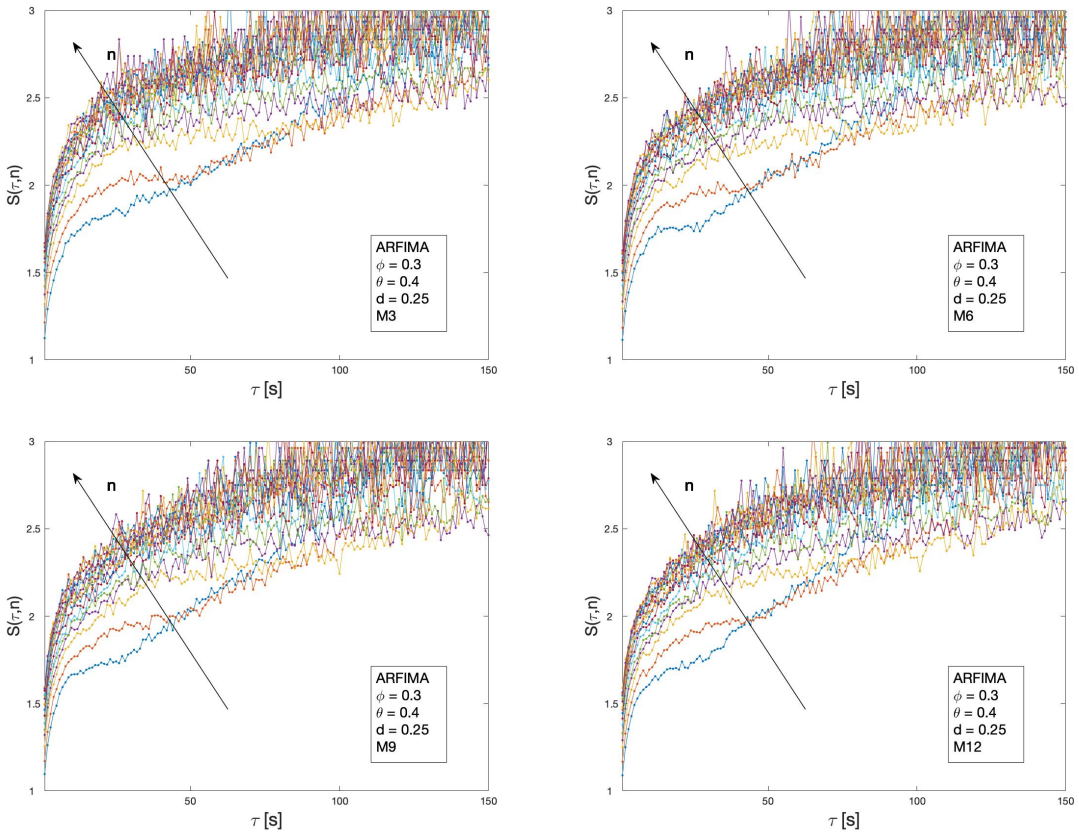


Figure 9: Cluster entropy results for series generated by means of ARFIMA processes. The series used to plot the entropy clusters was generated with the following parameters: $\phi = 0.3$, $d = 0.25$, $\theta = 0.4$. The time horizon analyzed in figures are M3, M6, M9 and M12

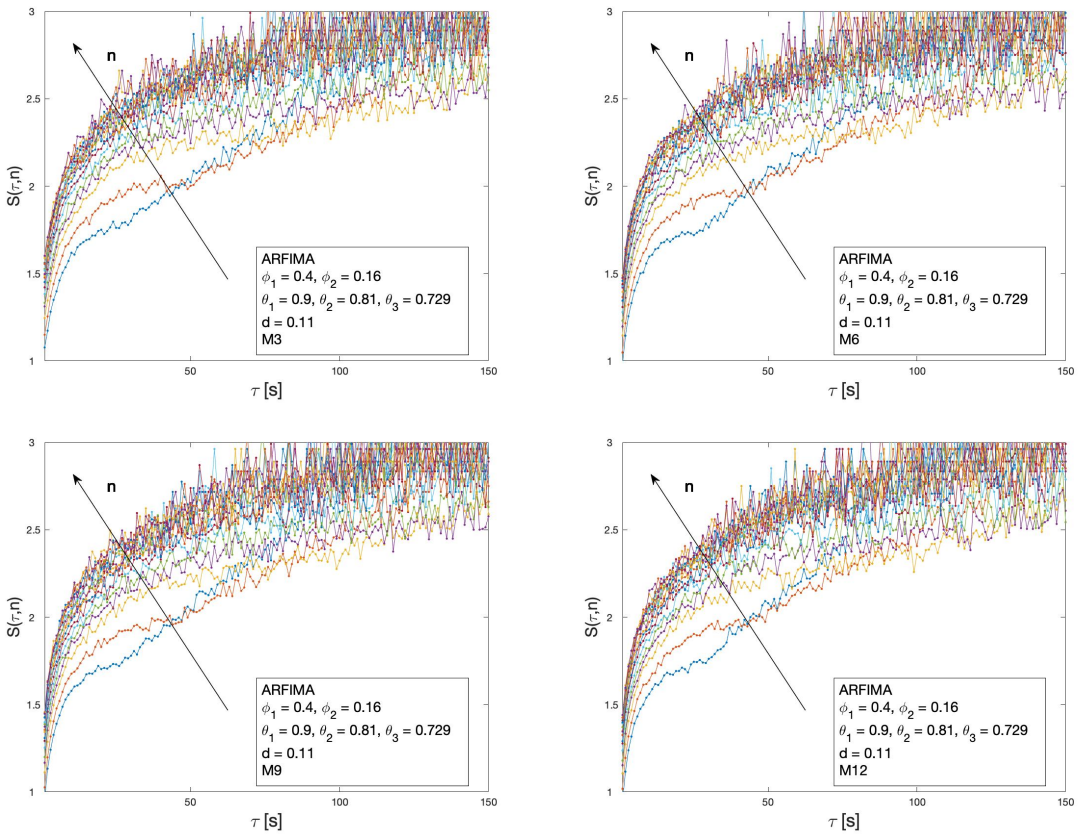


Figure 10: Cluster entropy results for series generated by means of ARFIMA processes. The autoregressive, moving average and differencing parameters are the following: $\phi_1 = 0.4$, $\phi_2 = 0.16$ and $\theta_1 = 0.9$, $\theta_2 = 0.81$, $\theta_3 = 0.729$ and $d = 0.11$. The time horizon analyzed in figures are M3, M6, M9 and M12

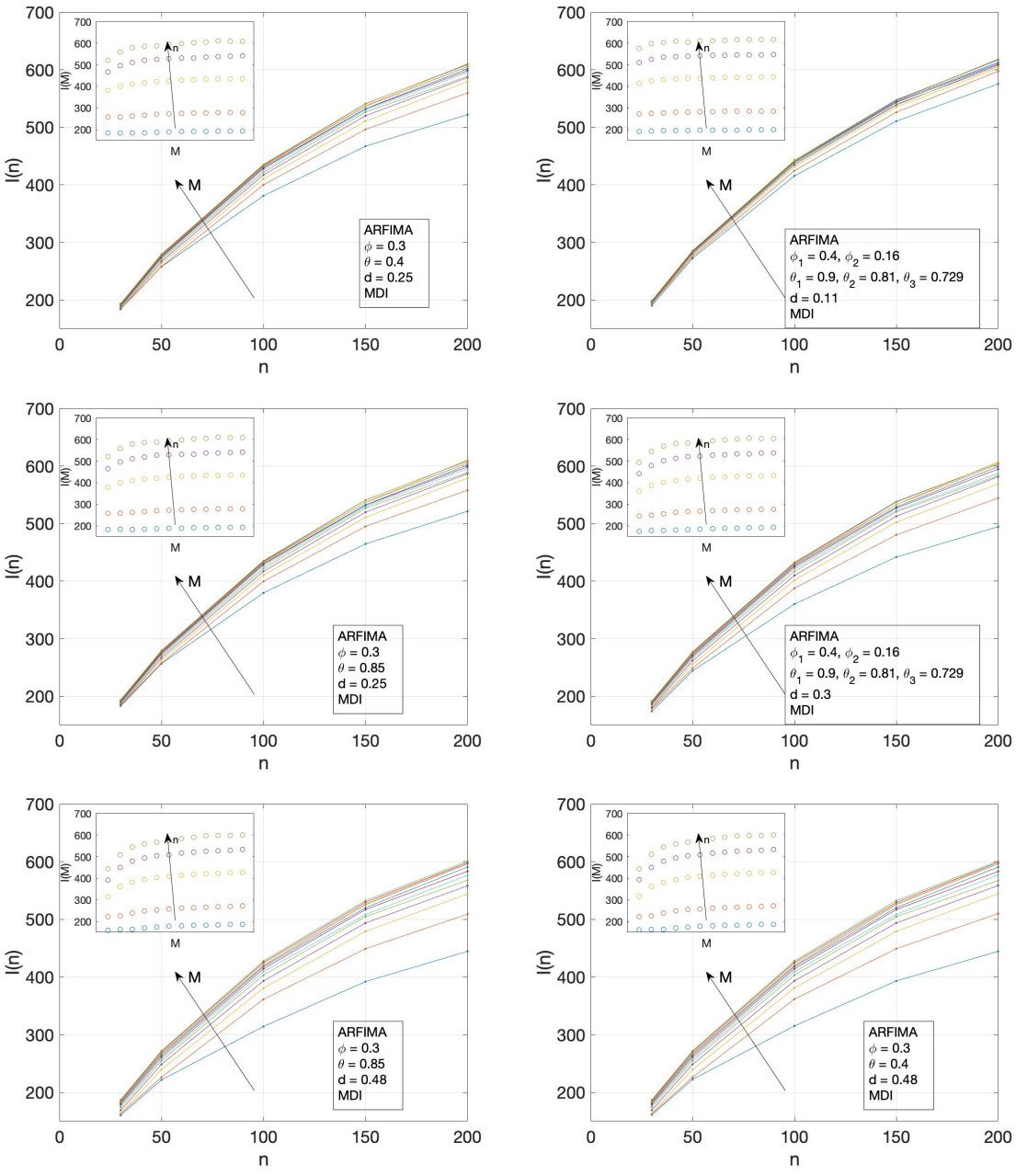


Figure 11: *Market Dynamic Index* plots for ARFIMA models. Panels (a) and (b) are referred to cluster entropy results in Figures 9, 10. Panels (c), (d), (e) and (f) are referred to other combination of parameters 2, 9, 6, 4 reported in Table 3.

5.6 Generalized Autoregressive Conditional Heteroskedasticity

Generalized autoregressive conditional heteroskedastic processes are used to model volatilities of time series (see Chapter 3 for a short overview on GARCH). GARCH series are generated by means of the MATLAB tool available at `https://it.mathworks.com/help/econ/garch.html`.

Figures 12, 35 and 13 illustrate cluster entropy curves for series obtained by means of GARCH models at different time horizons. The parameters set are the following. In Figure 12: $\omega = 0.1$, $\alpha = 0.475$ and $\beta = 0.1$. And in Figure 13: $\omega = 0.1$, $\alpha = 0.475$ and $\beta = 0.25$. Each plot refers to a different time horizon, namely M3, M6, M9 and M12. Several combination of parameters where studied (see Table 4), results are reported in Appendix A.6. All plots present a smooth logarithmic behavior for power-law correlated clusters. No significant differences can be observed neither over time, nor between different model specifications.

Figure 14 shows the market dynamic index calculated on GARCH series seen above. Values of $I(n, M)$ overlap at every moving average window n and time horizon M.

A number of other GARCH series were generated. Details on the specification analyzed are reported in Table 2 and results obtained for other set of parameters are reported in the Appendix A.6.

Model Number	α	β	ω
1	0.475	0.1	0.1
2	0.1	0.25	0.1
3	0.475	0.25	0.1

Table 4: Parameters choice for GARCH processes

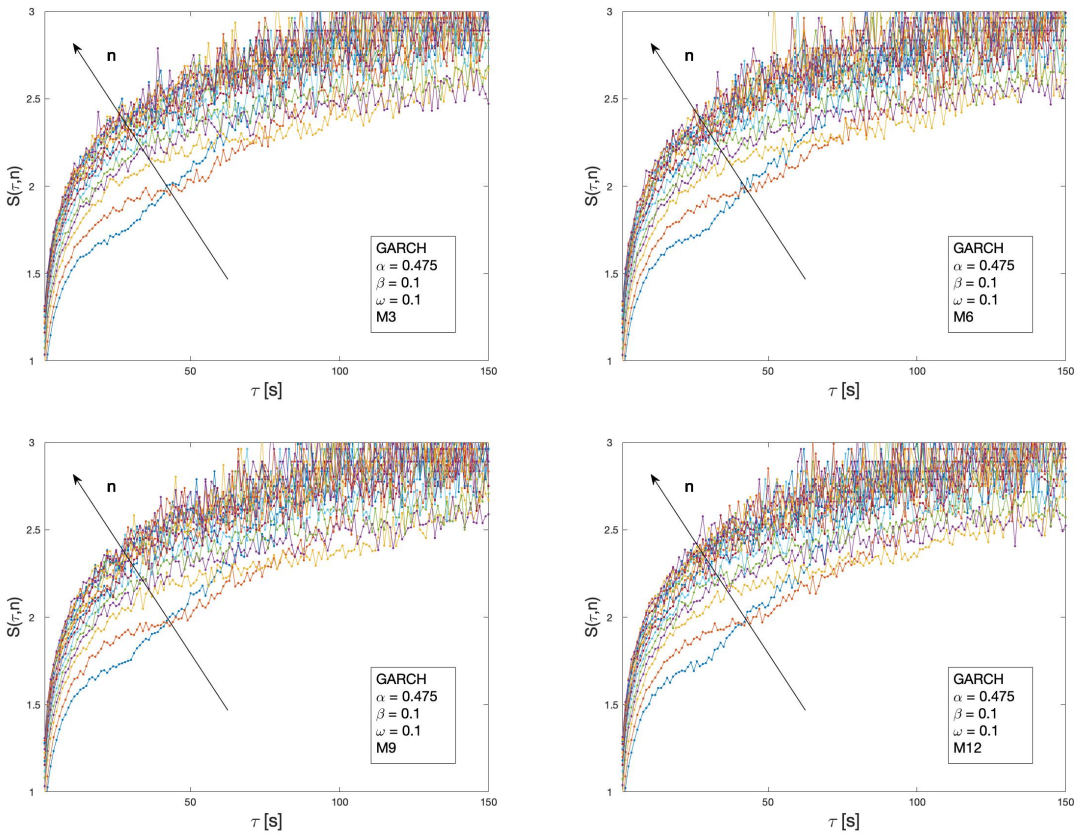


Figure 12: Cluster entropy results for series generated by means of GARCH processes. The series used to plot the entropy clusters are generated with the following parameters: $\omega = 0.1$, $\alpha = 0.475$ and $\beta = 0.1$. The time horizon analyzed in figures are M3, M6, M9 and M12

