

Family comes in all shapes and sizes.

— The Family Book, Todd Parr

To Ana

(and Matilde and João)

with endless Love...

ABSTRACT

In this work we consider the application of a plethora of Econophysics techniques to multivariate financial time series, particularly the Correlation matrix, the Forecastable Component Analysis, the Mutual Information, the Kullback-Leibler Divergence, the Approximate Entropy, the Distance Correlation and the Hurst exponent. The key idea was not to compare their differences but more to find their “joint strength” by combining their different views of time series. We applied these techniques to two different scenarios: one, more local, to 12 stocks quoted in the Portuguese Stock Market (PSI-20); the other one, more global, to 23 world stock markets. Also, we have studied and used “sliding windows” of different sizes. The motivation and importance of this kind of analysis relies on the well known multi-fractal behaviour that financial data exhibits.

We started by confirming some results found in literature, namely the ones from random matrix theory and the ones for the Hurst exponent. In this case, and based in previous results, we propose that the PSI-20 is becoming more mature. Distance correlation have shown to be a good complement to entropy measures like Mutual Information or Kullback-Leibler divergence. Approximate entropy, as a stand alone method, have shown potential complementarity with Distance correlation in the case of the stocks from PSI-20 index.

To our knowledge, it is the first time that energy statistics is applied to the PSI-20 data. It is interesting to note that this measure, and this is corroborated by Approximate entropy results, proposes two well defined behaviour for the PSI-20 stocks. One period, from 2000 to 2007, relatively calm, with low variation of Distance Correlation between stocks, and another period, from 2007 till now, much more agitated in what concerns this measure.

Unfortunately, we cannot say the same for the Distance Correlation results applied to the World Markets set. Nevertheless, we can find strong regional correlation for most of the markets. Some, but only a few, can be considered more global markets, with influence in all the others. There is, in that sense, a strong connection between the North-American markets and most of the European ones. That correlation has become higher since 2007, complementing the idea that the markets are more connected.

For Mutual Information or Kullback-Leibler Divergence the results are very sharp and we can clearly match high entropy values with real events. Some of them are only important for specific stocks or markets, but some others, more related to recession periods, are independent of a specific stock or market.

In general, a trend common to most markets is the progressive growing correlation over time. One possible reason to this is the progressive globalisation of markets, where the arbitrage opportunities are reduced due to more efficient markets. Also, the information we got from Hurst exponent was vital to confirm that stocks and markets are getting more and more mature, that is, less autocorrelated.

RESUMO

Neste trabalho consideramos a aplicação de algumas técnicas da Econofísica às séries financeiras temporais multivariadas, nomeadamente consideramos as técnicas das matrizes aleatórias como a matriz de correlação, as técnicas da análise de componentes, da informação mútua, da divergência de Kullback-Leibler, da entropia aproximada, da distância de correlação e do expoente de Hurst. A ideia fundamental não foi comparar as suas diferenças mas sim encontrar as suas “forças conjuntas” ao combinar a forma como cada técnica “vê” as séries temporais. Estas técnicas foram aplicadas em dois cenários distintos: um, mais local, a 12 ações cotadas no PSI-20, o índice da Bolsa portuguesa; o outro, mais global, foi aplicado a 23 mercados de diferentes países. Ainda, usou-se aqui uma técnica de cálculo por “janelas” temporais dado o conhecido comportamento multifractal dos dados financeiros.

Começamos por confirmar os resultados conhecidos da literatura para as matrizes aleatórias e para o expoente de Hurst. Neste último caso, e baseados nos resultados anteriores, propomos que o PSI-20 está a tornar-se um mercado mais maduro. A Distância de Correlação provou ser uma medida com boa complementaridade com medidas de entropia como a Informação Mútua ou a divergência de Kullback-Leibler. A Entropia Aproximada, por si só, mostrou uma boa complementaridade com a Distância de Correlação na aplicação às ações do PSI-20.

Que tenhamos conhecimento, é a primeira vez que a Distância de Correlação é aplicada ao PSI-20. É interessante notar que esta medida, e isto é corroborado pelos resultados da Entropia Aproximada, propõe dois períodos comportamentais bem definidos: um, de 2000 a 2007, com pequenas variações e valores também pequenos e outro, com grandes variações e com valores muito elevados de correlação entre as ações do PSI-20.

Contudo, esta observação não permanece quando aplicamos a mesma medida aos mercados mundiais. Todavia, encontramos correlações regionais fortes para a maior parte dos mercados. Alguns mercados, embora poucos, podem ser vistos como globais já que influenciam todos os outros. Neste sentido, é de referir a forte ligação dos mercados norte-americanos com os mercados europeus. Esta correlação continua a crescer desde 2007, ajudando a complementar a ideia de que os mercados estão mais ligados.

Para a Informação Mútua ou para a divergência de Kullback-Leibler os resultados são muito claros. Conseguimos ligar os valores mais elevados da entropia a acontecimentos reais. Uns, mais restritos, e portanto, influenciando apenas ações ou mercados pontuais; outros, mais globais, deixando a sua marca em todas as ações/mercados.

Em geral, uma tendência comum a todos os mercados é o aumento gradual temporal da correlação. Uma possível razão pode ter a ver com a progressiva globalização dos mercados, onde as oportunidades de arbitragem estão reduzidas devido ao facto dos mercados serem cada vez mais eficientes. A informação que obtivemos a partir do expoente de Hurst foi vital para confirmar a informação de que os mercados estão cada vez mais maduros, isto é, menos autocorrelacionados.

ACKNOWLEDGEMENTS

I owe, firstly, many thanks to my advisor, José Abílio Oliveira Matos, for being so helpful, patience, dedicated and committed to this project. Most of the time that I was lost, he was there to keep us up, was not his motto “Be Prepared”!

In second place I wish to thank my family, my teachers and some friends, not necessarily by this order of importance::

- To the scouts from my Group in Guimarães (an endless list started by Alexandre, Ernesto, Manel, Miguel and Samuel) for, most of the times without knowing, keeping me up;
- To Ricardo Gama for his friendship, even at distance, from the times since the Master degree;
- To some of my teachers, particularly Prof. Eduardo Laje and my master thesis advisor, Prof. Silvio Gama, from whom, without no pain, I got some of the most important lessons in my life;
- To my colleagues from IPG, particularly A. Martins, C. Rosa, J.C. Miranda, P. Costa and P. Vieira, for helping me to keep up my scientific motivation, for, at some times, their hospitality or for, at other times, just sharing meals and/or coffees;
- To my nephew and nieces, particularly my godsons Francisca and Dinis, but also Beatriz and Carolina, for their joy and life;
- To my grandfather, António Augusto Cordeiro Rodrigues, for reminding me all the time to accomplish this purpose;
- To my parents, Sr. Salgado and D. Conceição, and my mother-in-law, D. Isabel, for their continuous love, concern, support and understanding;
- To my beloved Ana, Matilde and João, for being unique and precious, for their love, joy, patience and... for everything!, and without whom all this effort would seem totally senseless.

CONTENTS

1	INTRODUCTION	1
1.1	Motivation	1
1.2	Econophysics	1
1.2.1	Brief history	2
1.2.2	Why Econophysics?	4
1.2.3	Current Econophysics efforts	5
1.3	Objectives	6
1.4	Contributions	6
1.5	Thesis Outline	7
2	DEFINITIONS AND BACKGROUND	9
2.1	Setting the Stage	9
2.1.1	Data and models	9
2.1.2	Financial time series analysis	10
2.1.3	Random Walk Hypothesis and the Brownian Motion	11
2.1.4	Stylized empirical facts	12
2.1.5	Market Crashes or “When things go terribly wrong”	14
2.2	Stochastic Processes	19
2.2.1	Random variables	19
2.2.2	Stochastic processes	20
2.3	Random Matrix Theory	21
2.3.1	Returns statistics	22
2.3.2	The correlation matrix	23
2.3.3	Eigenvalues and eigenvectors	24
2.4	Component Analysis	29
2.4.1	Principal Component Analysis	29
2.4.2	Independent Component Analysis	30
2.4.3	Forecastable Component Analysis (ForeCA)	32
2.5	Entropy	33
2.5.1	Definition	34
2.5.2	Entropy different incantations	35
2.5.3	Mutual Information	37
2.5.4	Kullback-Leibler Divergence	37
2.5.5	Approximate Entropy	38
2.6	Energy Statistics	39
2.6.1	Definitions	40
2.6.2	Properties	42
2.6.3	Brownian Covariance	43
2.7	Fractional Brownian Motion	44
2.8	Other Methods	46
2.9	Methodologies	47
2.9.1	Data Analysis Methodology	47
2.9.2	Computational Methodology	48

3	DATA	51
3.1	Data Considerations	51
3.2	Data Sets	52
3.2.1	PSI-20 set	52
3.2.2	World Markets set	54
3.3	Events of interest	55
4	PORTUGUESE STANDARD INDEX (PSI-20) ANALYSIS	57
4.1	PSI-20 Index	57
4.1.1	PSI-20 evolution	57
4.1.2	A random PSI-20	58
4.2	Dynamic analysis of PSI-20 using sliding windows	59
4.2.1	Step size decision	59
4.2.2	Window size decision	60
4.3	Results	63
4.3.1	Random Matrix	63
4.3.2	Component Analysis	66
4.3.3	Entropy	69
4.3.4	Distance Correlation	71
4.3.5	Hurst Exponent	73
4.4	Concluding Remarks	75
5	WORLD MARKETS ANALYSIS	77
5.1	Introduction	77
5.2	Results	77
5.2.1	Random Matrix	77
5.2.2	Component Analysis	80
5.2.3	Entropy	83
5.2.4	Distance Correlation	86
5.2.5	Hurst Exponent	97
5.3	Concluding Remarks	99
6	CONCLUSIONS AND FUTURE WORK	101
6.1	Conclusions	101
6.2	Future work	103
A	DATA	105
A.1	PSI-20 Stocks	106
A.2	Markets	118
B	CATALOGUE OF RESULTS	141
B.1	Markets Index versus Crisis Dates	142
B.2	Distance Correlation for PSI-20	145
C	PACKAGE DESCRIPTION	149
C.1	Hash	149
C.2	PerformanceAnalytics	149
C.3	Zoo	150
C.4	Pracma	150
C.5	Energy	151
C.6	Lattice	151

c.7	Xts	152
c.8	xtsExtra	152
c.9	entropy	152
c.10	ForeCA	153
D	SOFTWARE	155
D.1	Markets Matrix code	155
D.2	Returns code	156
D.3	Eigenvalues code	157
D.4	Approximate Entropy code	159
D.5	Distance Correlation code	160
D.6	Plots code	161
D.7	Kullback-Leibler Divergence code	164
D.8	Mutual Information code	165
D.9	ForeCa code	166
D.10	Marchenko-Pastur code	166
	BIBLIOGRAPHY	169

LIST OF FIGURES

Figure 1	NBER Recession dates	17
Figure 2	Alternative recession dates	18
Figure 3	Schematic representation of ICA	31
Figure 4	PSI-20 from 2000 to 2014	57
Figure 5	Real vs Random PSI-20 returns.	58
Figure 6	Real versus Random PSI-20 close values	58
Figure 7	PSI-20 returns time series and their distribution.	59
Figure 8	Distance Correlation values for different steps	60
Figure 9	DCor values for different “sliding” windows size	61
Figure 10	Markets DCor values for different “sliding” windows size	61
Figure 11	Markets ApEn values for different “sliding” windows size	62
Figure 12	Theoretical versus Real stocks eigenvalues density	63
Figure 13	Evolution of stocks eigenvalues ratio	65
Figure 14	Evolution of stocks weighted eigenvalues ratio	66
Figure 15	ForeCA stocks components	67
Figure 16	ForeCA stocks global results	68
Figure 17	MI for PSI-20 stock pairs	69
Figure 18	KL_{Div} for PSI-20 stock pairs	70
Figure 19	$ApEn$ for PSI-20 stocks	71
Figure 20	$DCov$ for PSI-20 stock pairs	72
Figure 21	$DCov$ for PSI-20 stock pairs	72
Figure 22	PSI-20 fluctuation function	73
Figure 23	Hurst exponent for PSI-20 stocks	74
Figure 24	Theoretical versus Real eigenvalues densities	78
Figure 25	World Markets Ratio λ_1/λ_3 versus λ_1/λ_2	78
Figure 26	Real vs Weighted Eigenvalues Ratios	79
Figure 27	Real vs Random Eigenvalues Ratios	79
Figure 28	ForeCA world markets Components	81
Figure 29	ForeCA global world markets results	82
Figure 30	MI for World markets pairs	83
Figure 31	KL_{Div} for World markets pairs	84
Figure 32	Approximate Entropy for European markets	85
Figure 33	Approximate Entropy for non-European markets	85
Figure 34	Distance Correlation for the ASX_HSI pair	86
Figure 35	Distance Correlation for the BSES_N_HSI pair	86
Figure 36	Distance Correlation for the HSI_NIK pair	87
Figure 37	Distance Correlation for the KOSPI_NIK pair	87
Figure 38	Distance Correlation for the AEX_ATX pair (60 days window width)	88
Figure 39	Distance Correlation for the AEX_STOXX pair	88
Figure 40	Distance Correlation for the ATX_IBEX pair	89
Figure 41	Distance Correlation for the ATX_PSI pair	89

Figure 42	Distance Correlation for the ATX_STOXX pair	90
Figure 43	Distance Correlation for the CAC_STOXX pair	90
Figure 44	Distance Correlation for the CAC_DJI pair	90
Figure 45	Distance Correlation for the DAX_IBEX pair	91
Figure 46	Distance Correlation for the DAX_SPY pair	91
Figure 47	Distance Correlation for the FTSE_PSI pair	92
Figure 48	Distance Correlation for the FTSE_MIB pair	92
Figure 49	Distance Correlation for the FTSE_MERVAL pair	93
Figure 50	Distance Correlation for the BVSP_MERVAL pair	94
Figure 51	Distance Correlation for the MERVAL_MXX pair	94
Figure 52	Distance Correlation for the DJI_FTSE pair	95
Figure 53	Distance Correlation for the DJI_IXIC pair	95
Figure 54	Distance Correlation for the IXIC_MXX pair	96
Figure 55	Distance Correlation for the SPY_STOXX pair	96
Figure 56	Hurst exponent for European markets	97

LIST OF TABLES

Table 1	Major XX century events for global markets.	14
Table 2	Major XXI century events for global markets.	15
Table 3	PSI-20 set business sectors	52
Table 4	PSI-20 set top-ten classification	53
Table 5	PSI-20 stock splits	53
Table 6	World Markets Set	54
Table 7	PSI-20 Set Correlation Matrix	64
Table 8	Descriptive statistics for stocks eigenvalues ratio	65
Table 9	ForeCA stocks results	66
Table 10	Hurst exponent for PSI-20 stocks	74
Table 11	ForeCA world markets results	80
Table 12	Hurst exponent for world markets	98

LISTINGS

Listing 1	Markets Matrix calculation code	155
Listing 2	Returns calculation code	156
Listing 3	Eigenvalues calculation code	157
Listing 4	Approximate Entropy calculation code	159
Listing 5	Distance Correlation calculation code	160

Listing 6	Plots representation code	161
Listing 7	Kullback-Leibler Divergence calculation code	164
Listing 8	Mutual Information calculation code	165
Listing 9	Forecastable Component Analysis calculation code	166
Listing 10	Marchenko-Pastur calculation code	167

INTRODUCTION

“Le marché, à son insu, obéit à une loi qui le domine: la loi de la probabilité.”¹
(Bachelier, Théorie de la spéculation)

Recent turmoil in world’s economy, and more particularly in Europe, brought back the feeling of tragedy to our lives and raised more questions than we can help out to answer. It is now clear, at least for some rational minds, that there is an emergency to understand the “laws” beneath financial markets, our new “lords”.

This introductory Chapter presents the motivation to study this subject and a brief introduction, a framework and an historical perspective of Econophysics.

1.1 MOTIVATION

Newton, after loosing 20000£ (twenty thousand British Pounds) on the “South Sea Bubble”, said that it was more difficult to model the madness of people than the motion of planets. This statement remains probably true after 200 years. And, if being true, is the search for better modelling of the economy and finance fields the answer to Newton’s anger?

To answer this question we must, firstly, ask the right questions. What drives, for instance, the movements of a financial time series?

There are several possible answers to this question. Physicists and mathematicians can work with empirical data and construct phenomenological theories. The quantitative nature of pure sciences allows a degree of abstraction when analysing series of numbers. One other answer is that Statistical Physics and Applied Mathematics have useful approaches to deal with collective dynamics in systems. These can be seen in such areas as biomedical signals, earthquakes, networks, traffic or river flow analysis, amongst others. One last possible answer is that we believe that it is possible to go through economical and financial questions using some of the well established ideas of mathematics and physics.

But, what can we learn from other fields of science that can help us to achieve a broader understanding of the questions in other scientific fields? Can, as to say, the atomic nucleus or the laws of nature, in some sense, be of some help to understand the stock markets?

This is, in a broader sense, the framework that moved our attention to the financial time series subject.

1.2 ECONOPHYSICS

Although interest in economic and financial subjects is as old as natural sciences studies, only in the last twenty years a respectable quantity of physicists and mathematicians

¹ The market, without knowing it, obeys a law which overwhelms it: the law of probability.

have driven their attention to economic and financial subjects. This has given birth to a new page in the book of Nature called “Econophysics”. This neologism, after the words “Economics” and “Physics”, was first introduced by H. E. Stanley in his talk title in a conference on Statistical Physics in Kolkata (Calcutta) in 1995 [Stanley, 1996], in an effort to put some attention on the increasing number of papers about stocks and markets written by physicists.

According to Mantegna and Stanley [2000], “the word *Econophysics* describes the present attempts of a number of physicists to model financial and economic systems using paradigms and tools borrowed from *Theoretical and Statistical Physics*”. Indeed, physicists have been applying concepts and methodologies of Statistical Physics (e.g., scaling, universality, disordered and self-organized systems) to describe such complex systems as economic or financial systems, because most approaches based on the fundamentals of Physics perceive financial/economic phenomena as complex evolving systems. This is due to the multiple interacting components exhibited by the inherent time series, like stock market indices or inflation rates.

In particular, these systems are expressed in the light of their statistical properties. In this way, their principles (microscopic models, scaling laws) are used to develop models to explain the corresponding behaviour. Econophysics is a result of a combination of methodology (from the Complex Systems theory), of numerical tools (from computational physics) and of empirical data (from economic and financial fields) [Roehner, 2004].

1.2.1 Brief history

The connection and interplay between physics and economy is about 5 hundred years old. In fact, the relationship between Physics and Economics, or in a larger view, between Physics and the Social Sciences, dates back to XVI century. Starting from Copernicus and later Halley, mostly known by their work as astronomers, who, respectively, studied the behaviour of the inflation and derived the foundations of life insurance.

Literature is full of examples of famous physicists involvement in economic or financial problems. Daniel Bernoulli introduced the idea of utility to describe people’s preferences (1738). Pierre-Simon Laplace, in his “*Essai philosophique sur les probabilités*” pointed out that events that might seem random and unpredictable in Economics can be quite predictable and can be shown to obey simple laws (1812).

The first known attempt to describe this new branch of knowledge is due to Adolphe Quetelet, who in 1835 named it “Social Physics”, when studying the existence of patterns in data sets ranging from economic to social problems, amplifying the ideas from Laplace [Roehner, 2010]. This idea was raised up again by Ettore Majorana, [Majorana, 1942], almost one hundred years later, in 1938, in his works on the analogy between statistical laws in Physics and in Social Sciences (see also, Mantegna [2005] and Mantegna [2006]).

Although Econophysics has emerged from the urge of describing economic or financial phenomena by means of applying methods from the science of Physics, it is worth to note that the first power-law ever discovered, a most commonly distribution evidenced in Physics (power-laws have received considerable attention in physics because they indicate scale free behaviour and are characteristic of critical or nonequilibrium

phenomena), was originally observed in Economics by Vilfredo Pareto [Pareto, 1897], when analysing the income distribution among the population. Pareto also found that large values in these distributions follow *universal scaling behaviour* independent of the countries considered.

Almost at the same time, Bachelier [1900] proposed the first theory of market fluctuation, five years before Einstein's famous paper on Brownian motion [Einstein, 1905], in which Einstein derived the partial differential heat/diffusion equation governing Brownian motion and estimated the size of molecules. Specifically, Bachelier gave the distribution function for the Wiener stochastic process – the stochastic process underlying Brownian motion – linking it mathematically with the diffusion equation. It is thus telling that the first theory of the Brownian motion was developed to model financial asset prices in speculative markets! These two examples illustrate that the relation between both sciences is bi-directional and not a one-way route, as one might believe, a fact that must be considered when studying this subject.

Poincaré (1854-1912), Bachelier's thesis advisor, pointed the possibility of unpredictability in a nonlinear dynamical system, establishing the foundations of the chaotic behaviour. Ironically, Poincaré, who did not appreciate Bachelier's results, made himself a large impact on real complex systems as one of the discoverers of chaotic behaviour in dynamical systems.

Jan Tinbergen, who studied physics with Paul Ehrenfest at Leiden University, won the first Nobel Prize in Economics in 1969 for having developed and applied dynamic models for the analysis of economic processes.

One of the most revolutionary development in the theory of speculative prices since Bachelier's initial work, is the Mandelbrot's hypothesis that price changes follow a Lévy stable distribution (see Nolan [2001]) rather than a Gaussian one. In fact, Mandelbrot [1963] and Fama [1965], independently, pointed out that the empirical return distributions are fundamentally different because they are fat-tailed and more peaked compared to the Normal distribution [2]. Based on daily prices in different markets, Mandelbrot and Fama found that a stable Lévy distribution served much better as a model to the empirical return distributions (see also, Koponen [1995] or Shlesinger et al. [1995] or Mantegna and Stanley [1994]). This result suggested that short-term price changes were not well-behaved since most statistical properties are not defined when the variance does not exist. Later, using more extensive data, the decay of the distribution was shown to be fast enough to provide finite second moment.

However, during the following decades, only a few physicists, such as Kadanoff in 1971 and Montroll and Badger in 1974, had an interest in research into social or economic systems [Chakarborti et al., 2011].

And one of the causes to this turn, the next major factor changing the Gaussian view of the world, was the advent and massification of computers. First, changing the speed and the range of financial transactions drastically. Second, the economies and markets started to watch each other more closely, since computer possibilities allowed for collecting exponentially more data. In this way, several non trivial couplings started to appear in economical systems, leading to nonlinearities. Nonlinear behaviour and overestimation of the Gaussian principle for fluctuations were responsible for the Black Monday Crash in 1987. That shock had, however, a positive impact visualizing the importance of the non-linear effects.

Poincaré established the foundations of the chaotic behaviour. The study of chaos turned out to be a major branch of theoretical physics (see [Mandelbrot \[1977\]](#) and [Mandelbrot \[1982\]](#)). For a beautiful and colourful presentation see [Peitgen et al. \[1992\]](#). More recently chaos theory turned to economy.

It was not until the 1990s that physicists started seriously turning to this interdisciplinary subject. Nowadays studies of chaos, self-organized criticality, cellular automata and neural networks are seriously taken into account, as economical and financial tools.

1.2.2 *Why Econophysics?*

When addressing the need for a new discipline that merges Physics and Economy two main reasons prevail:

1. The limitations of the traditional approach of Economics/Finance;
2. The advantages of the empirical method used in Physics.

In the limitations side we must include the Efficient Market Hypothesis (EMH), by [Fama \[1970\]](#), whose basis is the random walk hypothesis, with independent and identically distributed increments. Despite its popularity, this principle is strongly controversial and has been successively questioned, since it represents a idealization that can hardly be verified. It states, in simple words, that the price variation is random as a result of the activity of the traders who attempt to make profit (arbitrage opportunities); the application of their strategies induces a feedback dynamic in the market, randomising the stock-price. In fact, the idea that markets are rational, from which this theory departs, is a theoretical construction that can be easily violated.

Another example stands from the no risk-less Capital Asset Pricing Model (CAPM), by [Black and Scholes \[1973\]](#), which cannot be applied if investors differ in their expectations and if they cannot borrow limitless amount of money at the same interest rate. Also, we could include in this side the so called rationality of economic agents.

In the advantages side, we must refer that the appeal from Physics relies on the methodology frequently applied, mainly focused on an experimental basis, which makes the crucial difference between these disciplines. Physicists have learned to be suspicious about axioms and models. If empirical observation is incompatible with the model, the model must be reviewed or discarded, even if it is conceptually beautiful or mathematically convenient.

In reality, markets are not efficient, humans tend to be over-focused in the short term and blind in the long term, and errors get amplified through social pressure and herding, ultimately leading to collective irrationality, panic and crashes. Free markets can be, in this sense, actually more like bad tempered or wild markets. It would seem to be foolish to believe that the market can impose its own self-discipline.

To sum up, we may say, following [Stanley \[1999\]](#), that the interest of physicists in economic and financial fields, also coined as “statistical finance” is due to three main factors:

1. Economic fluctuations affect everybody, which means that their implications are ubiquitous;

2. Methods and concepts developed in the study of fluctuation systems might yield new results;
3. Existence of large data sets in economic/financial domain, which in some cases contains hundreds of millions of events.

1.2.3 *Current Econophysics efforts*

It has been proven that reliance on models based on incorrect axioms has clear and tremendous effects. For example, the Black-Scholes model [Black and Scholes, 1973] assumes that price changes have a Gaussian distribution, i.e. the probability of extreme events is deemed negligible. Unwarranted use of this model on stock markets led to the October 1987 crash. Ironically, it is the very use of this crash-free Black-Scholes model that “crashed” the market!

In the recent sub-prime crisis of 2008 also, the problem lay in part in the development of structured financial products that packaged sub-prime risk into seemingly respectable high-yield investments. The models used to price them were fundamentally flawed: they underestimated the probability of the multiple borrowers would default on their loans simultaneously. In other words, these models again neglected the possibility of a global crisis, even as they contributed to triggering one. Surprisingly, there is no framework in classical economics to understand wild markets, even though their existence is so obvious to the layman. Physicists, on the other hand, have developed several models allowing one to understand how small perturbations can lead to wild effects. The theory of complexity, developed in the physics literature over the last thirty years, shows that although a system may have an optimum state (such as a state of lowest energy), this is sometimes so hard to identify that the system in fact never settles there.

This three key ideas presents briefly some of the current efforts in Econophysics [Bentes, 2010]:

- Statistical characterization of the stochastic process of price changes of a financial asset: this is an active area, and attempts are ongoing to develop the most satisfactory stochastic model describing all the features encountered in empirical analyses. One important accomplishment in this area is an almost complete consensus concerning the finiteness of the second moment of price changes. This has been a long standing problem in finance, and its resolution has come about because of the renewed interest in the empirical study of financial systems.
- Development of a theoretical model that is able to encompass all the essential features of real financial markets. Several models have been proposed, and some of the main properties of the stochastic dynamics of stock price are reproduced by these models as, for example, the leptokurtic ‘fat-tailed’ non-Gaussian shape of the distribution of price differences. Parallel attempts in the modelling of financial markets have been developed by economists.
- Time correlation of a financial series. The detection of the presence of a higher-order correlation in price changes has motivated a reconsideration of some beliefs of what is termed technical analysis.

1.3 OBJECTIVES

The main objective of this work is to apply Econophysics techniques derived from Information and Random Matrix Theories in the study of financial data. The Econophysics techniques applied in this work are twofold: measures of “disorder”/complexity and measures of coherence (for a discussion of coherence and persistence in the scope of financial time series see Ausloos [2001]). The measures of “disorder” and complexity are the different forms of entropy (as defined by Shannon [1948], Rényi [1961], Theil [1967], Tsallis [1988] or Schreiber [2000]). Measures of coherence can be obtained from Random Matrix Theory such as the covariance matrix (see financial applications by Plerou et al. [2000] or Laloux et al. [2000]).

The main focus of this thesis is placed, then, on a plethora of measures for the following reasons:

1. They allow us to predict how the market indices will evolve;
2. They add to the portfolio of techniques used to study financial time series;
3. They allow us to characterise the specific features of each market index;
4. They are measures of how markets perceive risk.

Each technique captures different *nuances* of the signal evolution. The use of different tools at the same times allow us to have more confidence in the obtained results, avoiding the several pitfalls of using a single technique.

This work carries several types of analyses, from entropy to correlation matrix analysis between different stocks or markets indices. All analyses were performed on daily data from Portuguese PSI-20 stocks and on worldwide markets indices. The daily indices were used as benchmarks for the different stocks or markets studied. Only world markets indices and stock prices from Portuguese Stock Market were used but it should be noted that the same techniques are applicable to other type of financial assets data.

We hope that the combination of both families of techniques gives a complementary view of the data in order to search for early warning information and for signs of information transfer by measuring in a quantitative way the transfer of information between stocks or markets.

1.4 CONTRIBUTIONS

The main contributions of this thesis are:

1. All of the seven methods applied have shown interesting and complementary features so that we can not discard none of these methods.
2. Distance Correlation have shown to be a good complement to entropy measures like Mutual Information or Kullback-Leibler Divergence.
3. Approximate Entropy, as a stand alone method, have shown potential complementarity with Distance Correlation in the case of PSI-20 stocks.
4. Hurst Exponent results were vital to confirm that stocks and markets are getting more and more mature, that is, less autocorrelated.

1.5 THESIS OUTLINE

This thesis is organized as follows:

- Chapter 2 provides a background to some mathematical tools needed, particularly those concerned with Random Matrix Theory (RMT), their eigenvalue analysis and the calculation of the correlation coefficients as the elements of the correlation matrix; also, provides background for those tools related with component analysis like Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Forecastable Component Analysis (ForeCA) and their definition and application to financial time series, namely the entropy and mutual information concepts; finally, some background is given in relatively new tools like the Approximate Entropy and the Energy Statistics and an more old tool like the Hurst Exponent;
- Chapter 3 considers the data used in this thesis;
- Chapter 4 characterizes the PSI-20, Portuguese stock market, and applies the methods defined in Chapter 2; also, some concluding remarks are exposed;
- in Chapter 5 are applied the methods defined in Chapter 2 to a vast number of World markets indices; also, again, some concluding remarks are highlighted;
- finally, Chapter 6 draws the conclusions about the use of these methods in financial time series and propose some work to be done in future studies.

In order to keep this text clear and readable, some subjects and results, although interesting, have been placed in Appendix.

DEFINITIONS AND BACKGROUND

"A very small cause which escapes our notice determines a considerable effect that we cannot fail to see, and then we say that the effect is due to chance." - Henri Poincaré

In this chapter are presented and defined, with mathematical rigour, the tools used in this thesis. Since the main interest is the study of financial time series we start with stochastic processes, firstly developed in the scope of Statistical Physics. Following, are introduced the techniques derived from Random Matrix Theory, Component Analysis, Entropy and Information Theory and Energy Statistics. At the end of the chapter are presented the data and computational methodologies used with these techniques.

2.1 SETTING THE STAGE

Although we must take into account that human beings and particles may behave in a significantly different manner, there is an obvious temptation to create an analogy between economic phenomena (considered a result of the interaction among many heterogeneous agents) and Statistical Mechanics. So, when we talk about basic tools of Econophysics, we are talking about probabilistic and statistical methods often taken from Statistical Physics and/or from Applied Mathematics.

2.1.1 Data and models

There are, generally, two main routes to problem solving in science:

- to use a model and, from there, study the real data to infer the consequences;
- to look at the data and from there infer a model.

The approach followed in Econophysics is typically the second one, that is, to look first at the data and then to get the best model that describes it. This empirical overview of the data tends to be a first approximation to study a subject. Despite this approach, one of the implicit goals of Econophysics, is to merge these two routes and make a bridge between Econophysics and Economics: data are only useful within an interpretative framework.

As with other complex systems, economics, and especially finance has lots of data available. To analyse these data, we have to summarise and reduce them to manage their complexity. In this work we will consider equally spaced data but with one day time interval, which will be named a trading day. The frequency of data must be taken into account because of the granularity effect, that is, as we can see from the literature, measures for different scales yield different results.