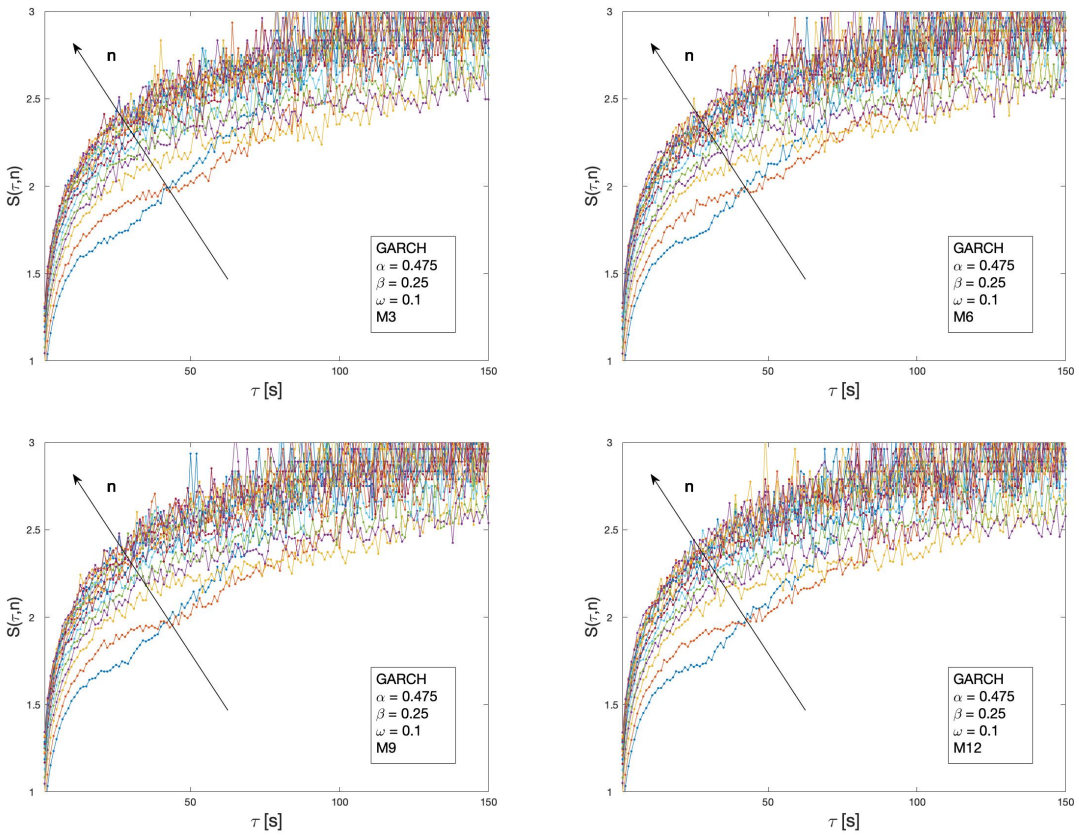


Figure 13: Cluster entropy results for series generated by means of GARCH processes. The series used to plot the entropy clusters are generated with the following parameters: $\omega = 0.1$, $\alpha = 0.475$ and $\beta = 0.25$. The time horizon analyzed in figures are M3, M6, M9 and M12

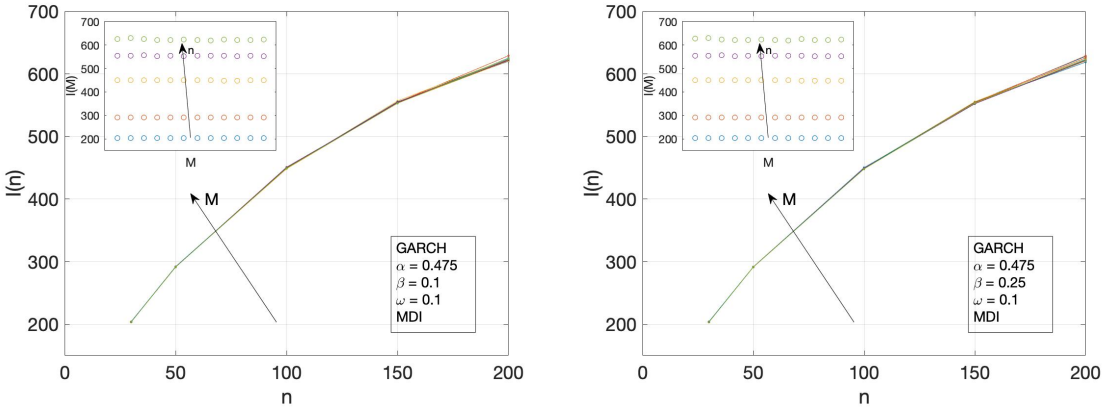


Figure 14: *Market Dynamic Index for GARCH(1,1) models.* The model used to generate the series has the following parameters: in (a) $\alpha = 0.475$ and $\beta = 0.1$ and in (b) $\alpha = 0.475$ and $\beta = 0.25$.

6 Discussion and Conclusions

In this section, the main results of the analysis of the cluster entropy and Market Dynamic Index are discussed and compared with those obtained for real-world financial markets.

As previously discussed, cluster entropy behavior is described by Equation (93). Hence, power-law correlated clusters, i.e. clusters with length $\tau < n$, are expected to follow a logarithmic behavior, regardless of the moving average window n, and therefore independently on the partitioning method used. On the other hand, exponentially correlated clusters, i.e. clusters with length $\tau \geq n$, are expected to follow a linear behavior described by the term τ/n, which does depend on the moving average window n and whose slope decreases as $1/n$. Cumulative information measures are useful to summarize key information in a single numerical index. The Market Dynamic Index is thus obtained by means of Equation (95) from cluster entropy results. It summarizes the information present in the entropy curves at different time horizons M and moving average windows n. We assume that power-law correlated clusters can be considered as the information carriers. Hence, those clusters are considered into the calculation of the Market Dynamic Index.

In this thesis it has been shown that differences emerge from the comparison between cluster entropy results obtained on artificial series models and real-world market series. Figures show cluster entropy results for the following processes: Fractional Brownian Motion (Figure 2), Geometric Brownian Motion (Figure 3), Cox-Ingersoll-Ross (Figure 5), Hull-White-Vasicek (Figure 7) and GARCH (Figures 12 and 13). All cluster entropy curves exhibit a smooth logarithmic behavior for power-law correlated clusters, independently on the model and on the time horizon.

As one can see in Appendix A.1, cluster entropy results for series generated by means of Fractional Brownian Motion processes with Hurst exponent $0 < H < 0.5$ (anticorrelated FBMs) do not present any horizon dependence. Conversely, Fractional Brownian Motion series with $0.5 < H < 1$ (positively correlated FBMs) do show some horizon dependence. However, as it will be clarified below, Fractional Brownian Motion series do not fully reproduce financial markets behavior.

A significant horizon dependence emerges for series generated by ARFIMA processes, as one can note by observing Figures 9, 10. Given a set of parameters, power-law correlated cluster entropy frequently deviates from the ideal logarithmic path and vary over different horizons. The magnitude of such deviations is strongly dependent on the differencing parameter d, which quantifies the correlation degree of the series. Thus, cluster entropy for series generated by ARFIMA process exhibit horizon dependence as observed in real world financial markets.

Thus, in conclusion, the Market Dynamic Index summarizes the results obtained on cluster entropy. In particular, series that do not embed persistence, namely those generated by means of Fractional Brownian Motion, Geometric Brownian Motion, Cox-Ingersoll-Ross, Hull-White-Vasicek and GARCH processes, do not present any horizon dependence. As one can see in Figures 2 (bottom), 4, 6, 8, 14, Market Dynamic Index curves overlap for each

moving average window n and time horizon M. Additionally, a comparison over series obtained by fully independent processes shows that the index $I(n, M)$ assumes approximately equal values independently on the process used to generate the series.

Results of the Market Dynamic Index of persistent series generated by Fractional Brownian Motion are reported in Appendix A.1. Anti-persistent series, i.e. series generated with Hurst exponent $0 < H < 0.5$, show overlapping values of the Market Dynamic Index and therefore do not present any horizon dependence. Conversely, positively correlated series, i.e. series characterized by $0.5 < H < 1$, do present a significant degree of horizon dependence. However, series generated by Fractional Brownian Motion fail at reproducing financial markets with such a large Hurst exponent. To generate an horizon dependence comparable with that found in financial markets an unrealistic persistence must be chosen.

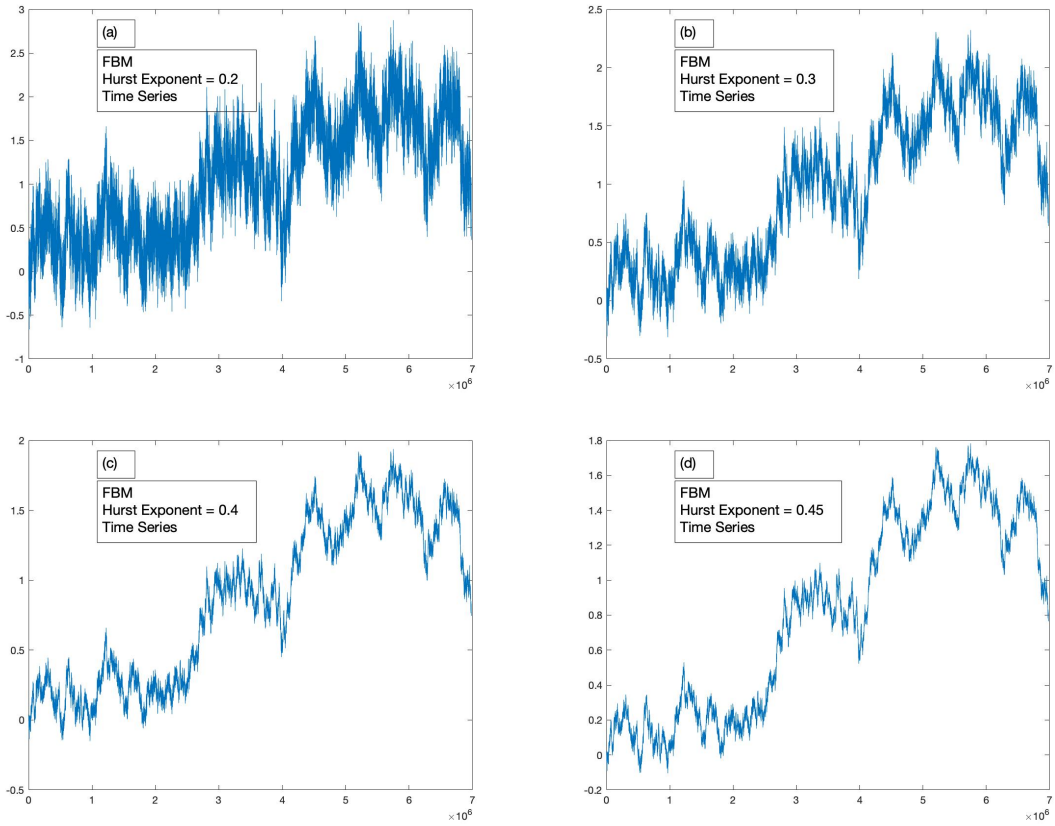
On the other hand, Market Dynamic Index for series generated by ARFIMA processes are reported in 11. When the differencing parameter d is $0 < d < 0.2$ Market Dynamic Index curves are n-invariant for small values of n, but horizon dependence emerges at larger n. When $0.2 < d < 0.5$ Market Dynamic Index curves show a significant horizon dependence even at small n. Therefore, according to the choice of the differencing parameter d, series generated by ARFIMA processes can simulate financial markets dynamics very closely.

It must be then concluded that cluster entropy behavior is deeply linked to positive persistence and long-range correlation embedded in the series. In real-world financial series horizon dependence deviates from the case of absolutely random series, such as those generated by means of stochastic differential equations. Therefore, as suggested by many recent studies, contrary to the traditional financial theories, the hypothesis of efficient markets and random walk behavior of financial stochastic processes do not hold. Thence, the integration performed by the Market Dynamic Index on cluster entropy allows to obtain this result in a cumulative, and then particularly robust, manner. Moreover, as shown on NASDAQ and DJIA, where the former is a diversified stock market with a high degree of heterogeneity and the latter is an index representative of a chosen set of industrial stocks, the Market Dynamic Index can quantify intrinsic market heterogeneity.

Further investigation are needed in order to achieve deeper insights and shed light on the characteristic behaviour shown by the moving average cluster entropy and how these relates to the intrinsic properties of financial markets.

A Appendixes

A.1 Fractional Brownian Motion



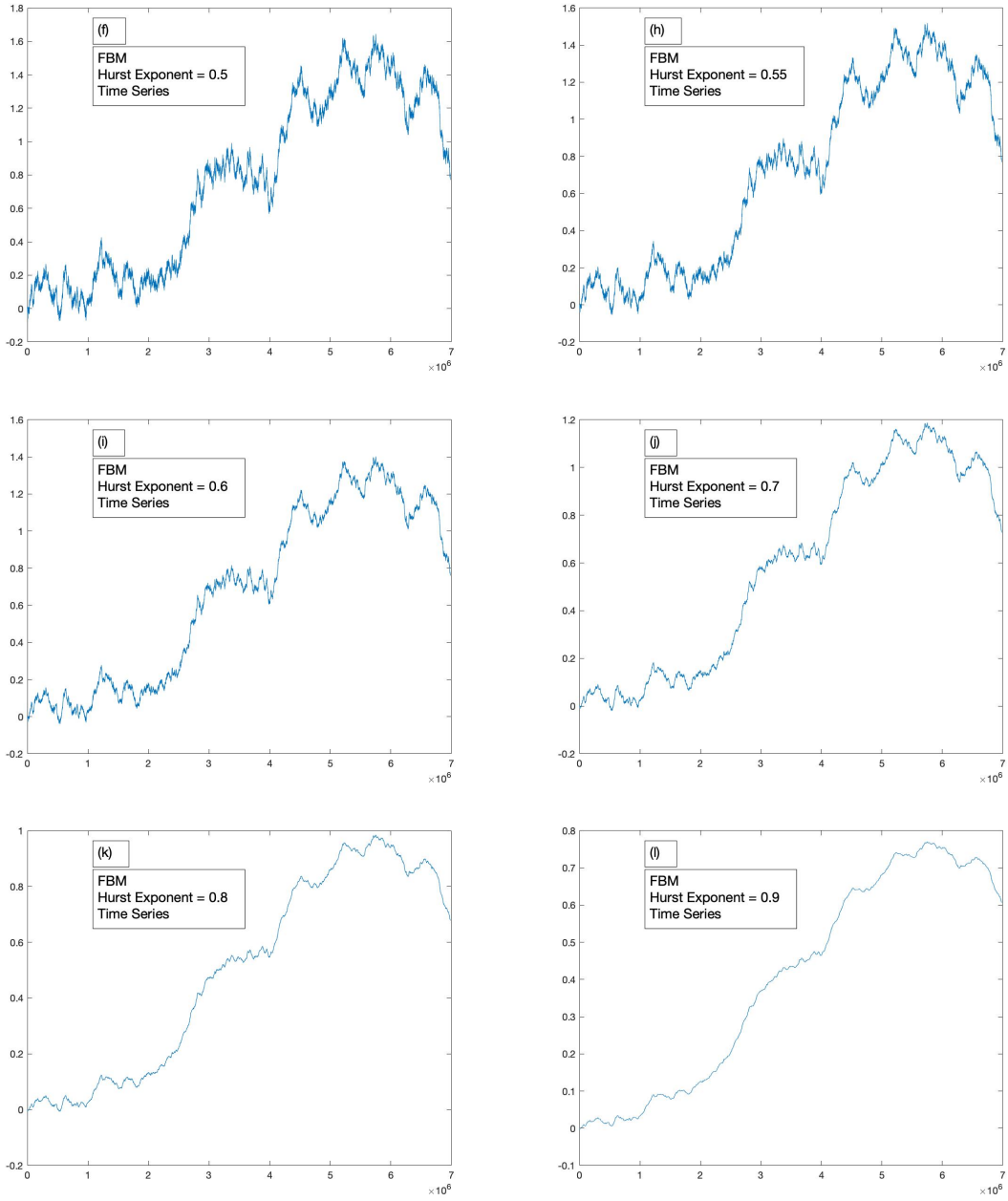


Figure 15: Time Series for series generated by Fractional Brownian Motion with Hurst exponent $0.1 < H < 0.9$. Results for $H = 0.5$ are reported in Chapter 5 and repeated here for the sake of completeness.

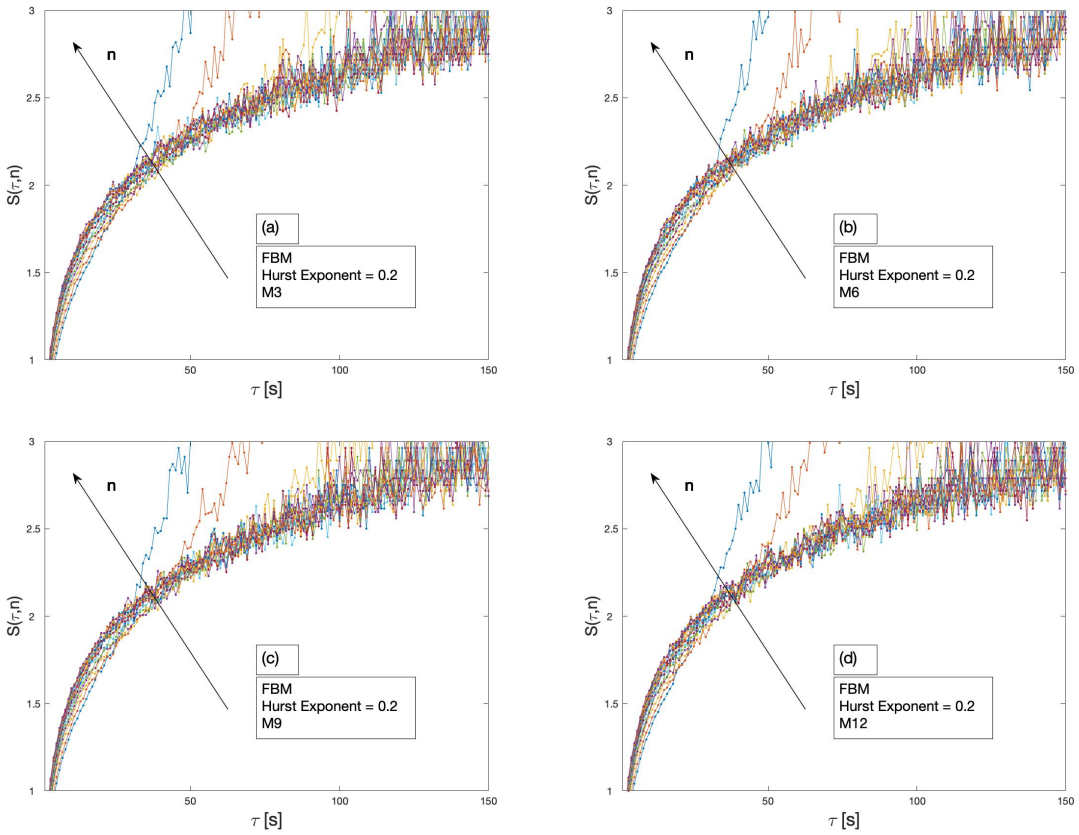


Figure 16: Cluster entropy for series generated by means of Fractional Brownian Motion with $H = 0.2$. The time horizon analyzed in figures (a), (b), (c) and (d) is M3, M6, M9 and M12 respectively.

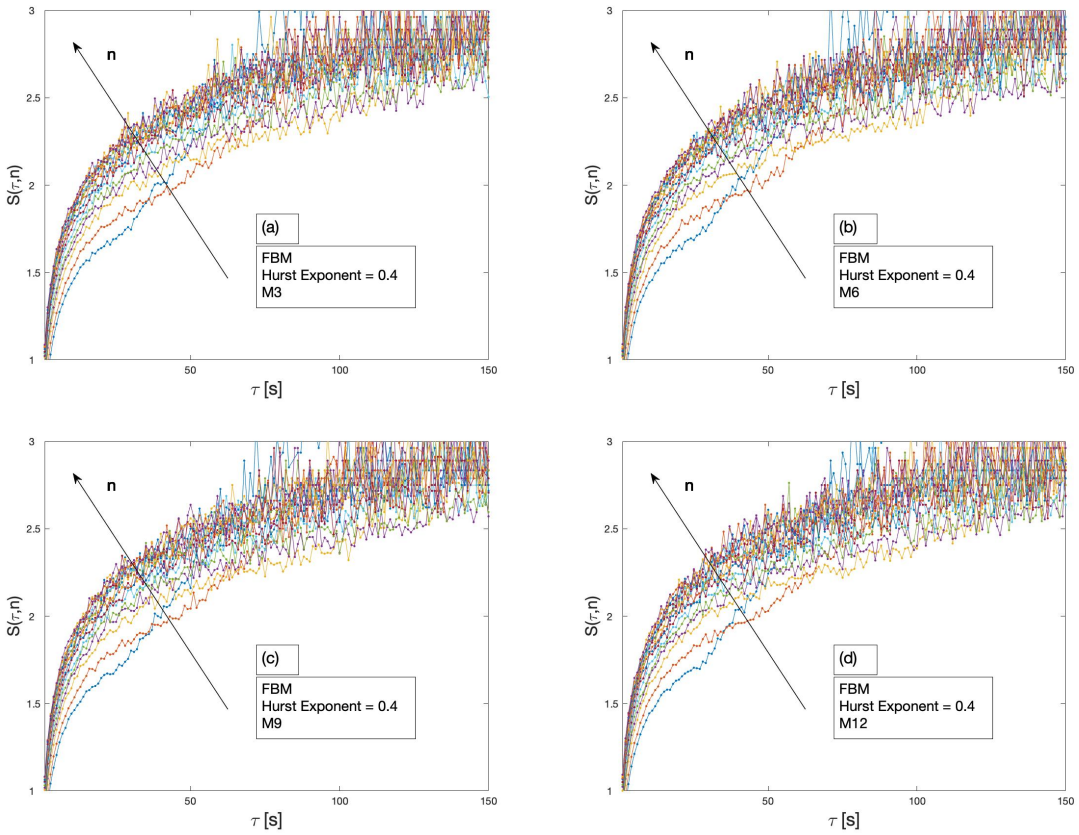


Figure 17: Cluster entropy for series generated by means of Fractional Brownian Motion with $H = 0.4$. The time horizon analyzed in figures (a), (b), (c) and (d) is M3, M6, M9 and M12 respectively.

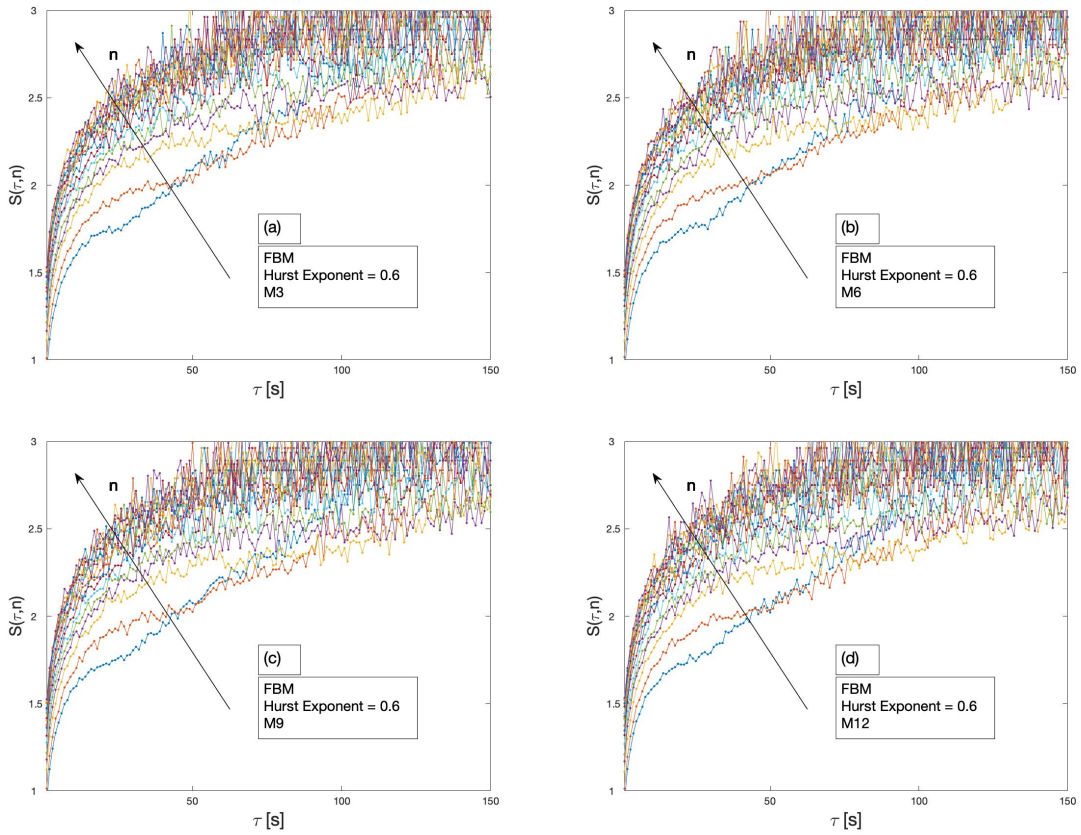


Figure 18: Cluster entropy for series generated by means of Fractional Brownian Motion with $H = 0.6$. The time horizon analyzed in figures (a), (b), (c) and (d) is M3, M6, M9 and M12 respectively.

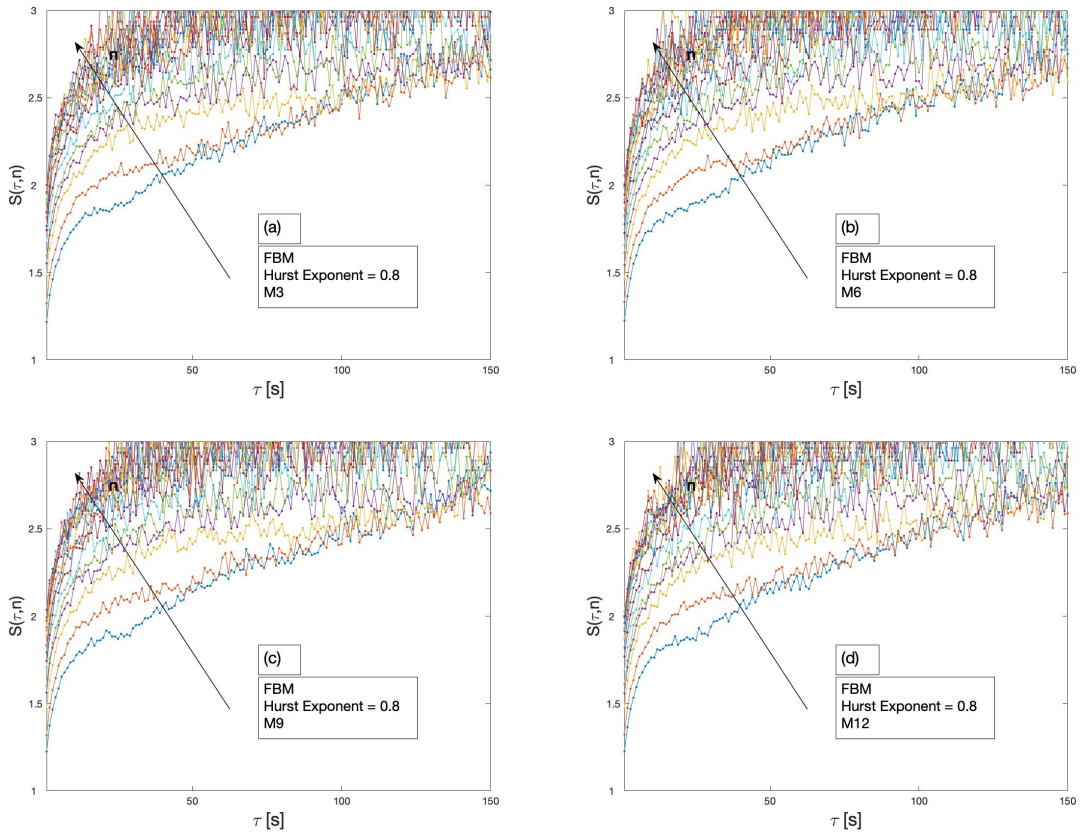
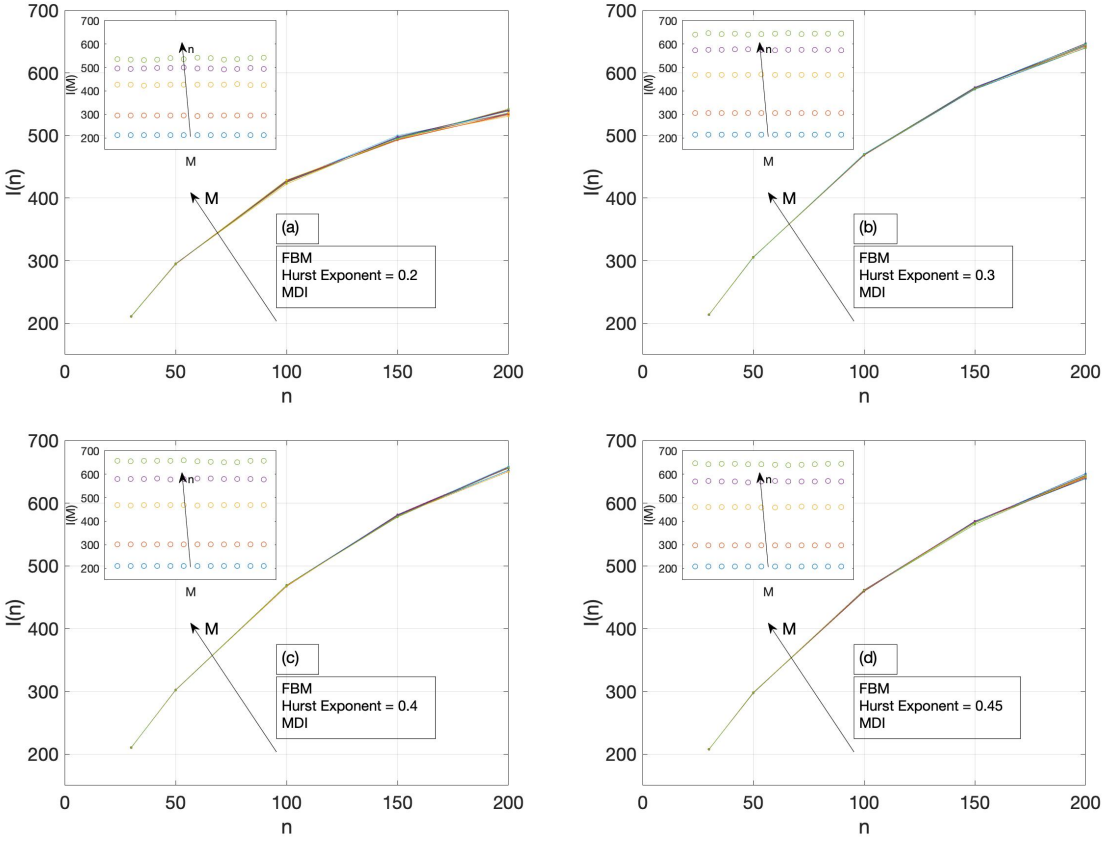


Figure 19: Cluster entropy for series generated by means of Fractional Brownian Motion with $H = 0.8$. The time horizon analyzed in figures (a), (b), (c) and (d) is M3, M6, M9 and M12 respectively.