2.1.2 Financial time series analysis

When studying financial time series the aim is to “understand” them with the ultimate goal to “predict” them (for a good reference on the subject follow [Tsay \[2005\]](#), or, more general, [Chatfield \[2003\]](#)). By this understanding we mean one of these two views:

- to model in a mathematical way the time series, that is to say, to represent reality using appropriate mathematical formulae;
- to find a set of plausible causes interesting enough to explain the time series behaviour.

Also, our starting point includes the common idea that financial time series are intrinsically non-stationary.

In Econophysics, it is not usual to study the original financial series. This approach has its drawbacks, although. The one that comes first to mind is that we cannot study stationarity, that is, the long term information. The focus, instead, goes to a transformed quantity (as in the financial literature) named *one-day returns*. Sometimes these are called log-returns to distinguish them from a similar quantity without the logarithm being applied, $\frac{x_i - x_{i-1}}{x_{i-1}}$. In what follows in this work, returns means always the log-returns. The main reason to use the log-returns has to do with the additive process associated to the time series. For an asset, that is, any good to which we can give a price, with an associated time series x we have the following definition:

Definition 1. Let x_i be the value of a time series x at time i . *Returns* are defined as:

$$\eta_i = \log \frac{x_i}{x_{i-1}}, \quad (1)$$

where η_i is the return at time step i . Since x_i are asset values, they are positive and thus the returns are always well defined. The use of the ratio between two consecutive values makes the quantity dimensionless and the use of logarithms gives a different sign to gains and losses.

The distribution of returns was first modelled for bonds, [Bachelier \[1900\]](#), as a Normal distribution,

$$P(r) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{r^2}{2\sigma^2}} \quad (2)$$

where σ^2 is the variance of the distribution.

Returns can be used to compare different series, to search for patterns both exclusive to some series only or for the whole group of series. We can, also, use them to give us a new perception of the involved correlations.

Also, of interest to a better understanding of the following sections, is the definition of financial volatility. Volatility, σ , corresponds to standard deviation and is a measure for the variation of a price of a financial instrument over time.

Definition 2. The annualized volatility σ is the standard deviation of the financial instrument’s yearly logarithmic returns.

Therefore, if the daily logarithmic returns of a stock have a standard deviation of σ_d and the time period of returns is P , the annualized volatility is

$$\sigma = \frac{\sigma_d}{\sqrt{P}}. \quad (3)$$

The Equation (3) converts returns or volatility measures from one time period to another assuming a particular underlying model or process because it is an extrapolation of a random walk, or Wiener process, whose steps have finite variance. More generally, though, for natural stochastic processes, the precise relationship between volatility measures for different time periods is more complicated. Some use the Lévy stability exponent α to extrapolate natural processes:

$$\sigma_T = T^{1/\alpha} \sigma. \quad (4)$$

If $\alpha = 2$ we get a Wiener process scaling relation [Mandelbrot, 1963].

2.1.3 Random Walk Hypothesis and the Brownian Motion

“What if the time series were similar to a random walk?”, or, “It is possible to predict future price movements using the past price movements?” are long asked questions by experts and laymen.

Another view of the complexity/disorder is the (fractional) Brownian motion, that appeared in Bachelier PhD thesis, in 1900, [Bachelier, 1900], when studying the Paris Stock Exchange as a way to describe the evolution of the financial assets. Louis Bachelier, who firstly proposed a theory of stock market fluctuations, reached the conclusion that “the mathematical expectation of the speculator is zero” and described this condition as a “fair game”. He gave the distribution function the name for what is now known as the Wiener stochastic process (the stochastic process that underlies Brownian Motion) linking it mathematically with the diffusion equation. Feller [1968], called it the Bachelier-Wiener process. This work states that the second order moments of the increments of a heat/diffusion process scale as

$$E \{ (X(t_2) - X(t_1))^2 \} \propto |t_2 - t_1|, \quad (5)$$

where X is the stochastic process under study.

Henri Poincaré, Bachelier’s advisor, observed that “*M. Bachelier has evidenced an original and precise mind [but] the subject is somewhat remote from those our other candidates are in the habit of treating*”.

Nevertheless, his thesis anticipated many of the mathematical discoveries made later by Wiener and Markov, and outlined the importance of such ideas in today’s financial markets, stating that “*it is evident that the present theory solves the majority of problems in the study of speculation by the calculus of probability*”.

Later, works from Hurst in the 50’s and Mandelbrot in the 60’s gave rise to the fractional Brownian motion, a generalization of the Brownian motion, firstly described by Bachelier. The Hurst exponent has become an important estimation sign of the financial data disorder or complexity. These two concepts, entropy and fractional Brownian motion, provide a measure of financial data disorder or complexity [Matos et al., 2006].

In the seventies, Black, Scholes and Robert Morton, [Black and Scholes, 1973], following the ideas of Osborne [1959], Osborne [1977] and Samuelson [1973], modelled the share price as a stochastic process known as a geometric Brownian motion. They also established the isomorphism between the standard deviation of the fluctuations in price of a financial instrument and investment risk. Nowadays, a modern version of Bachelier's theory is still routinely used in financial literature. This theory predicts a Gaussian probability distribution for stock-price fluctuations. The random walk hypothesis, with independent and identically distributed increments, is the basis of the Efficient Market Hypothesis Fama [1970], as we stated in Chapter 1.

Present in Econophysics is the conviction about scaling arguments coming from the study of systems in critical states (see, for instance, Mantegna and Stanley [1995], Cont et al. [1997] or Di Matteo et al. [2005]). The empirical study of those distributions led also to the analysis of distributions of economic shocks, growth rate variations, firm and city sizes. In all these measures scaling laws were found, thus giving confidence that the same type of analysis could be applied to the study of the distributions used to characterise complex systems.

2.1.4 *Stylized empirical facts*

Physicists interest in analysing financial data has been to find common or universal regularities in the time series (a different approach from those of the economists doing traditional statistical analysis of financial data). The results of their empirical studies showed that the apparently random variations in time series share some statistical properties which are interesting, non-trivial and common for various values and time periods. These are called stylized empirical facts.

The concept of "stylized facts" was introduced in macroeconomics around 1960 by Nicholas Kaldor, who advocated that a scientist studying a phenomenon "should be free to start off with a stylized view of the facts". In his work, Kaldor [1957] isolated several statistical facts characterizing macroeconomic growth over long periods and in several countries, and took these robust patterns as a starting point for theoretical modelling. This expression has thus been adopted to describe empirical facts that arose in statistical studies of financial time series and that seem to be persistent across various time periods, places, markets or assets.

Stylized facts are, then, obtained by taking a common denominator among the properties observed in different markets and financial instruments. By doing so, one gains in generality but tends to lose in precision of the statements one can make about asset returns. Indeed, stylized facts are usually formulated in terms of qualitative properties of asset returns and may not be precise enough to distinguish among different parametric models Cont [2001]. One can find many different lists of these facts in several reviews (see Bollerslev et al. [1994] or Cont [2001]).

1. Absence of autocorrelations: linear autocorrelations of asset returns are often insignificant, except for very small intra-day time scales (20 minutes) for which micro-structure effects come into play. The auto-correlation of log returns rapidly decays to zero for $\tau \geq 15$ minutes, which supports the Efficient Market Hypothesis. When

τ is increased, weekly and monthly returns exhibit some auto-correlation but the statistical evidence varies from sample to sample.

2. Heavy/Fat tails: the distribution of returns seems to display a power-law or Pareto-like tail, with a tail index which is finite, between 2 – 5 for most data sets studied [Gabaix et al., 2003]. This excludes stable laws with infinite variance and the normal distribution. However, the precise form of the tails is difficult to determine as Mandelbrot [1963] pointed out. The Gaussian/Normal distribution is a special case of the more general Lévy distributions, and is often used as an approximation to log-normal distributions. In contrast, these distributions display power-law decay in the tails and this is related to the fractal nature of financial data [Higushi, 1988], where uni-fractal processes, such as fractional Brownian motion [Mantegna and Stanley, 2000, Bouchaud and Potters, 2003] and simple multi-fractal processes (see [Lux, 2004] and Calvet and Fisher [2002]) have been considered for financial data. The "fat tails" can only be obtained by "nonperturbative" methods, mainly by numerical ones, since they contain the deviations from the usual Gaussian approximations [Nolan, 2006].
3. Gain/loss asymmetry: one observes large draw downs in stock prices and stock index values but not equally large upward movements.
4. Aggregational Gaussianity: as one increases the time scale t over which returns are calculated, their distribution looks more and more like a normal distribution, meaning that the shape of the distribution is not the same at different time scales. The fact that the shape of the distribution changes with τ makes it clear that the random process underlying prices must have non-trivial temporal structure.
5. Intermittency: returns display, at any time scale, a high degree of variability. This is quantified by the presence of irregular bursts in time series of a wide variety of volatility estimators.
6. Volatility clustering: different measures of volatility display a positive autocorrelation over several days, which quantifies the fact that high-volatility events tend to cluster in time, and decays roughly as a power law with an exponent between 0.1 and 0.3. Price fluctuations are not identically distributed and the properties of the distribution, such as the absolute return or variance, change with time. To sum up, large changes tend to be followed by large changes, and analogously for small changes.
7. Existence of nonlinear correlation: Abhyankar et al. [1997] found nonlinear dependence in the four important stock-market indices. Also, Ammermann and Patterson [2003] have shown that nonlinear dependencies play a significant role in the returns for a broad range of financial time series (see <http://finance.martinsewell.com/stylized-facts/nonlinearity/> for more details).
8. Conditional heavy tails: even after correcting returns for volatility clustering, the residual time series still exhibit heavy tails. However, the tails are less heavy than in the unconditional distribution of returns.

9. Slow decay of autocorrelation in absolute returns: the autocorrelation function of absolute returns decays slowly as a function of the time lag, roughly as a power law with an exponent $\beta \in [0.2, 0.4]$. This is sometimes interpreted as a sign of long-range dependence.
10. Leverage effect [Reigneron et al., 2011]: most measures of volatility of an asset are negatively correlated with the returns of that asset.
11. Volume/volatility correlation: trading volume is correlated with all measures of volatility.
12. Asymmetry in time scales: coarse-grained measures of volatility predict fine-scale volatility better than the other way round.

One important question is to what extent these stylized empirical facts are relevant to empirical studies in finance.

2.1.5 Market Crashes or “When things go terribly wrong”

The ultimate purpose of this thesis, as stated in Chapter 1, is to find information pieces that can give us some light of how the markets evolve to crashes. These crashes are not so rare as a layman can sometimes account for (for an explanatory reading follow Ball [2006]). For that reason, it can be instructive to recall some of the most important events (see Table 1) that affected markets from the XX century.

Date	Events	Description
1929 to 1938	Great Depression	Stock market crash and banking collapse (43 and 13 months duration respectively)
1953 to 1954	Post Korean War	poor government policies and high interest rates (10 months)
1973 to 1975	Oil Crisis	quadrupling of oil price by OPEC and high government spending due to Vietnam War (16 months)
1979 to 1980	Energy Crisis	Iranian revolution increases oil price
1982 to 1983	Recession	tight monetary policy in the U.S. to control inflation and sharp correction to overproduction
1988 to 1992	Recession	general recession in commodity prices
1991	Japanese recession	collapse of a real estate bubble halts Japan growth
1997	Asian financial crises	collapse of the Thai currency inflicts damage on many Asian economies

Table 1: Major XX century events for global markets.

XXI Century Crashes

In Table 2 are displayed a list of major events that have affected international markets in the XXI century.

Date	Events
2000/03	DotCom crash
2001/09/11	Terrorist attack (New York)
2002/05	Stock Market Downturn
2003/12	General Threat level raised
2004/03/11	Terrorist attack (Madrid)
2005/12/08	European Central Bank first warning
2007/08/09	Global liquidity shortage
2008/02/17	Northern Rock (UK) goes public
2008/09/07	Fannie Mae and Freddie Mac put in Government protection
2008/09/15	Lehman Brothers Bankruptcy
2010/04/23	Greece financial support
2010/11/21	Ireland financial support
2011/04/06	Portugal financial support
2013/03	Cyprus financial support

Table 2: Major XXI century events for global markets.

Despite all the dates presented in Table 2, it will be presented in more detail two specific events that turned to be global: the DotCom Bubble and the Housing Bubble and Credit Crisis.

Let us, firstly, start with bubbles and crashes. A bubble is defined to occur when investors put so much demand on a stock that they drive the price beyond accuracy or rationality usually determined by the performance of that stock. A crash is defined as a significant drop in the total value of a market, historically attributable to the popping of a bubble, creating a situation where the majority of investors are trying to flee the market at the same time. Attempting to avoid more losses, investors during a crash are panic selling, hoping to unload their declining stocks onto other investors. This panic selling contributes to the declining market, which eventually crashes and affects everyone. Typically crashes in the stock market have been followed by a depression.

Now let us look in more detail at the two financial “disasters” of the XXI century.

DotCom Bubble (Silicon Valley, United States - March 11, 2000 to October 9, 2002)

This bubble was a result of the popularization of the Internet in 1995. From nothing, an international market was created. This “new economy” was the home for a huge number of speculators, that did not took a look to the business plan of the companies they were investing in. Some of them worth millions and were made of “nothing”. After some time of illusion, some companies started to report huge losses. It was the end of

an era. During this period, the Nasdaq Composite lost 78% of its value as it fell from 5046.86 to 1114.11.

Housing Bubble (United States and Britain) and Credit Crisis (around the World) (2007-2009)

This bubble was a result of diverse factors. Following the bursting of the DotCom bubble and the recession of the early 2000s, the Federal Reserve kept short-term interest rates low for an extended period of time. This period coincided, in the United States, with a housing boom. People began to view their homes as a "piggy bank". As home prices soared and many home owners "stretched" to make their mortgage payments, the possibility of a collapse grew. However, the true extent of the danger was hidden because so many mortgages had been turned into AAA-rated securities.

When the long held belief that home prices do not decline turned out to be inaccurate we saw large losses for banks and other financial institutions. These losses spread to other asset classes, fuelling a crisis of confidence in the health of many of the world's largest banks. Events reached their climax with the bankruptcy of Lehman Brothers in September 2008, which resulted in a credit freeze that brought the global financial system to the brink of a collapse.

The credit crisis and accompanying recession caused unprecedented volatility in financial markets around the world. Stocks fell 50% or more from their highs through March 2009 before rallying more than 50% once the crisis began to ease. During this period, the S&P 500 declined 57% from its high in October 2007 of 1576 to its low in March 2009 of 676 (see Beattie [2013]).

Recession dates

When studying periods of crisis it is interesting to note that it is not easy to decide when a period of crisis happens. Here, we follow the The National Bureau of Economic Research (NBER), www.nber.org, which is the largest Economics research organization in the United States.

NBER is a private non-profit research organization "committed to undertaking and disseminating unbiased economic research among public policy makers, business professionals, and the academic community."

The main information obtained for this work from NBER is the start and end dates for recessions in the United States. In the XXI century, NBER proposed the following recessions:

- March, 2001 to November, 2001
- December, 2007 to June, 2009

In Figure 1 the two XXI century recession periods, according to NBER, are depicted in blue against two of the markets indices. It is interesting to note that there is an obvious relationship between markets evolution and those recession periods.

It seems, also, fair to say that the first recession period was not so noticeable in non North American or European Markets, as we can see from Merval or Straits indices.

This may indicate that the markets are going global or it is only a question of recession “intensity”? A complete catalogue of results is resumed in Appendix B.

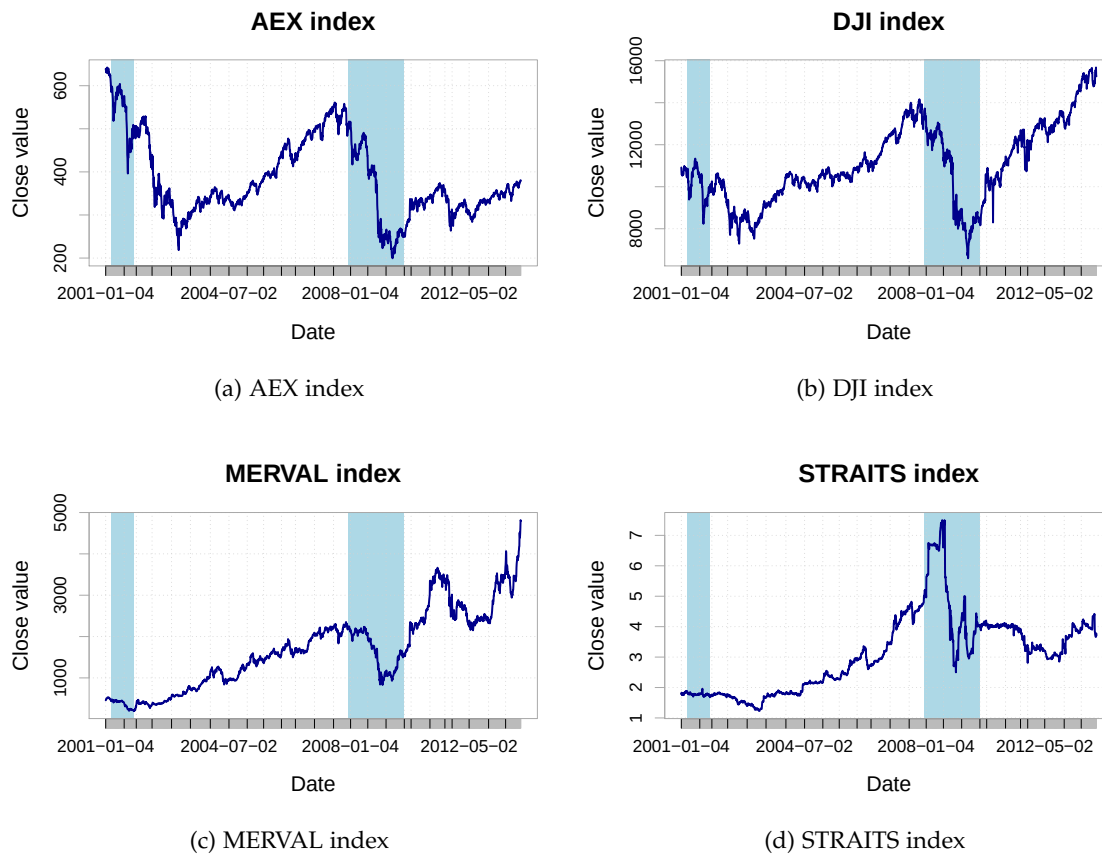


Figure 1: NBER Recession dates

As stated before, not only NBER proposes recession periods. For instance, the Centre for Economic Policy Research (CEPR), an european organism, www.cepr.org, has a different view on recession periods. Concerning Europe and the XXI century, the following recession periods were proposed:

- 1st quarter of 2008 until 2nd quarter of 2009,
- 3rd quarter of 2011 and still going on.

It is fair to say that in the last six quarters Europe changed, experiencing very little growth, but still not strong enough to give CEPR a motive to propose an end to recession started in 2011.

Now, just for a comparative point of view, in Figure 2 it is possible to observe two different recessions periods for the United States: on the right side is the NBER recession proposal and on the left side is another organization proposal. The differences have more significance for the first recession period.

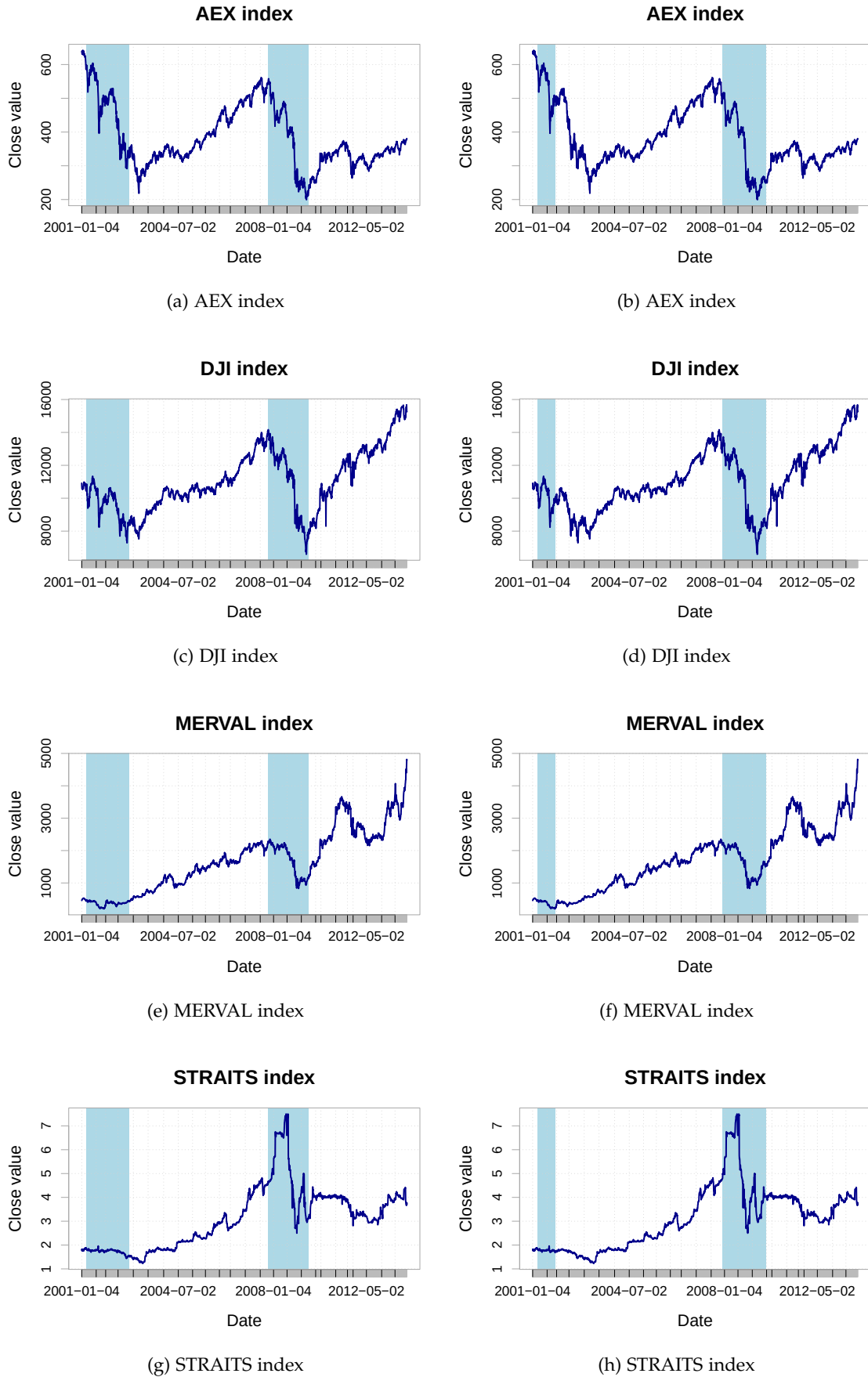


Figure 2: Alternative recession dates

2.2 STOCHASTIC PROCESSES

The theory of Stochastic Processes is generally referred to as the "dynamical" part of probability theory, where we study a collection of random variables from the point of view of their interdependence and limiting behaviour. This theory can be formulated in very different ways, like, for instance, a random walk model, a Fokker-Planck type equation or a Langevin equation (for a statistical point of view see [Lindsey \[2004\]](#)). We can apply a stochastic process whenever we have a process developing in time and controlled by probabilistic laws [\[Parzen, 1999\]](#).

In this context, it is interesting to note that many elements of the theory of stochastic processes, were first developed in connection with the study of fluctuations and noise in physical systems and financial data ([Bachelier \[1900\]](#), [Einstein \[1905\]](#)). Some systems can present unpredictable chaotic behaviour due to dynamically generated internal noise. Either stochastic or chaotic, noisy processes represent the rule rather than an exception in nature [\[Chakarborti et al., 2007\]](#).

All the stochastic processes that will be considered in this work are indexed by time. The notation used in this section follows the one used in [Papoulis \[1985\]](#).

2.2.1 Random variables

The expression random variable is in a way misleading and actually an historical accident, as a random variable is not a variable, but rather a function that maps events to real numbers.

Definition 3. Let $\mathcal{A}$ be a σ -algebra and Ω the space of events relative to the experiment. A function $X : (\Omega, \mathcal{A}) \rightarrow \mathbb{R}$ is a *random variable* if for every subset $A_r = \{\omega : X(\omega) \leq r\}$, $r \in \mathbb{R}$, the condition $A_r \in \mathcal{A}$ is satisfied.

1. A random variable X is said to be *discrete* if the set $\{X(\omega) : \omega \in \Omega\}$ (i.e. the range of X) is countable;
2. A random variable Y is said to be *continuous* if it has a cumulative distribution function which is absolutely continuous.

One useful definition is the expected value of a random variable, as it gives what we should expect if we repeat the process over and over.

Definition 4. Consider a discrete random variable X . The *expected value*, or *expectation*, of X , denoted $E\{X\}$, is the weighted average of all possible values of X by their corresponding probabilities, i.e. $E\{X\} = \sum_x x f_X(x)$ ($f_X(x)$ is the probability function of X). If X is a continuous random variable, then $E\{X\} = \int_x x f_X(x) dx$ ($f_X(x)$ is the probability density function of X).

Note that if the corresponding sum or integral does not converge, the expectation does not exist. One example of this situation is the Cauchy random variable.

Definition 5. Going further in the definitions, let X and Y be two random variables, then the *covariance* of X and Y is

$$C_{X,Y} = E\{(X - E\{X\})(Y - E\{Y\})\}. \quad (6)$$

If $X = Y$ then we get the *variance* of X :

$$\text{Var}_X = C_{X,X}. \quad (7)$$

The *standard deviation* of the random variable X is the square root of variance

$$\sigma_X = \sqrt{\text{Var}_X}. \quad (8)$$

The *correlation coefficient* of two random variables X and Y is

$$r_{X,Y} = \frac{C_{X,Y}}{\sigma_X \sigma_Y}, \quad (9)$$

where σ_X and σ_Y are the standard deviations of two stock return series. It is a common measure of the dependence between the return series of the two stocks. The elements of the correlation matrix are restricted to the domain $-1 \leq c_{ij} \leq +1$: for $0 < c_{ij} \leq +1$ the stocks are correlated (in a positive way), for $-1 \leq c_{ij} < 0$ the stocks are anti-correlated (correlated in a negative way), and for $c_{ij} = 0$ the stocks are uncorrelated. The cross-correlation defined above calculates the dependence between the return series in the whole period of the sample data.

2.2.2 Stochastic processes

Definition 6. Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. A *stochastic process* is a collection $\{X(t) \mid t \in T\}$ of random variables $X(t)$ defined on $(\Omega, \mathcal{F}, \mathbf{P})$, where T is a set, called the index set of the process. T is usually (but not always) a subset of $\mathbb{R}$. One can also think of a stochastic process as a function $X = (X(t, \omega))$ in two variables: $t \in T$ and $\omega \in \Omega$, such that for each t , $X_t(\omega) := X(t, \omega)$ is a random variable on $(\Omega, \mathcal{F}, \mathbf{P})$. Given any t , the possible values of $X(t)$ are called the states of the process at t . The set of all states (for all t) of a stochastic process is called its state space. If T is discrete, then the stochastic process is a discrete-time process. If T is an interval of $\mathbb{R}$, then $\{X(t) \mid t \in T\}$ is a continuous-time process. If T can be linearly ordered, then t is also known as the time.

Let $X(t)$ and $Y(t)$ be stochastic processes, with $t \in T$ and T being the index set.

Definition 7. The *mean* $\eta(t)$ of $X(t)$ is the expected value of the random variable $X(t)$

$$\eta_X(t) = E\{X(t)\}. \quad (10)$$

The *cross-correlation* of two processes $X(t)$ and $Y(t)$ is

$$R_{XY}(t_1, t_2) = E\{X(t_1)Y(t_2)\}. \quad (11)$$

The *autocorrelation* $R(t_1, t_2)$ of $X(t)$ is the expect value of the product $X(t_1)X(t_2)$

$$R(t_1, t_2) = E\{X(t_1)X(t_2)\}. \quad (12)$$

The *cross-covariance* of two processes $X(t)$ and $Y(t)$ is

$$C_{XY}(t_1, t_2) = E\{X(t_1)Y(t_2)\} - \eta_X(t_1)\eta_Y(t_2). \quad (13)$$

The *autocovariance* $C(t_1, t_2)$ of $X(t)$ is the covariance of the random variables $X(t_1)$ and $X(t_2)$

$$C(t_1, t_2) = R(t_1, t_2) - \eta(t_1)\eta(t_2). \quad (14)$$

The ratio

$$r(t_1, t_2) = \frac{C(t_1, t_2)}{\sqrt{C(t_1, t_1)C(t_2, t_2)}} \quad (15)$$

is the *correlation coefficient* of the process $X(t)$.

2.3 RANDOM MATRIX THEORY

The R/S, DFA and Geometric Brownian Motion methods that will be considered in Section 2.7 are suitable for analysing univariate data. But, as the stock-market data are essentially multivariate time-series data, it is worth to look for other instruments. Also, in the multivariate signal processing problem, one key issue might be when instabilities occur in signal patterns and how we might determine if the fluctuations are damped, remain at low level, or combine in some way as to cause a major event, e.g. a market crash. Crashes are also interesting since the market dynamics changes during the event (see Mendes et al. [2003], Araújo and Louçã [2006]).

Random matrix theory (RMT) is concerned with the study of large-dimensional matrices, in particular with their eigenvalues, eigenvectors and singular values, whose entries are sampled according to known probability densities. The interest in random matrices appeared in the context of multivariate statistics with the works of Wishart and Hsu in the 30's, but it was only in the 50's, with Wigner (Wigner [1955] and Wigner [1958]), who introduced random matrix ensembles and derived the first asymptotic result although in the context of nuclear physics. It seems that the problem of interpreting the correlations among large amounts of spectroscopic data on the energy levels, whose exact nature is unknown, is similar of interpreting the correlations among different stocks returns. Therefore, with the minimal assumption of a random Hamiltonian, given by a real symmetric matrix with independent random elements, a series of predictions can be made.

In 1967, a seminal paper by Marchenko and Pastur [Marchenko and Pastur, 1967] on the spectrum of empirical correlation matrices gave birth to many interesting applications in very different contexts. However, its central objective, as a new statistical tool to analyse large dimensional data sets, only became fully relevant more recently, when the computational storage and handling of huge amounts of data became common to almost all human activity. In fact, the correlations among stock returns have also been addressed by means of the random matrix theory. The quest for the causes that explain the dynamics of N quantities in a financial context, say for instance, the daily returns of the different stocks of the PSI-20, brought a great development to this subject.

2.3.1 Returns statistics

As stated before, in Econophysics the focus goes to returns. As already know, their distribution is not Gaussian and has fat tails, decaying as a power law. The empirical probability distribution function of the returns on short time scales (from high frequency data to a few days, where we still can assume that the returns have zero mean) can be satisfactory fit by a Student-t distribution [Bouchaud and Potters, 2003]:

$$P(r) = \frac{1}{\sqrt{\pi}} \frac{\Gamma\left(\frac{1+\mu}{2}\right)}{\Gamma\left(\frac{\mu}{2}\right)} \frac{a^\mu}{(r^2 + a^2)^{\frac{1+\mu}{2}}}, \quad (16)$$

where a is related to the variance of the distribution, $\sigma^2 = a^2 / (\mu - 2)$, and μ moves in the interval $[3, 5]$ (Plerou et al. [1999], Gopikrishnan et al. [1999]). On longer time scales, from a few weeks to months, the returns distribution approaches a Gaussian [Bouchaud and Potters, 2003]. However, we have to point out two restrictions:

1. The returns cannot be used as independently drawn Student random variables, that is to say, returns are far from being considered independent and identically distributed (i.i.d.) random variables: from empirical evidence, it is known that asset returns are clearly not independent as they exhibit certain patterns;
2. Because of their nature there is diminishing predictability of data that are further away from the present. In other words, the volatility of financial returns is itself a dynamical variable over time, having a broad distribution of characteristic frequencies.

Formally, the returns at time t can be represented by the product of a volatility component σ^t and a directional component ζ^t [Bouchaud and Potters, 2003]:

$$r^t = \sigma^t \zeta^t, \quad (17)$$

where, for instance, the ζ^t are such that now are i.i.d. random variables with unit variance and σ^t is a positive random variable with both fast and slow components. Or vice-versa, because, in fact, a Student-t variable can be written as in Equation (17) where the ζ is Gaussian and σ is an inverse Gamma random variable. Indeed, σ^t and ζ^t cannot be considered independent. From the literature (see Bouchaud and Potters [2003] for a review) we know that when considering stock markets, negative past returns tend to increase future volatilities and vice-versa: this is the “leverage” effect, coined by Black in 1976, which tells us that the average of quantities such as $\zeta^t \sigma^{t+\tau}$ is negative when $\tau > 0$. But, going back to Equation (17) and considering the first assumption, the slow part of σ^t is actually a long memory process such that its correlation function decays as a slow power-law of the time lag τ :

$$\overline{\sigma^t \sigma^{t+\tau}} - \bar{\sigma}^2 \propto \tau^{-\nu}, \nu \sim 0.1 \quad (18)$$

In the more general case of a multivariate distribution of returns there is a need to extend these previous results to a multivariate ambient, where there are N correlated stocks and a joint distribution of simultaneous returns $\{r_1^t, r_2^t, \dots, r_N^t\}$. All marginals of

this joint distribution must resemble the Student-t distribution, Equation (16), and it must be compatible with the true correlation matrix of the returns:

$$C_{ij} = \int \prod_k [dr_k] r_i r_j P(r_1, r_2, \dots, r_N). \quad (19)$$

This previous result, Equation (19), leads us to the “copula specification problem” in quantitative finance, that is, a multivariate probability distribution of N random variables u_i all having a uniform marginal probability distribution in $[0, 1]$. Further developments about this “copula specification problem” are out of the scope of this thesis.

2.3.2 The correlation matrix

“Correlation” is defined as “a relation existing between phenomena or things or between mathematical or statistical variables which tend to vary, be associated, or occur together in a way not expected on the basis of chance alone”¹.

When we discuss about correlations in stock prices, we are interested in the relations between variables such as close prices and transaction volumes, for instance, and more importantly how these relations affect the nature of the statistical distributions which govern the prices variation in the time series.

We pay, now, our attention to the estimation of the correlations between the price movements of different assets (for a recent review, [Frahm and Jaekel \[2008\]](#)). Denoting by T the total number of observations of each of the N quantities, say, thinking about stock returns, T is the total number of trading days in the sampled data.

The realization of the i th quantity ($i = 1, \dots, N$) at “time” t ($t = 1, \dots, T$) will be r_i^t . Now, the normalized $T \times N$ matrix of returns, denoted as $\mathbf{X}$, will be: $X_{ti} = \frac{r_i^t}{\sqrt{T}}$. If we want to characterize the correlations between these quantities, the simplest form is to compute the Pearson estimator of the correlation matrix:

$$E_{ij} = \frac{1}{T} \sum_{t=1}^T r_i^t r_j^t \equiv (\mathbf{X}^T \mathbf{X})_{ij}, \quad (20)$$

where $\mathbf{E}$ is the empirical correlation matrix, most probably different from the “true” correlation matrix $\mathbf{C}$:

$$\rho_{ij}^t = \frac{\langle r_i^t r_j^t \rangle \langle r_i^t \rangle \langle r_j^t \rangle}{\sqrt{[\langle r_i^{t2} \rangle - \langle r_i^t \rangle^2] [\langle r_j^{t2} \rangle - \langle r_j^t \rangle^2]}}, \quad (21)$$

where the $\langle \dots \rangle$ gives a time average over the consecutive trading days included in the return vectors. These correlation coefficients fulfill the condition $-1 \leq \rho_{ij} \leq 1$ and form an $N \times N$ correlation matrix $\mathbf{C}_t$, which serves as the basis of further analyses.

Apart for dimensionality, correlation and covariance are very similar concepts.

¹ In Merriam-Webster Online Dictionary. Retrieved July 31, 2014, from <http://www.merriam-webster.com/dictionary/correlations>

We also present, here, the covariance matrix with variable weights at time T , over an horizon M , $\sigma^T(M)$, that is given by:

$$\sigma_{ij}^T(M) = \frac{\sum_{s=0}^M W_s r_{i,T-s} r_{j,T-s}}{\sum_{s=0}^M W_s}, \quad (22)$$

where $r_{i,t}$ is the value of return r_i at time t , and W_s is the weight given for the covariance at delay s , (time $T - s$).

The weight vector, $\mathbf{W}$, can be used to have decreasing components since higher weights are attributed to moments closer to the time being analysed. One example traditionally used and the same that is used in this work is $W_i = R^i$, with $0 < R < 1$. Then we have $\sum_{s=0}^T W_{T-s} = \frac{R^T}{1-R}$, and W_i corresponds to a geometric series. Typical values (see [Litterman and Winkelmann \[1998\]](#)) are $R = 0.9$ and $T = 20$.

Some interesting studies using correlation matrix forecasts of financial asset returns have been done in financial risk management ([Embrechts et al. \[2002\]](#) and [Bouchaud and Potters \[2003\]](#)). In market maturities, [Matos et al. \[2006\]](#) and [Sharkasi et al. \[2006a\]](#), studied the behaviour of eigenvalues of the covariance matrices around crashes and also studied the ratio of the dominant (first eigenvalue) to the sub-dominant (second eigenvalue) for emerging and mature markets. Their results showed that mature markets react to crashes in a different way than emerging ones which, as suggested before, take longer to recover than mature markets. Their investigation also suggests that the second largest eigenvalue may thus be expected to provide additional information on market movements.

In more recent years, there are increasing works concentrated on the variation of the cross correlations between market equities over time. [Di Matteo et al. \[2010\]](#) have investigated the evolution of the correlation structure among 395 stocks quoted on the U.S. equity market from 1996 to 2009, in which the connected links among stocks are built by a topologically constrained graph approach. They found that the stocks have increased correlations in the period of larger market instabilities. [Fenn et al. \[2011\]](#) have used the RMT method to analyse the time evolutions of the correlations between the market equity indices of 28 geographical regions from 1999 to 2010, and they also observe the increase of the correlations between several different markets after the credit crisis of 2007-2008.

2.3.3 Eigenvalues and eigenvectors

The empirical determination of a correlation matrix is a difficult task. If one considers N assets, the correlation matrix contains $N(N-1)/2$ mathematically independent elements, which must be determined from N time-series of length T . If T is not very large compared to N , then generally the determination of the covariances is noisy, and therefore the empirical correlation matrix is to a large extent random. The smallest eigenvalues of the matrix are the most sensitive to this ‘noise’. But the eigenvectors corresponding to these smallest eigenvalues determine the minimum risk portfolios in Markowitz theory [[Laloux et al., 2000](#)]. It is thus important to distinguish “signal” from “noise” or, in other words, to extract the eigenvectors and eigenvalues of the correlation matrix

containing real information (those important for risk control), from those which do not contain any useful information and are unstable in time.

It is, then, useful to compare the properties of an empirical correlation matrix to a “null hypothesis” - a random matrix which arises, for instance, from a finite time-series of strictly uncorrelated assets. Deviations from the random matrix case might then suggest the presence of true information.

The eigenvalues and eigenvectors of random matrices approach a well-defined functional form in the limit when N tends to infinity. It is then possible to compare the distribution of empirically determined eigenvalues to the distribution that would be expected if the data were completely random. Obtaining the difference between $\mathbf{E}$ and $\mathbf{C}$ was really the goal of the Marchenko and Pastur effort [Marchenko and Pastur, 1967]. This difference may be found considering the ratio between N and T :

$$q = \frac{N}{T}. \quad (23)$$

- If N and T are about the same order, that is, $q \sim O(1)$, then $\text{Tr}\mathbf{E}^{-1} = \text{Tr}\mathbf{C}^{-1} / (1 - q)$ [Bouchaud and Potters, 2011].
- If N is small compared to T , then we expect that the Pearson estimator $\mathbf{E}$ is close to its “true” value and so a good estimator of $\text{Tr}\mathbf{C}^{-1}$ is $\text{Tr}\mathbf{E}^{-1}$. This is the case when $q \rightarrow 0$, where we get the “true” density of the eigenvalues.
- In the opposite, the asymptotic limit, the spectrum of the eigenvalues (their empirical density) is mostly distorted when compared to the “true” density. When $T, N \rightarrow \infty$ the spectrum has some degree of universality with respect to the distribution of the r_i^t ’s.

The correlation matrix defined in Equation (20) is a $N \times N$ symmetric matrix and so we can diagonalize it. This is the beginning of the relationship between Random Matrix Theory and the Principal Component Analysis.

Three Classical Results

The asymptotic behaviour of random matrices attracted more attention and it was quickly realized that this behaviour is often independent of the distribution of the entries. Furthermore, the limiting distribution typically takes non-zero values only on a bounded interval, displaying sharp edges.

Until recently, the majority of the results established were concerned with the spectra, or eigenvalue distributions, of such matrices. But now, the study of the eigenvectors of random matrices also starts to become relevant. Of interest are both the global regime, which refers to statistics on the entire set of eigenvalues, and the local regime, concerned with spacings between individual eigenvalues. In this thesis, we will briefly consider the three classical results and their behaviour in these regimes:

1. Wigner’s semicircle law for the eigenvalues of symmetric or Hermitian matrices;
2. the Marchenko-Pastur law for the eigenvalues of sample covariance matrices;
3. the Tracy-Widom distribution for the largest eigenvalue of Gaussian unitary matrices.

Wigner's semicircle law, for example, can be considered universal in the sense that the eigenvalue distribution of a Symmetric or Hermitian matrix with i.i.d. entries, properly normalized, converges to the same density regardless of the underlying distribution of the matrix entries. Also, in this asymptotic limit, the eigenvalues are almost surely supported on the interval $[-2, 2]$, illustrating the sharp edges behaviour mentioned before. Historically, results such as Wigner's semicircle law, were initially discovered for specific matrix ensembles and later were extended to more general classes of matrices. As another example, the circular law for the eigenvalues of a non-symmetric matrix with i.i.d. entries was initially established for Gaussian entries in 1965, but only in 2008 was it fully expanded to arbitrary densities. From a practical standpoint, the benefits of universality are clear, given that the same result can be applied to a vast class of problems.

Sharp edges are also important for practical applications. Here, the hope is to use the behaviour of random matrices to separate signals from noise. In such applications, the finite size of the matrices of interest poses a problem when adapting asymptotic results valid for matrices of infinite size. Nonetheless, an eigenvalue that appears significantly outside of the asymptotic range is a good indicator of non-random behaviour.

The spectral properties of random matrices are one interesting application of the Central Limit Theorem. In fact, and just considering the simplest ensemble of random matrices, the one where all elements of the matrix $\mathbf{H}$ are i.i.d. random variables and the only constraint being the matrix symmetry ($H_{ij} = H_{ji}$), in the limit of very large matrices, the distribution of its eigenvalues has universal properties, which can be considered independent of the distribution of the elements of the matrix. So, let us consider a square symmetric matrix $\mathbf{H}$, $N \times N$. The statistics of the eigenvalues λ_α of large random matrices, in particular the density of eigenvalues $\rho(\lambda)$, is defined as:

$$\rho_N(\lambda) = \frac{1}{N} \sum_{\alpha=1}^N \delta(\lambda - \lambda_\alpha), \quad (24)$$

where λ_α are the eigenvalues of the $N \times N$ symmetric matrix $\mathbf{H}$ under study and δ is the Dirac function. We will need the "resolvent" $G(\lambda)$ of the matrix $\mathbf{H}$, defined as:

$$G_{ij}(\lambda) = \left(\frac{1}{\lambda \mathbf{I} - \mathbf{H}} \right)_{ij}, \quad (25)$$

where $\mathbf{I}$ is the identity matrix. The trace of $G(\lambda)$, using the eigenvalues of $\mathbf{H}$, is:

$$\text{Tr} G(\lambda) = \sum_{\alpha=1}^N \frac{1}{\lambda - \lambda_\alpha}. \quad (26)$$

And the deduction goes through (see, for a full explanation, [Bouchaud and Potters, 2003]), until we get

$$\rho(\lambda) = \frac{1}{2\pi\sigma^2} \sqrt{4\sigma^2 - \lambda^2}, \quad |\lambda| \leq 2\sigma \quad (27)$$

which is the "semi-circle" law derived by Wigner in the late fifties of the XX century.

In finance we often see correlation matrices $\mathbf{C}$, which are positive definite. $\mathbf{C}$ can be written as $\mathbf{C} = \mathbf{H}\mathbf{H}^T$, where $\mathbf{H}^T$ designates the transpose. As $\mathbf{H}$ is, generally, a rectangular matrix of size $M \times N$ where M is the assets number and N the observations days,

then $\mathbf{C}$ will be $M \times M$. If $N = M$ then to get the eigenvalues from $\mathbf{C}$ we just need to obtain them from $\mathbf{H}$: $\lambda_C = \lambda_H^2$, that is, $\rho(\lambda_C) d\lambda_C = 2\rho(\lambda_H) d\lambda_H$, and, by Equation (27),

$$\rho(\lambda_C) = \frac{1}{2\pi\sigma^2} \sqrt{\frac{4\sigma^2 - \lambda_C}{\lambda_C}}, \quad 0 \leq \lambda_C \leq 4\sigma^2 \quad (28)$$

However, usually $N \neq M$, then we can obtain similar formula if we consider that in the limit $N, M \rightarrow \infty$,

$$\rho(\lambda_C) = \frac{Q}{2\pi\sigma^2} \frac{\sqrt{(\lambda_{max} - \lambda_C)(\lambda_C - \lambda_{min})}}{\lambda_C} \quad (29)$$

and

$$\lambda_{min}^{max} = \sigma^2 \left(1 + 1/Q \pm 2\sqrt{1/Q}\right) \quad (30)$$

with a ratio $Q = \frac{N}{M} \gg 1$, $\lambda_C \in [\lambda_{min}, \lambda_{max}]$ and σ^2 being the variance of the elements of $\mathbf{C}$.

From Equation (29), and taking into attention that $N \rightarrow \infty$, we can predict the following:

- A. The lower “edge” of the spectrum is positive (except the case $Q = 1$ where $\lambda_{min} = 0$ and therefore it diverges); for the other cases there is no eigenvalue between 0 and λ_{min} . Near this edge the density of the eigenvalues exhibits a sharp maximum;
- B. The density of eigenvalues vanishes above a certain upper edge λ_{max} .

We can treat Equation (25) in a more general way. We will need to define the “resolvent” $G_H(z)$ of the matrix $\mathbf{H}$, most well known by Stieltjes transform, as:

$$G_H(z) = \frac{1}{N} \text{Tr} \left[(z\mathbf{I} - \mathbf{H})^{-1} \right], \quad (31)$$

where z is a complex number and $\mathbf{I}$ is the identity matrix. Then, the eigenvalues spectrum would be,

$$\rho_N(\lambda) = \lim_{\epsilon \rightarrow 0} \frac{1}{\pi} \Im (G_H(\lambda - i\epsilon)), \quad (32)$$

with $\Im$ being the imaginary part of the complex number. When N tends to infinity, in the limit, we almost surely have a unique and well defined density $\rho_\infty(\lambda)$ [Bouchaud and Potters, 2011]. This asymptotic result, under certain conditions, can be used to describe the eigenvalue density of a single instance. This is probably the cause to RMT great success.

Eigenvalues in literature

In the last fifteen years, several authors have been applying RMT in a tentative to understand the structure of financial correlation matrices in such a highly random setting.

For a first lecture on the problematic Galluccio et al. [1998] will do. Plerou et al. [1999] shown that for the correlation matrix of 406 companies in the S&P index, on daily data, from 1991 to 1996, only seven out of the 406 eigenvalues were clearly significant with respect to a random null hypothesis, that is, the statistics of the most of the eigenvalues of the correlation matrix calculated from stock return series agree with the predictions

of random matrix theory, but with deviations for a few of the largest eigenvalues, and their corresponding eigenvectors.

This was also observed in other studies: Laloux et al. [1999], Laloux et al. [2000], Plerou et al. [2000], Plerou et al. [2001], Plerou et al. [2002], Sharifi et al. [2004] and Wilcox and Gebbie [2004]. Also, in these studies, the correlation (or covariance) matrices of financial time series appeared to contain such a large amount of noise that the eigenvalue structure could essentially be regarded as random.

However, some previous studies, see as an example [Gopikrishnan et al., 1999], have focused only on the largest eigenvalue with no attention paid to the others.

Extended work by [Plerou et al., 1999] was conducted to explain information contained in the deviating eigenvalues, which revealed that the largest eigenvalue corresponds to a market wide influence to all stocks and the remaining deviating eigenvalues correspond to conventionally identified business sectors. This also suggested that it is possible to improve estimates by setting the insignificant eigenvalues to zero, mimicking a common noise-reduction method used in signal processing.

Wilcox and Gebbie [2004] examined the composition of all the eigenvalues of ten years of Johannesburg Stock Exchange. The authors concluded that the leading, that is, the first three, eigenvalues may be interpreted in terms of independent trading strategies with long range correlations indicating a role not just for one but also for a small number of the dominant eigenvalues. This means that only a few of the larger eigenvalues might carry collective information.

All these results strongly suggest that eigenvalues of correlation matrix falling under the Marchenko-Pastur distribution contain no genuine information about the financial markets. Hence, one should systematically filter out such noise from the correlations for more accurate estimations of, for instance, future portfolio risk. Following Wilcox and Gebbie [2004], Sharkasi et al. [2006a] we will consider the three larger eigenvalues and its respective eigenvectors as carrying meaningful information.

Further, Kwapien et al. [2005] investigated the distribution of eigenvalues of correlation matrices for equally-separated time windows with respect to the German DAX in order to study, quantitatively, the relation between stock price movements and properties of the distribution of the corresponding index motion. They reported that the importance of an eigenvalue is related to the correlation strength of different stocks, which means that the more aggregated the market behaviour, the larger the first eigenvalue (the maximum eigenvalue).

In this context, another relevant study is the one done by Drozd et al. [2007] with a comparison between empirical data and random matrix theory.

Dynamics of the top eigenvector

The Wigner and the Marchenko-Pastur ensembles are in some sense maximally random as no prior information about the matrices is assumed. But, for stock markets, it is intuitive that stocks are sensitive, for example, to global news about the economy. So, we must have some, at least one, common factor to all stocks. A reasonable null-hypothesis is that the true correlation matrix is:

$$C_{ii} = 1, C_{ij} = \bar{\rho}, \forall i \neq j. \quad (33)$$

This corresponds to add a rank one perturbation matrix to the empirical correlation matrix with one large eigenvalue $N\bar{\rho}$ and $N - 1$ zero eigenvalues. When $N\bar{\rho} \gg 1$, the empirical correlation matrix will also have a large eigenvalue close to $N\bar{\rho}$.

But, what happens when $N\bar{\rho}$ is not very large compared to unity? That case was solved in great detail in 2005 (Bouchaud and Potters [2011]). There it was considered a more general case where the true correlation matrix has k special eigenvalues, called “spikes”. So, in general, financial covariance matrices are such that a few large eigenvalues are well separated from the “bulk”, where all other eigenvalues reside. So, again, we expect to have a large eigenvalue $\lambda_{max} \approx N\bar{\rho}$ when stocks are correlated on average. The associated eigenvector is the so-called “market mode”, that is to say, in a first view, all stocks move in the same direction.

Plerou et al. [1999] and Plerou et al. [2002] found that the distribution of eigenvector components for the eigenvectors corresponding to the eigenvalues outside the RMT bound displayed systematic deviations from the RMT prediction and that these “deviating eigenvectors” were stable in time. They analyzed the components of the deviating eigenvectors and found that the largest eigenvalue corresponded to an influence common to all stocks.

Their analysis of the remaining deviating eigenvectors showed distinct groups, whose identities corresponded to conventionally-identified business sectors. The important question, here, is then if and if yes how do these λ_{max} and $\vec{V}_{max}$ behave in time.

2.4 COMPONENT ANALYSIS

Reducing the parameter space is a commonly used approach for successfully modelling multivariate time series, because the number of parameters involved increases quickly with the dimension of the series.

Several methods are available to perform dimension reduction, including the canonical correlation analysis (CCA) of Box and Tiao [1977], the factor models of Peña and Box [1987], the independent components analysis (ICA) of Back and Weigend [1997], and the principal components analysis (PCA) of Stock and Watson [2002]. These methods seek linear combinations that have certain characteristics useful in model building: for instance, the CCA produces linear combinations that rank from the most predictable to the least predictable.

2.4.1 Principal Component Analysis

PCA invention is attributed to Karl Pearson (1901) who created this as an analogue of the principal axes theorem in mechanics; it was later independently developed and named by Harold Hotelling in the 1930s. The method is mostly used as a tool in exploratory data analysis and for making predictive models.

In fact, PCA is closely related to RMT, since it is also done through eigenvalue decomposition of the correlation (or covariance) matrix of the return series. This method uses an orthogonal transformation to convert a set of possible correlated returns into several uncorrelated components, which are ranked by their explanatory power for the total variance of the system.

As an example, Meric and Meric [1997] applied the Box M method and PCA to test whether or not the correlation matrices before and after the international crash of 1987 were significantly different. Their results showed that there are significant alterations in the co-movements of the studied markets and that the benefits of international diversification for the European markets decreased markedly after this crash.

Definition

PCA is defined as a statistical procedure that by means of an orthogonal transformation converts a set of observations of (possibly correlated) variables into a set of linearly uncorrelated variables called principal components. This transformation is defined in such a way that the first principal component has the largest possible variance. The remaining components have the highest variance possible under the constraint that they are orthogonal (uncorrelated with) to the preceding components. Principal components are guaranteed to be independent if the data set is jointly normally distributed.

PCA is considered the simplest of the true eigenvector-based multivariate analyses. Its main objective, as stated above, is to decompose the fluctuations of the quantity r_i^t into uncorrelated components of decreasing variance. This quantity can be written in terms of the eigenvalues λ_α and the eigenvectors $\vec{V}_\alpha$ as:

$$r_i^t = \sum \sqrt{\lambda_\alpha} V_{\alpha,i} \epsilon_\alpha^t, \quad (34)$$

where $V_{\alpha,i}$ is the i -th component of $\vec{V}_\alpha$ and ϵ_α^t are uncorrelated (for different α 's) random variables of unit variance. This PCA decomposition is quite useful in some situations like the one with a dominant eigenvalue. Then, as a good approximation of the dynamics of the N variables r_i we have:

$$r_i^t \approx \sqrt{\lambda_1} V_{1,i} \epsilon_1^t. \quad (35)$$

So, the $V_{\alpha,i}$ can be physically interpreted as being the weights of the different stocks $I = 1, \dots, N$. Also, typically in stock markets, the largest eigenvalue is called the “market mode” and corresponds, in a portfolio view, to invest equally on all stocks, $V_{1,i} = 1/\sqrt{N}$.

PCA algorithms

PCA algorithms use only second order statistical information, so the higher order statistical information provided by non-Gaussian signals is not required or used. PCA algorithms can be either implemented with standard, or “batch”, algorithms or with on-line algorithms. Examples of on-line or “neural” PCA algorithms include Baldi and Hornik [1989] and Oja [1989].

2.4.2 Independent Component Analysis

The method known as independent component analysis (ICA) is also named as blind source separation (Heraut and Jutten [1986], Jutten and Heraut [1991] and Common [1994]). The central assumption is that an observed multivariate time series (such as daily stock returns) reflect the reaction of a system (such as the stock market) to a

few statistically independent time series. ICA seeks to extract out these independent components as well as the mixing process. ICA can be expressed in terms of the related concepts of entropy [Bell and Sejnowski, 1995], mutual information [Amari et al., 1996], contrast functions [Common, 1994] and other measures of the statistical independence of signals [Back and Weigend, 1997].

In financial context, ICA was proposed for the first time by Moody and Wu [1996] to separate the observational noise from the true price in a foreign exchange rate time series. Concerning the PSI-20 index a very interesting study using ICA is Dionisio et al. [2006].

ICA denotes, then, the process of taking a set of measured signal vectors and extracting from them a (new) set of statistically independent vectors called the independent components or the sources. They are estimates of the original source signals which are assumed to have been mixed in some prescribed manner to form the observed signals.

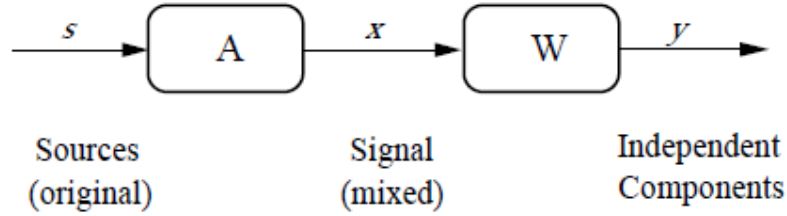


Figure 3: Schematic representation of ICA

The original sources are mixed through matrix to form the observed signal. The demixing matrix transforms the observed signal into the independent components. Figure 3 shows the most basic form of ICA.

Now, we present the basic ICA model according to the formal definition given by Common [1994].

Definition

ICA assumes that the observed data are generated by a set of unobserved components that are independent. Let $\mathbf{x}_t = (x_{1t}, x_{2t}, \dots, x_{mt})^T$ be the m -dimensional vector of stationary time series, with $E[\mathbf{x}_t] = \mathbf{0}$ and $E[\mathbf{x}_t \mathbf{x}_t^T] = \Gamma_x(0)$ being positive definite. It is assumed that $\mathbf{x}_t$ is generated by a linear combination of r ($r \leq m$) latent factors. That is,

$$\mathbf{x}_t = \mathbf{A} \mathbf{s}_t, \quad t = 1, 2, \dots, T \quad (36)$$

where $\mathbf{A}$ is an unknown $m \times r$ full rank matrix, with elements a_{ij} that represent the effect of s_{jt} on x_{it} , for $i = 1, 2, \dots, m$ and $j = 1, 2, \dots, r$ and $\mathbf{s}_t = (s_{1t}, s_{2t}, \dots, s_{rt})^T$ is the vector of unobserved factors, which are called independent components (ICs).

It is assumed that $E[\mathbf{s}_t] = \mathbf{0}$, $\Gamma_s(0) = E[\mathbf{s}_t \mathbf{s}_t^T] = \mathbf{I}_r$, and that the components of $\mathbf{s}_t$ are statistically independent. Let $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T)$ be the observed multivariate time series. The problem is to estimate both $\mathbf{A}$ and $\mathbf{s}_t$ from only $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T)$. That is, ICA looks for an $r \times m$ matrix, $\mathbf{W}$, such that the components given by

$$\hat{\mathbf{s}}_t = \mathbf{W}\mathbf{x}_t, \quad t = 1, 2, \dots, T \quad (37)$$

are as independent as possible. However, previous assumptions are not sufficient to enable us to estimate $\mathbf{A}$ and $\mathbf{s}_t$ uniquely, and it is required that no more than one independent component be normally distributed. From Equation (36) we have:

$$\Gamma_{\mathbf{x}}(0) = E[\mathbf{x}_t \mathbf{x}_t^T] = \mathbf{A}\mathbf{A}^T \quad (38)$$

$$\Gamma_{\mathbf{x}}(\tau) = E[\mathbf{x}_t \mathbf{x}_{t-\tau}^T] = \mathbf{A}\Gamma_{\mathbf{s}}(\boldsymbol{\theta})\mathbf{A}^T, \quad \tau \geq 1. \quad (39)$$

All of the dynamic structure of the data therefore comes through the unobserved components.

ICA algorithms

ICA algorithms may use higher than 2 order statistical information for separating the signals (see, for example, [Cardoso \[1989\]](#) and [Common \[1994\]](#)). For this reason non-Gaussian signals (or at most, one Gaussian signal) are normally required for ICA algorithms based on higher order statistics. ICA algorithms based on second order statistics have also been proposed ([Belouchrani et al. \[1997\]](#)).

The earliest ICA algorithm that we are aware of and one which started much interest in the field is that proposed by [Heraut and Jutten \[1986\]](#). Since then, various approaches have been proposed in the literature to implement ICA. These include: minimizing higher order moments ([Cardoso \[1989\]](#)) or higher order cumulants ([Cardoso and Souloumiac \[1993\]](#)), maximization of mutual information of the outputs or maximization of the output entropy ([Bell and Sejnowski \[1995\]](#)), minimization of the Kullback-Leibler divergence between the joint and the product of the marginal distributions of the outputs ([Amari et al. \[1996\]](#)).

ICA algorithms are typically implemented in either off-line (batch) form or using an on-line approach.

2.4.3 Forecastable Component Analysis (ForeCA)

Data reduction (DR) techniques are often applied to multivariate time series $\mathbf{X}_t$, hoping that forecasting on the lower dimensional space $\mathbf{S}_t$ is more accurate, simpler and more efficient than the usual techniques. For instance, standard DR techniques such as PCA or ICA, do not explicitly address forecastability of the sources. That rises the interrogation: just because a signal has high variance does not mean it is easy to forecast.

Here, we introduce Forecastable Component Analysis (ForeCA), another dimension reduction technique for temporally dependent signals, following [Goerg \[2013\]](#). Based on a new forecastability measure, ForeCA finds an optimal transformation to separate a multivariate time series into a forecastable and an orthogonal white noise space.

Definition 8. For a second-order stationary process y_t , let

$$\begin{aligned}\Omega : y_t &\rightarrow [0, \infty] \\ \Omega(y_t) &= 1 - \frac{H_{s,a}(y_t)}{\log_a(2\pi)} = 1 - H_{s,2\pi}(y_t).\end{aligned}\tag{40}$$

be the forecastability of y_t , with

$$H_{s,a}(y_t) := - \int_{-\pi}^{\pi} f_y(\lambda) \log_a f_y(\lambda) d\lambda \tag{41}$$

being the differential entropy of the spectral density of y_t , $f_y(\lambda)$, and $a > 0$ the logarithm base.

About $\Omega(y_t)$ properties we can say that it satisfies:

- $\Omega(y_t) = 0$ if and only if y_t is white noise, that is, a random signal with constant power spectral density;
- invariant to scaling and shifting, that is, $\Omega(ay_t + b) = \Omega(y_t)$ for $a, b \in \mathbb{R}$, $a \neq 0$;
- max sub-additivity for uncorrelated processes, that is

$$\Omega(ax_t + \sqrt{1 - \alpha^2}y_t) \leq \max\{\Omega(x_t), \Omega(y_t)\},$$

if $Ex_t y_s = 0$ for all $s, t \in \mathbb{Z}$; equality if and only if $\alpha \in \{0, 1\}$.

The goal, here, is to find a linear combination of a multivariate second-order stationary time series $\mathbf{X}_t$, that makes $y_t = w^T \mathbf{X}_t$ as forecastable as possible. Based on the previous definition we can state the ForeCA optimization problem:

$$\max_w \Omega(w^T \mathbf{X}_t) = \max_w \left(1 + \frac{\int_{-\pi}^{\pi} f_y(\lambda) \log_a f_y(\lambda) d\lambda}{\log_a(2\pi)} \right) \tag{42}$$

subject to $w^T \Sigma_{\mathbf{X}} w = 1$. Proof details can be followed in [Goerg \[2013\]](#).

2.5 ENTROPY

The early notion of entropy as a measure of disorder comes from the work of Clausius in the 19th century, where entropy provided a way to state the second law of Thermodynamics as well as a definition of temperature. This law postulates that the entropy of an isolated system tends to increase continuously until it reaches its equilibrium state. Later, around 1900, within the framework of Statistical Physics established by Boltzmann and Gibbs, it was defined as a statistical concept.

In 1948, entropy found its way in engineering and mathematics, through the works of Shannon in information theory and mathematics and of Kolmogorov in probability theory. [Shannon \[1948\]](#) gave a new meaning to entropy in the context of Information Theory, relating entropy with the absence/presence of information in a given message.

The theoretical ground of entropy proved to be fertile and “only” twenty six years ago, [Tsallis \[1988\]](#) generalized again the concept of entropy, introducing the idea of

non-extensive entropy although this idea was already present in Rényi's work in the 60's [Rényi, 1961]. Significant research has been done ever since with Shannon entropy providing the general framework for the treatment of equilibrium systems where short/space/temporal interactions dominate.

Entropy, one of the early ideas behind thermodynamics that later led the way to the emergence of Statistical Physics, has been shown to be pervasive and, perhaps surprisingly, well suited to crossing disciplinary boundaries, giving an easier interpretation to the previously defined concept of topological entropy. The influence of thermodynamics was such that it lent its name to the thermodynamical formalism by Bowen and Ruelle [Ruelle, 2004].