I(n)
n
M
I(M)
(a)
FBM
Hurst Exponent = 0.2
MDI
(b)
FBM
Hurst Exponent = 0.3
MDI
(c)
FBM
Hurst Exponent = 0.4
MDI
(d)
FBM
Hurst Exponent = 0.45
MDI

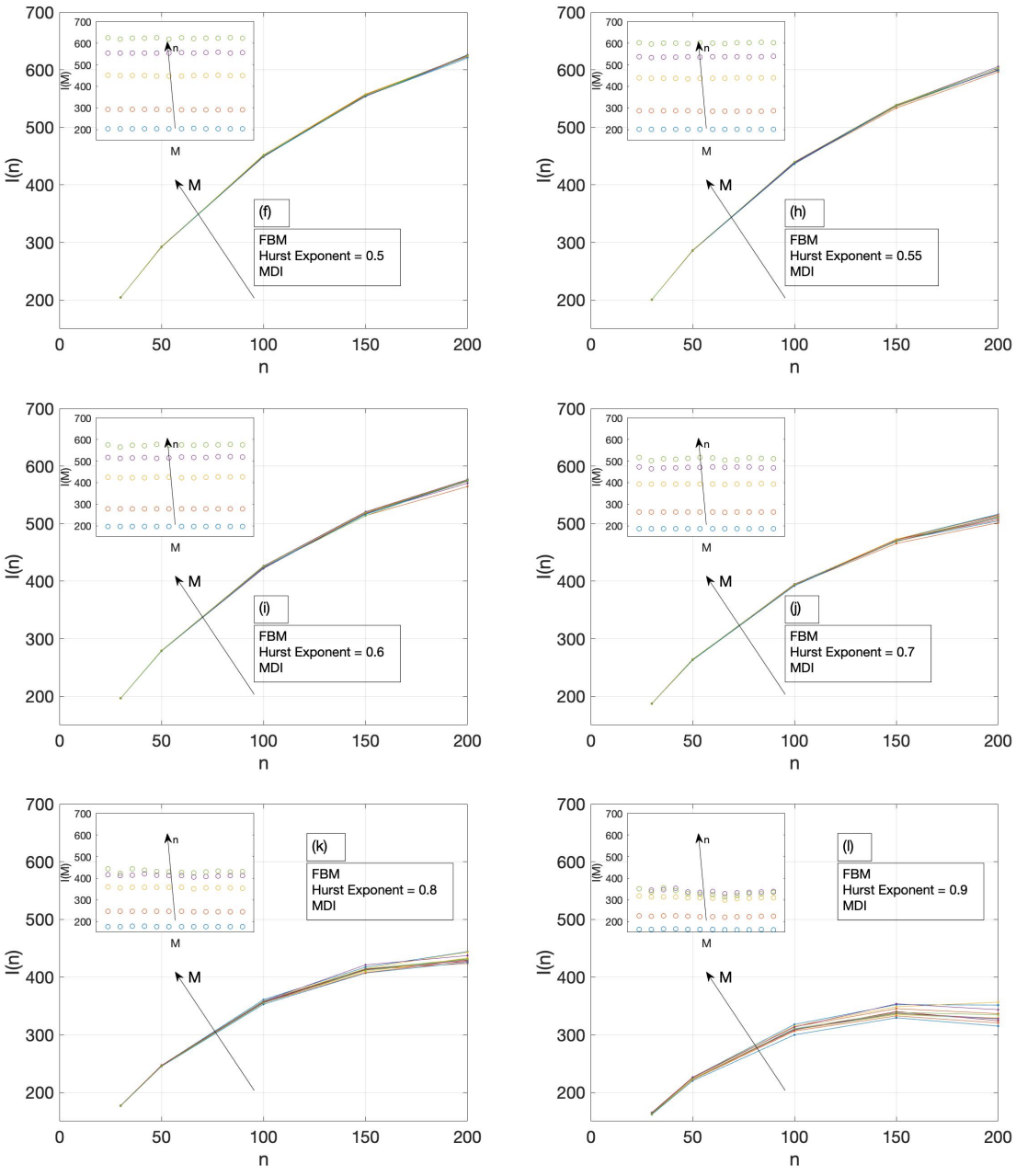


Figure 20: Market Dynamic Index for series generated by Fractional Brownian Motion with Hurst exponent $0.1 < H < 0.9$. Results for $H = 0.5$ are reported in Chapter 5 and repeated here for the sake of completeness.

A.2 Geometric Brownian Motion

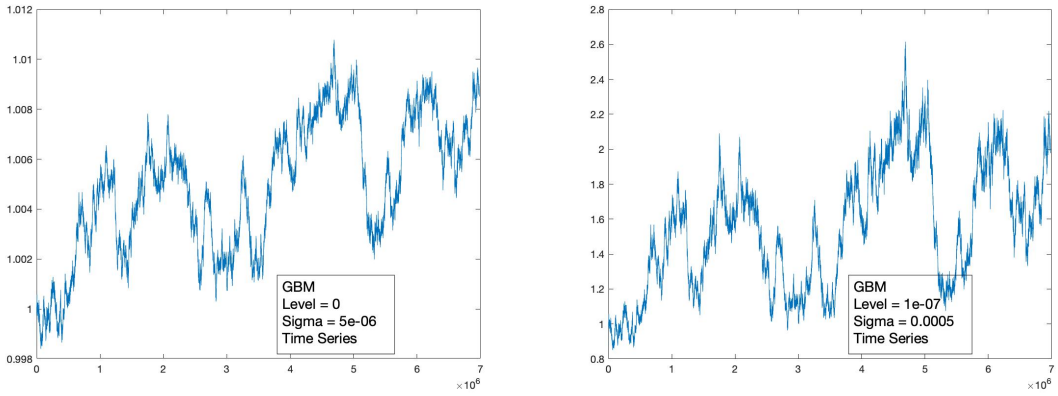


Figure 21: Time Series generated by means of Geometric Brownian Motion processes. To generate the series the following parameters are used: (a) $r = 0$ and $\sigma = 5 \cdot 10^{-6}$ (b) $r = 1 \cdot 10^{-7}$ and $\sigma = 5 \cdot 10^{-4}$. Results in the right panel are reported in Chapter 5 and repeated here for the sake of completeness.

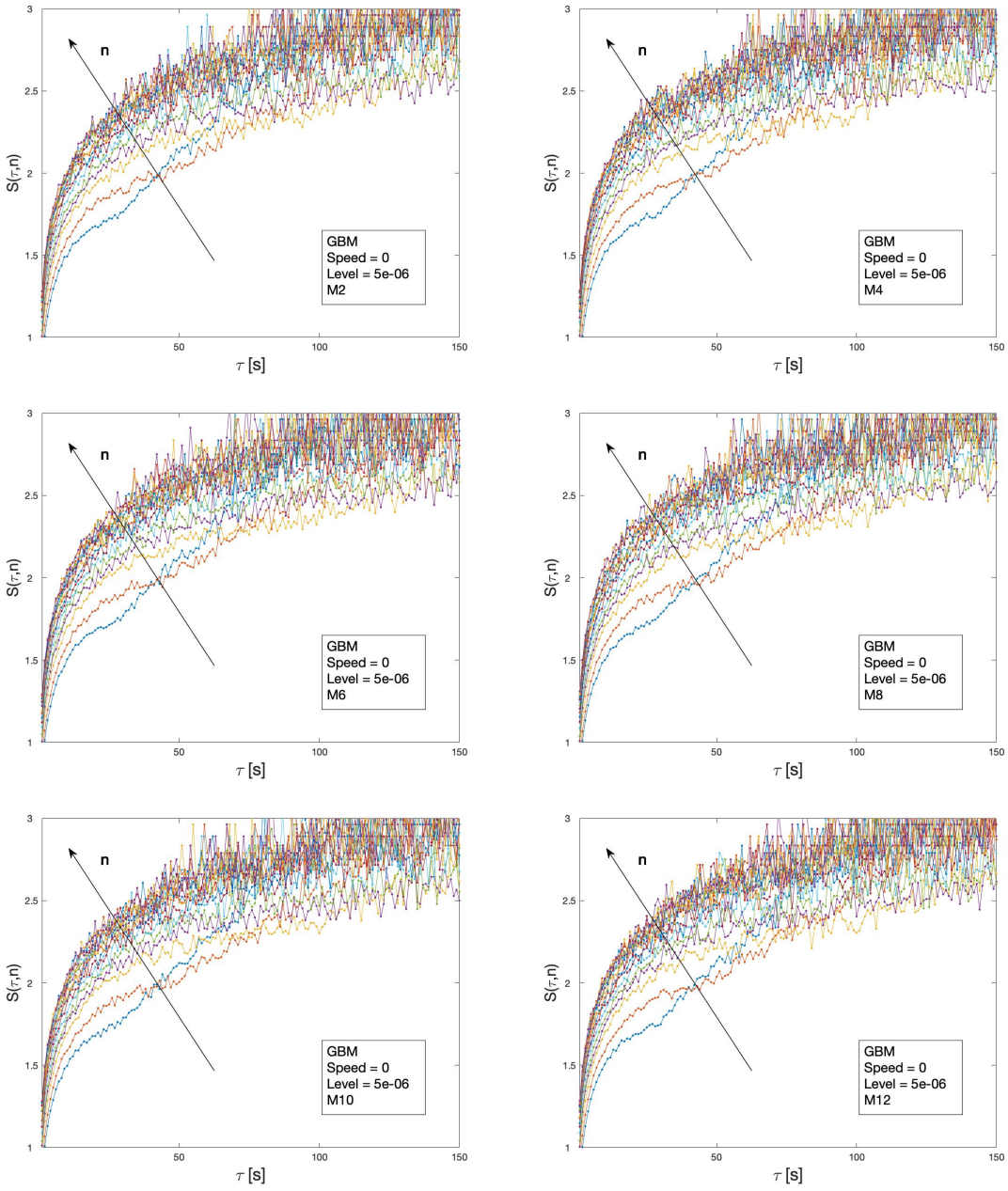


Figure 22: Cluster entropy for series generated by means of Geometric Brownian Motion processes. To generate the series the following parameters are used: $r = 0$ and $\sigma = 5 \cdot 10^{-6}$. Time horizons reported are M2, M4, ..., M12.

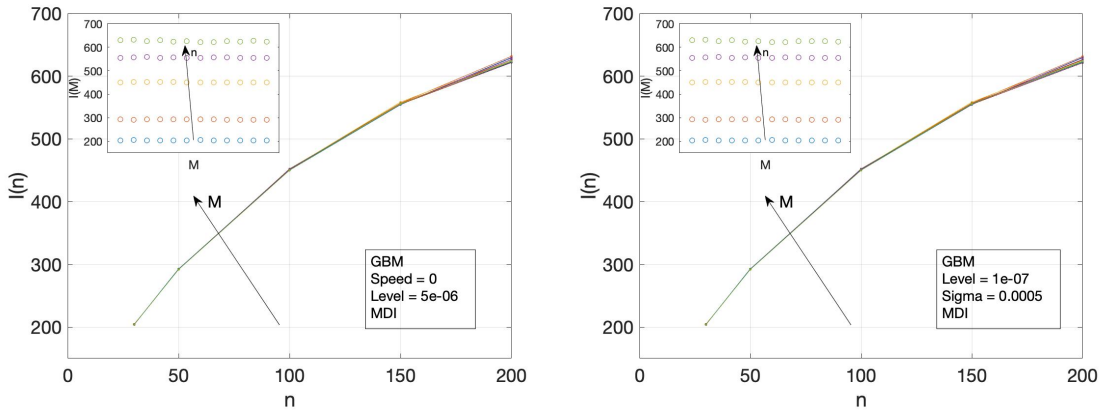


Figure 23: Market Dynamic Index for series generated by means of Geometric Brownian Motion processes. To generate the series the following parameters are used: (a) $r = 0$ and $\sigma = 5 \cdot 10^{-6}$ (b) $r = 1 \cdot 10^{-7}$ and $\sigma = 5 \cdot 10^{-4}$. Results in the right panel are reported in Chapter 5 and repeated here for the sake of completeness.

A.3 Cox-Ingersoll-Ross

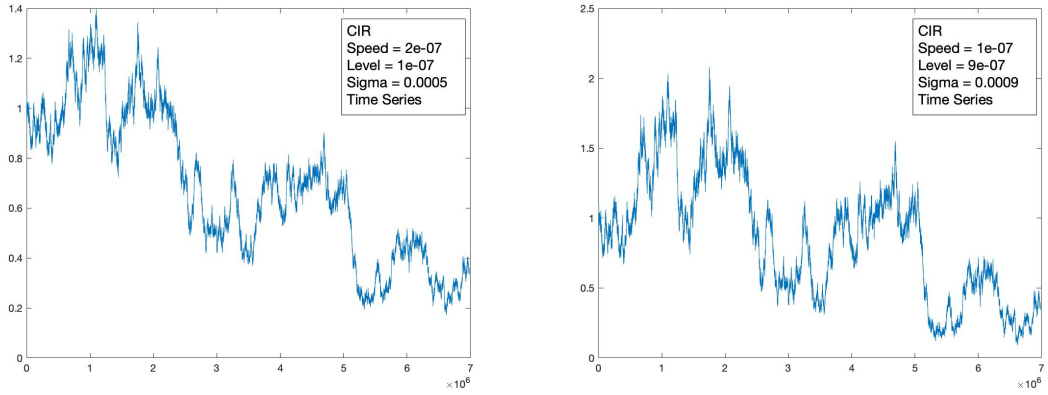


Figure 24: Time Series generated by means of Cox-Ingersoll-Ross processes. To generate the series the following parameters are used: (a) $s = 2 \cdot 10^{-7}$, $r = 1 \cdot 10^{-7}$ and $\sigma = 5 \cdot 10^{-4}$ (b) $s = 1 \cdot 10^{-7}$, $r = 9 \cdot 10^{-7}$ and $\sigma = 9 \cdot 10^{-4}$. Results in the left panel are reported in Chapter 5 and repeated here for the sake of completeness.

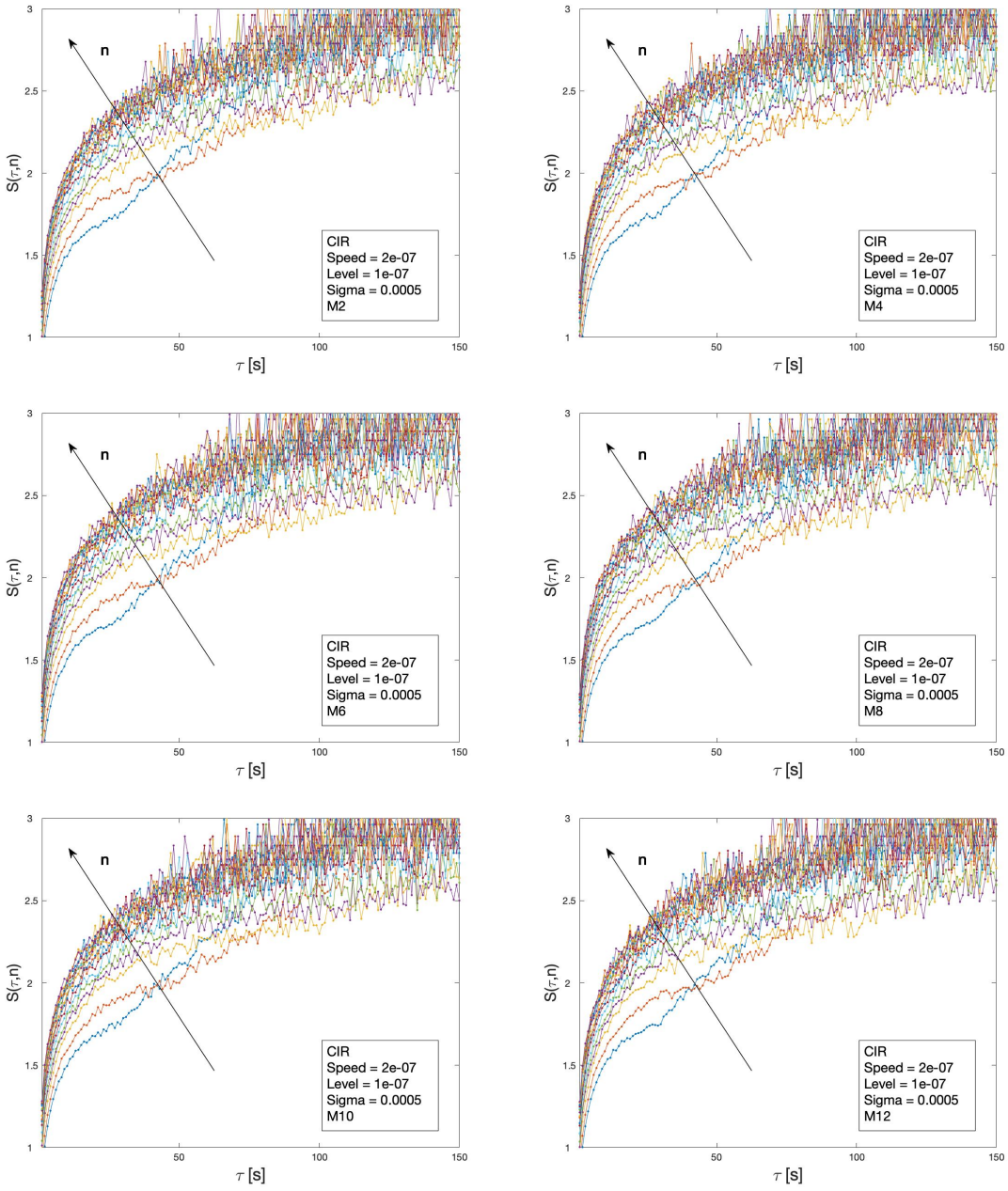


Figure 25: Cluster entropy for series generated by means of Cox-Ingersoll-Ross processes. To generate the series the following parameters are used: $s = 2 \cdot 10^{-7}$, $r = 1 \cdot 10^{-7}$ and $\sigma = 5 \cdot 10^{-4}$. Time horizons reported are M2, M4, ..., M12.

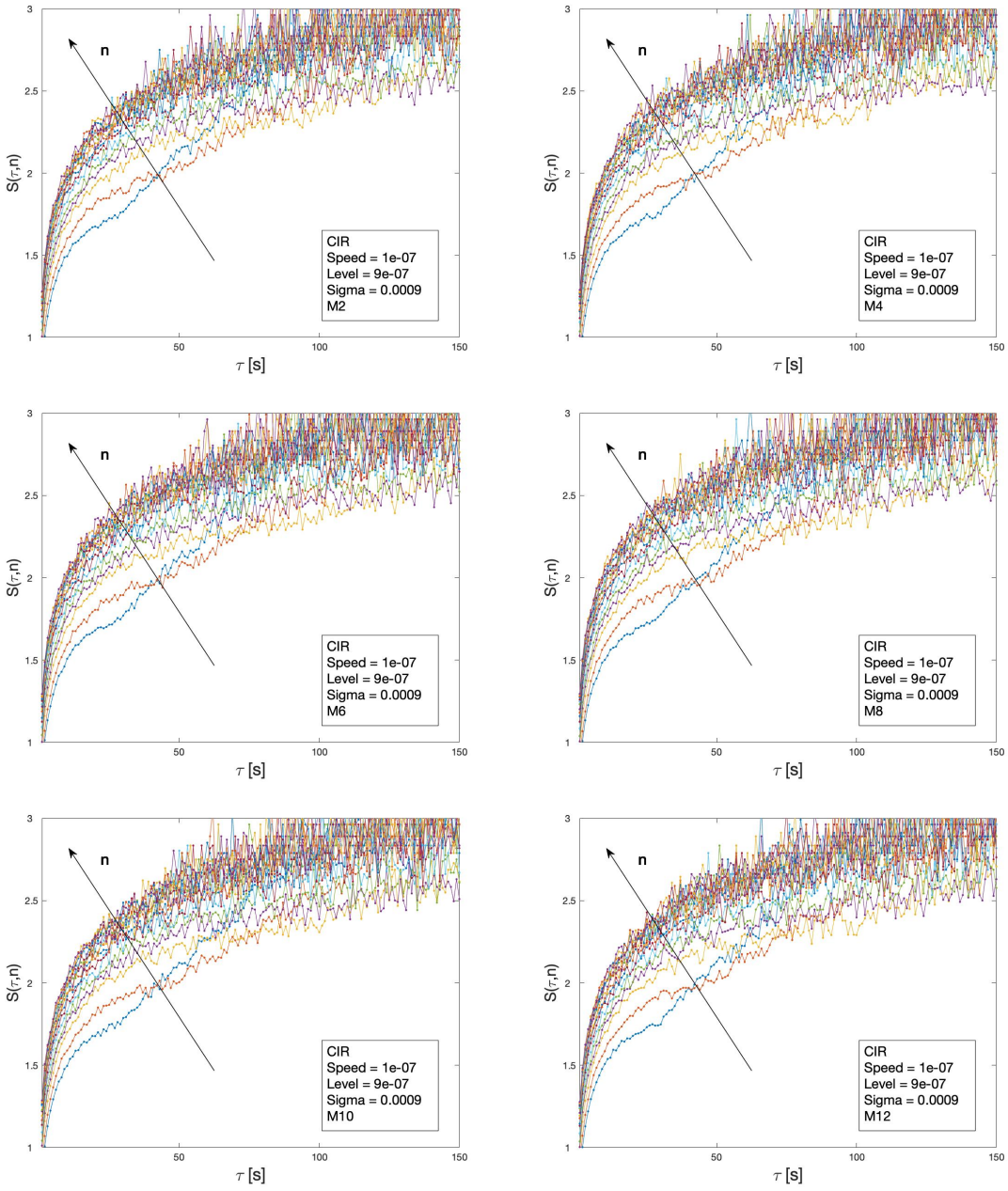


Figure 26: Cluster entropy for series generated by means of Cox-Ingersoll-Ross processes. To generate the series the following parameters are used: $s = 1 \cdot 10^{-7}$, $r = 9 \cdot 10^{-7}$ and $\sigma = 9 \cdot 10^{-4}$. Time horizons reported are M2, M4, ..., M12.

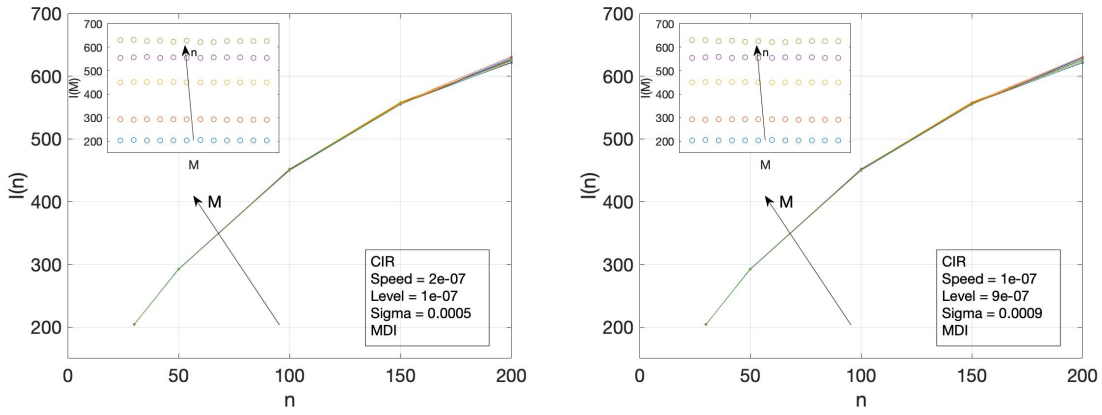


Figure 27: Market Dynamic Index for series generated by means of Cox-Ingersoll-Ross processes. To generate the series the following parameters are used: (a) $s = 2 \cdot 10^{-7}$, $r = 1 \cdot 10^{-7}$ and $\sigma = 5 \cdot 10^{-4}$ (b) $s = 1 \cdot 10^{-7}$, $r = 9 \cdot 10^{-7}$ and $\sigma = 9 \cdot 10^{-4}$. Results in the left panel are reported in Chapter 5 and repeated here for the sake of completeness.

A.4 Hull-White-Vasicek

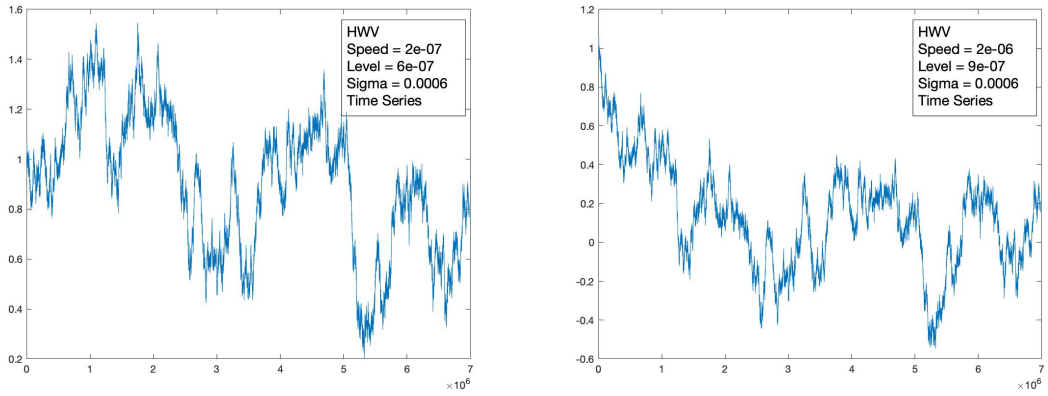


Figure 28: Time Series generated by means of Hull-White-Vasicek processes. To generate the series the following parameters are used: (a) $s = 2 \cdot 10^{-7}$, $r = 6 \cdot 10^{-7}$ and $\sigma = 6 \cdot 10^{-4}$ (b) $s = 2 \cdot 10^{-6}$, $r = 9 \cdot 10^{-7}$ and $\sigma = 6 \cdot 10^{-4}$. Results in the right panel are reported in Chapter 5 and repeated here for the sake of completeness.

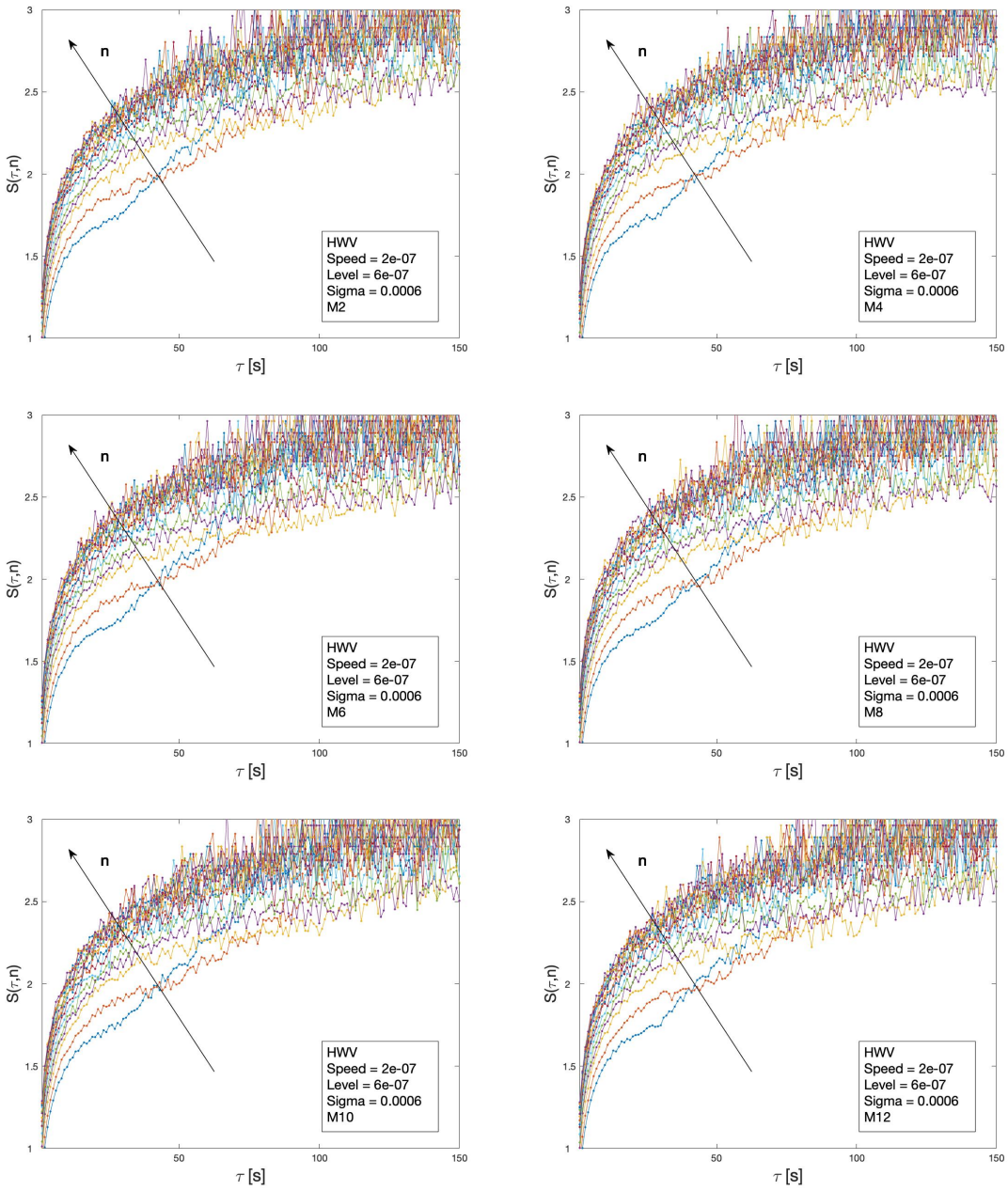


Figure 29: Cluster entropy for series generated by means of Hull-White-Vasicek processes. To generate the series the following parameters are used: $s = 2 \cdot 10^{-7}$, $r = 6 \cdot 10^{-7}$ and $\sigma = 6 \cdot 10^{-4}$. Time horizons reported are M2, M4, ..., M12.