The idea here is to apply entropy concepts to financial time series. For a good starting point follow Maasoumi and Racine [2002]. For a more general thermodynamic approach see McCauley [2003].

2.5.1 Definition

Definition 9. Let X be a discrete random variable on a finite set $\mathcal{X} = \{x_1, \dots, x_n\}$, with a probability distribution function $p(x) = P(X = x)$. The entropy $H(X)$ of X is defined as

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log p(x). \quad (43)$$

Higher entropy implies less predictability, which seems to be the case for all financial markets. If we apply the previous definition to a continuous time series, e.g. financial, we have to partition the signal into k symbols, in order to complete the partition we need to choose the length of the words we will be using, say size m . The Shannon entropy for symbol sequences, with an alphabet of k symbols and block length m , gets a particular form [Kantz and Schreiber, 2004].

Before presenting the formula it is necessary a short introduction on how to code the sequences. We have k^m possible sequences, we can associate any integer number j , such that $0 \leq j < k^m$, with its digit representation on base k as $j = (j_{m-1}j_{m-2} \dots j_1j_0)_k$, where each digit $0 \leq j_i < k$ for $0 \leq i < m$. We can then associate a probability p_j to each of these sequences.

Definition 10. The Shannon entropy for blocks of size m for an alphabet of k symbols is

$$\tilde{H}(m) = - \sum_{j=0}^{k^m-1} p_j \log p_j, \quad (44)$$

the entropy of the source is then

$$\tilde{h} = \lim_{m \rightarrow \infty} \frac{\tilde{H}(m)}{m}. \quad (45)$$

This definition is attractive for several reasons: it is easy to calculate and it is well defined for a source of symbol strings. In the particular case of returns, if we choose a symmetrical partition we know that half of the symbols represent losses and half of the symbols represent gains. If the sequence is predictable, we have the same losses and

gains sequences repeated every time, the entropy will be lower; if however all sequences are equally probable the uncertainty will be higher and so it will be the entropy. Entropy is thus a good measure of uncertainty.

This particular method has problems (the entropy depends on the choice of encoding) as it is not a unique characteristic for the underlying continuous time series. Also, since the number of possible states grows exponentially with m , after a short number of sequences, in practical terms it will become difficult to find a sequence that repeats itself. This entropy is not invariant under smooth coordinate changes, both in time and encoding. Also, the entropy shows a different behaviour for odd and even k if we have a large bulk in the centre of the distribution, as it usually happens for financial time series. These are strong handicaps for its adoption into financial time series study.

2.5.2 Entropy different incantations

But, Shannon entropy is only the entrance door to entropy world. In fact, many systems do not satisfy the simplifying assumptions of ergodicity and independence. Due to the prevalence of these phenomena, several entropy measures were derived. Among them, a most popular one is Tsallis entropy, which constitutes itself as a generalized form of Shannon entropy. Despite the debate generated over its meaning, for which the profusion of several mathematical constructions has certainly played a central role, entropy is commonly understood as a measure of disorder, uncertainty, ignorance, dispersion, disorganization, or even, lack of information.

More recently, an econometric meaning has been given to entropy, while considering that the entropy of an economic system is a measure of the ignorance of the researcher who knows only some moments values representing the underlying population. Besides its multiples applications, entropy has started to be perceived as a consistent alternative to the standard-deviation, when assessing stock market volatility.

The underlying rationality is that, as a more generalized measure, entropy is able to capture uncertainty regardless of the kind of the empirical distribution evidenced by the data. This is especially so, as it is widely recognized that returns are usually non-normally distributed, where the application of the standard-deviation turns out to be unsatisfactory. Entropy, as a function of many moments of the probability distribution function, considers much more information than the standard-deviation. Some of the main potentialities of this measure are:

- It can be defined either for quantitative or qualitative observations;
- Whereas entropy depends on the potential number of states of a distribution it is a result of the specific weight of each state;
- The information value is related to the respectively distribution function.

2.5.2.1 Order- q Rényi entropies

A series of entropy-like quantities, the order- q Rényi entropies [Rényi, 1961], characterise the amount of information which is needed in order to specify the value of an observable with a certain precision [Kantz and Schreiber, 2004].

Definition 11. Let $\mathcal{P}_\epsilon$ be a partition of disjoint boxes $\mathcal{P}_j$, of size $\text{length} \leq \epsilon$, over the support of measure μ . If we consider $\mu(\mathcal{P}_j) = p_j$ then

$$\tilde{H}_q(\mathcal{P}_\epsilon) = \frac{1}{1-q} \log \sum_j p_j^q \quad (46)$$

is the q -order Rényi entropy for the partition $\mathcal{P}_\epsilon$.

Note for $q = 1$ we have to apply the l'Hopital rule where we get

$$\tilde{H}_1(\mathcal{P}_\epsilon) = -p \sum_j p_j \log p_j. \quad (47)$$

$\tilde{H}_1(\mathcal{P}_\epsilon)$ is thus the Shannon entropy as defined in Equation (43). In contrast to the other Rényi entropies is additive, i.e., if the probabilities can be factorised into independent factors, the entropy of the joint process is the sum of the entropies of the independent processes.

2.5.2.2 Kolmogorov-Sinai entropy

The Rényi entropies gain even more relevance when they are applied to transition probabilities, Equation (45). We apply the same reasoning as before: apply a partition $\mathcal{P}_\epsilon$ on the dynamic range of the observable, and introduce the joint probability $p_{i_1, i_2, \dots, i_m}$ that at an arbitrary time n the observable falls into the interval I_{i_1} , at time $n+1$ fall into interval I_{i_2} , and so on.

Definition 12. The block entropies of block size m is

$$H_q(m, \mathcal{P}_\epsilon) = \frac{1}{1-q} \log \sum_{i_1, i_2, \dots, i_m} p_{i_1, i_2, \dots, i_m}^q. \quad (48)$$

The order- q entropies are then

$$h_q = \sup_{\mathcal{P}} \lim_{m \rightarrow \infty} \frac{1}{m} H_q(m, \mathcal{P}_\epsilon) \Leftrightarrow h_q = \sup_{\mathcal{P}} \lim_{m \rightarrow \infty} h_q(m, \mathcal{P}_\epsilon), \quad (49)$$

where

$$h_q(m, \mathcal{P}_\epsilon) := H_q(m+1, \mathcal{P}_\epsilon) - H_q(m, \mathcal{P}_\epsilon), \quad h_q(0, \mathcal{P}_\epsilon) = H_q(0, \mathcal{P}_\epsilon). \quad (50)$$

In the original sense only h_1 was called the Kolmogorov-Sinai entropy [Kolmogorov, 1958, Sinai, 1959], but since the idea is the same, the name was extended to cover all the other Rényi entropies.

Kolmogorov and Sinai were the first to consider correlations in time in information theory. The limit $q \rightarrow 0$ gives the topological entropy h_0 . As D_0 , the fractal dimension of the support of the measure, just counts the number of non-empty boxes in partition, h_0 gives just a measure of the different orbits, not of their relative importance as we get with h_1 .

Another extension of entropy, related with Rényi entropies, is Tsallis non extensive entropy [Tsallis, 1988], with applications to economics described in Tsallis et al. [2003].

2.5.3 Mutual Information

Gaussian processes can be completely defined by second order statistics, namely the mean and the variance, but when talking about non-Gaussian processes higher order statistics are needed.

We will make use of second order statistics Correlation Coefficient and the high order statistics known as Mutual Information (MI) to measure the dependency between two random variables. In fact, the Mutual Information, though hard to compute, is a natural measure of the independence between random variables. MI accounts for the whole dependency structure and not only the covariance.

We can define the Mutual Information by the entropies $H(X)$, $H(Y)$ and $H(X, Y)$ (see for example Papoulis [1985]):

$$MI(X; Y) = H(X) - H(X|Y) \quad (51)$$

$$H(X|Y) = H(X, Y) - H(Y) \quad (52)$$

$$MI(X; X) = H(X). \quad (53)$$

Mutual Information is always non-negative and zero if and only if the variables are statistically independent.

2.5.4 Kullback-Leibler Divergence

Following the 1951 classical paper of S. Kullback and R.A. Leibler entitled "On information and sufficiency" [Kullback and Leibler, 1951] it is presented the Kullback-Leibler divergence. Kullback and Leibler were concerned with the statistical problem of discrimination, by considering a measure of the "distance" or "divergence" between statistical populations in terms of their measure of information.

For independent signals, the joint probability can be factorized into the product of the marginal probabilities. Therefore, the independent components can be found by minimizing the Kullback-Leibler divergence, or distance, between the joint probability and marginal probabilities of the output signals [Amari et al., 1996].

Hence, the goal of finding statistically independent components can be expressed in several ways: look for a set of directions that factorize the joint probabilities and, then, find a set of "interesting" directions with minimum mutual information. Where the mutual information between variables vanish, they are statistically independent.

The goal of finding interesting directions is similar to projection pursuit (Friedman and Tukey [1974] and Huber [1985]). In the knowledge discovery and data mining community the term "interestingness" (Ripley [1996]) is also used to denote unexpectedness (Silberschatz and Tuzhilin [1996]).

Assuming that H_i , $i = 1, 2$, is the hypothesis that x was selected from the population whose density function is f_i , $i = 1, 2$, then we define

$$\log \frac{f_1(x)}{f_2(x)} \quad (54)$$

as the information in x for discriminating between H_1 and H_2 .

In their seminal paper (Kullback and Leibler [1951]), they have denoted by $I(1, 2)$ the mean information for discrimination between H_1 and H_2 per observation from f_1 , i.e.,

$$I(1, 2) = KL_x(f_1, f_2) = \int f_1(x) \log \frac{f_1(x)}{f_2(x)}. \quad (55)$$

This quantity, in Equation (55) is called the Kullback-Leibler divergence and is denoted by $KL(f_1, f_2)$, despite the fact that, originally, Kullback and Leibler denoted

$$J(1, 2) = KL(f_1, f_2) + KL(f_2, f_1) \quad (56)$$

as the divergence between f_1 and f_2 .

Now, let us consider some properties of Kullback-Leibler divergence:

- $KL(f_1, f_2) \geq 0$ with $KL(f_1, f_2) = 0$ if and only if $f_1(x) = f_2(x)$ almost everywhere;
- $KL(f_1, f_2) \neq KL(f_2, f_1)$, that is, $KL(f_1, f_2)$ is not symmetric;
- $KL(f_1, f_2)$ is additive for independent random events: $KL_{xy}(f_1, f_2) = KL_x(f_1, f_2) + KL_y(f_1, f_2)$, being X and Y independent variables;

For most densities f_1 and f_2 , $KL(f_1, f_2)$ needs to be computed numerically. One exception is when f_1 and f_2 are both Gaussian distributions.

In the univariate case, the Kullback-Leibler divergence between two Gaussian distributions p, q with means μ_1, μ_2 and variances σ_1^2, σ_2^2 , is given by

$$KL(p, q) = \log \frac{\sigma_1}{\sigma_2} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}. \quad (57)$$

In the multivariate case, the Kullback-Leibler divergence between multivariate Gaussian distributions p, q is given by:

$$KL(p, q) = 0.5 \left[\log \left(\frac{\det(\Sigma_2)}{\det(\Sigma_1)} \right) + \text{tr} \left(\Sigma_2^{-1} \Sigma_1 \right) + (\mu_2 - \mu_1)^T \Sigma_2^{-1} (\mu_2 - \mu_1) - N \right], \quad (58)$$

with mean vectors μ_1, μ_2 and covariance matrices Σ_1, Σ_2 .

2.5.5 Approximate Entropy

The Approximate Entropy (*ApEn*) method is an information theory based estimate of the complexity of a time series introduced by Steve Pincus [Pincus, 1991], formally based on the evaluation of joint probabilities, in a way similar to the entropy of Eckmann and Ruelle [Eckman and Ruelle, 1985]. The original motivation and main feature, however, was not to characterize an underlying chaotic dynamics, rather to provide a robust model-independent measure of the randomness of a time series of real data, possibly - as it is usually in practical cases - from a limited data set affected by a superimposed noise.

ApEn has been used by now to analyse data obtained from very different sources. See, for instance, Ho et al. [1997]. These authors point some weaknesses to *ApEn*, namely its

strong dependence on sequence length and its poor self-consistency (i.e., the observation that $ApEn$ for one data set is larger than $ApEn$ for another for a given choice of parameters should, but does not, hold true for other parameters choices).

Given a sequence of N numbers $\{u(j)\} = \{u(1), u(2), \dots, u(N)\}$, with equally spaced times $t_{j+1} - t_j \equiv \Delta t = \text{const}$, one first extracts the sequences with embedding dimension m , that is, $x(i) = \{u(i), u(i+1), \dots, u(i+m-1)\}$, with $1 \leq i \leq N - m + 1$. The $ApEn$ is then computed as

$$ApEn = \Phi^m(r) - \Phi^{m+1}(r), \quad (59)$$

where r is a real number representing a threshold distance between series, and the quantity $\Phi^m(r)$ is defined as

$$\Phi^m(r) = \langle \ln [C_i^m(r)] \rangle = \frac{\sum_{i=1}^{N-m+1} \ln [C_i^m(r)]}{N - m + 1}. \quad (60)$$

Here $C_i^m(r)$ is the probability that the series $x(i)$ is closer to a generic series $x(j)$ with $(j \leq N - m + 1)$ than the threshold r ,

$$C_i^m(r) = \frac{N[d(i, j) \leq r]}{N - m + 1}, \quad (61)$$

with $N[d(i, j) \leq r]$ the number of sequences $x(j)$ close to $x(i)$ less than r . As definition of distance between two sequences, the maximum difference (in modulus) between the respective elements is used,

$$d(i, j) = \max_{k=1,2,\dots,m} (|u(j+k-1) - u(i+k-1)|). \quad (62)$$

For a somewhat more mathematical presentation of this subject see [Rukhin \[2000\]](#). Only more recently this method has been introduced to financial time series ([Pincus and Kalman \[2004\]](#) and [Pincus \[2008\]](#)).

2.6 ENERGY STATISTICS

Energy statistics and energy distance are concepts developed by [Székely et al. \[2007\]](#) and were born in the more broad field of independence [[Bakirov et al., 2006](#)]. Energy statistics is based on the notion of potential energy as presented by Newton. Statistical observations are like heavenly bodies governed by a statistical potential energy which is zero only when an underlying statistical null hypothesis is present. In this way, energy statistics are functions of distances between statistical observations.

Distance correlation is a recent multivariate dependence coefficients approach to the problem of measuring the dependence between random vectors, even if they are arbitrary and/or not of equal dimension. The pertinence of this measure to this work relies on the fact that an interesting approach to measure complicated dependence structures in multivariate data (see, for instance, [Embrechts et al. \[2002\]](#) or [Feuerverger \[1993\]](#)) is to study their vectors independence.

2.6.1 Definitions

Energy distance was introduced in 1985 and is a (statistical) distance between probability distributions. If X and Y are independent random vectors in $\mathbb{R}^d$ with cumulative distribution functions F and G respectively, then the energy distance between these distributions is:

$$D(F, G) = 2E\|X - Y\| - E\|X - X'\| - E\|Y - Y'\| \quad (63)$$

where X, X' and Y, Y' are independent and identically distributed. $D(F, G) = 0$ if and only if X and Y are identically distributed.

Later, Székely et al, based on this energy statistics, developed the concept of distance covariance (dCov) as the square root of

$$v_n^2 = \frac{1}{n^2} \sum_{k,l=1}^n A_{kl} B_{kl}, \quad (64)$$

where A_{kl} and B_{kl} are linear functions of the pairwise distance between sample elements.

The distance correlation goes beyond the classical Pearson product-moment correlation, ρ , when in the multivariate environment because the diagonal covariance matrix generated implies independence but it is not a sufficient condition for independence. Over the years other methods have been proposed, and one of them, most notably proposed by Rényi called maximal correlation.

For all distributions with finite first moments, the distance correlation R generalizes the idea of correlation in, at least, two ways:

1. $R(X, Y)$ is defined for X and Y in arbitrary dimensions;
2. $R(X, Y) = 0$ characterizes independence of X and Y .

This coefficient $R(X, Y)$ satisfies $0 \leq R(X, Y) \leq 1$ and $R(X, Y) = 0$ only if X and Y are independent. In this way distance covariance and distance correlation provide a natural extension of Pearson product-moment covariance $\sigma_{X,Y}$ and correlation ρ .

Let X in $\mathbb{R}^p$ and Y in $\mathbb{R}^q$ be random vectors, where p and q are positive integers. We will also denote f_X as the characteristic function of X , f_Y as the characteristic function of Y and $f_{X,Y}$ as the joint characteristic function of X and Y . X and Y are independent if and only if $f_{X,Y} = f_X f_Y$, in what concerns characteristic functions. So, it is a natural idea to try to find a suitable norm to measure the distance between $f_{X,Y}$ and $f_X f_Y$.

Székely and Rizzo [2009] defined a measure of dependence

$$v^2(X, Y; w) = \|f_{X,Y}(t, s) - f_X(t) f_Y(s)\|_w^2, \quad (65)$$

that is,

$$v^2(X, Y; w) = \int_{\mathbb{R}^{p+q}} |f_{X,Y}(t, s) - f_X(t) f_Y(s)|^2 w(t, s) dt ds, \quad (66)$$

with a suitable choice of an arbitrary positive weight function $w(t, s)$ so that this measure of dependence is analogous to classical covariance, but with the property that $v^2(X, Y; w) = 0$ if and only if X and Y are independent.

Definition 13. The distance covariance (dCov) between random vectors X and Y with finite first moments (that is $E\|X\|_p < \infty$ and $E\|Y\|_q < \infty$) is the non-negative number $\nu(X, Y)$ defined by

$$\nu^2(X, Y) = \|f_{X,Y}(t, s) - f_X(t) f_Y(s)\|^2, \quad (67)$$

where t and s are vectors.

Similarly,

Definition 14. Distance variance (dVar) is defined as the square root of $\nu^2(X) = \nu^2(X, X) = \|f_{X,X}(t, s) - f_X(t) f_X(s)\|^2$. By definition of the norm $\|\cdot\|$, it is clear that $\nu(X, Y) \geq 0$ and $\nu(X, Y) = 0$ if and only if X and Y are independent.

We can now define distance correlation.

Definition 15. The distance correlation (dCor) between random vectors X and Y with finite first moments is the non-negative number $R(X, Y)$ defined by

$$R^2(X, Y) = \begin{cases} \frac{\nu^2(X, Y)}{\sqrt{\nu^2(X)\nu^2(Y)}}, & \nu^2(X)\nu^2(Y) > 0; \\ 0, & \nu^2(X)\nu^2(Y) = 0. \end{cases} \quad (68)$$

Remains the problem of the calculus of these quantities. To define the distance dependence statistics we consider a random sample $(\mathbf{X}, \mathbf{Y}) = \{(X_k, Y_k) : k = 1, \dots, n\}$ of n i.i.d random vectors (X, Y) from the joint distribution of the random vectors X and $\mathbb{R}^p$ and Y and $\mathbb{R}^q$. Then to compute the Euclidean distance matrices $(a_{kl}) = (|X_k - X_l|_p)$ and $(b_{kl}) = (|Y_k - Y_l|_q)$ we define $A_{kl} = a_{kl} - \bar{a}_{k.} - \bar{a}_{.l} + \bar{a}_{..}$, $k, l = 1, \dots, n$, where

$$\bar{a}_{k.} = \frac{1}{n} \sum_{l=1}^n a_{kl}, \quad \bar{a}_{.l} = \frac{1}{n} \sum_{k=1}^n a_{kl}, \quad \bar{a}_{..} = \frac{1}{n^2} \sum_{k,l=1}^n a_{kl}. \quad (69)$$

Similarly we define $B_{kl} = b_{kl} - \bar{b}_{k.} - \bar{b}_{.l} + \bar{b}_{..}$, $k, l = 1, \dots, n$.

Definition 16. The non-negative sample distance covariance $\nu_n(\mathbf{X}, \mathbf{Y})$ and sample distance correlation $R_n(\mathbf{X}, \mathbf{Y})$ are defined by

$$\nu_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2} \sum_{k,l=1}^n A_{kl} B_{kl}, \quad (70)$$

and

$$R_n^2(\mathbf{X}, \mathbf{Y}) = \begin{cases} \frac{\nu_n^2(\mathbf{X}, \mathbf{Y})}{\sqrt{\nu_n^2(\mathbf{X})\nu_n^2(\mathbf{Y})}}, & \nu_n^2(\mathbf{X})\nu_n^2(\mathbf{Y}) > 0; \\ 0, & \nu_n^2(\mathbf{X})\nu_n^2(\mathbf{Y}) = 0, \end{cases} \quad (71)$$

respectively, and where the sample distance variance is defined by

$$\nu_n^2(\mathbf{X}) = \nu_n^2(\mathbf{X}, \mathbf{X}) = \frac{1}{n^2} \sum_{k,l=1}^n A_{kl}^2. \quad (72)$$

2.6.2 Properties

Here, we will show some properties taken from the theorems in Székely and Rizzo [2009] and from previous results in Székely et al. [2007].

Theorem 17. *If $(\mathbf{X}, \mathbf{Y})$ is a sample from the joint distribution of (X, Y) , then $\nu_n^2(X, Y) = \|f_{X,Y}^n(t, s) - f_X^n(t) f_Y^n(s)\|^2$.*

We must remark that this result is an alternative way of calculating Equation (70) but, as stated in the literature, a much harder and time consuming way.

Theorem 18. *If $E|X|_p < \infty$ and $E|Y|_q < \infty$, then almost surely $\lim_{n \rightarrow \infty} \nu_n(\mathbf{X}, \mathbf{Y}) = \nu(X, Y)$.*

Corollary 19. *If $E(|X|_p + |Y|_q) < \infty$, then almost surely $\lim_{n \rightarrow \infty} R_n^2(\mathbf{X}, \mathbf{Y}) = R^2(X, Y)$.*

Theorem 20. *For random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ such that $E(|X|_p + |Y|_q) < \infty$, the following properties hold:*

- (i) $0 \leq R(X, Y) \leq 1$, and $R = 0$ if and only if X and Y are independent.
- (ii) $\nu(X) = 0$ implies that $X = E[X]$, almost surely.
- (iii) If X and Y are independent, then if $\nu(X + Y) \leq \nu(X) + \nu(Y)$. Equality holds if and only if one of the random vectors X or Y is constant.

Proof of this last statement can be found in Székely and Rizzo [2009].

Theorem 21. (i) $\nu(\mathbf{X}, \mathbf{Y}) \geq 0$.

(ii) $\nu(\mathbf{X}, \mathbf{Y}) = 0$ if and only if every sample observation is identical.

(iii) $0 \leq R_n(X, Y) \leq 1$.

(iv) $R_n(X, Y) = 1$ implies that the dimensions of the linear subspaces spanned by $\mathbf{X}$ and $\mathbf{Y}$ respectively are almost surely equal, and if we assume that these subspaces are equal, then in this subspace $\mathbf{Y} = a + b\mathbf{X}C$ for some vector a , non-zero real number b and orthogonal matrix C .

When considering that (X, Y) has a bivariate normal distribution, there is a deterministic relation between R and $|\rho|$.

Theorem 22. *If X and Y are standard normal, with correlation $\rho = \rho(X, Y)$, then:*

(i) $R(X, Y) \leq |\rho|$,

(ii) $R^2(X, Y) = \frac{\rho \arcsin \rho + \sqrt{1-\rho^2} - \rho \arcsin(\rho/2) - \sqrt{4-\rho^2} + 1}{1 + \pi/3 - \sqrt{3}}$,

(iii) $\inf_{\rho \neq 0} \frac{R(X, Y)}{|\rho|} = \lim_{\rho \rightarrow 0} \frac{R(X, Y)}{|\rho|} = \frac{1}{2(1 + \pi/3 - \sqrt{3})^{1/2}} \cong 0.89066$.

2.6.3 Brownian Covariance

To define Brownian covariance, let W be a two-sided one-dimensional Brownian motion/Wiener process with expectation zero and covariance function

$$|s| + |t| - |s - t| = 2 \min(s, t), \quad t, s \geq 0. \quad (73)$$

Comparing to the standard Wiener process, this is twice the covariance.

Definition 23. The Brownian covariance or the Wiener covariance of two real-valued random variables X and Y with finite second moments is a non-negative number defined by its square

$$\omega^2(X, Y) = Cov_W^2(X, Y) = E[X_W X'_W Y_W Y'_W], \quad (74)$$

where (W, W') does not depend on (X, Y, X', Y') .

It is interesting to note that if in Cov_W we replace W by the identity function, id , then $Cov_{id}(X, Y) = |Cov(X, Y)| = |\sigma_{X,Y}|$, the absolute value of Pearson's product-moment covariance. While the standardized product-moment covariance, Pearson correlation (ρ), measures the degree of linear relationship between two real-valued variables, we shall see that standardized Brownian covariance measures the degree of all kinds of possible relationships between two real-valued random variables.

We will extend now the definition of $Cov_W(X, Y)$ to random processes in higher dimensions. If X is an $\mathbb{R}^p$ -valued random variable, and $U(s)$ is a random process defined for all $s \in \mathbb{R}^p$ and independent of X , define the U -centered version of X by

$$X_U = U(X) - E[U(X) | U], \quad (75)$$

whenever the conditional expectation exists.

Definition 24. If X is an $\mathbb{R}^p$ -valued random variable, Y is an $\mathbb{R}^q$ -valued random variable and $U(s)$ and $V(t)$ are arbitrary random processes defined for all $s \in \mathbb{R}^p$, $t \in \mathbb{R}^q$, then the (U, V) covariance of (X, Y) is defined as the non-negative number whose square is

$$Cov_{U,V}^2(X, Y) = E[X_U X'_U Y_V Y'_V], \quad (76)$$

whenever the right-hand side is non-negative and finite.

In particular, if W and W' are independent Brownian motions with covariance function as Equation (73) on $\mathbb{R}^p$ and $\mathbb{R}^q$ respectively, the Brownian covariance of X and Y is defined by

$$\omega^2(X, Y) = Cov_W^2(X, Y) = Cov_{W,W'}^2(X, Y). \quad (77)$$

Similarly, for random variables with finite variance the Brownian variance is

$$\omega(X) = Var_W(X) = Cov_W(X, X). \quad (78)$$

Definition 25. The Brownian correlation is defined as

$$\text{Cor}_W(X, Y) = \frac{\omega(X, Y)}{\sqrt{\omega(X)\omega(Y)}} \quad (79)$$

whenever the denominator is not zero; otherwise $\text{Cor}_W(X, Y) = 0$.

We finish this part with the surprising result from the next theorem.

Theorem 26. *For arbitrary $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ with finite second moments*

$$\omega(X, Y) = \nu(X, Y).$$

To summarise the results from [Székely et al. \[2007\]](#), distance covariance and distance correlation are natural extensions and generalizations of classical Pearson covariance and correlation in possibly three ways.

1. In one direction, the ability to measure linear association to all types of dependence relations was extended;
2. In another direction, the bivariate measure to a single scalar measure of dependence between random vectors in arbitrary dimension was also extended;
3. In addition to the obvious theoretical advantages, there are the practical advantages that dCov and dCor statistics are computationally simple and applicable in arbitrary dimension not constrained by sample size.

Probably dCov is not the only possible or the only reasonable extension with the above mentioned properties, but this extension was received as a natural generalization of Pearson's covariance in the sense that the covariance of random vectors was defined with respect to a pair of random processes, and if these random processes are i.i.d. Brownian motions, which is a very natural choice, then we arrive at the distance covariance; on the other hand, if we choose the simplest non-random functions, a pair of identity functions (degenerate random processes), then we arrive at Pearson's covariance. To sum up, distance correlation extends the properties of classical correlation to multivariate analysis and the general hypothesis of independence.

2.7 FRACTIONAL BROWNIAN MOTION

Two of the most important and simple models of probability theory and financial econometrics are the random walk and the Martingale theory. They assume that the future price changes only depend on the past price changes. Their main characteristic is that the returns are uncorrelated.

But are they truly uncorrelated or are there long-time correlations in the financial time series? This question has been studied especially since it may lead to deeper insights about the underlying processes that generate the time series (see, for instance, [Lo \[1991\]](#), [Ding et al. \[1993\]](#) and [Harvey \[1993\]](#) or, for a more recent review, [Doukhan et al. \[2003\]](#)).

Depending on the scientific field there are, typically, more than ten measures to quantify the long-time correlations. In the financial literature we find two methods: the Rescaled Range analysis (R/S) and the detrended fluctuation analysis (DFA). For further details see [Taqqu et al. \[1995\]](#).

In the 50's, Hurst, while analysing hydrological flows, proposed a single exponent to characterise time variation in time series [Hurst, 1951]. This approach is a generalisation of Brownian motion later called fractional Brownian motion [Mandelbrot and Van Ness, 1968], and is characterised by a single exponent, called Hurst exponent. Another way of estimating the Hurst exponent was introduced via DFA by Peng et al. [1994] while studying DNA patterns and their characteristics.

In order to measure the strength of trends or “persistence” in different processes, the rescaled range (R/S) analysis to calculate the Hurst exponent can be used. One studies the rate of change of the rescaled range with the change of the length of time over which measurements are made. We divide the time series ξ_t of length T into N periods of length τ such that $N\tau = T$. For each period $i = 1, 2, \dots, N$ containing τ observations, the cumulative deviation is

$$X(\tau) = \sum_{t=(i-1)\tau+1}^{i\tau} (\xi_t - \langle \xi \rangle_t), \quad (80)$$

where $\langle \xi \rangle_t$ is the mean within the time-period and is given by

$$\langle \xi \rangle_t = \frac{1}{\tau} \sum_{t=(i-1)\tau+1}^{i\tau} \xi_t. \quad (81)$$

The range in the $i - th$ time period is given by $R(\tau) = \max X(\tau) - \min X(\tau)$, and the standard deviation is given by

$$S(\tau) = \left[\frac{1}{\tau} \sum_{t=(i-1)\tau+1}^{i\tau} (\xi_t - \langle \xi \rangle_t)^2 \right]^{1/2}. \quad (82)$$

Then $R(\tau) / S(\tau)$ is asymptotically given by a power-law

$$R(\tau) / S(\tau) = k\tau^H \quad (83)$$

where k is a constant and H the Hurst exponent.

In general, “persistent” behaviour with fractal properties is characterized by a Hurst exponent $0.5 < H \leq 1$, random behaviour by $H = 0.5$ and “anti-persistent” behaviour by $0 \leq H < 0.5$.

Usually the Equation (83) is rewritten in terms of logarithms, $\log(R(\tau) / S(\tau)) = H \log(\tau) + \log(k)$, and the Hurst exponent is determined from the slope.

In the DFA- n method, the time-series ξ_t of length T is first divided into N non-overlapping periods of length τ such that $N\tau = T$. In each period $i = 1, 2, \dots, N$ the time-series is first fitted through a polynomial function $z_n(t) = a_n t^n + a_{n-1} t^{n-1} + a_0$, called the local trend. In this thesis we use a quadratic function $n = 2$ as our fit function. Then it is detrended by subtracting the local trend, in order to compute the fluctuation function,

$$F(\tau) = \left[\frac{1}{\tau} \sum_{t=(i-1)\tau+1}^{i\tau} (\xi_t - \langle \xi \rangle_t)^2 \right]^{1/2}. \quad (84)$$

The function $F(\tau)$ is re-computed for different box sizes τ (different scales) to obtain the relationship between $F(\tau)$ and τ [Kantelhardt et al., 2001].

A power-law relation between $F(\tau)$ and the box size τ , $F(\tau) \sim \tau^\alpha$, indicates the presence of scaling. The scaling or “correlation exponent” α quantifies the correlation properties of the signal. If

- $\alpha = 0.5$: the signal is uncorrelated (white noise);
- $\alpha > 0.5$: the signal is anti-correlated;
- $\alpha < 0.5$: there are positive correlations in the signal.

For a recent application considering Hurst exponent applied to financial time series, follow Gomes [2012].

2.8 OTHER METHODS

Despite the methods or techniques considered in previous sections, it is useful to say that they not close all the existing techniques. So, in this section we consider other interesting techniques but that are not going to be applied in this research.