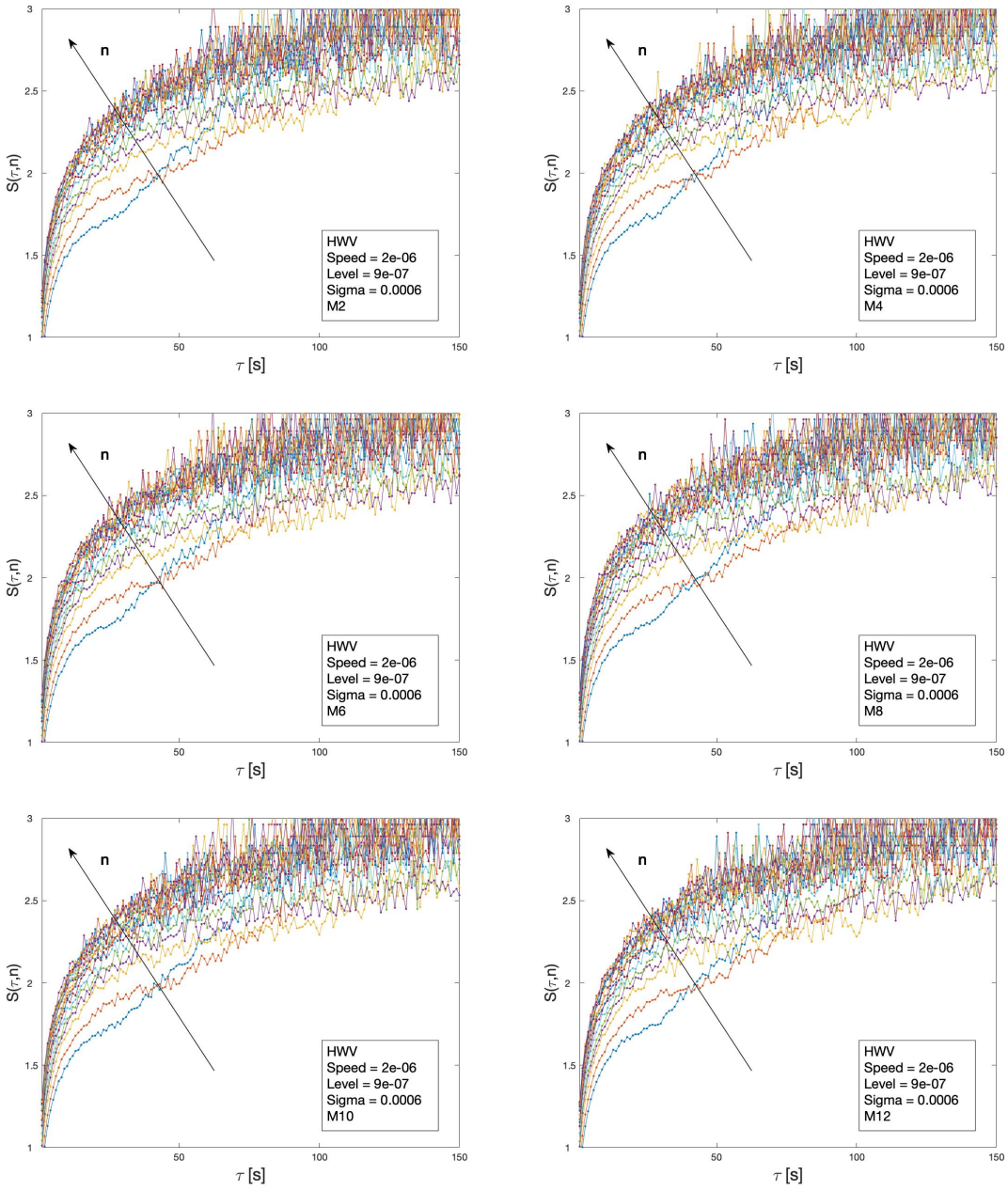


Figure 30: Cluster entropy for series generated by means of Hull-White-Vasicek processes. To generate the series the following parameters are used: $s = 2 \cdot 10^{-6}$, $r = 9 \cdot 10^{-7}$ and $\sigma = 6 \cdot 10^{-4}$. Time horizons reported are M2, M4, ..., M12.

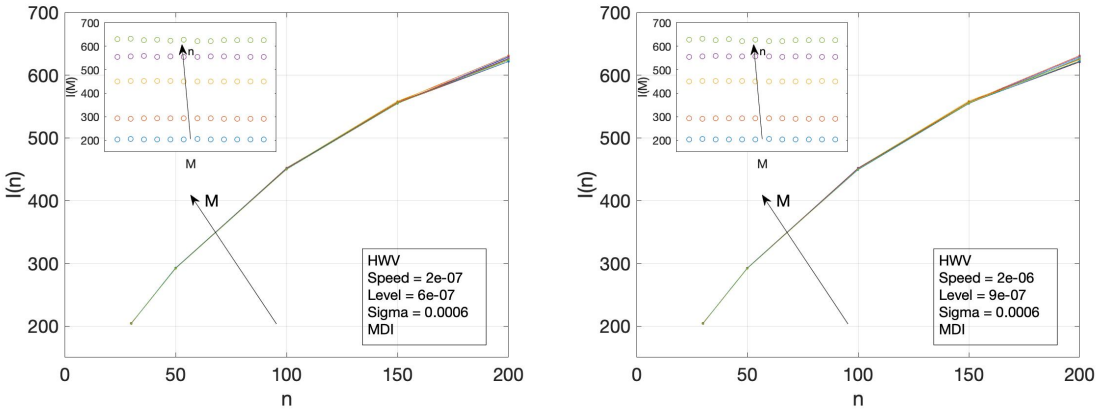


Figure 31: Market Dynamic Index for series generated by Hull-White-Vasicek processes. To generate the series the following parameters are used: (a) $s = 2 \cdot 10^{-7}$, $r = 6 \cdot 10^{-7}$ and $\sigma = 6 \cdot 10^{-4}$ (b) $s = 2 \cdot 10^{-6}$, $r = 9 \cdot 10^{-7}$ and $\sigma = 6 \cdot 10^{-4}$. Results in the right panel are reported in Chapter 5 and repeated here for the sake of completeness.

A.5 Autoregressive Fractionally Integrated Moving Aaverage

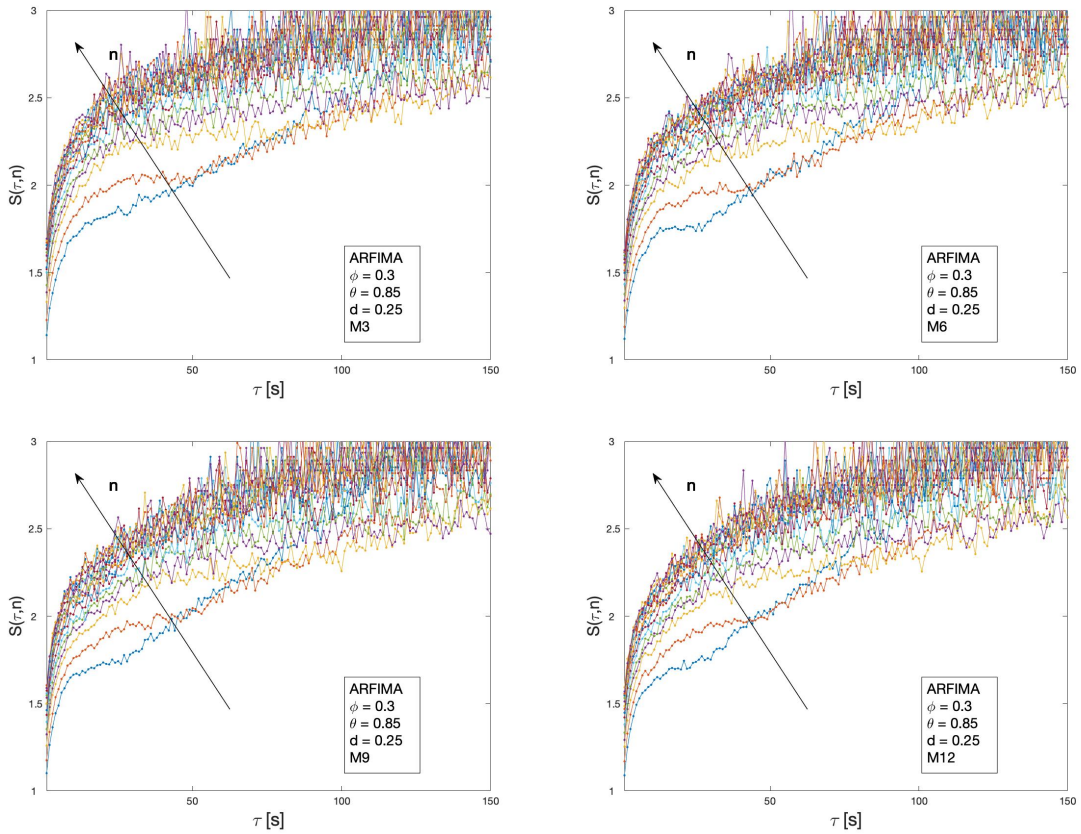


Figure 32: Entropy clusters for series generated by means of ARFIMA processes. The series used to plot the entropy clusters was generated with the following parameters: $\phi = 0.3$, $d = 0.25$, $\theta = 0.85$. Plots (a), (b), and (c) shows curves for time horizons M3, M6, M9 and M12 respectively.

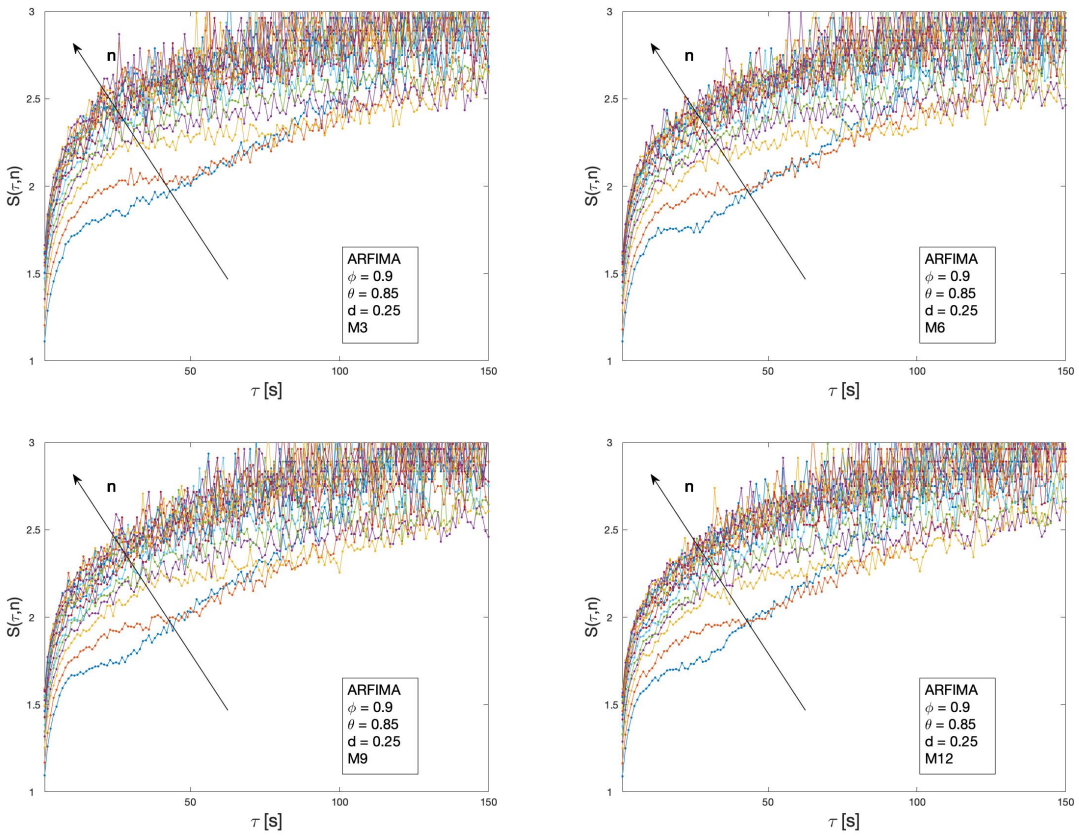


Figure 33: Entropy clusters for series generated by means of ARFIMA processes. The series used to plot the entropy clusters was generated with the following parameters: $\phi = 0.9$, $d = 0.25$, $\theta = 0.85$. Plots (a), (b), and (c) shows curves for time horizons M3, M6, M9 and M12 respectively.

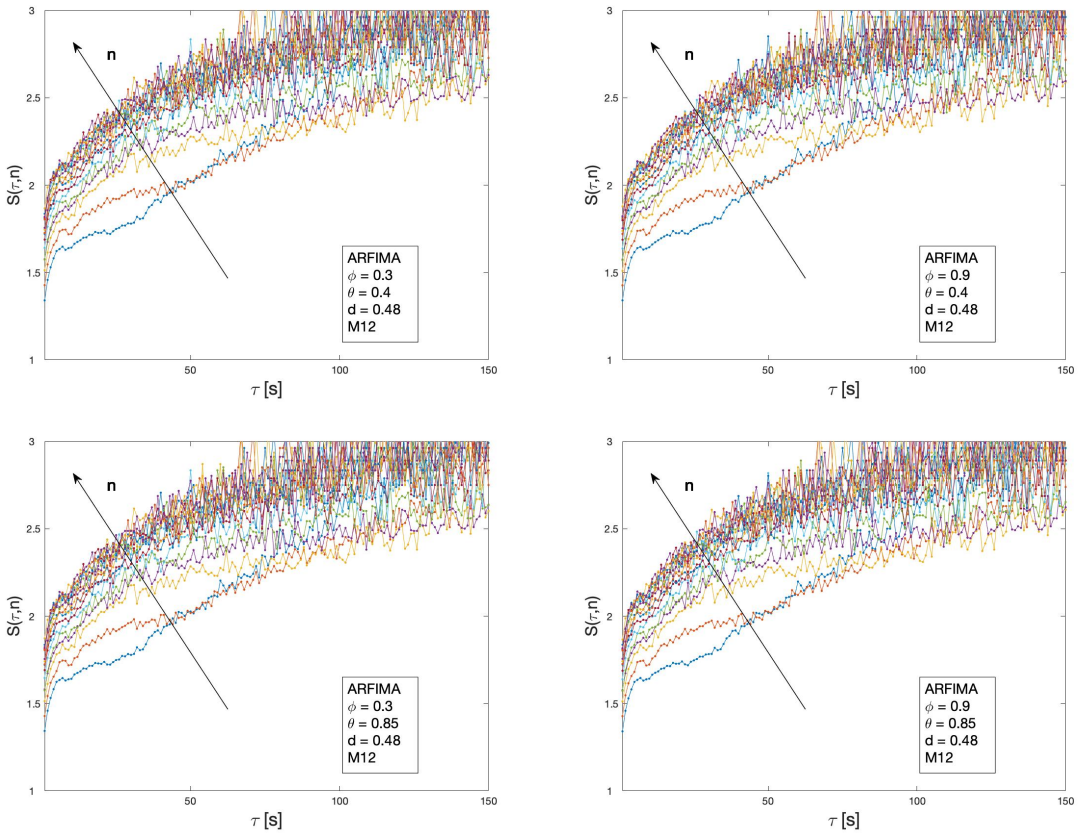


Figure 34: Entropy clusters for series generated by means of ARFIMA processes. The series used to plot the entropy clusters was generated with the following parameters: (a) $\phi = 0.3$ and $\theta = 0.4$, (b) $\phi = 0.9$ and $\theta = 0.4$, (c) $\phi = 0.3$ and $\theta = 0.85$, (d) $\phi = 0.9$ and $\theta = 0.85$. Plots (a), (b), and (c) shows curves for time horizons M3, M6, M9 and M12 respectively.

A.6 Generalized Autoregressive Conditional Heteroskedasticity

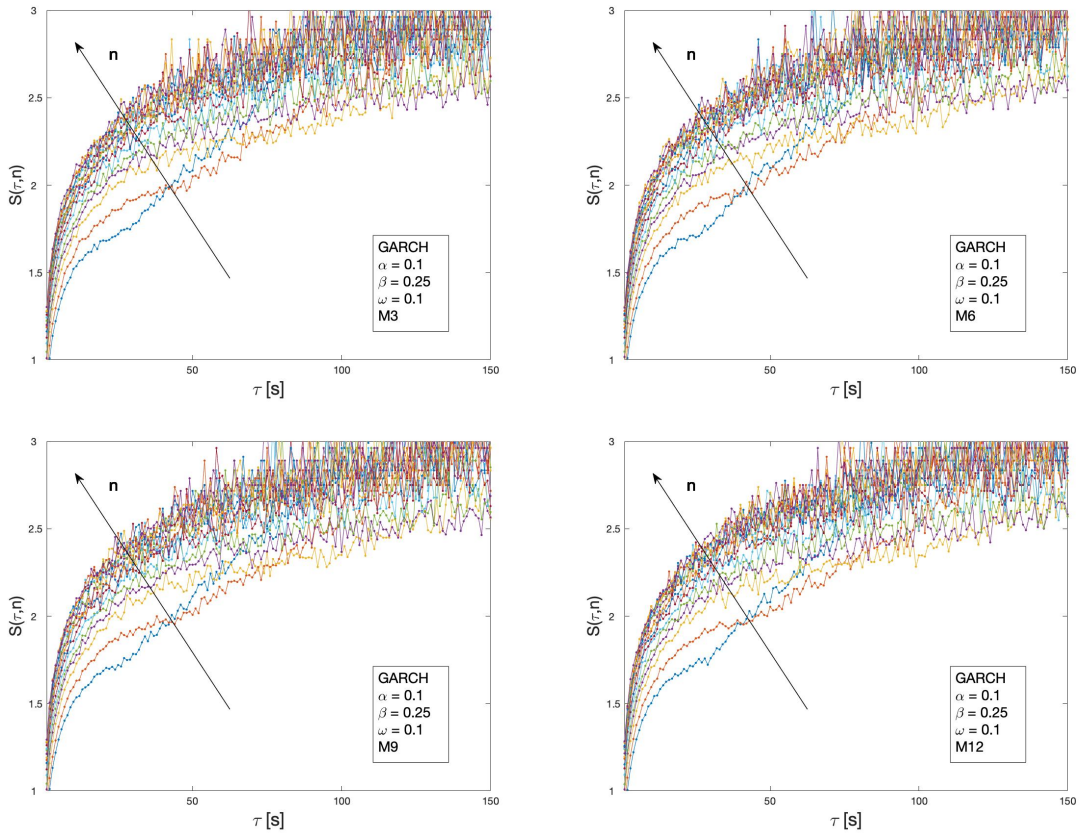


Figure 35: Cluster entropy for series generated by means of GARCH processes. The series used to plot the entropy clusters are generated with the following parameters: $\omega = 0.1$, $\alpha = 0.1$ and $\beta = 0.25$. The time horizon analyzed in figures (a), (b), (c) and (d) is M3, M6, M9 and M12 respectively.

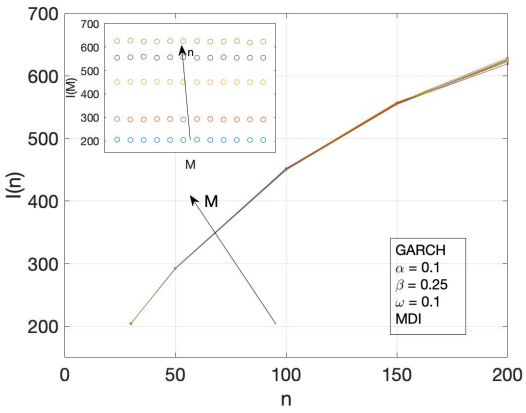


Figure 36: Market Dynamic Index for series generated by GARCH processes. The series used to plot the entropy clusters are generated with the following parameters: $\omega = 0.1$, $\alpha = 0.1$ and $\beta = 0.25$.

A.7 Market Data

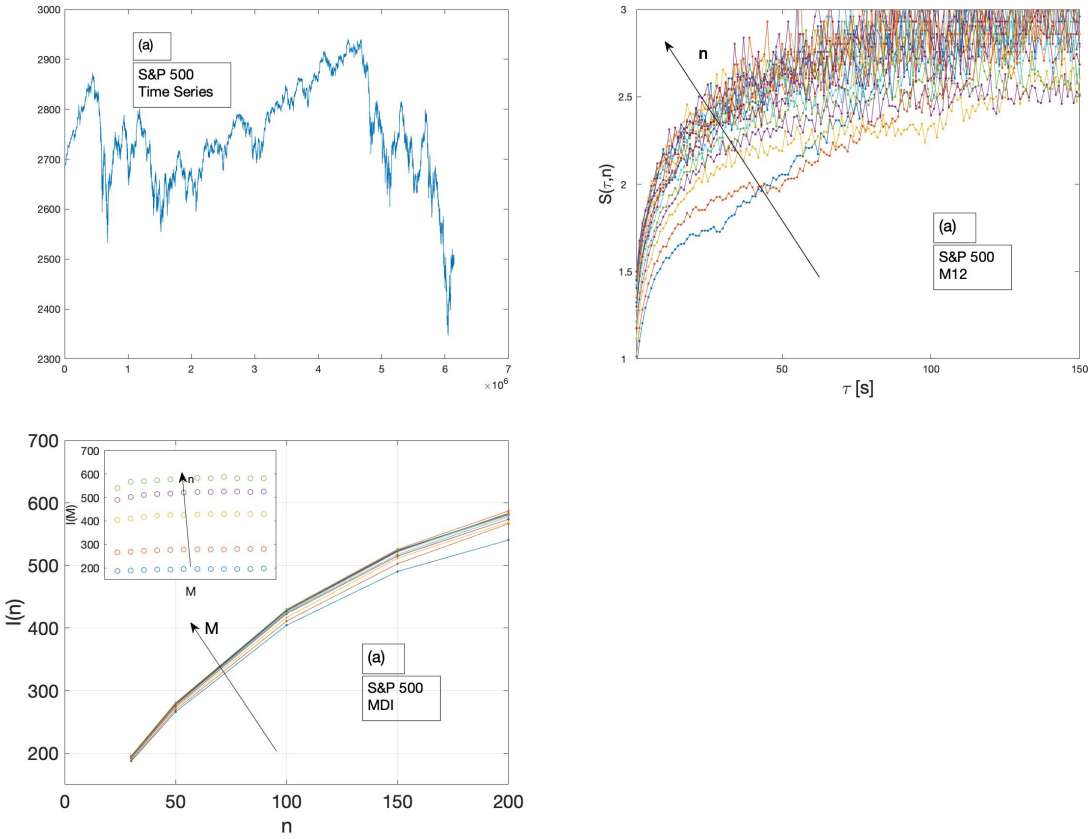


Figure 37: Time series, cluster entropy curves (M12) and Market Dynamic Index for S&P500. Results are reported from Ponta and Carbone (2019)

A.8 MATLAB Scripts

Listings

```
close
clear
tic

% %%%%%%%%%%%%%%%%%%%%%%%%I N S T R U C T I O N S %%%%%%%%%%%%%%%%%%%%%%%%
% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% 0 Check the output directory
% 1 Modify "ModelName".
% 2 Modify parameters "one", "two", "three".
% 3 Modify cell array "SeriesInformation".
% 4 Modify generating function.
% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% User Input
% First, choose the raw data to take series dimensions from (length of each
% time horizon). Second, choose the model in "ModelName".


[N RawDataLengths] = loadsamplinglengths('CCMP');
ModelName = 'FBM'
Label = '(l)'
H = 0.9

SeriesInformation{1,1} = ModelName;
SeriesInformation{2,1} = ['Hurst Exponent = ', num2str(H)];
createandsetdirectory(ModelName);
save(['/Users/pietromurialdo/BigData/Results/', ModelName '/', '
    SeriesInformation'], 'SeriesInformation')

y = fbmgen(N, H);
save(['/Users/pietromurialdo/BigData/Results/', ModelName '/', 'timeseries.
    mat'], 'y')


% To let this section work, move the series you want to draw figures of in
% the folder in [/Users/pietromurialdo/BigData/Results/, ModelName].


COMPUTE_ANALYSIS(ModelName, y, RawDataLengths);



PLOT_TIMESERIES(ModelName, Label);
for i = 1:12
    PLOT_ENTROPY_FIG(ModelName, Label, i);
end
```

```
PLOT_MDI(ModelName, '(h)');
```

Listing 1: Main Script

```
function [N, periodLengths] = loadsamplinglengths(RealSeriesName)

% the directory for real data is assumed to be constant.

inputDir = ['/Users/pietromurialdo/BigData/RawIndexData/'];

tmp = load([inputDir, RealSeriesName, 'RawLengths', '.mat']);
periodLengths = tmp.DataLength;
N = periodLengths{12};
clear tmp
end
```

Listing 2: Load Sampling Lengths

```
function [Z] = fbmgen(N, H)

seed = 123456789;
Z = fbmwoodchan(N, H, 'seed', seed);
end
```

Listing 3: FBM Generating Script

```
function COMPUTE_ANALYSIS(ModelName, TimeSeries, RawDataLengths)

%
%
%
%
%
% Definition and evential creation of destination directory. The directory
% is specific to the family model level. That means that simulation many
% combination of parameters for "GBM" will always be saved in the same
% directory. Then the user must deal with the output at each iteration. The
% model used is specifi

directory = ['/Users/pietromurialdo/BigData/Results/', ModelName, '/'];

window = [30,50,100,150,200:100:1500];


```

```
[sampledYList, SampledSeriesLength] = seriessampling(TimeSeries,
    RawDataLengths);
save([directory, 'SampledSeriesLength.mat'], 'SampledSeriesLength');

entropyMtx = cell(12,1);
for i = 1:12
    filename = [directory, 'EntropyMtx-M', num2str(i)];
    entropyMtx{i,1} = entropy(sampledYList{i}, window, filename);
end

filename = [directory, 'MarketDynIndex'];
[intS] = mdindex(entropyMtx, window, filename);

end
```

Listing 4: Main Analysis

```
function PLOT_TIMESERIES(ModelName, Label)


directory = ['/Users/pietromurialdo/BigData/Results/', ModelName, '/'];

tmp = load([directory, 'SeriesInformation.mat'])
info = tmp.SeriesInformation;
clear tmp

tmp = load([directory, 'timeseries.mat'])
TimeSeries = tmp.y;
clear tmp

length(TimeSeries)
max(TimeSeries)
min(TimeSeries)

plot(TimeSeries);

% text(0.7*length(TimeSeries), 0.5*(max(TimeSeries) + min(TimeSeries)), '
    Time Series', 'fontsize', 15);
% for i = 1:length(info)
%     text(0.7*length(TimeSeries), 0.5*(max(TimeSeries) + min(TimeSeries))
    - (0.5*i)*(max(TimeSeries) + min(TimeSeries)),...
%         info{i}, 'fontsize',15);
% end

dim = [.7, .3, .1, .1];
for i = 1:length(info)
    string{i} = info{i};
end
```

```
string{i + 1} = 'Time Series';
annotation('textbox', dim, 'String', string, 'FitBoxToText', 'on', '
    fontsize', 14);

if ischar(Label)
    dim = [.25, .6, .3, .3];
    annotation('textbox', dim, 'String', Label, 'FitBoxToText', 'on', '
    fontsize', 14);
end

print('-djpeg', [directory, Label, 'TimeSeries']);
close

end
```

Listing 5: Plot of Time Series

```
function PLOT_ENTROPY_FIG(SeriesModelName, Label, TimeHorizonIndex)

%
%
%
%
% Definition of basic variables. It is assumed a major directory containing
% all results; further specific directories are addressed in the variable
% "SeriesModelName", which specify the model used to generate the series.
% "window" is assumed constant; in case of modifications, nested functions
% and linked functions must be checked.


directory = ['/Users/pietromurialdo/BigData/Results/', SeriesModelName, '/'
    ];
window = [30,50,100,150,200:100:1500];

%
%
%
%
% Load the sampled series length used to normalize entropy curves.

tmp = load([directory, 'SampledSeriesLength.mat']);
pathsLength = tmp.SampledSeriesLength;
clear tmp

%
%
%
%
```

```
%
% Load info used to generate the artificial data. Since Directory and info
% must match, eventual mismatches are indications of mistakes.

tmp = load([directory, 'SeriesInformation.mat'])
info = tmp.SeriesInformation;
clear tmp

%
%
%
%
% Load entropy matrix according to the model specified and the index passed
% as input

tmp = load([directory, 'EntropyMtx-M', num2str(TimeHorizonIndex), '.mat'])
paths = tmp.entropyMatrix;
clear tmp

%
%
%
%
% First block of code plots entropy curves figures.
%
% Second block writes figure specification inside the plot according to
% string vector Info.
%
% Third block saves figure
%
%
% %%%%%%%%%%
% %%%%%%%%%%
%
%
%
% Plot entropy curves

indexFig = (1:length(window))';
for k = 1:length(window)
    figure(9000);
    hold on; %fissa la figura in uso in modo che tutte le modifiche future
    vengano aggiunte alla figura corrente.
    grid off; %toglie la griglia.
    box on; %chiude gli assi a inquadrettare la figura.

    plot(paths{indexFig(k)}(:,1), ...
        - log(paths{indexFig(k)}(:,2) ./ (0.75 * pathsLength)), ...
        '.-');

```

```
    fig = gca;
    fig.XScale = 'lin';
    fig.YScale = 'lin';
    fig.XLim = [1 150];
    fig.XTick = [50 100 150];
                fig.YLim = [4 12];
                fig.YTick = [4:2:12];
                fig.YTickLabels = [1:0.5:3];
%     fig.YLim = [4 13]; %%%
%     fig.YTick = [4:2:12];
%     fig.YTickLabels = [1:0.5:3];
    xlabel('\tau [s]', 'fontsize', 18)
    ylabel('S(\tau,n)', 'fontsize', 16) %Cosa significa '[a.u.]'?

    freccia = annotation('arrow');
    freccia.Position = [0.45 0.3 -0.27 0.55];
    freccia.LineWidth = 0.75;
    text(22,11,'n','fontsize', 16)
end

%
%
%
%
% Figure specifications

dim = [.7, .3, .1, .1];
for i = 1:length(info)
    string{i} = info{i};
end
string{i + 1} = ['M', num2str(TimeHorizonIndex)];
annotation('textbox', dim, 'String', string, 'FitBoxToText', 'on', '
    fontsize', 14);

if ischar(Label)
    dim = [.5, .15, .3, .3];
    annotation('textbox', dim, 'String', Label, 'FitBoxToText', 'on', '
    fontsize', 16);
end

%
%
%
%
% Save Figure


filename = [directory, SeriesModelName, 'EntropyCurvesM', num2str(
    TimeHorizonIndex)];
%hgsave(filename)
```

```
print('-djpeg',filename)

close

end
```