NETWORKS Networks have been studied at an early stage in the history of mathematics. For example, the well known problem of Königsberg bridges was solved by Euler in the 17th century. More recently, it is worth to consider the work of Erdős and Rényi [1959]. Yet only recently, with the enormous growth in computer power, some of those problems have been looked at again from a different viewpoint. Examples of these types of networks or other novel methods where networks are applied to the study of time series, include small worlds and scale free networks (see, for instance, Newman [2003]).

AGENT BASED SYSTEMS The analogy between cellular automata, with simple laws that rule the interaction between neighbours, and economical systems, with all agents individually seeking profit maximisation, has led to the use of agent based systems. The agents are autonomous entities that live and interact among them usually by neighbourhood relations.

The set of ingredients for modelling markets are:

1. a large number of independent agents participate in a market;
2. each agent has alternatives in making decisions;
3. the aggregate activity results in a market price, which is known to all;
4. agents use public price history to make their decisions.

Bonanno et al. [2001] consider that the financial markets show several levels of complexity that may occurred for being systems composed by agents that interact nonlinearly between them. These authors, proposed also that the traditional models of asset pricing (Capital Asset Pricing Model (CAPM) and Arbitrage Pricing Theory (APT)) failed because the basic assumptions of these models are not verified empirically.

For a recent review of the use of agent based systems in Econophysics see [Ausloos \[2006\]](#). Another type of agent based systems is that related to Game Theory where we can find several well known cases like the prisoner's dilemma and the minority game.

COPULAS The copula problem, describing the dependence between random variables, gave a big number of possible structures of financial asset correlations, but these seemed to be chosen more for mathematical convenience than for plausible underlying mechanisms, which created the generalized idea that these copulas were in fact very unnatural.

There is, however, a very interesting exception that is a natural extension of the mono-variate Student-t distribution that has a clear financial interpretation [[Bouchaud and Potters, 2011](#)]. For a personal view on the application of copulas to finance see [Embrechts \[2009\]](#).

WAVELETS Wavelets properties, namely the method flexibility in handling very irregular data series, the capacity of representing the data without knowing the underlying structure and the capacity to locate in time regime shifts and shocks made this one of the most interesting methods in financial time series. For an extended reading see [Vuorenmaa \[2005\]](#) and [Sharkasi et al. \[2006b\]](#)

TURBULENCE AND THE OMORI LAW Another striking resemblance that unfolds when analysing stock market volatility is its resemblance with the turbulence in fluids. [Mantegna and Stanley \[2000\]](#) addresses this as follows: "In turbulence, one ejects energy at a large scale by, e.g., stirring a bucket of water, and then one observes the manner in which the energy is transferred to successively smaller scales.

In financial systems 'information' can be injected into the system on a large scale and the reaction to this information is transferred to smaller scales – down to individual investors". This resemblance was introduced before by [Mandelbrot \[1972\]](#) and then by the same authors ([Mantegna and Stanley \[1996\]](#), [Mantegna and Stanley \[1997\]](#)) and later reviewed by [Sornette \[2002\]](#).

Moreover, the Omori law for seismic activity after major earthquakes has equally proved to be useful when understanding large crashes in stock markets [[Lillo and Mantegna, 2003](#)].

Other applications concerning applications of concepts of Physics to financial markets, such as, the diffusion anomalous systems, whose general framework can be provided by the nonlinear Fokker-Planck equation, could be developed.

There is, indeed, a great deal of other empirical research using methods and analogies borrowed from Physics that space limitations prevent us to describe any further (see, for example [Lee and Stanley \[1988\]](#), [Mandelbrot et al. \[1997\]](#) or [Bartolozzi et al. \[2006\]](#)).

2.9 METHODOLOGIES

2.9.1 Data Analysis Methodology

We are interested in studying the dynamic variation of the stocks/markets correlations evolving with time t , so we will look at the correlations calculated over a sliding or rolling window. We will create a time-evolving sequence of correlation matrices by

rolling the time window of T returns (there is one return for each time step) through the full data set.

The choice of T is a compromise between excessively noisy and excessively smoothed correlation coefficients [Onnela et al., 2003] and is usually chosen such that $Q = T/N = 1$ [Fenn et al., 2011].

Also, it must be taken in consideration the type of data we are dealing with. In this work it could be interesting to study sizes T of the rolling window to be $T = 20$, $T = 60$, $T = 120$ and $T = 240$ trading days, that is, approximately 1, 3, 6, and 12 months of data, because these sizes have financial meaning, namely the quarterly, semester and annual company results presentation.

Equation (22) is applied to calculate the correlation coefficients over a subset of return series within the rolling window $[t - T + 1, t]$. For instance, the correlations in the first sliding window are computed by the return series within $[1, T]$ and $[2, T + 1]$ for the following rolling window.

By only shifting the time window by five data point, there is a significant overlap in the data contained in consecutive windows. This approach enables us to track the evolution of the stocks/markets correlations and to identify time steps at which there were significant changes in the correlations.

2.9.2 Computational Methodology

The purpose of this Section is to introduce some of the computational methodology used in this thesis. The choice of computational tools and techniques applied in this work is almost as important as the mathematical formulation since the results are based on their discriminating application and they serve as a basis for characterising the work.

Knowledge and Data availability

Internet has not only brought more comprehensive search but has realised new ways for people to coordinate and share scientific work. Two good examples are the access to pre-prints from others scientists or the access to the financial data available from sources like Yahoo Finance (finance.yahoo.com) or 4-traders (www.4-traders.com) .

Free Software

Universities were some of the first places to adopt the Internet, and for long time academic centres were both its major users and its backbone. The Internet has allowed development of new tools, with email and the Web being two of the best known examples.

New methods for transfer of information promoted the emergence, in 1984, of the Free Software movement. Free Software existed before this date, initially sharing software was the rule that later became the exception.

The Free Software Foundation created the GNU project, designed to create a Free Software derivative of UNIX. At the same time a license was developed to legally uphold the ideals of Free Software. That license is called Gnu Public License (www.gnu.org/licenses/gpl-2.0.html) and it forms the cornerstone of the Free Software move-

ment. The software projects presented here (Appendix D) are released under this license (version 3).

Use of free software

A consequence of using Free Software is that programs can be ported everywhere. In this case this implies many Operating Systems, although naturally the tools are easiest to set up in the environment in which they have been developed.

Reproducibility of results

All results should be possible to be reproduced easily. This usually entails the use of scripts to drive the different parts of the analysis.

Redundant methods

In order to avoid single failure points every effort has been made to implement all methods using at least two different implementations. This in itself does not guarantee the correctness of the results but does increase our confidence in them.

One other technique coming from software development is “Unit testing”. The idea here is that tests for the code are written first, then the code itself. There is an analogy with mathematical systems in that one of the methods we use is the identification of invariants (quantities that remain unchanged over a given range of operations).

Unit testing advocates the writing of tests where we compare the empirical result to that expected based on known cases, in order to ensure the correctness of the code at hand.

Languages and libraries

Tools described are general and not restricted to implementation of any particular technique; they allow and encourage the creation and use of libraries related to the problems studied.

An important distinction between different languages relates to their libraries, whether the standard library or available add-ons. Both languages referenced later benefit from a wide range of libraries that clearly constitutes its major advance over other similar solutions.

LateX

This document was written in LyX (www.lyx.org), that builds over LateX (Knuth [1984] and L^AT_EX [1986]).

R language and R packages

R (<http://www.r-project.org>) is a free implementation of the S language. S, from Statistics, was primarily developed at AT&T Bell Laboratories to be a language oriented towards Statistics.

The repository of available packages (almost all of which are Free Software) can be found in R homepage CRAN (Comprehensive R Archive Network, <http://cran.r-project.org>).

In this work the following packages were used:

- hash (version 2.2.6);
used to implement a data structure in the .csv data files.
- performanceAnalytics (version 1.4.3541);
used for statistical calculation and for data plotting.
- zoo (version 1.7-11);
used to order the indexed Close values.
- pracma (version 1.7.7);
used for Approximate Entropy calculations.
- energy (version 1.6.2);
used for Distance Correlation calculations.
- lattice (version 0.20-29);
used for data plotting.
- xts (version 0.9-7);
used for data plotting.
- xtsExtra (version 0.0-1)
used for data plotting.
- entropy (version 1.2.0);
used for Kullback-Leibler and Mutual Information calculations.
- ForeCA (version 0.1);
used for Forecastable Component Analysis calculations.

More details about these packages can be found in Appendix C.

Finally, some support to activity on using R can be followed in R Studio, www.rstudio.com, used in this work, or R Metrics, www.rmetrics.org (see Würtz [2004]).

My companion prattled away about Cremona fiddles and the difference between a Stradivarius and an Amati. "You don't seem to give much thought to the matter at hand" [the Lauriston Garden murder], I said, interrupting Holmes' musical disquisition. "No data yet," he answered. "It is a capital mistake to theorize before you have all the evidence. It biases the judgement." Sir Arthur Conan Doyle, A Study in Scarlet (1886)

The purpose of this Chapter is to introduce and explain the data sets used in this thesis. Two data sets are used: the PSI-20 set and the World Markets set. Each necessary component of the PSI-20 stocks or World Markets indices has its own .csv file

All the data on the respective market indices is public and came from Yahoo Finance (finance.yahoo.com) and 4-Traders (www.4-traders.com) with a major concern for coherence of the data sources used.

Also, the daily Close value as the value for the day has been considered to obviate any time zone difficulties.

3.1 DATA CONSIDERATIONS

Empirical data

Though different kinds of financial time series were being recorded and studied for decades, the scale changed about 20 years ago. The advent of computers and automation of the stock exchanges and financial markets has lead to the explosion of the amount of data recorded.

Nowadays, all transactions on a financial market are recorded tick-by-tick, i.e. every event on a stock is recorded with a time stamp defined up to the millisecond, leading to huge amounts of data.

For example, the empirical database Reuters Datascope Tick History (RDTH) database, today records roughly 25 gigabytes of data per trading day [Tilak, 2012]. Prior to this tremendous increase in recording market activity, statistics were computed mostly with daily data.

Simulated data

It is often not possible to study certain effects using empirical data. For example, it is very difficult to find empirical data with a certain value of auto-correlation, or perfect Gaussian distribution. Also, the results obtained by analysis of empirical data sometimes need to be compared against a benchmark.

In such situations, artificial data can be simulated according to required specifications. Simulated data can also serve as reliable benchmarks.

3.2 DATA SETS

3.2.1 *PSI-20 set*

The PSI-20 set is formed by twelve stocks that were obtained from the PSI-20 Index, which is a price index calculation based on 20 stocks obtained from the universe of Portuguese companies listed to trade on the Main Market and was designed to become the underlying element of futures and options contracts.

The choice criteria were two:

- the availability of data in the period 2001-2014, to maximize the days where all the stocks were in the market;
- the best PSI-20 representation, that is, stocks from almost all the sectors and from different importance.

In Table 3 are summarized the stocks used with their respective business sector. Data and summary statistics on the markets studied are recorded and are presented in Appendix A.

Abrev.	Stock Name	Sector
^BES	Banco Espírito Santo	Financial Services
^BPI	Banco Português de Investimento	Financial Services
^EDP	Energias de Portugal	Electricity
^JMT	Jerónimo Martins	Distribution
^EGL	Mota-Engil	Construction
^NBA	Novabase	Technological Services
^PTI	Portucel	Paper
^PTC	Portugal Telecom	Telecommunications
^SEM	Semapa	Paper
^SONC	Sonae Com	Telecommunications
^SON	Sonae SGPS	Distribution
^ZON	Zon Optimus	Media

Table 3: PSI-20 set business sectors

The data used in this study are the close values and its log returns from these 12 stocks and cover the period common to all stocks from January 25, 2001 to September 13, 2013 for a total of 3362 observations.

For a more close look to PSI-20 stocks degree of importance, based on their stock market capitalization, we can see in Table 4 their “top ten” classification between 2000 and 2013. As we can see, from the 12 chosen stocks, only sensibly half are represented in this top ten. The idea, here, was to choose representative stocks. It is also possible to analyse particular stock “movements” in this classification but this is out of scope of this study.

Position	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
1st		PTC	PTC	PTC	PTC	PTC	EDP	EDP	EDP				JMT	
2nd	PTC				EDP	EDP	PTC			EDP	PTC	JMT		JMT
3rd	EDP	EDP	BES	EDP					PTC	PTC	EDP	EDP	EDP	EDP
4th	BES	BES	EDP	BES	BES	BES	BES	PTC			JMT		BES	BES
5th								BES		BES		PTC		
6th					ZON		BPI		BES	JMT			PTC	
7th			BPI	ZON		ZON		BPI			BES	BES	PTI	PTC
8th	BPI	SON		BPI	BPI	BPI			JMT					ZON
9th		BPI	ZON	SON	SON	SON	SON	SON				PTI	SON	PTI
10th	ZON		SON	JMT	JMT	JMT	ZON	JMT	BPI	BPI	PTI		BPI	SON

Table 4: PSI-20 set top-ten classification

3.2.1.1 Stock splits and other corrections

In order to obtain correct data we needed to study the stocks history, namely the stock splits. Stock splits are conceptually a simple corporate event that consists in the division of each share into a higher number of shares of smaller par value. These operations have long been a part of financial markets.

Abrev.	Stock-Split			Rights Issues			Exceptions
	Date	Last Price	Next Price	Date	LP	NP	Goal
^BES				2000-Jul-11	25.70	17.35	
				2002-Feb-06	14.35	11.40	
				2006-Apr-27	15.00	11.59	
				2009-Mar-19	5.54	3.65	
				2012-Apr-16	1.05	0.65	
^BPI				2000-Oct-30	3.99	3.82	take over threat (BCP)
				2006-Mar-13	4.24	5.33	
				2008-Jun-20	2.92	2.81	
^EDP	2000-Jul-17	17.95	3.64				
^JMT	2007-May-28	22.00	4.54	2004-Jun-08	9.78	8.64	
^EGL	2001-Jan-23	8.35	1.66	2000-Aug-07	11.30	11.40	
^NBA							
^PTI	2001-Jan-22	7.35	1.44	2001-Sep-04	0.91	0.90	
^PTC							
^SEM	2000-Sep-14	19.98	3.96				
^SONC							
^SON	2000-Jun-21	50.61	9.65	2005-12-27	1.22	0.95	spin-off Sonae Industria
^ZON				2005-Jun-14	3.38	6.77	social capital reduction

Table 5: PSI-20 stock splits

Portugal, for instance, witnessed 26 of these operations from 1999 (the year the Euro was introduced) to June 2003 essentially due to a legislative change that took place when the corporate law was adapted for the change from Escudo to Euro [Pereira and Cutelo,

2010]. Stock splits are associated with positive abnormal returns in the short run (around the announcement dates and ex-dates).

If a company has undergone stock splits over its lifetime, comparing historical stock prices to those of the present day would not accurately reflect performance. For this reason, we must compare split-adjusted share prices.

For discerning and analysing the real performance of the stock, it is standard to adjust the old prices to reflect the splits. In other words, we have to find the present equivalent of the past prices. In Table 5 are shown the main operations concerning the twelve PSI-20 stocks studied. This information is partially adapted from Pereira and Cutelo [2010].

3.2.2 World Markets set

The choice of the markets used in this study was driven by the goal of studying major markets across the world in an effort to ensure that tests and conclusions could be as general as possible.

In Table 6 we summarise the markets used in this study.

Data and summary statistics on the markets studied are recorded and are presented in Appendix A.

Abrev.	Index Name	Country	Region
^AEX	Amsterdam Exchange Index	Netherlands	Europe
^ASX	Australian Securities Exchange	Australia	Asia/Pacific
^ATX	Austrian Traded Index	Austria	Europe
^BSESN	Bombay Stock Exchange	India	Asia/Pacific
^BVSP	Bovespa - Bolsa de Valores de S. Paulo	Brazil	America
^CAC	Compagnie des Agents de Change	France	Europe
^DAX	Deutscher Aktien Index	Germany	Europe
^DJI	Dow Jones Industrial Average	United States	America
^FTSE	Footsie	United Kingdom	Europe
^HSI	Hang Seng Index	Hong Kong	Asia/Pacific
^IBEX	Índice Bursátil Espanol	Spain	Europe
^IXIC	Nasdaq Composite	United States	America
^JKSE	Jakarta Stock Exchange - Composite Index	Indonesia	Asia/Pacific
^KOSPI	Seoul Composite	South Korea	Asia/Pacific
^MERVAL	Mercado de Valores de Buenos Aires	Argentina	America
^MIB	Milano Italia Borsa	Italy	Europe
^MXX	IPC - Mexican Stock Exchange Index	Mexico	America
^NIK	Nikkei Tokyo	Japan	Asia/Pacific
^PSI20	Portuguese Stock Index	Portugal	Europe
^SPY	S&P 500	United States	America
^SSMI	Swiss Market	Switzerland	Europe
^STOXX	DJ Euro Stoxx 50		Europe
^STRAITS	Straits Times	Singapore	Asia/Pacific

Table 6: World Markets Set

We have considered here the major and most active markets worldwide from America (North and South), Asia/Pacific, Africa and Europe. The data used in this work are the

daily Close values for these 23 markets obtained from January 2, 2001 to September 25, 2013.

In the chapters that follow when we refer the values for markets and/or compare them we are actually comparing the (log-) return of the chosen index for that market. This decision was made in order to simplify the language.

Subsequently, we obtained the “common data”, i.e., the subset of days where all the markets are open, excluding local holidays and periods where the transaction of any market was suspended. Regardless these strict criteria, the data used in this work make for a total of 2965 common daily Close values.

3.3 EVENTS OF INTEREST

As noted in Chapter 2, Section 2.9.1, a sliding window approach will be used to analyse and calculate the values for the different measures for the data sets. This will help us to confine the search for “early warning signs” to a few windows before and after the events of interest.

Also, some “neutral” events are going to be explored using the same methodology in order to perform a comparative analysis.

The chosen events of interest are the recession dates proposed by NBER (see Sub-Section 2.1.5 in Section 2.1 in Chapter 2). So, we are going to look in more detail the following periods:

- from 14-02-2001 until 09-11-2001, the first XXI recession and the respective before and after recession periods: from 04-01-2001 until 13-02-2001 and from 12-11-2001 until 17-01-2002;
- from 16-11-2007 until 17-06-2009, the second XXI recession and the respective before and after recession periods: from 02-08-2007 until 14-11-2007 and from 18-06-2009 until 09-09-2009;

These before and after periods were chosen to be, approximately, about 20% each of the total recession period. This criterion was due to the availability of the data (mainly for the before recession period).

For the “neutral” periods we considered the following two:

- from 19-02-2004 until 26-08-2004, the first neutral period and the respective before and after neutral periods: from 08-01-2004 until 18-02-2004 and from 27-08-2004 until 08-10-2004;
- from 07-06-2011 until 13-03-2013, the second XXI neutral period and the respective before and after neutral periods: from 30-12-2010 until 26-05-2011 and from 14-03-2013 until 25-06-2013;

In the next two chapters the techniques presented in Chapter 2 will be applied to the data sets presented in this chapter.

PORTUGUESE STANDARD INDEX (PSI-20) ANALYSIS

“One of the funny things about the stock market is that every time one person buys, another sells, and both think they are astute”. William Feather

In this chapter we will apply the mathematical tools presented/described in Chapter 2 to the PSI-20 data set. Let us start by presenting some of the features of this index.

4.1 PSI-20 INDEX

The Portuguese Stock Index PSI-20 is the national benchmark index, reflecting the price evolution of the 20 largest most liquid assets selected from the set of companies listed on the Portuguese Main Market. The rules for construction of PSI-20 are published [PSI \[2003\]](#), but can be summarised briefly as giving a different weight to each asset belonging to the index, such that no asset has more than 20% of the total weight. PSI-20 had its beginning in January 4th, 1993.

Figure 4 shows the PSI-20 index evolution from January 24, 2000 to September 25, 2013.

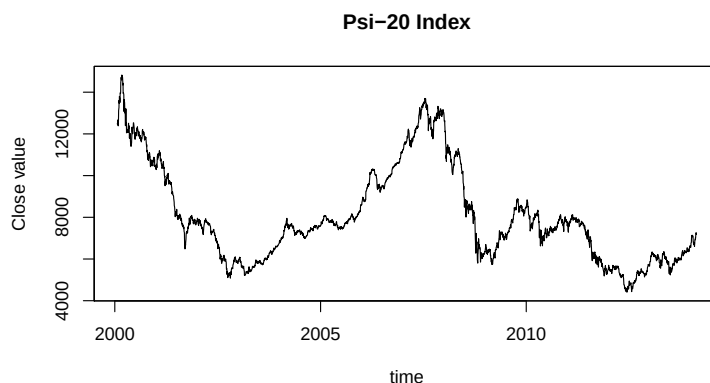


Figure 4: PSI-20 from 2000 to 2014

4.1.1 PSI-20 evolution

After the 2000 peak (roughly corresponding to the dotcom bubble burst), we essentially assist to a decline in the index value until the end of 2002. Additionally, the sub-sample period January 2, 2001 to November 23, 2001 was characterized by a climate of economic and political instability in Europe and United States due to the high value of the Dollar against the Euro, the Israel-Palestinian conflict, and the terrorist attacks on September 11, 2001 and the subsequent climate of uncertainty, with negative impacts on the financial markets, including the Portuguese stock market.

In this period the PSI-20 index declined by 24,42 per cent. Between 2002 and 2007 we assisted to world markets recovery, but in 2008, with the mortgage and sub-prime crises, the world markets in general, and PSI-20 in particular, went down once again.

Some ups and downs are found between 2009 and 2011, with the market/investors probably still “astonished” with what had happened before. In the first quarter of 2011 another fall, a period coincident with the international assistance program applied to Portugal. Finally, from the beginning of the second quarter of 2012 we are having some recovery signals in the PSI-20 index.

4.1.2 A random PSI-20

Now, we generated a shuffled data by randomly reordering the full return time series for the PSI-20 index. This process destroys the temporal correlations between the return time series but preserves the distribution of returns for each series as we can see in Figure 5.

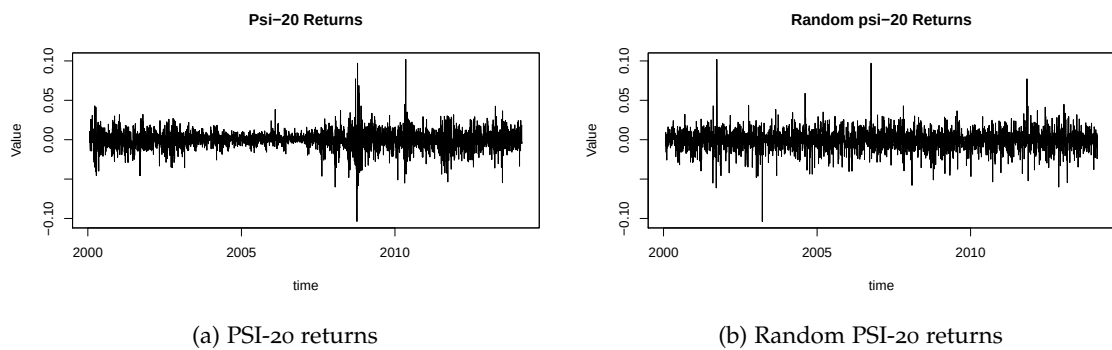


Figure 5: Real vs Random PSI-20 returns.

To try to highlight interesting features in the correlations, we compare the real PSI-20 close values to a corresponding distribution for randomly shuffled returns (a random PSI-20 close values). For a visual comparing between these markets we present Figure 6.

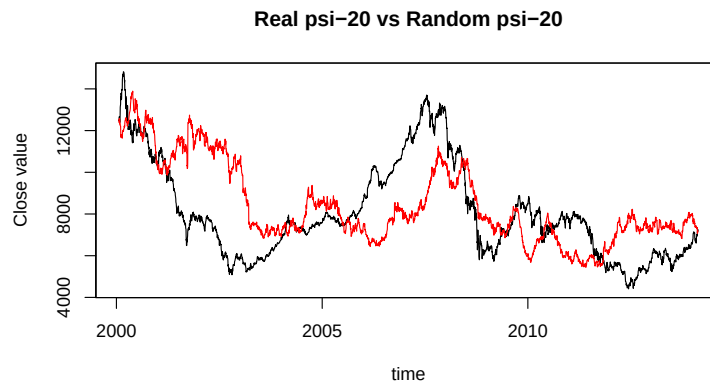


Figure 6: Real versus Random PSI-20 close values

As we are going to work all the time with returns, now we show their values along time and their distribution (see Figure 7).

According to [Rege et al. \[2013\]](#) the distribution of the returns of the PSI-20 exhibits much higher kurtosis and extreme values than the Normal distribution do. They also found that the best fit is provided by the Student t and the Generalized Hyperbolic distributions.

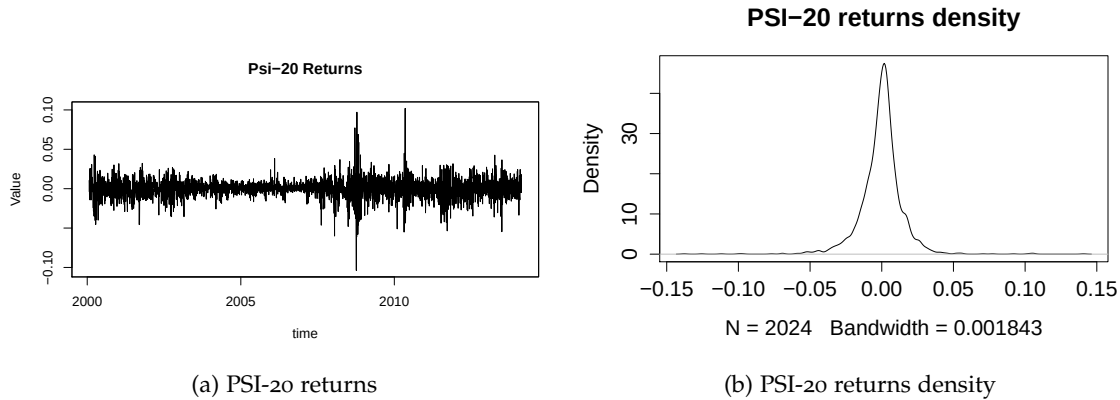


Figure 7: PSI-20 returns time series and their distribution.

A broader and earlier study reaching the same conclusions but applied to a “World Market Index” was done by [Fergusson and Platen \[2006\]](#).

4.2 DYNAMIC ANALYSIS OF PSI-20 USING SLIDING WINDOWS

In Section 2.9.1 a sliding/rolling windows approach was introduced. The nature of the approach (i.e. based on the interval characterisation) means that we can apply these techniques to different intervals of fixed size (20, 60 and 120 points, corresponding, approximately, to 1 month, 3 months and 6 months of data).

Each one of these sub-intervals is characterised by different results. The purpose of this analysis on different scales is to test the dependence of the results on the granularity of the data, since we expect different behaviours at different scales for financial time series.

4.2.1 Step size decision

The first analysis was done on the step size, that is, the number of data points used to “slide” the window.

To illustrate this, we consider, for instance, Figure 8 where are shown the Distance Correlation window values versus the window step size for the PSI-20 stocks BES and BPI. These results serve only, at this stage, for comparison terms. Each point represents the Distance Correlation value in the centre of a sliding window, moved along the series.

We can see for all the calculated steps (5, 10 and 20), that the Distance Correlation values remain essentially the same. So, this is not a distinguishable criterion to have into account.

Eventually, the more readable value is for the 20 steps case.

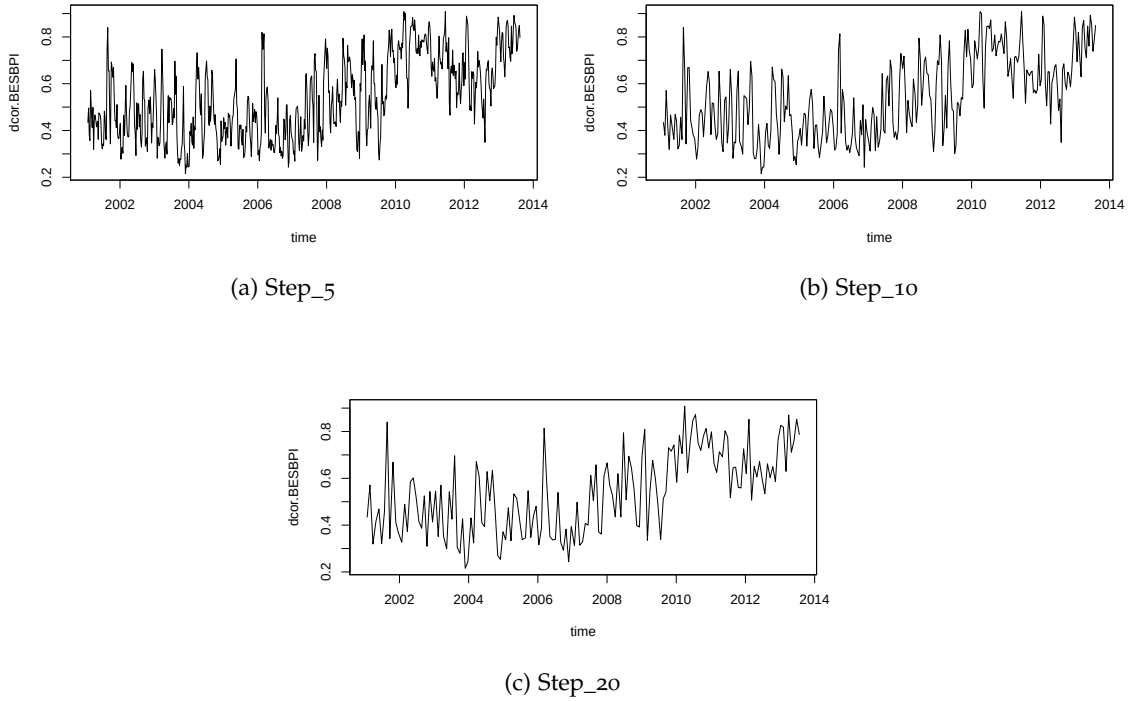


Figure 8: Distance Correlation values for different steps

4.2.2 Window size decision

The other studied criterion is the window size. Does the results, in general, remain the same despite the size of the window? Taking into account the recommendation by Fenn et al. [2011], the size should be $Q \sim O(1)$ that is to say $T = 12$. On the other side, we are talking about companies, so, $T = 60$ represent approximately 3 months of data, and this is a relevant period with almost all the companies presenting quarterly reports.

Example 1

In Figure 9 it is possible to compare the effect of having two different size sliding windows. The 20 days window gives higher Distance Correlation values but it is harder to read than the 60 day one. It is notable that the Distance Correlation value goes down as the window size goes bigger (see Figure 9). Are we losing relevant information by choosing one or another size? A possible answer can be pointed later when we will try to identify the events corresponding to peaks or valleys.

Example 2

For another example (see Figure 10), the same happens if we consider the World Markets set. It can be seen for the different sliding windows that the Distance Correlation values

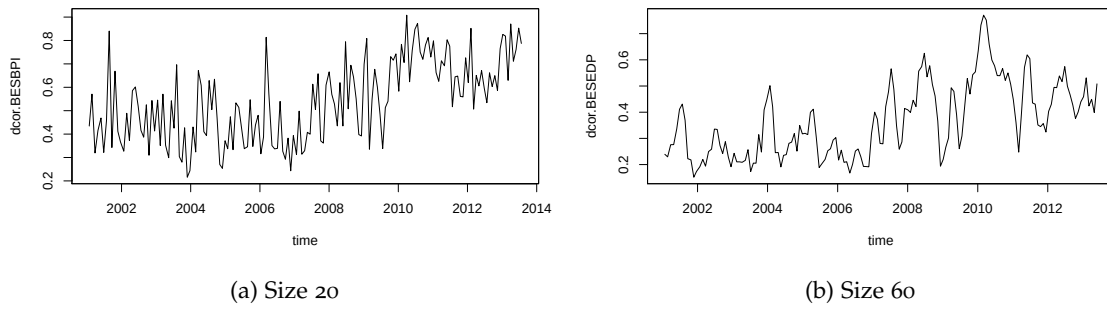


Figure 9: DCor values for different “sliding” windows size

between AEX and ASX suffer significantly as the window size gets bigger. Eventually, the more readable values are for the 120 sliding window, but for this case the Distance Correlation is more smoother and weaker than the previous sizes.

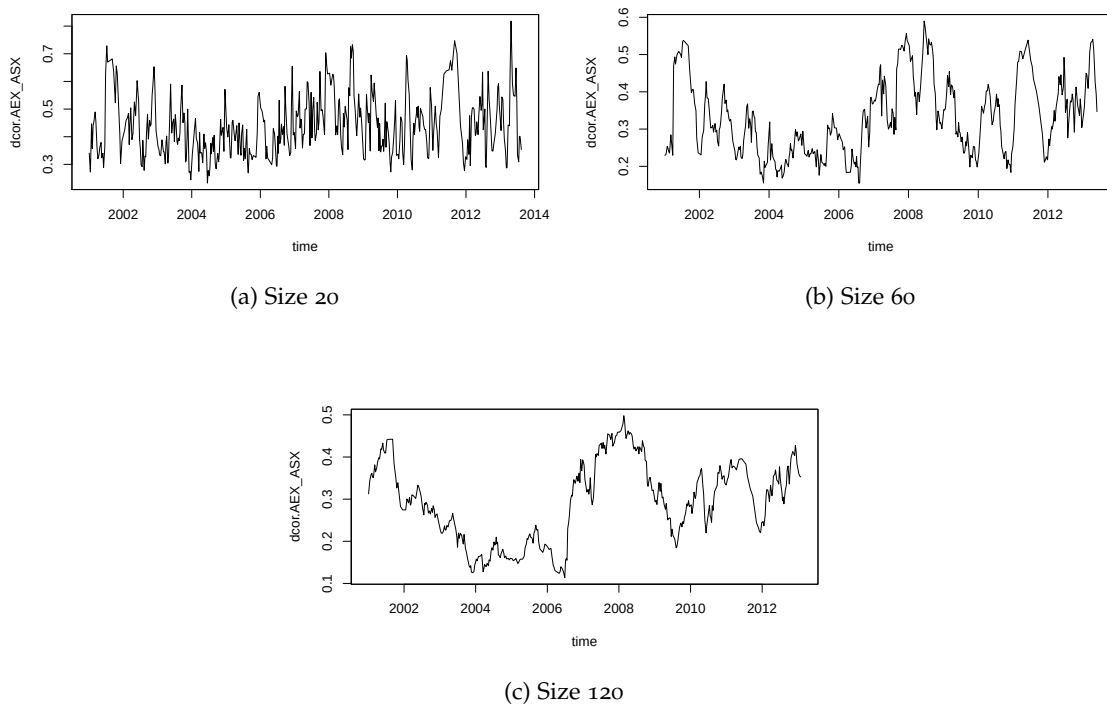


Figure 10: Markets DCor values for different “sliding” windows size

Despite that, for instance, it is easier to understand what happens to the correlation between these two markets. We can, roughly, define three typical behaviours for this relationship: the first, corresponding to periods of world crisis, between 2000 and mid 2001 and between nearly 2007 and 2008, where the correlation goes up; the second, corresponding to non-crisis periods, between mid 2001 and late 2006 and between 2008 and

nearly 2010, where the correlation goes down; the third, from 2010, where the correlation seems to go up, although with some breaks in the meantime.

Example 3

The results concerning different window widths deserve some more considerations. We can see, as an example, the Approximate Entropy for AEX in Figure 11. From these three plots it is clear that $ApEn$ gets quite a lot bigger as the window width becomes bigger. On the other hand, the results become smoother and with them also the variation becomes more clear. Despite obtaining higher entropy values as the sizes gets bigger, the relative difference between those entropy values is shrinking.

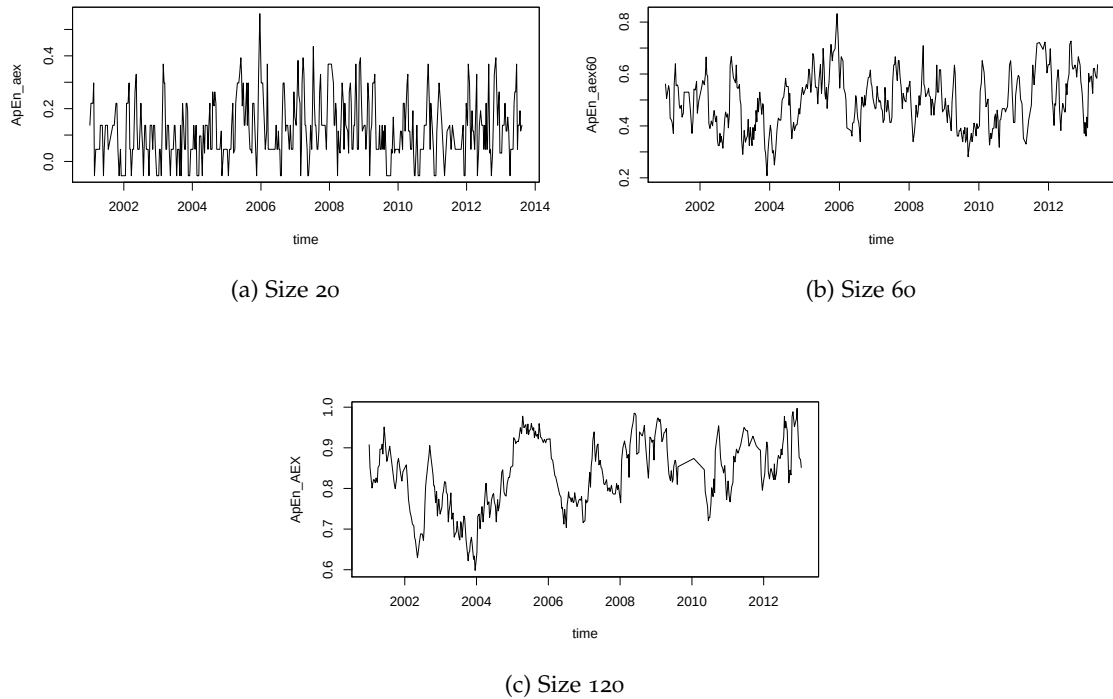


Figure 11: Markets $ApEn$ values for different “sliding” windows size

It is, for instance, easier to distinguish the peaks and the valleys. We can, roughly, define six typical behaviours for this market: the first, corresponding in part to periods of world crisis, between 2001 and 2004; the second from 2004 to mid 2005, a fast growing period, followed by a fast descending period, from mid 2005 to mid 2006; then another growing period from mid 2006 to 2009 followed by another descending period from 2009 until almost 2010; the last, from 2010, where the entropy seems to go up, although with some breaks in the meantime.

In conclusion, the window size criterion is important in what concerns the measured values because these values depend on the size of the window chosen. So, in the next Section the results will be presented using 20 days window and/or 60 days window depending on their readability.

4.3 RESULTS

The Econophysics tools presented in Chapter 2 are here applied to the Portuguese Standard Index PSI-20. PSI-20 index whose main characteristics are described in Appendix A.

The Portuguese case was chosen both for:

- a) regional relevance;
- b) relatively little previous studies;
- c) its relevance as a showcase both as an emerging young/mature market and its relevance to discuss features on the techniques presented.

This initial application is the forerunner and constitutes the main test for the World Markets set, analysed in the next Chapter.

4.3.1 Random Matrix

For the PSI-20 set we consider $3362/5 = 672$ samples by sequentially sliding a window of $T = 20$ days by 5 days (roughly one month). For each period, we look at the empirical correlation matrix of the $N = 12$ stocks during that period. The quality factor is therefore $Q = T/N = 20/12 = 1.67$.

4.3.1.1 Marchenko-Pastur band

In order to perform a study with random matrices we started by comparing the real eigenvalues density with the theoretical one as proposed by Marchenko and Pastur [1967] (see Figure 12). It is clear that several eigenvalues leak out of the Marchenko-Pastur band, even after taking into account the Tracy-Widom tail, which have a width given by $\sqrt{q}\lambda_+^{2/3}/N^{2/3} \approx 0.02$ which is very small in this case. The eigenvectors corresponding to these eigenvalues were explored in several works as we can see in Bouchaud and Potters [2011].

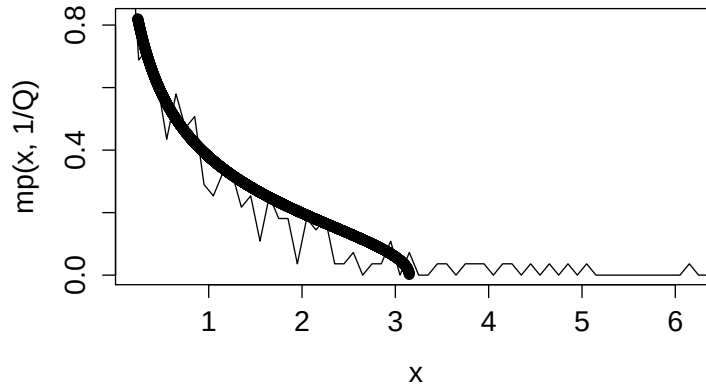


Figure 12: Theoretical versus Real stocks eigenvalues density

4.3.1.2 Correlation Matrix

Calculating the total Correlation Matrix for the time series using the Statistical Software R, we obtain for the 12 stocks the results shown in Table 7.

BES	BPI	EDP	EGL	JMT	NBA	PTC	PTI	SEM	SON	SONC	ZON
1.00	0.84	0.80	0.45	0.12	0.64	-0.00	0.02	0.39	0.09	0.54	0.47
	1.00	0.75	0.52	0.21	0.68	0.24	0.10	0.40	-0.06	0.49	0.33
		1.00	0.61	0.04	0.49	-0.04	0.28	0.36	0.07	0.50	0.36
			1.00	0.06	0.42	0.03	0.30	0.26	-0.00	0.45	0.18
				1.00	0.26	0.27	0.15	0.28	0.43	0.52	0.50
					1.00	0.04	-0.04	0.48	0.15	0.35	0.05
						1.00	0.09	0.19	0.17	-0.04	-0.07
							1.00	0.09	0.38	0.24	0.35
								1.00	0.21	0.25	0.21
									1.00	0.18	0.29
										1.00	0.60
											1.00

Table 7: PSI-20 Set Correlation Matrix

The Correlation Matrix, (see Table 7) confirms some empirical ideas and results from the literature we had about the stocks, namely that the first and the second ones, BES and BPI, are highly correlated, which is not a surprise as these two stocks are from the financial sector.

More surprisingly is the high correlation between each of these two and the third one, EDP that comes from electrical/energy sector. Interestingly there are no negative correlations between the stocks, probably because none of the business sectors presents are antagonist.

The eighth, PTI, seems to be the one less correlated globally, which is a surprise namely to what concerns SEM, a company from the same sector. The eleventh, SON, seems to be the one most well correlated globally. Probably not a surprise due to their more global presence in the business world.

4.3.1.3 Eigenvalues

Now, we will calculate and visualize (see Figure 13) the evolution of the ratio between the highest three eigenvalues and their relationship for the twelve stocks.

From Figure 13 it is understandable that the ratio between the highest eigenvalue and the third highest one, named λ_1/λ_3 , is generally higher than the ratio between the highest eigenvalue and the second one, named λ_1/λ_2 , as it was expected. Also, they are in a way correlated because the general framework between peaks and valleys does not differ at all.

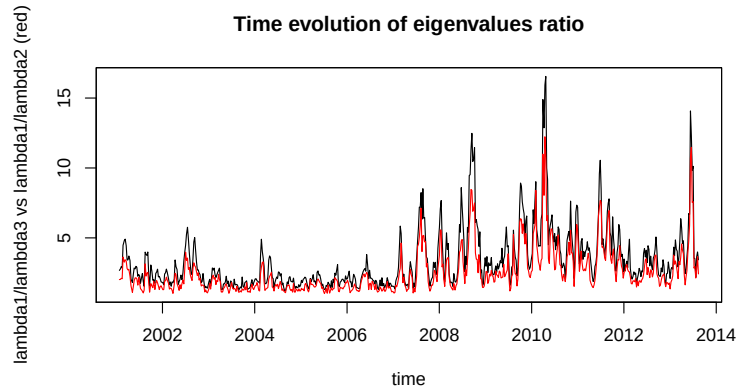


Figure 13: Evolution of stocks eigenvalues ratio

It is possible to calculate some statistics for these two ratios (Table 8). It is interesting to note the almost equal Skewness and Kurtosis. Also, it is worth to refer the maximum values: λ_1 reaches more than 16 times λ_3 value and reaches more than 12 times λ_2 value.

	λ_1/λ_3	λ_1/λ_2
Minimum	1.22	1.02
Quartile 1	1.92	1.48
Median	2.68	2.05
Arithmetic Mean	3.38	2.55
Geometric Mean	3.01	2.29
Quartile 3	4.13	3.02
Maximum	16.55	12.23
Stdev	2.17	1.60
Skewness	2.33	2.35
Kurtosis	7.48	7.40

Table 8: Descriptive statistics for stocks eigenvalues ratio

Looking closer at the Figure 13 we can observe that these ratios reached the highest values in the last 7 years. We can propose a division between a relatively stable period from 2000 to 2007, with the maximum ratios reaching the value 5, and a quite unstable period from 2007 until present, with more than 15 peaks above the value 5. The challenge, now, is to find relevant financial information that could explain these peaks.

We also did some calculations using a weighted covariance matrix (with parameters $R = 0.9$ and an horizon of 20 trading days). The values obtained suggest that there is no noticeable difference between a real covariance matrix and a weighted one (see Figure 14).

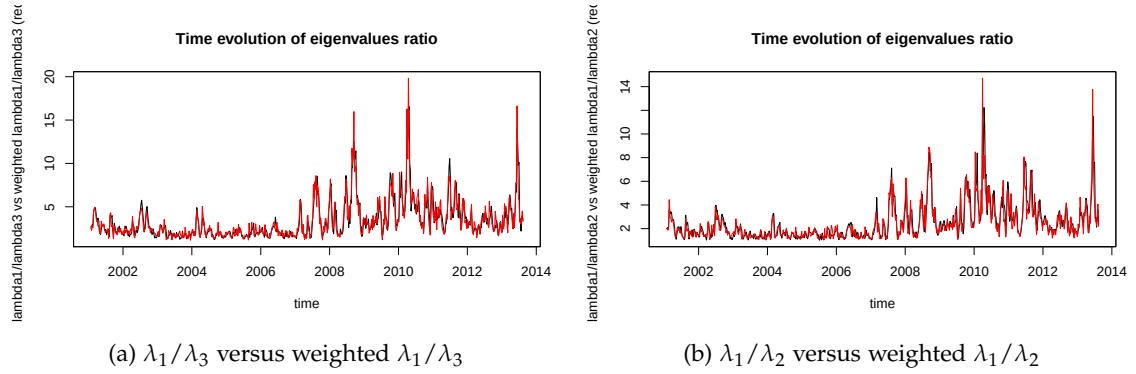


Figure 14: Evolution of stocks weighted eigenvalues ratio

4.3.2 Component Analysis

4.3.2.1 Forecastable Components (ForeCA)

ForeCA is a novel dimension reduction technique for temporally dependent signals. Contrary to other popular dimension reduction methods, such as PCA or ICA, ForeCA explicitly searches for the most “forecastable” signal. The measure of forecastability $\hat{\Omega}$ is based on negative Shannon entropy of the spectral density of the transformed signal.

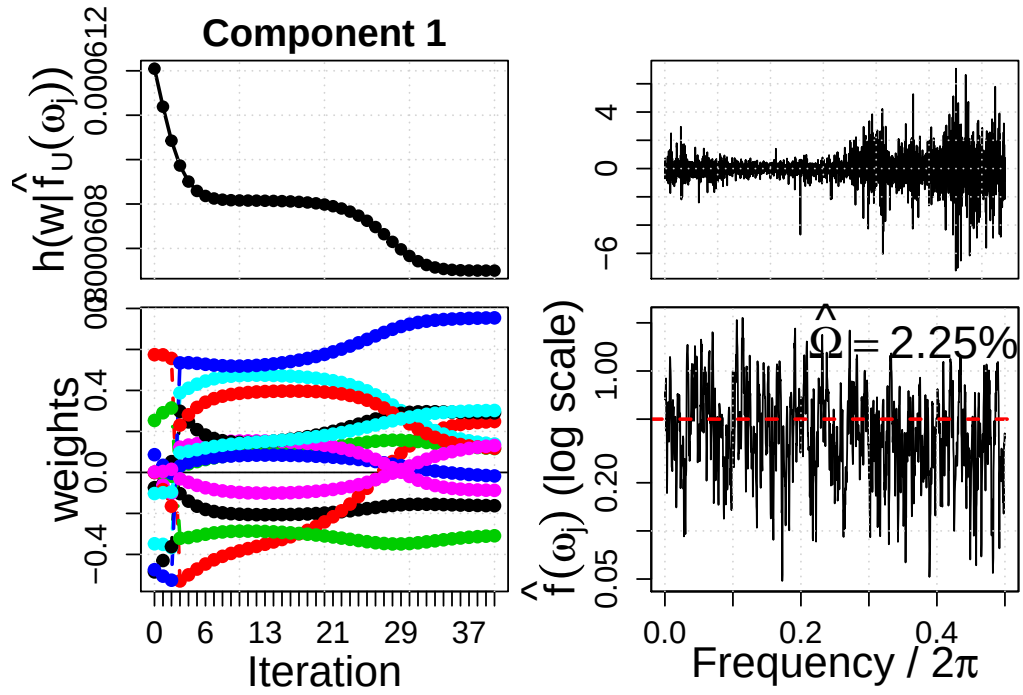
In Table 9 are shown the global forecastability results using this technique. We can “read” that the most predictable signal would be BES and the less one would be SEM.

BES	BPI	EDP	EGL	JMT	NBA	PTC	PTI	SEM	SON	SONC	ZON
2.06	1.55	1.31	1.46	1.54	1.56	1.37	1.44	1.20	1.28	1.28	1.46

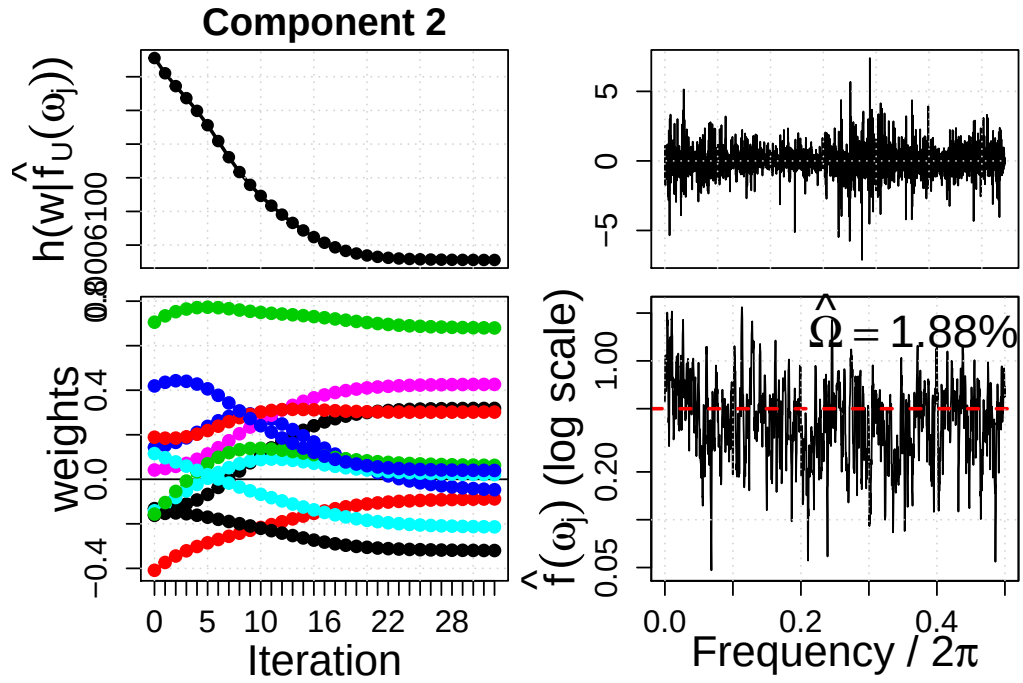
Table 9: ForeCA stocks results

In Figure 15 it is possible to visualize from top to bottom and from left to right: the component values, the values variation, the weights iteration and the spectral density estimation (smoothed). In respect to the last value, $\hat{\Omega}$, the forecastability, the values are in line to others found in financial time series Goerg [2013].

Also, in Figure 16, it is shown a biplot between the two components and the forecastability and the white noise for both components. Also, we can appreciate the forecastability values for the 12 PSI-20 stocks, whose numerical value was already shown in Table 9. It is interesting to note the almost absence of white noise, being PTI the relevant exception.

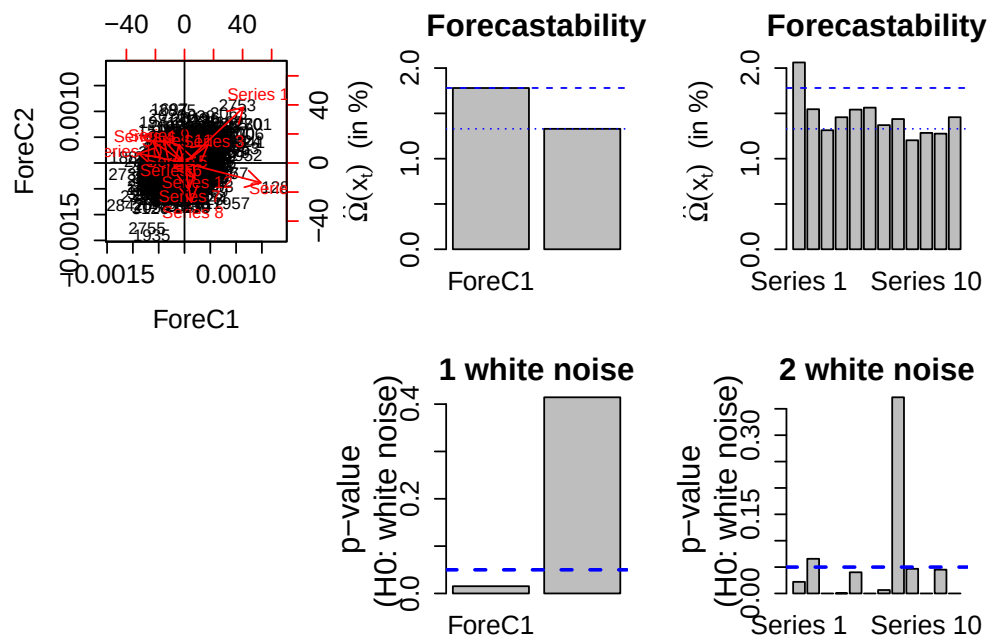


(a) ForeCA component 1



(b) ForeCA component 2

Figure 15: ForeCA stocks components



4.3.3 Entropy

4.3.3.1 Mutual Information

The Mutual Information between the stocks set was calculated using an R library called “entropy”.

We got abnormal values, the peaks, during 2001 and during 2008-2009, which corresponds to the first and second recession periods although the first recession period is not so notorious in the BES-BPI case (see Figure 17).

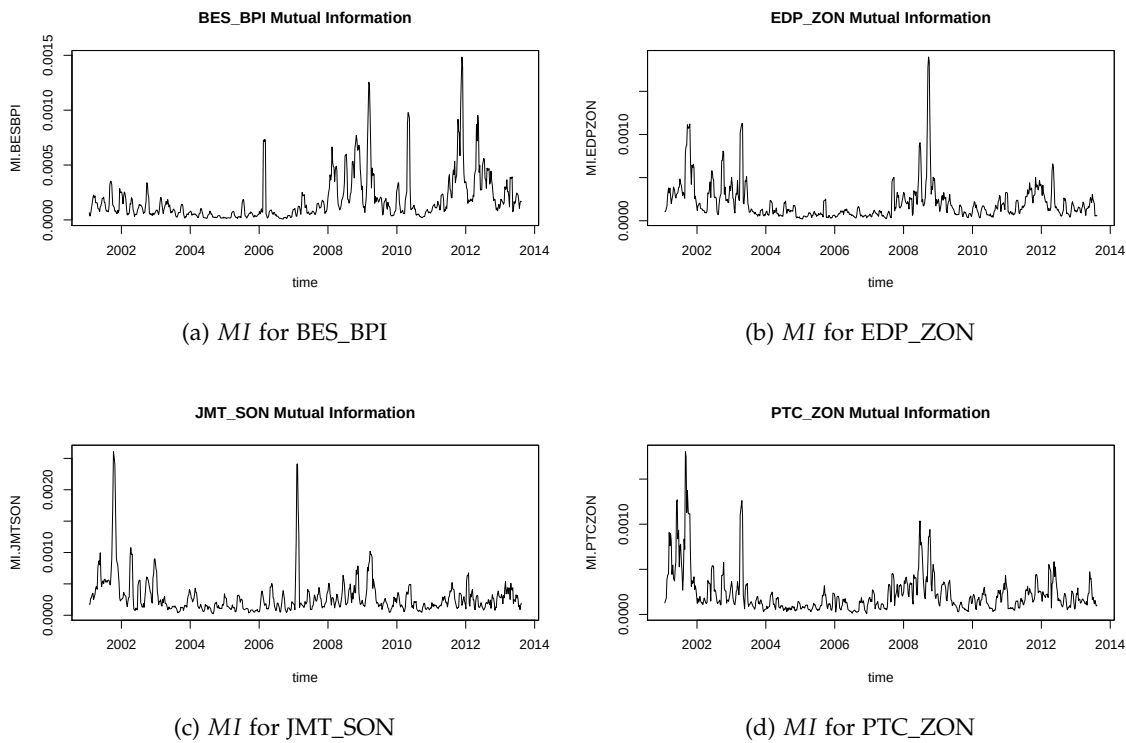


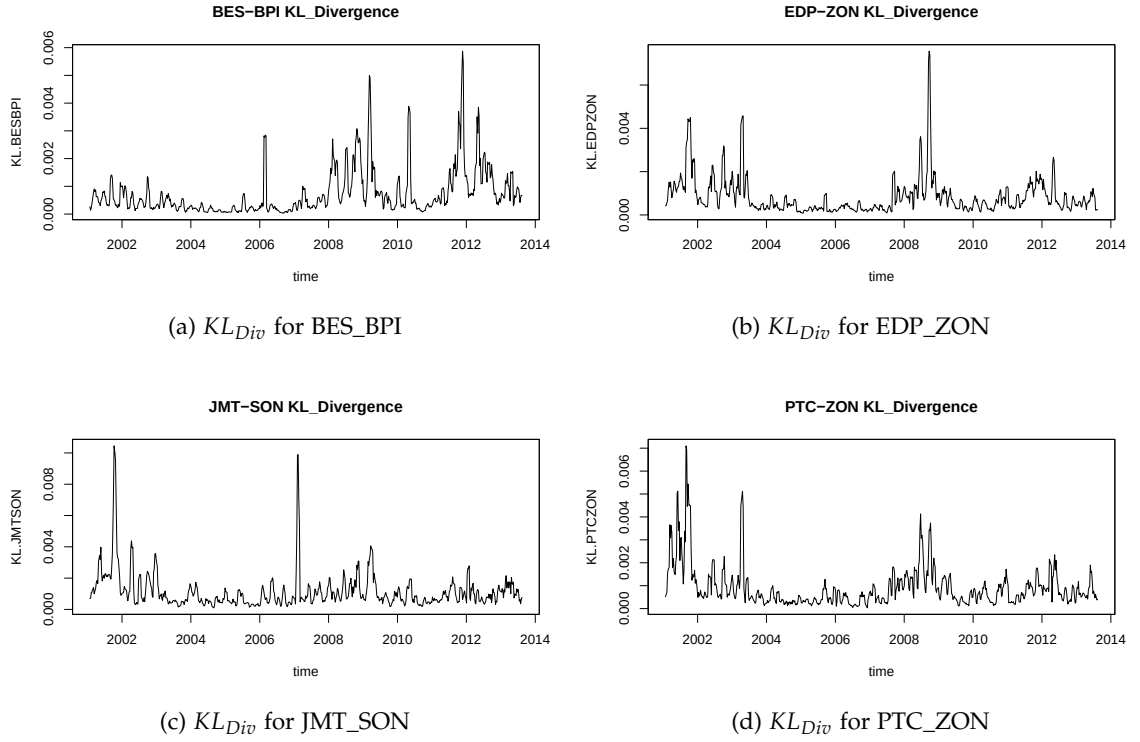
Figure 17: MI for PSI-20 stock pairs

Also, it is interesting to see that in the BES-BPI case we can find a peak in the first quarter of 2006, related to the aborted take-over attempt by Banco Comercial Português over BPI, and that from the second recession period until now there are some peaks due, probably, to the fact that this second recession became a financial system crisis bringing turbulence over financial institutions.

In the EDP-ZON and PTC-ZON cases there is a common peak in the first quarter of 2003 that we attribute to the split of PT Multimedia (now known by ZON) from PT. For the comparative periods proposed in Chapter 3, namely 2004 and from 2011 until 2013, there are no interesting peaks, apart from the one reported before for the BES-BPI case.

4.3.3.2 Kullback-Leibler divergence

The Kullback-Leibler divergence for the stocks set was calculated using an R library called “entropy” and are shown in Figure 18.

Figure 18: KL_{Div} for PSI-20 stock pairs

The results are almost the same as the ones obtained for the Mutual Information. This is probably due to the fact that these two measures are very similar. So, the conclusions extracted for the Mutual Information technique can be adopted to the Kulback-Leibler divergence technique conclusions.

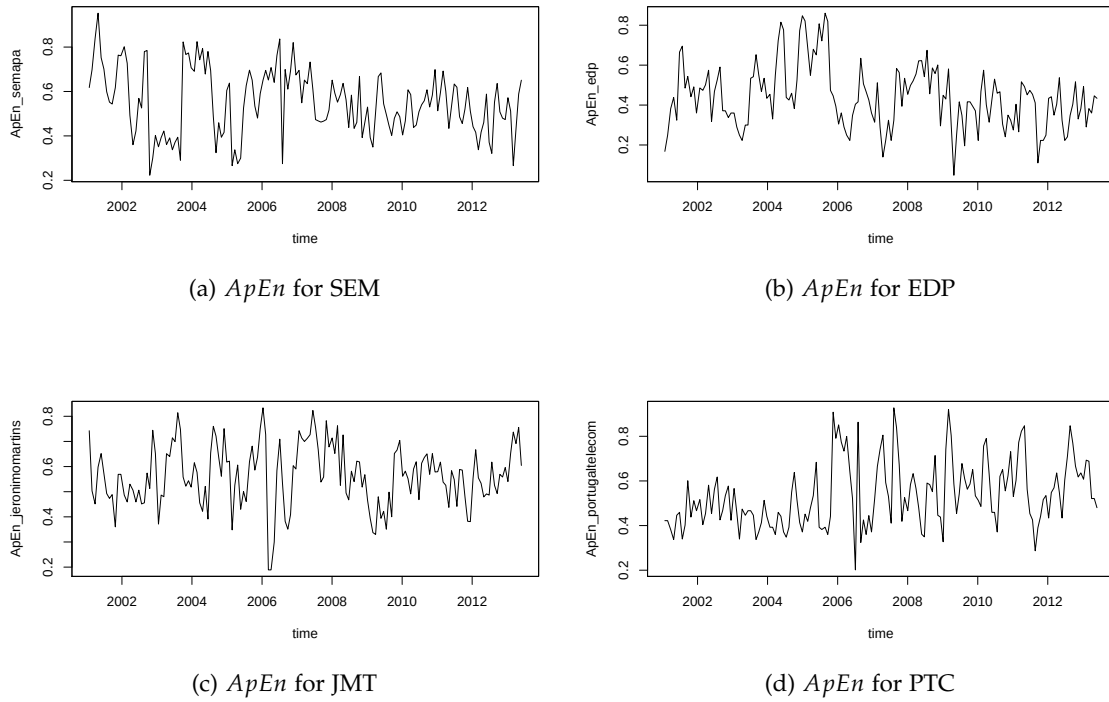
4.3.3.3 Approximate Entropy

Approximate Entropy ($ApEn$) was proposed and is being used as a measure of systems complexity. In this way, $ApEn$ is a “regularity statistic” that quantifies the unpredictability of fluctuations in a time series. Intuitively, then, the presence of repetitive patterns of fluctuation in a time series should render it more predictable than a time series in which such patterns are absent.