Listing 6: Plot of Cluster Entropy Curves

```
function PLOT_MDI(SeriesModelName, Label)



% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

directory = ['/Users/pietromurialdo/BigData/Results/', SeriesModelName, '/'
    ];
window = [30,50,100,150,200:100:1500];


% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


tmp = load([directory, 'SeriesInformation.mat']);
info = tmp.SeriesInformation;
clear tmp

% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

tmp = load([directory, 'MarketDynIndex.mat']);
integral = tmp.mdi{1,1}
clear tmp


% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%




windowIndex=[];
newWindow = [30,50,100,150,200];
fontsize2 = 18;
font_sz = 18;

for i = 1:length(newWindow)
    tmpIndex = find(window == newWindow(i));
    windowIndex = [windowIndex; tmpIndex];
```

```
    clear tmpIndex
end


% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
%
%
%


figure;
hold on; box on; grid on;

for i = 1:12
    plot(newWindow, - integral(i,:), '.-');
end

ylabel('I(n)' ,'fontsize', fontsize2);
xlabel('n' ,'fontsize', fontsize2);
set(gca, 'xlim', [0 200]);
set(gca, 'ylim', [150 700]);
set(gca, 'fontsize', font_sz);
ta = annotation('arrow');
ta.Position = [0.5 0.2 -0.15 0.3];
ta.LineWidth = 0.5;
text(63,400,'M','fontsize',16);
%figlegend(SeriesModelName, 125, 250);

% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
%
%
% Figure specifications

dim = [.7, .3, .1, .1];
for i = 1:length(info)
    string{i} = info{i};
end
string{i + 1} = 'MDI';
annotation('textbox', dim, 'String', string, 'FitBoxToText', 'on', '
    fontsize', 14);

if ischar(Label)
    dim = [.5, .3, .3, .3];
    annotation('textbox', dim, 'String', Label, 'FitBoxToText', 'on', '
    fontsize', 16);
end

% create smaller axes in top right, and plot on it
```

```
axes('Position',[0.2 .6 .3 .3]);
box on
hold on

for i = 1:length(windowIndex)
    plot([1:1:12], - integral(1:12, i), 'o', 'MarkerSize', 5);
end

xticks([]);
ylabel('I(M)','fontsize',11);
xlabel('M','fontsize',11);
set(gca,'fontsize',11);
set(gca,'xlim',[0 13]);
set(gca, 'ylim', [150 700]);
set(gca,'Color',[255/256 255/256 239/256]);
ta1 = annotation('arrow');
ta1.Position = [0.35 0.63 -0.015 0.22];
ta1.LineWidth = 0.5;
text(6.3,580,'n','fontsize',11);




% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
%
%
%


nomefile2 = [directory, 'MDI'];

%hgsave(nomefile2)
print('-djpeg', nomefile2)

close

end
```

Listing 7: Plot of Market Dynamic Index

```
function [yList, minSampledLength] = seriessampling(oneSeries, seriesLength
    )

for i = 1:length(seriesLength)
    len(i,1) = seriesLength{i};
    newSeries{i} = oneSeries(1:seriesLength{i});
```

```
end

minLength = min(len);
for i = 1:length(len)
    samplingFreq{i} = round(len(i)/minLength);
    sampledSeries{i} = newSeries{i}(1:samplingFreq{i}:end);
    newLength(i) = length(sampledSeries{i});
end

minSampledLength = min(newLength);
clear newLength
for i = 1:12
    yList{i,1} = sampledSeries{i}(1:minSampledLength);
end
end
```

Listing 8: Series Sampling

```
function [entropyMatrix] = entropy(y, movAvgWindow, filename)

%la funzione lavora su una serie alla volta e restituisce una matrice nnx2,
%dove nn indica il numero di finestre di media mobile. la prima colonna
%indica la dimensione del cluster tau e la seconda la freq assoluta
%corrispondente.

N = length(y);
entropyMatrix = cell(length(movAvgWindow), 1);

for ii=1:3
    DMA_type = ii;
    [y_tilde_vect_cell, sigma_2_DMA_vect, sigma_DMA_vect] = DMA(y,
    movAvgWindow, N, DMA_type);
    [ProbSomma] = computeClusterProbability(y, y_tilde_vect_cell, 1);
    ProbSommaAllDMA{:,ii} = ProbSomma'; % ogni colonna contiene la matrice
    ProbSomma, che    una nnx2, per un diverso tipo di media mobile.
    clear ProbSomma
end

for kkk = 1:length(movAvgWindow)
    vtemp = [ProbSommaAllDMA{1}{kkk};ProbSommaAllDMA{2}{kkk};
    ProbSommaAllDMA{3}{kkk}];
    tau_Entropia =unique(vtemp(:,1));
    for kk=1:length(tau_Entropia)
        idxE = find(vtemp(:,1)==tau_Entropia(kk));
        if(length(idxE)==3)
            Entropia(kk) = sum(vtemp(idxE,2))/3;
        elseif(length(idxE)==2)
            Entropia(kk) = sum(vtemp(idxE,2))/2;
```

```
        else
            Entropia(kk) = vtemp(idxE,2);
        end
    end
    entropyMatrix{kkk,1} = [tau_Entropia,Entropia'];
    clear vtemp tau_Entropia Entropia

end

save([filename, '.mat'], 'entropyMatrix');

end
```

Listing 9: Entropy Calculation

```
function [integralS, integralD, integralTot] = mdindex(entropyMtx,
    movAvgWindow, filename)

%ATTENZIONE! Le linee commentate riguardano il calcolo di intergrali nei
%casi di cluster di dimensione maggiore della finestra di MA usata per
%calcolarli. Es. Dopo il calcolo dell'entropia, si ottiene una cell array
%dove ogni cella contiene dati riferiti a una data finestra di MA. Ognuna
%di queste celle contiene due colonne, la prima con la dimensione del
%cluster, la seconda con la freq assoluta del cluster stesso. Per certe
%serie,    possibile che non siano stati trovati cluster di dimensione
%maggiore o uguale a quella della finestra di MA usata per individuarli.
%Pertanto, il calcolo degli integrali per cluster di dimensioni maggiori
%della finestra di MA usata per individuarli, d   errore quando questi
%cluster non esistono.


windowIndex=[];
newWindow = [30,50,100,150,200];

for i = 1:length(newWindow)
    tmpIndex = find(movAvgWindow == newWindow(i));
    windowIndex = [windowIndex; tmpIndex];
    clear tmpIndex
end

for i = 1:length(entropyMtx)

    for ii = 1:length(newWindow)
        entropyMtx{i}{windowIndex(ii)}(:,1);
        idxff_s = find(entropyMtx{i}{windowIndex(ii)}(:,1) <= newWindow(ii)
    ); % seleziono cluster di dim. minore di newWindow(ii)
        %         idxff_d = find(entropyMtx{i}{windowIndex(ii)}(:,1) >
    newWindow(ii));
```

```
31
32          dati_entropia{i,ii} = [entropyMtx{i}{windowIndex(ii)}(:,1), -log(
    entropyMtx{i}{windowIndex(ii)}(:,2))];
33
34          tmp_s = cumsum(dati_entropia{i,ii}(idxff_s,2));
35          %           tmp_d = cumsum(dati_entropia{i,ii}(idxff_d,2));
36          %           tmp_tot = cumsum(dati_entropia{i,ii}(:,2));
37          integralS(i,ii)=tmp_s(end);
38          %           integralD(i,ii)=tmp_d(end);
39          %           integralTot(i,ii)=tmp_tot(end);
40      end
41  end
42  mdi{1,1} = integralS;
43  % mdi{1,2} = integralD;
44  % mdi{1,3} = integralTot;
45  save([filename, '.mat'], 'mdi');
46  end
```

Listing 10: MDI Calculation

```
1  function[y_tilde_vect_cell, sigma_2_DMA_vect, sigma_DMA_vect] = DMA(y,nv,N,
       DMA_type)
2
3  if(DMA_type==1)
4      [y_tilde_vect_cell, sigma_2_DMA_vect, sigma_DMA_vect] = DMA_backward(y,
       nv,N);
5  elseif(DMA_type==2)
6      [y_tilde_vect_cell, sigma_2_DMA_vect, sigma_DMA_vect] = DMA_centered(y,
       nv,N);
7  else
8      [y_tilde_vect_cell, sigma_2_DMA_vect, sigma_DMA_vect] = DMA_forward(y,
       nv,N);
9  end
10 end
```

Listing 11: DMA Main Script

```
1  function [y_tilde_vect_cell, sigma_2_DMA_vect, sigma_DMA_vect] =
       DMA_backward(y,nv,N)
2
3  sigma_DMA_vect=[];
4
5  sigma_2_DMA_vect=[];
6  N = length(y);
7  c=0;
8
```

```
for ni = 1:length(nv)
    n=nv(ni);
    y_tilde_vect=[];

    for i = n+1:length(y)

        y_tilde = (sum(y(i-n+1:i)))/n;

        y_tilde_vect = [y_tilde_vect;y_tilde];

    end

    c=c+1;
    y_tilde_vect_cell{c}=y_tilde_vect;

    sigma_2_DMA = sum((y(n+1:end)-y_tilde_vect(1:end)).^2)/(N-n);

    sigma_2_DMA_vect = [sigma_2_DMA_vect;sigma_2_DMA];

    sigma_DMA_vect = [sigma_DMA_vect;sqrt(sigma_2_DMA)];

    clear y_tilde y_tilde_vect sigma_2_DMA
end
```

Listing 12: Backward DMA

```
function [y_tilde_centrale_vect_cell, sigma_2_DMA_vect_centrale,
    sigma_DMA_vect_centrale] = DMA_centered(y,nv,N)

sigma_2_DMA_vect_centrale=[];
sigma_DMA_vect_centrale=[];
c=0;

for ni = 1:length(nv)
    n=nv(ni);

    y_tilde_centrale_vect=[];

    if(mod(n,2)==0)

        start_i = n/2+1;

        end_i=length(y)-(n/2);

    else
        start_i = (n-1)/2+1;
        end_i=length(y)-((n-1)/2);
    end
```

```
22
23      for i= start_i:end_i
24          if(mod(n,2)==0)
25              y_tilde_centrale_t = [0.5 *(y(i-(n/2)))+0.5 *(y(i+(n/2)))+ sum(
    y((i-((n-2)/2)):(i+((n-2)/2))))];
26
27              y_tilde_centrale = y_tilde_centrale_t/n;
28          else
29              y_tilde_centrale_t = sum(y((i-((n-1)/2)):(i+((n-1)/2))));
30              y_tilde_centrale = y_tilde_centrale_t/n;
31          end
32
33
34          y_tilde_centrale_vect =[y_tilde_centrale_vect;y_tilde_centrale];
35
36      end
37      c=c+1;
38
39      y_tilde_centrale_vect_cell{c}=y_tilde_centrale_vect;
40
41      if(mod(n,2)==0)
42          sigma_2_DMA_centrale = sum((y((n/2)+1:end-((n/2)))-
    y_tilde_centrale_vect(1:end)).^2)/(N-n);
43      else
44          sigma_2_DMA_centrale = sum((y((n-1)/2+1:end-((n-1)/2))-
    y_tilde_centrale_vect(1:end)).^2)/(N-n);
45      end
46
47
48
49
50      sigma_2_DMA_vect_centrale = [sigma_2_DMA_vect_centrale;
    sigma_2_DMA_centrale];
51
52
53      sigma_DMA_vect_centrale = [sigma_DMA_vect_centrale;sqrt(
    sigma_DMA_vect_centrale)];
54
55
56      clear y_tilde_centrale y_tilde_centrale_vect sigma_2_DMA_centrale
    y_tilde_centrale_t
57  end
```

Listing 13: Centered DMA

```
1   function [y_tilde_destra_vect_cell, sigma_2_DMA_vect_destra,
    sigma_DMA_vect_destra] = DMA_forward(y,nv,N)
2
```

```
3
4  sigma_2_DMA_vect_destra=[];
5  sigma_DMA_vect_destra = [];
6  c=0;
7
8  for ni = 1:length(nv)
9      n=nv(ni);
10
11     y_tilde_destra_vect=[];
12
13     start_i = 1;
14
15     end_i=length(y)-n;
16
17     for i= start_i:end_i
18         y_tilde_destra = (sum(y(i:(i+n-1))))/n;
19
20
21         y_tilde_destra_vect =[y_tilde_destra_vect;y_tilde_destra];
22
23     end
24     c=c+1;
25
26     y_tilde_destra_vect_cell{c}=y_tilde_destra_vect;
27
28
29     sigma_2_DMA_destra = sum((y(1:end-n)-y_tilde_destra_vect(1:end)).^2)/(N
       -n);
30
31     sigma_2_DMA_vect_destra = [sigma_2_DMA_vect_destra;sigma_2_DMA_destra];
32
33
34     sigma_DMA_vect_destra = [sigma_DMA_vect_destra;sqrt(
       sigma_DMA_vect_destra)];
35     clear y_tilde_destra y_tilde_destra_vect sigma_2_DMA_destra
36 end
```

Listing 14: Forward DMA

```
1  function [ProbSomma] = computeClusterProbability(serie,y_media_mobile_cell,
       iii)
2
3  for k=1:length(y_media_mobile_cell) %lunghezza pari al numero di finestre
       di MA
4
5      Prezzo_MediaMobile = (serie(end-length(y_media_mobile_cell{k})+1:end)-
       y_media_mobile_cell{k}); %differenza tra serie reale e sua media mobile
6
```

```
7       segno_p = sign(Prezzo_MediaMobile(1)); %segno della prima diff
8
9       count = 1;
10      durata = [];
11      for x=2:length(Prezzo_MediaMobile)
12          segno_c= sign(Prezzo_MediaMobile(x));
13          if(segno_p == segno_c)
14              count = count +1;
15          else
16              durata = [durata; count];
17              count = 1;
18              segno_p = segno_c;
19          end
20      end
21
22
23      tau = unique(durata); %lista di tutte le durate possibili, ordinate in
        crescente
24      if ~isempty(tau)
25          for i =1:length(tau)
26              ripetizioni(i) = length(find(durata==tau(i))); %per ogni durata
        , restituisce il numero di ripetizioni
27          end
28
29          probabilita{k,iii}=[tau,ripetizioni']; %k scorre con la
        corrispondete finestra di MA
30          clear tau durata ripetizioni Prezzo_MediaMobile
31
32
33      end
34  end
35
36
37  %'size della matrice probabilit :'
38  a=size(probabilita);
39  for n = 1:a(1) %per ogni riga della matrice probabilit ; ossia per ogni
        finestra di ma.
40      tau_tot=[];
41      ripetizioni_tot=[];
42      for j =1:a(2) %solo 1 in teoria
43          if ~isempty(probabilita{n,j})
44              tau_t =  probabilita{n,j}(:,1);
45              ripetizioni_t = probabilita{n,j}(:,2);
46
47              for i = 1:length(tau_t)
48                  idx = find(tau_tot==tau_t(i));
49
50                  if isempty(idx)
51                      tau_tot=[tau_tot;tau_t(i)];
52                      ripetizioni_tot= [ripetizioni_tot;ripetizioni_t(i)];
```

```
                else
                    ripetizioni_tot(idx) =ripetizioni_tot(idx)+
    ripetizioni_t(i);
                end
            end
        end
    end

    [tau_tot_ord I] = sort(tau_tot);
    ripetizioni_tot_ord= ripetizioni_tot(I);

    ProbSomma{n}= [tau_tot_ord,ripetizioni_tot_ord]; % dove n    l'indice
    delle finestre di MA.

    clear tau_tot ripetizioni_tot tau_t ripetizioni_t ripetizioni_tot_ord
    tau_tot_ord
end
```

Listing 15: Cluster Probability Computation

References

J. Bryant. *Entropy Man.* VOCAT International, 2015.

A. Carbone. Information measure for long-range correlated sequences: the case of the 24 human chromosomes. *Scientific Reports*, 2013.

A. Carbone, G. Kaniadakis, and A. M. Scarfone. Tails and ties. *The European Physical Journal B*, 57(2):121–125, 2007.

J. Franke, W. K. Härdle, and C. M. Hafner. *Statistics of Financial Markets: An Introduction.* Springer, 2015.

B. B. Mandelbrot and J. W. Van Ness. Fractional brownian motion, fractional noises and applications. *SIAM review*, 1968.

L. Ponta and A. Carbone. Information measure for financial time series: quantifying short-term market heterogeneity. *Physica A*, 2018.

L. Ponta and A. Carbone. Quantifying horizon dependence of asset prices: a cluster entropy approach. *arXiv preprint arXiv:1908.00257, 2019*, 2019.

C. Shannon. A mathematical theory of communication. *Bell. Syst. Tech. J.*, 1948.