$ApEn$ value reflects the likelihood that “similar” patterns of observations will not be followed by additional “similar” observations. A time series containing many repetitive patterns has a relatively small $ApEn$; a less predictable time series has a higher entropy value.

Our results suggests that the stock time series are highly unpredictable with significant $ApEn$ values variations during time as we can see in Figure 19.

The results are very irregular, nevertheless we can infer, by inspection, two distinct periods: one, from 2000 to 2008, with higher $ApEn$ variations and another, more calm, from 2009 to present. Obviously, no rule dominates alone, so we can observe a very interesting exception with PTC, being the lower $ApEn$ variations from 2000 to 2006.

Figure 19: $ApEn$ for PSI-20 stocks

A closer look, using the recession periods, tells us that the $ApEn$ has an atypical behaviour tendency, diminishing as the period goes through. The exceptions are in the first recession period for EDP and PTC (Figure 19).

4.3.4 Distance Correlation

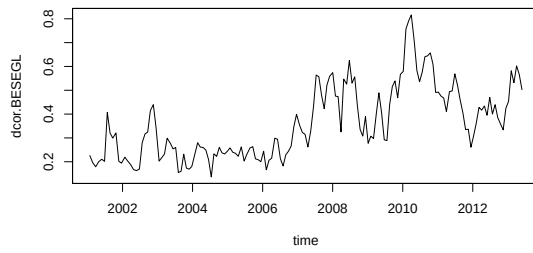
Here are presented the results obtained with Distance Correlation. In a general way, for most of the observed correlations the most striking fact seems so evident that we can propose a division between a relatively stable period from 2000 to 2007, with the maximum correlation values being well under the correlation values present in a quite unstable period from 2007 until present (see Figure 20).

The exception is Novabase (NBA) as we can see from Figure 21. One possible reason to this behaviour may be the fact that NBA was not a full-time PSI-20 stock between 2000 and 2014.

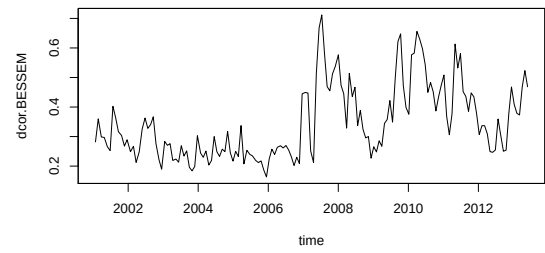
This division suggests by one hand that the magnitudes of the two recessions are quite distinct and that the time series are now much more correlated. This means that an important event will spread easily.

In the recession periods we see the Distance Correlation values going down with time, showing the same tendency already observed in Approximate Entropy.

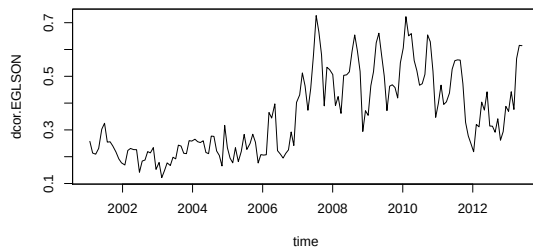
For a complete “catalogue” of results on PSI-20 please refer to the Appendix B.



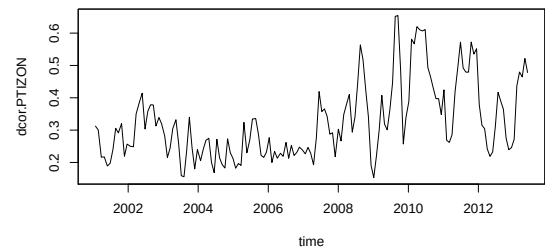
(a) Distance Correlation pair BES-EGL



(b) Distance Correlation pair BES-SEM

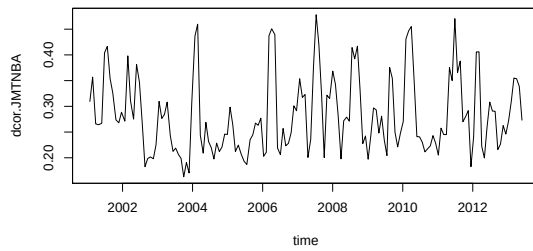


(c) Distance Correlation pair EGL-SON

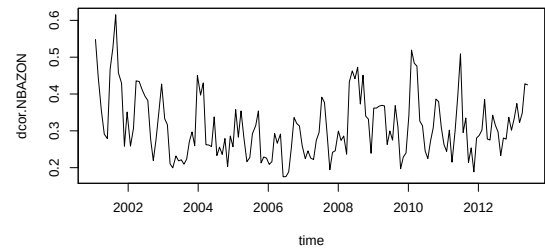


(d) Distance Correlation pair PTI-ZON

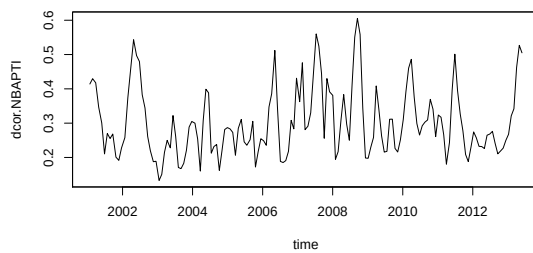
Figure 20: $DCov$ for PSI-20 stock pairs



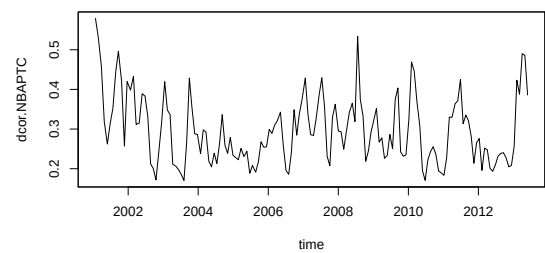
(a) Distance Correlation pair JMT-NBA



(b) Distance Correlation pair NBA-ZON



(c) Distance Correlation pair NBA-PTI



(d) Distance Correlation pair NBA-PTC

Figure 21: $DCov$ for PSI-20 stock pairs

4.3.5 Hurst Exponent

Here we present some results on PSI-20 data set for Hurst exponent calculated using detrended fluctuation analysis (DFA).

But, first of all, for the robustness and liability of the results let us show the fluctuation function (Figure 22) obtained for the PSI-20 index. The linear fit over all windows from all scales (see explanation in Section 2.7) gives a Pearson correlation coefficient of 0.998 and a standard-deviation (assuming the errors normally distributed) of 0.004 taken for the log-log results.

Hurst exponent is obtained by fitting a power law to the DFA function $\langle F(t) \rangle$ computed in the sliding window. Pearson Correlation coefficients are computed for the fit in each case.

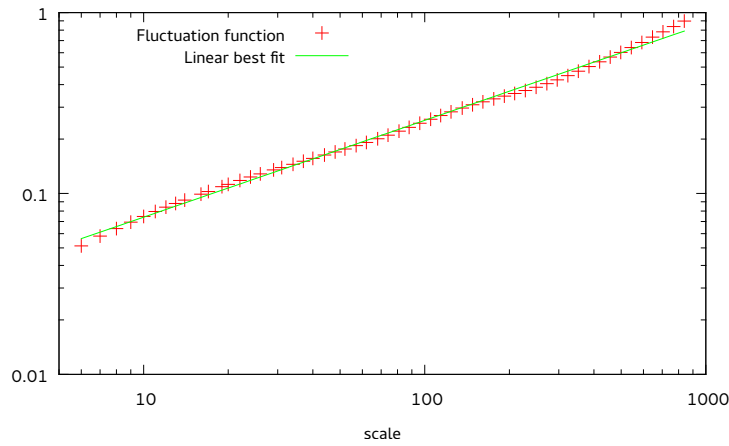


Figure 22: PSI-20 fluctuation function

Let us now consider in Figure 23 some Hurst exponent calculations for some PSI-20 stocks. Their values are, typically, around 0.5 and 0.7 meaning that there is a small long memory process present in these stocks. The correlation coefficient $r(t)$ is also plotted for each point revealing the quality of the fit where the H exponent is evaluated; in all graphics the correlation coefficient is near 1. All correlation coefficients, $r(t)$, may be seen to fall in the range 0.95 – 1, giving us confidence in the power law behaviour of $\langle F(t) \rangle$.

Of interest are the observed “abrupt valleys” in all four plots, namely the ones that are common for BES, BPI and PTC in the beginning of 2006. These, and all the other present “abrupt valleys” should have a event related meaning.

For a global Hurst exponent for the stocks we can view Table 10. It is noticeable that half of the Hurst exponents, H , are under or above 0.5, meaning that there is some diversity in stocks maturity and in independence from past results. EDP is the best example of a stock that does not follow trends, that is, have “anti-persistence” behaviour. Others examples could be SEM or even PTC, PTI and SON, all corresponding to classical business sectors. On the other hand we see NBA and SONC having the most “persistent” behaviour. These stocks correspond to technological companies, that is, belonging to a more “turbulent” business sector. The same can be said about BES and BPI, from the financial sector, another “turbulent” business sector.

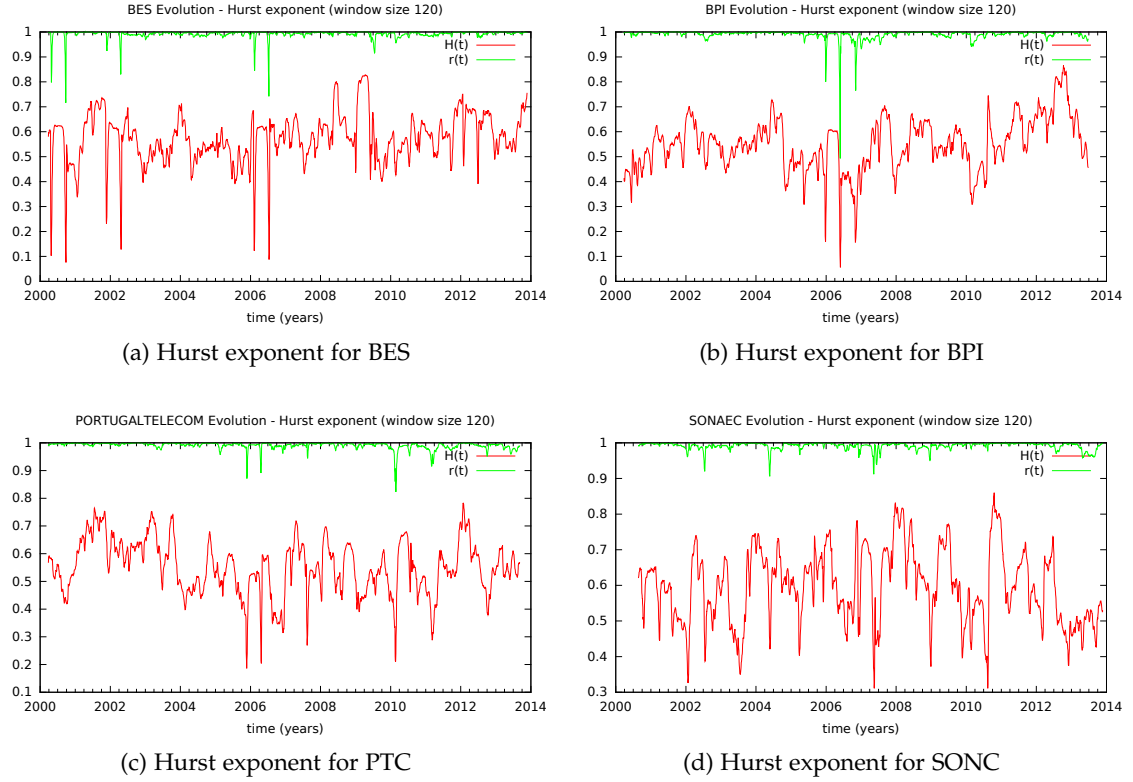


Figure 23: Hurst exponent for PSI-20 stocks

Stock	H	R	σ_H
$^{\wedge}\text{BES}$	0.525	0.998	0.00443
$^{\wedge}\text{BPI}$	0.53	0.999	0.00302
$^{\wedge}\text{EDP}$	0.392	0.975	0.0121
$^{\wedge}\text{JMT}$	0.505	0.999	0.00309
$^{\wedge}\text{EGL}$	0.495	0.999	0.00341
$^{\wedge}\text{NBA}$	0.567	0.998	0.0053
$^{\wedge}\text{PTI}$	0.472	0.991	0.00839
$^{\wedge}\text{PTC}$	0.462	0.997	0.00454
$^{\wedge}\text{SEM}$	0.437	0.992	0.00727
$^{\wedge}\text{SONC}$	0.559	0.999	0.00307
$^{\wedge}\text{SON}$	0.473	0.996	0.00581
$^{\wedge}\text{ZON}$	0.501	0.998	0.00469

Table 10: Hurst exponent for PSI-20 stocks

4.4 CONCLUDING REMARKS

In this chapter some results found in literature were confirmed, namely the ones from random matrix theory and the ones for Hurst exponent.

For Mutual Information or Kullback-Leibler Divergence the results are very sharp and a event related comparison was applied to find out the coincidences. This analysis has shown that we can match the more interesting values obtained with real events.

To our knowledge, it is the first time that energy statistics is applied to the PSI-20 data. It is interesting to note that this measure proposes two well defined behaviour for the PSI-20 stocks. One period, from 2000 to 2007, relatively calm, with low variation of Distance Correlation between stocks, and another period, from 2007 till now, much more agitated in what concerns this measure.

Nevertheless, besides the proposal that the stocks are much more correlated in this period, and that this happen because of the global recession, it is only possible to suggest that the Distance Correlation values tend to diminish after the most important event take place.

Distance Correlation proposal is complemented by Approximate Entropy. Also, this measure, proposes these two well defined periods. When, in periods of crisis, $ApEn$ becomes agitated with higher variations but also diminishing with time.

“I compare her (Fortune) to one of those raging rivers, which when in flood overflows the plains, sweeping away trees and buildings, bearing away the soil from place to place; everything flies before it, all yield to its violence, without being able in any way to withstand it; and yet, though its nature be such, it does not follow therefore that men, when the weather becomes fair, shall not make provision, both with defences and barriers, in such a manner that, rising again, the waters may pass away by canal, and their force be neither so unrestrained nor so dangerous. So it happens with fortune, who shows her power where valour has not prepared to resist her, and thither she turns her forces where she knows that barriers and defences have not been raised to constrain her.” Niccolò Machiavelli, The Prince , Chapter XXV

5.1 INTRODUCTION

In this chapter we will apply the mathematical tools presented in the Chapter 2 to the World Markets set. The data used in this study was taken from a set of worldwide market indices, enumerated in Chapter 3, and are constituted by the daily close values for the respective indices. As it is usual in this kind of analysis, the results come from the analysis of the returns $\eta_i = \log \frac{x_i}{x_{i-1}}$.

In Appendix A we can observe the returns for all the 23 markets. Looking at the returns helps us to look only to relative variation and not to absolute values. In fact, these markets are quite different in absolute values, as it can be seen.

5.2 RESULTS

Applying the techniques from Chapter 2 we reach a set of results that we will show and interpret in this Section.

5.2.1 Random Matrix

For this set we consider $2965/5 = 589$ samples by sequentially sliding a window of $T = 20$ days by 5 days (roughly one month calculated week by week). For each period, we look at the empirical correlation matrix of the $N = 23$ markets during that period. The quality factor is therefore $Q = T/N = 20/23 = 0.87$.

We started by comparing the real eigenvalues density with the theoretical one as proposed by Marchenko and Pastur [1967] (see Figure 24).

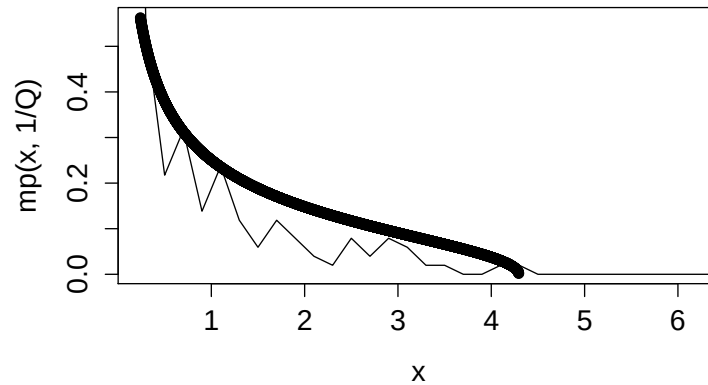


Figure 24: Theoretical versus Real eigenvalues densities

Next, just to support our confidence, we calculate and relatively compare the 3 highest eigenvalues from a subset of the World Markets set: the 9 European markets subset. It is fair to say that there is no special reason for choosing this subset.

Eigenvalues calculation

In Figure 25 we compare the relationship between the 3 major eigenvalues. We can generally say that the highest eigenvalue is getting higher over the time. It starts to be 3,3 to 5 times higher in the beginning of the XXI century and more recently became almost 10 to 15 times higher than the second. More recently, the difference between them is getting, again, smaller. From the second to the third highest we can infer a relationship of 2.

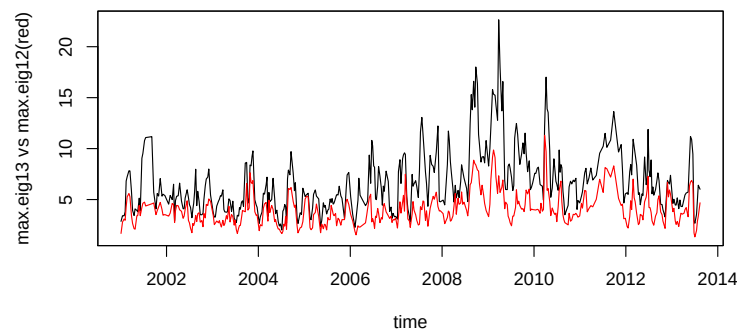


Figure 25: World Markets Ratio λ_1/λ_3 versus λ_1/λ_2

Weighted time series

In order to understand if there is any interest in considering, for the eigenvalues calculation, weighted time series (see Subsection 2.3.2 and Equation (22), we simulated and obtained the results shown in Figure 26.

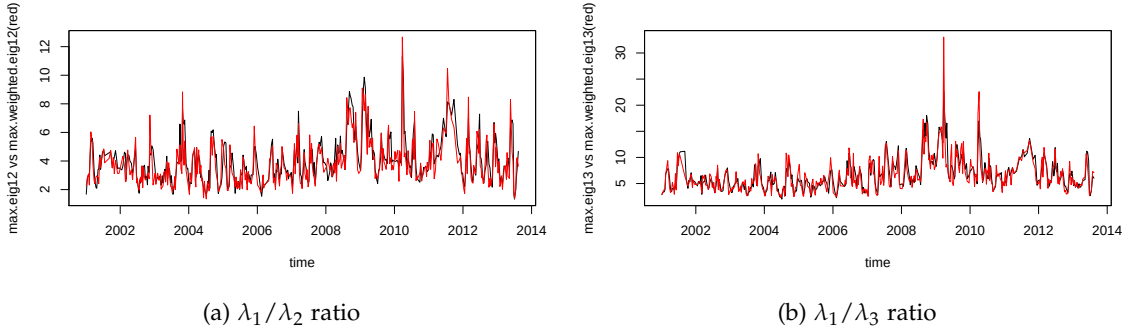


Figure 26: Real vs Weighted Eigenvalues Ratios

We can, with no doubt, say that there is no difference between considering a real market or a weighted market. In a way, this means that there is no memory and that the returns are independent from one step to another.

We did another simulation but for random markets. The result was what we were expecting, that is, the eigenvalues are more similar in a random market. And again for the third eigenvalue.

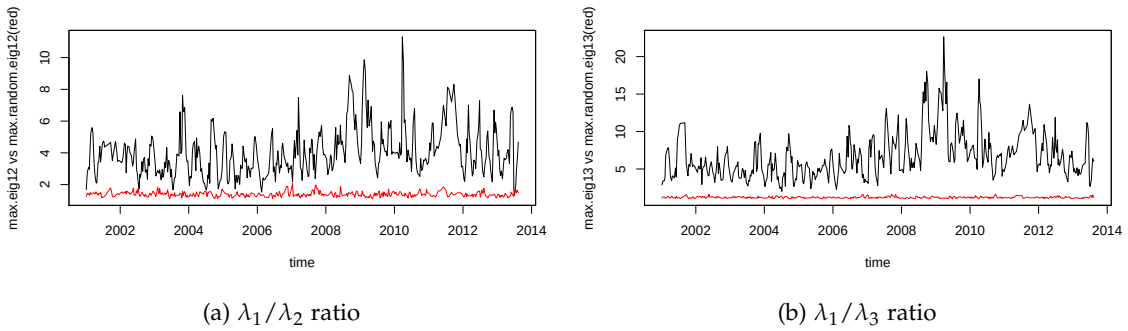


Figure 27: Real vs Random Eigenvalues Ratios

5.2.2 Component Analysis

Forecastable Components (ForeCA)

As said before, ForeCA is a novel dimension reduction (DR) technique for temporally dependent signals. The measure of forecastability $\hat{\Omega}$ is based on negative Shannon entropy of the spectral density of the transformed signal.

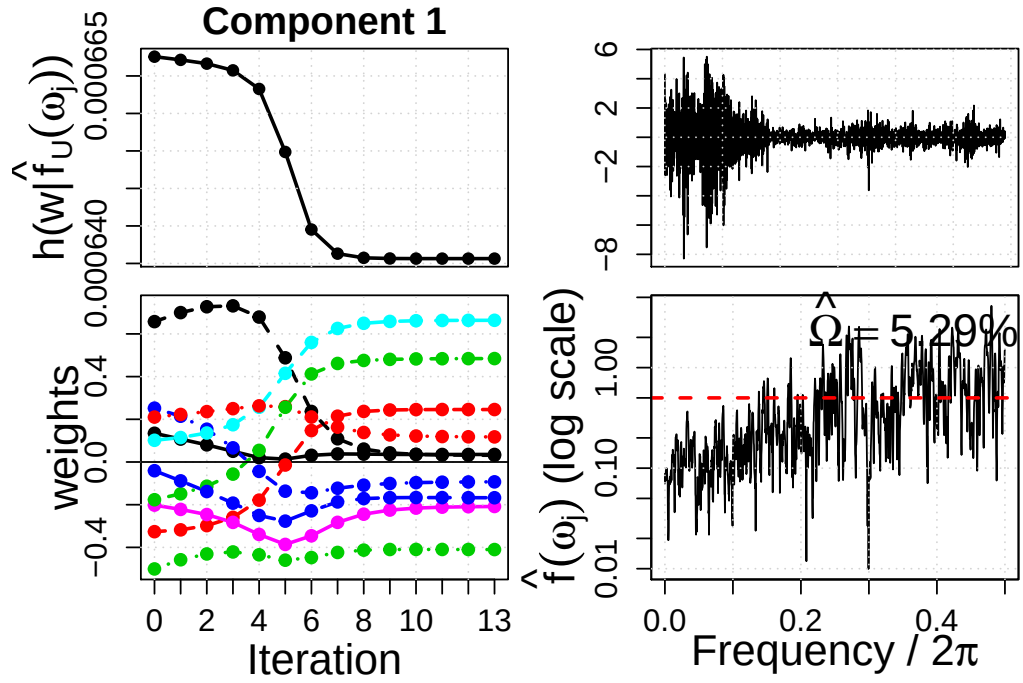
Here, we will show an example using only the European markets, a subset from the World Markets set. In Table 11 are shown the global forecastability results using this technique. We can “read” that the most predictable signal would be ATX and the less one would be CAC.

AEX	ATX	CAC	DAX	FTSE	IBEX	MIB	PSI-20	SSMI	STOXX
1.60	1.76	1.46	1.58	1.58	1.63	1.60	1.55	1.67	1.53

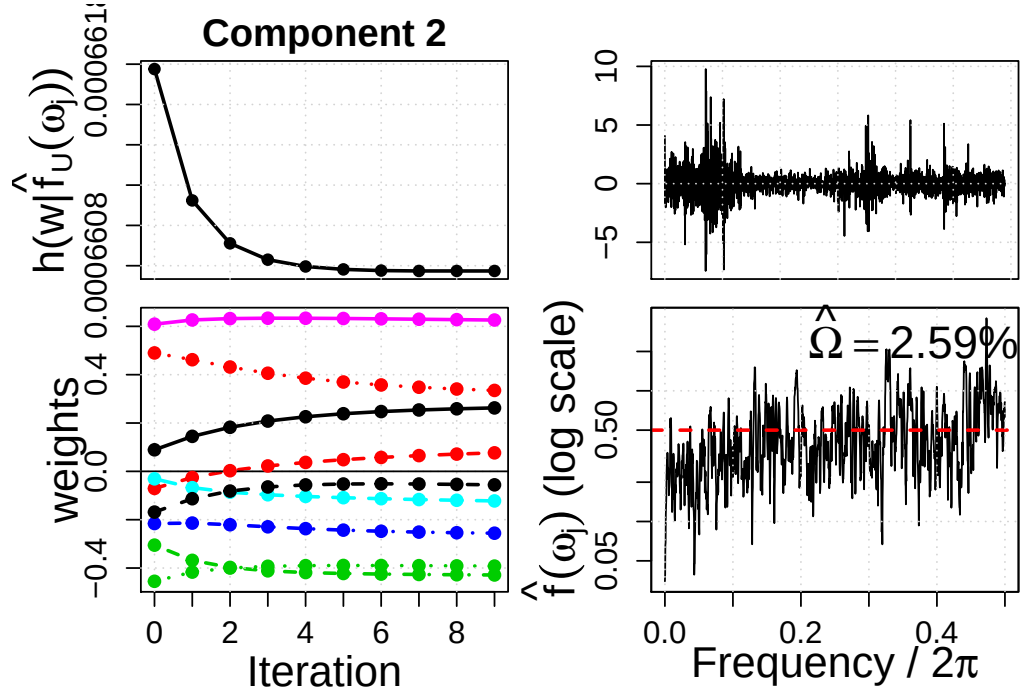
Table 11: ForeCA world markets results

In Figure 28 it is possible to visualize from top to bottom and from left to right: the component values, the values variation, the weights iteration and the spectral density estimation (smoothed). In respect to the last value, $\hat{\Omega}$, the forecastability, the values are in line to others found in financial time series Goerg [2013], although these market time series seems to be more predictable than the stocks time series, as we can infer by comparing the results obtained in Chapter 4 to those obtained here.

Also, in Figure 29, it is shown a biplot between the two components and the forecastability and the white noise for both components. Also, we can appreciate the forecastability values for the 10 European markets, whose numerical value was already shown in Table 11. It is interesting to note the almost absence of white noise. The exception is PSI-20 and in a minor scale, ATX and MIB.



(a) ForeCA component 1



(b) ForeCA component 2

Figure 28: ForeCA world markets Components

5.2.3 Entropy

Mutual Information

The Mutual Information between the World Markets set was calculated using an R library called “entropy”.

Our results suggests that the highest values observed in Figure 30, the peaks, have all correspondence to real events. First of all, they are concentrated in 2001 and 2007-2009, the recession periods.

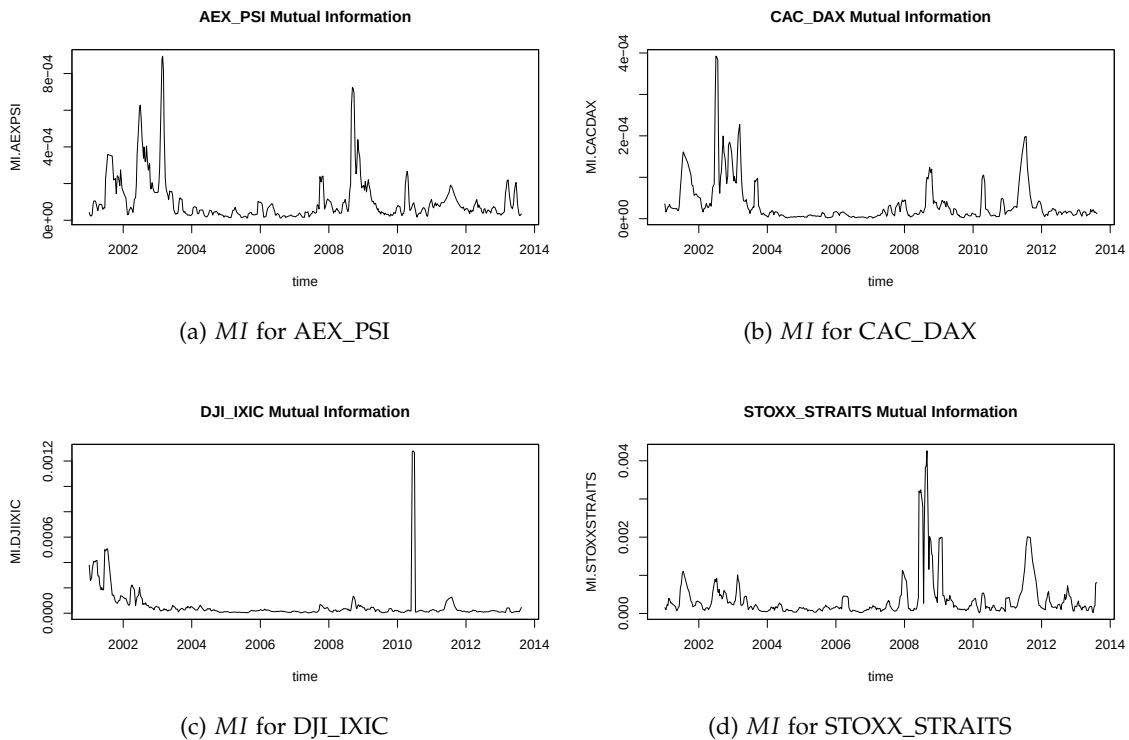


Figure 30: MI for World markets pairs

Despite this, two interesting exceptions must be taken into account. The first one is that the Mutual Information values for European markets remain for some time more slightly high after the recession periods. A tentative explanation can reside in the fact that these recession periods were defined for United States, not for Europe.

The second one is that we found a very pronounced value in mid-2010 in the DJI-IXIC case, two North-American markets. We relate this to the Dodd-Frank Wall Street Reform and Consumer Protection Act, which is “only” the biggest Wall Street reform since the Great Depression in the late 20’s of the XX century.

It is also worth to say that markets that does not seem to be geographically related, like STOXX and STRAITS show Mutual Information values 10 times higher than the values between geographically or commercially more related markets like DJI and IXIC or CAC and DAX.

Kullback-Leibler divergence

The Kullback-Leibler divergence for the World markets set was calculated using an R library called “entropy” and are shown in Figure 31.

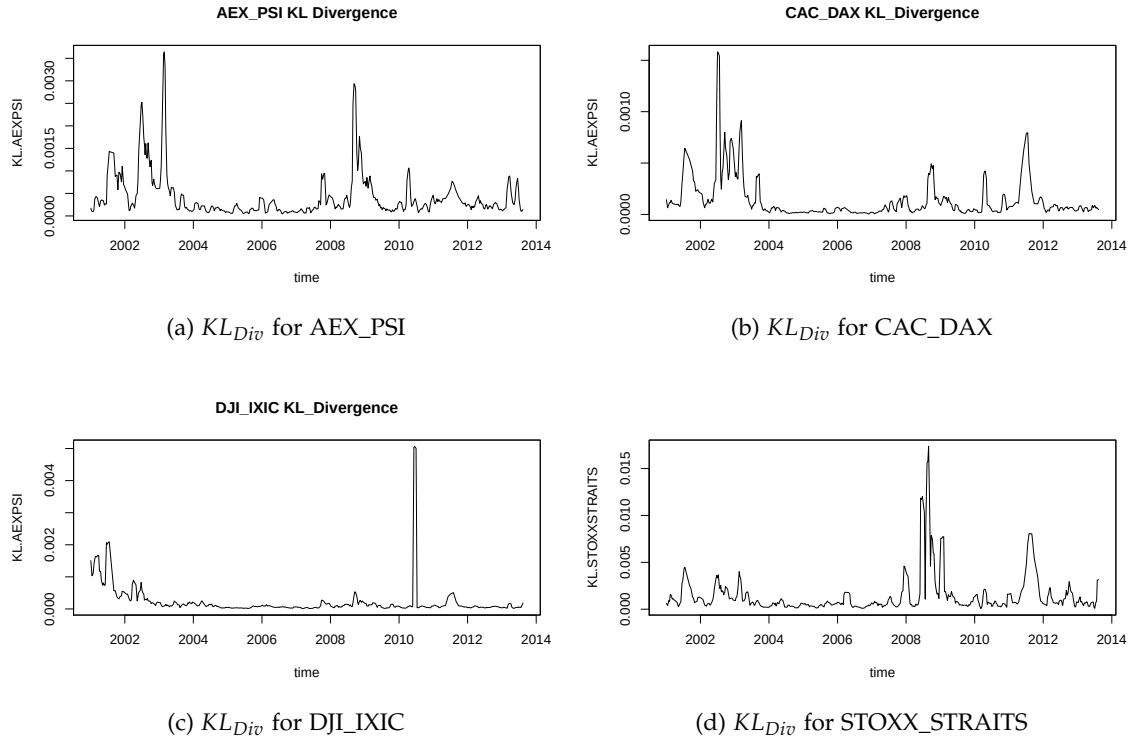


Figure 31: KL_{Div} for World markets pairs

The results are almost the same as the ones obtained for the Mutual Information. This is probably due to the fact that these two measures are very similar. So, the conclusions extracted for the Mutual Information technique can be adopted to the Kulback-Leibler divergence technique conclusions.

Approximate Entropy

Here are presented the results obtained with Approximate Entropy for World Markets set. To analyse possible regional patterns we dedicated some attention to European region dividing the results in European markets and non-European markets.

Our results suggests that all the time series seem highly unpredictable with significant $ApEn$ values variations during time as we can see in Figure 32 and Figure 33.

Despite this unpredictability $ApEn$ seems to peak at the beginning of recession periods and then goes down with time, although this is more notorious in the second one.

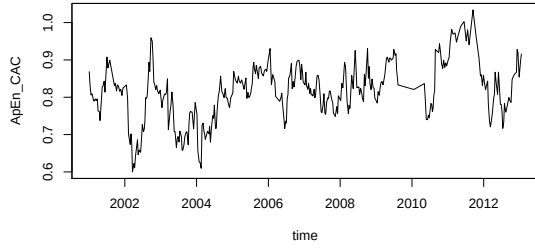
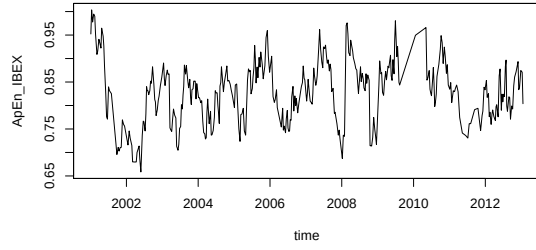
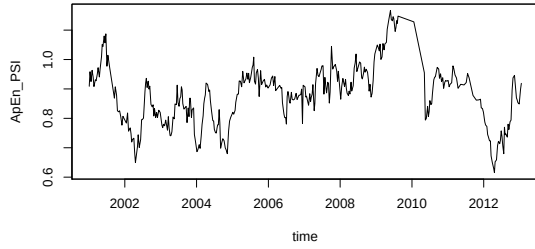
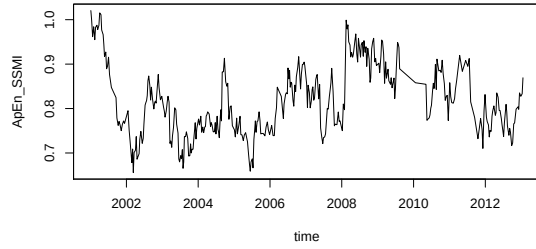
(a) *ApEn* for CAC(b) *ApEn* for IBEX(c) *ApEn* for PSI-20(d) *ApEn* for SSMI

Figure 32: Approximate Entropy for European markets

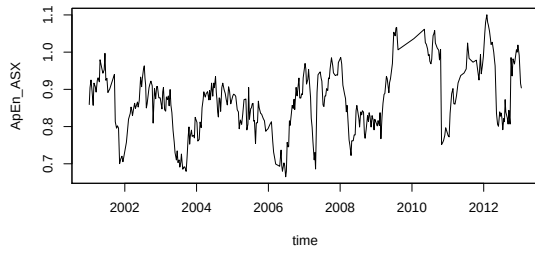
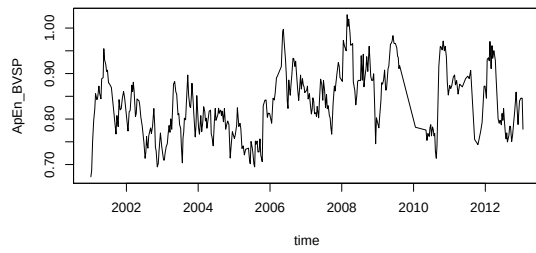
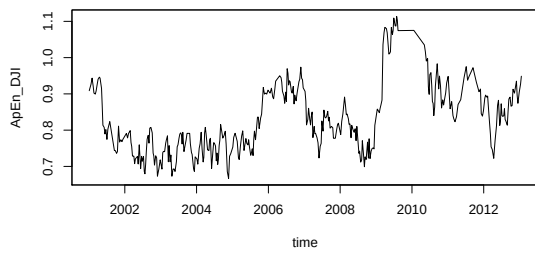
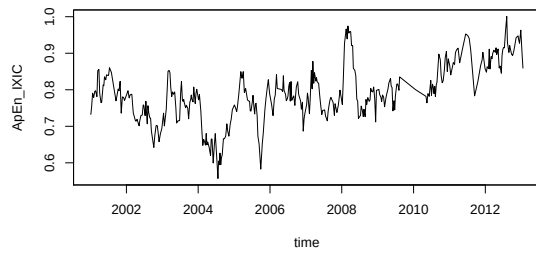
(a) *ApEn* for ASX(b) *ApEn* for BVSP(c) *ApEn* for DJI(d) *ApEn* for IXIC

Figure 33: Approximate Entropy for non-European markets

5.2.4 Distance Correlation

Here are presented some of the results obtained for Distance Correlation. For a complete “catalogue” of results concerning PSI-20 please refer to the Appendix B.

Asia-Pacific Markets

ASX

For the ASX market we can observe that there is no high correlation with any other market. Almost all the correlations goes between 0.3 and 0.7. As an example (see Figure 34) it is shown the correlation between ASX and HSI.

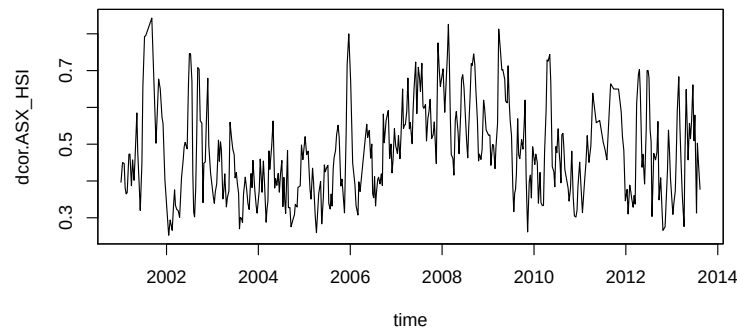


Figure 34: Distance Correlation for the ASX_HSI pair

BSESN

For this market we can only find a little different correlation relationship with the HSI market (Figure 35). The correlation goes up until 2008 and goes down from 2008 on, but does not leave the interval 0.3 to 0.7, apart from some peaks reaching 0.8 in 2008. For all

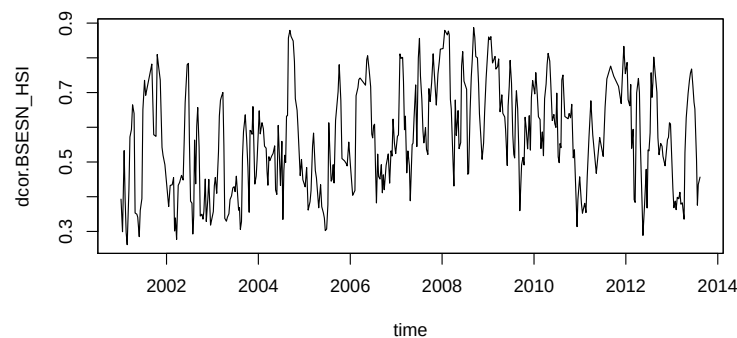


Figure 35: Distance Correlation for the BSESN_HSI pair

the other market it is not easy to find a pattern. Almost all the correlations are between 0.3 and 0.7 for most of the time series.

HSI, JKSE and NIK

For this market we can find interesting correlation relationship with the BSESN market, as commented before. Also, there are some pertinent comments on the correlation with some of the Asian markets: with NIK the correlation remains between 0.4 and 0.8 until 2007 (see Figure 36), but going down, and then, jumps to 0.5 to 0.8 and starts going down until now. The same transition in 2007 happens with other markets like JKSE but then remaining more “constant” before and after that year. For all the other markets it

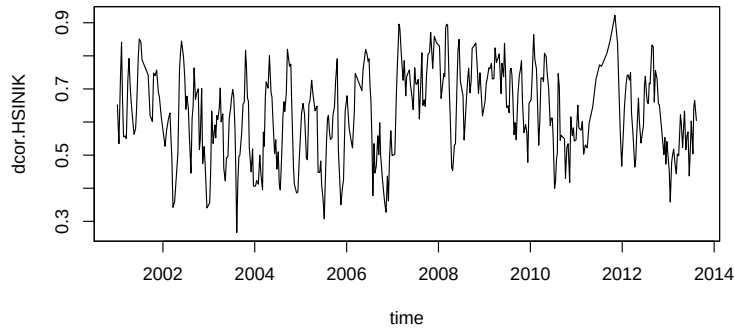


Figure 36: Distance Correlation for the HSI_NIK pair

is not easy to find a pattern. Almost all the correlations are between 0.3-0.7.

KOSPI

For the KOSPI market we can find a pertinent correlation with NIK in Figure 37. The correlation remains between 0.5 and 0.8 until 2007, and then, jumps to 0.6 to 0.9 between 2007 and 2011 and, after that, starts to oscillate in a no characteristic way.

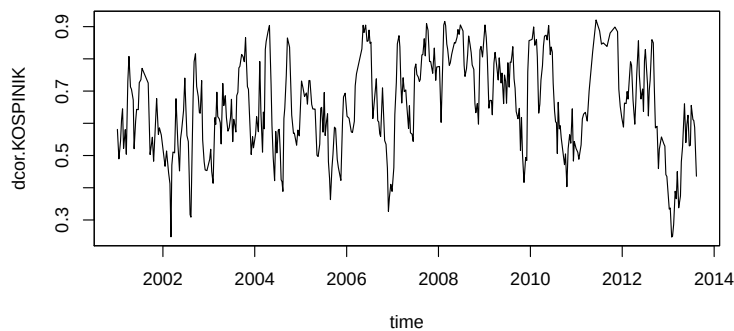


Figure 37: Distance Correlation for the KOSPI_NIK pair

European Markets

AEX

For the AEX market we can observe that there is a very high correlation with the other European markets, being the PSI-20 the exception, with correlation values typically 20% under. For the AEX_ATX pair it is possible to observe (see Figure 38) an interesting behaviour.

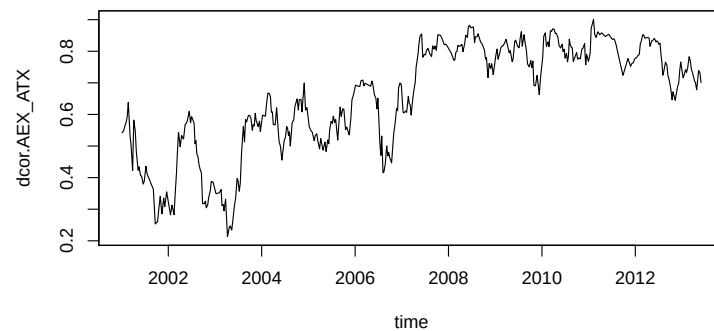


Figure 38: Distance Correlation for the AEX_ATX pair (60 days window width)

From 2007, corresponding to the crisis beginning, the correlation between these two markets grew from about 0.6 to 0.8, clearly showing more correlation. Apart from the European country markets there is only a very high correlation between AEX and STOXX, as we can see in Figure 39.

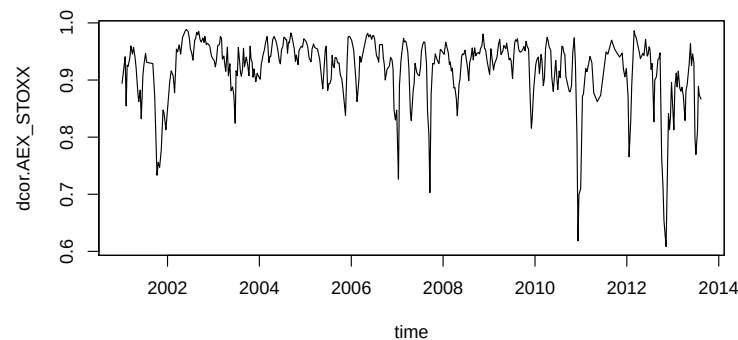


Figure 39: Distance Correlation for the AEX_STOXX pair

ATX

As AEX we can observe a very high correlation with the other European markets (for an example, see Figure 40), although only from 2008, jumping roughly from 0.5 to 0.8. In the PSI or SSMI case this jump also appears but fades quickly (see Figure 41).

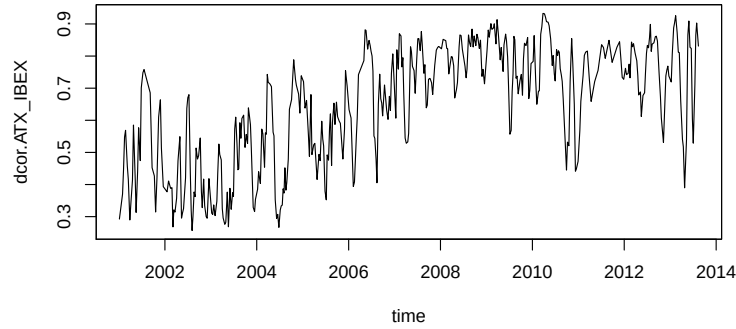


Figure 40: Distance Correlation for the ATX_IBEX pair

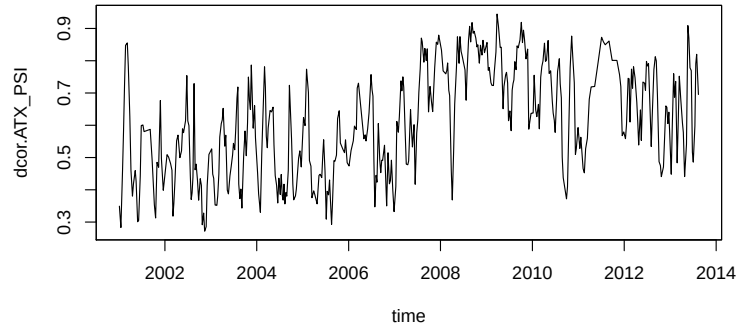


Figure 41: Distance Correlation for the ATX_PSI pair

Apart from the European country set, as with AEX, there is only a very high correlation between ATX and STOXX, but, again, only beginning in 2008 (Figure 42).

CAC

For the CAC market we can observe a very high correlation with the other European markets, from above 0.8, being the PSI-20 the only exception, with correlations varying between 0.5 and 0.8. Another interesting relationship is with STOXX (Figure 43).

We can also observe correlations between 0.5 and 0.8 for the relations with the North American subset (DJI, IXIC and SPY) and the Latin-American subset (BVSP, Merval and MXX). See, as an example, CAC versus DJI (Figure 44). For the other world markets we observe correlations between 0.4 and 0.8.

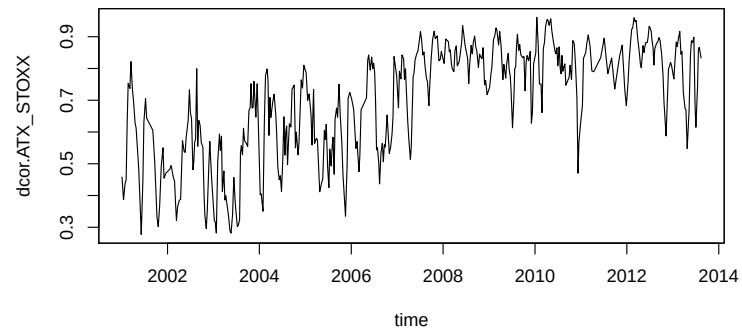


Figure 42: Distance Correlation for the ATX_STOXX pair

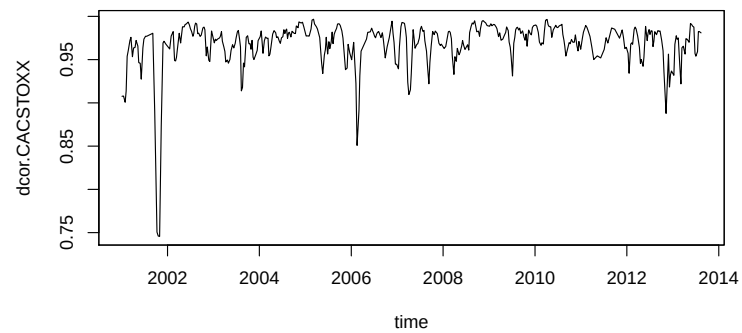


Figure 43: Distance Correlation for the CAC_STOXX pair

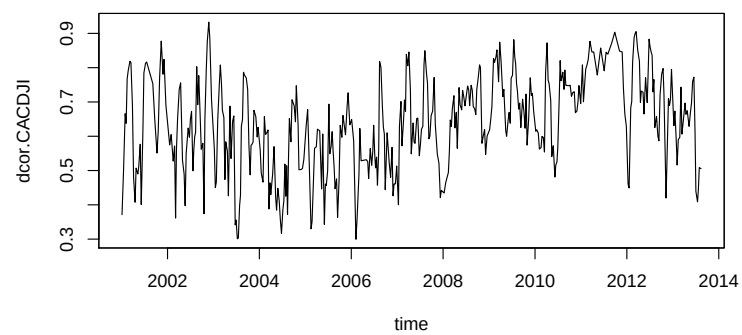


Figure 44: Distance Correlation for the CAC_DJI pair

DAX

For the DAX market we can observe a very high correlation with the other European markets, from above 0.8, being the exceptions the PSI-20, with correlations varying between 0.4 and 0.8 and the SSML, with correlations between 0.7 and 0.8. Another interesting relationship is with IBEX with the correlation jumping to 0.8 only from 2005 but going down more recently (Figure 45).

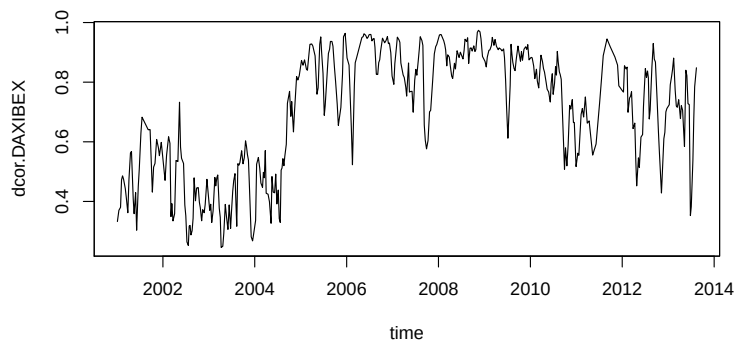


Figure 45: Distance Correlation for the DAX_IBEX pair

We can also observe correlations between 0.4 to 0.8 for the relations with the North American subset (DJI, IXIC and SPY) and the Latin-American subset (BVSP, Merval and MXX). See, as an example, DAX versus SPY (Figure 46). For the other world markets

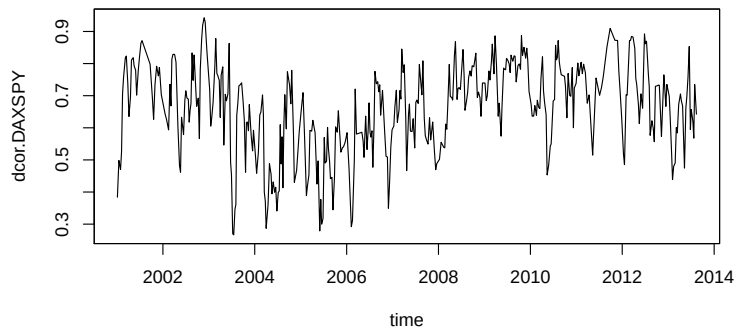


Figure 46: Distance Correlation for the DAX_SPY pair

we observe correlations between 0.3 and 0.7.

FTSE

For the FTSE market we can observe a very high correlation with the other European markets, from above 0.8, being the exceptions the PSI-20 as can be noted in Figure 47, with correlations varying between 0.4 and 0.8 (but varying in time).

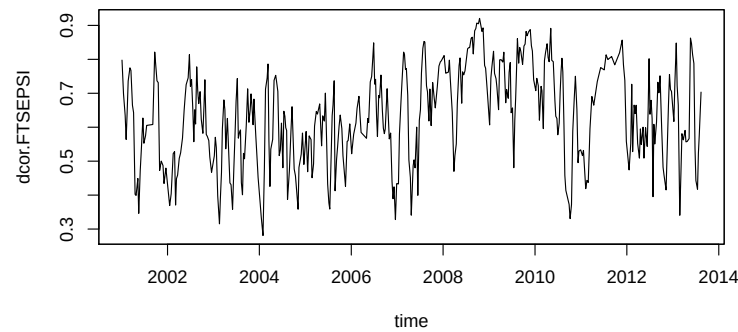


Figure 47: Distance Correlation for the FTSE_PSI pair

About FTSE and MIB, the correlation remains around 0.8 until 2011, and then, going down to 0.7 (see Figure 48). We observe the same interesting relationship with IBEX, as

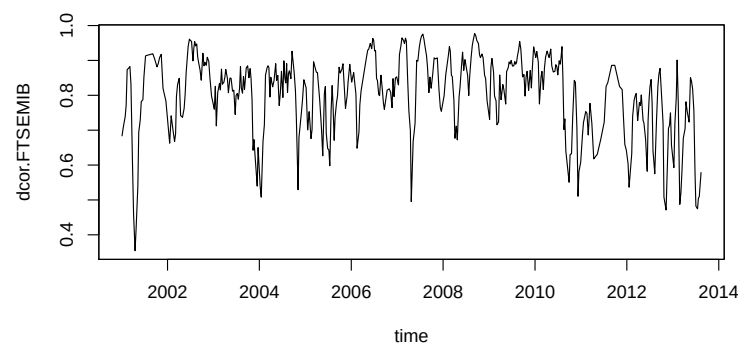


Figure 48: Distance Correlation for the FTSE_MIB pair

happened with DAX and IBEX, with the correlation jumping to 0.8 only from 2005 but then going down from 2011.

We can also observe correlations between 0.3 and 0.7 from the year 2000 until 2007 for the relations with the Latin-American subset (BVSP, Merval and MXX). More recently happens that the correlation goes up for correlations values around 0.7 from 2007 until 2012 and finally starting going down from 2012. See, for example the correlation with Merval (Figure 49). We can also observe correlations between 0.4 and 0.8 for the relations with the North American subset (DJI, IXIC and SPY), getting higher from 2007. For the other world markets we observe correlations between 0.3 and 0.7.

IBEX

For IBEX we can observe a very high correlation with the other European markets, from above 0.8, but only since 2005. The exceptions are the PSI and the SSMI. The first, because the 2005 jump is not so abrupt and because the correlation (apart from peaks) never goes higher than 0.8. The later because of the jump also being not so abrupt and because the

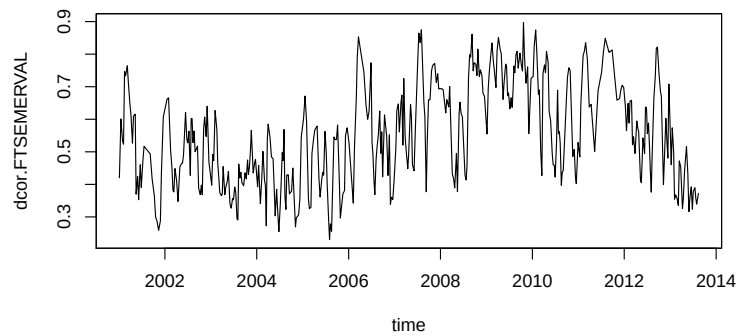


Figure 49: Distance Correlation for the FTSE_MERVAL pair

correlation stays around 0.8 only until 2011. From that year on the correlation starts to go down.

We can also observe correlations between 0.3 and 0.8 for the relations with the North American subset (DJI, IXIC and SPY) and with the Latin-American subset (BVSP, MERVAL and MXX), getting higher from 2007 and lower from 2011.

For the other world markets we observe correlations between 0.3 and 0.7.

MIB and SSMI

For MIB market we can observe a very high correlation with the other European markets and in a lower grade with the North American subset. Generally, we observe a diminishing correlation from 2011, for all the world markets. The correlations for these markets are, typically, between 0.3 and 0.7. We can apply to SSMI almost the same observations as we did for MIB market.

PSI-20 and STOXX

Nothing more relevant to say.

5.2.4.1 Latin-American Markets

BVSP

For the BVSP market we can observe that there is a high correlation, although variable, with the other five markets from North or Latin-America. As an example we show the correlation between BVSP and MERVAL (see Figure 50).

For the other seventeen world markets nothing interestingly different from the correlation variation between 0.3 and 0.7 can be observed.

MERVAL

For this market we can observe, with the other five markets from North or Latin-America, that there is a time varying correlation: between 0.3 and 0.7, from 2000 to 2006; going up, between 0.5 and 0.8, from 2006 to 2009; going up, again, between 0.7

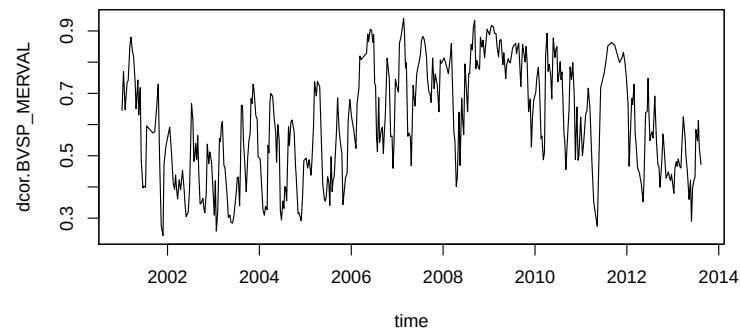


Figure 50: Distance Correlation for the BVSP_MERVAL pair

and 0.9, from 2009 to 2011; going down, quickly, from 2011 till now. As an example we show the correlation between MERVAL and MXX (Figure 51):

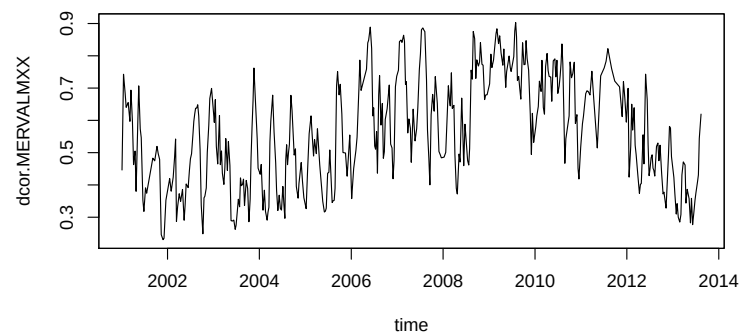


Figure 51: Distance Correlation for the MERVAL_MXX pair

For the European subset, there seems also to be a time varying correlation, although less intense, but similar to the one described above.

MXX

The observations are similar to those made for MERVAL market.

5.2.4.2 North American Markets

DJI

For this market, the correlation with PSI, MIB, IBEX is between 0.3 and 0.7 and a little bit higher with other European markets like SSMI, STOXX and FTSE (see Figure 52).

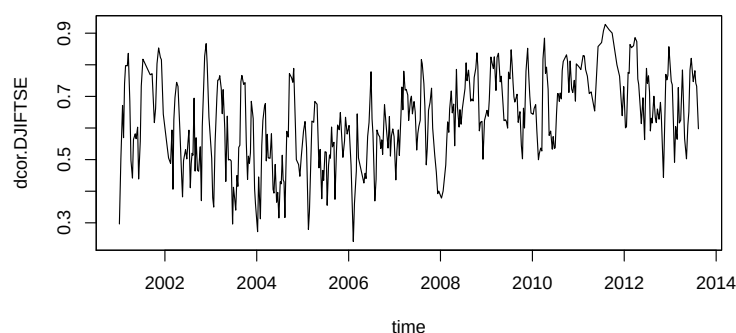


Figure 52: Distance Correlation for the DJI_FTSE pair

Apart from that, with the Latin American subset we can find correlation values similar to those found with that European ones. Finally, the correlation with the North American markets subset is very high. See, for example, Figure 53 about the correlation with IXIC.

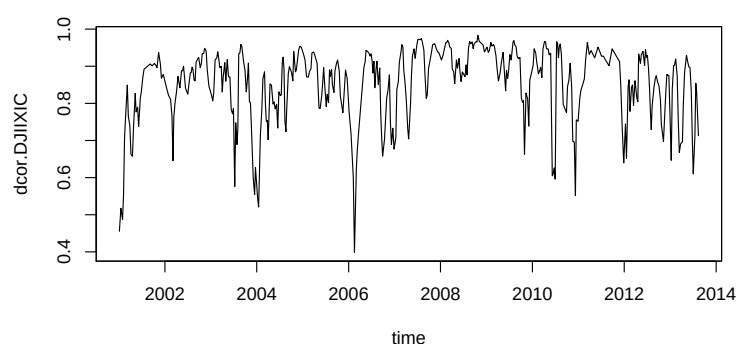


Figure 53: Distance Correlation for the DJI_IXIC pair

IXIC

For this market, about the correlation with the Latin American subset we can find a more varying correlation relationship than to the values found for the European ones (see Figure 54).

The correlation with the North American markets subset, as noted before, is very high.

SPY

For this market, about the correlation with the European subset we can find a varying correlation relationship: going down, between 0.4 and 0.8, from 2000 to 2005; going up,

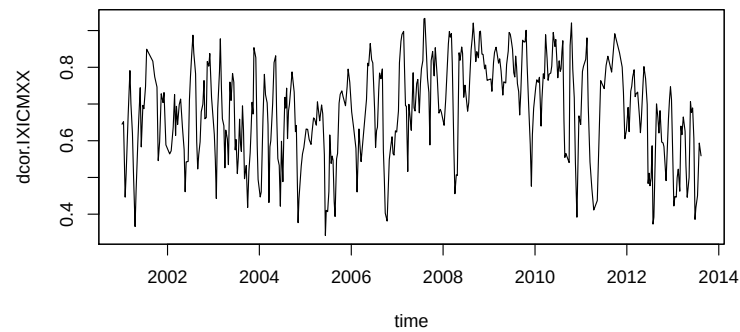


Figure 54: Distance Correlation for the IXIC_MXX pair

between 0.4 and 0.8, from 2005 to 2010; stable, between 0.6 and 0.8, from 2010 to 2012; going down from 2012 till now (Figure 55).

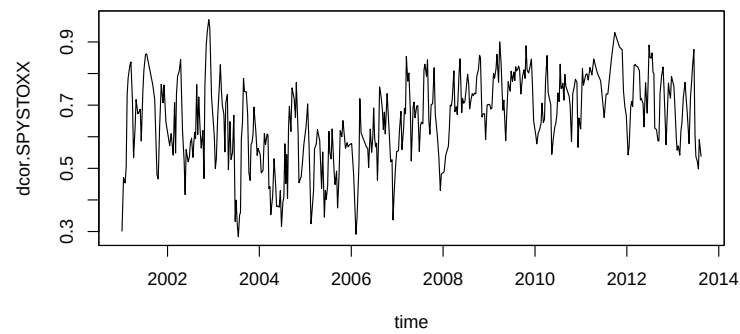


Figure 55: Distance Correlation for the SPY_STOXX pair

The correlation with the North American markets subset, as noted before, is very high.

5.2.5 Hurst Exponent

Let us now consider some Hurst exponent calculations for some world markets. We start analysing a subset of some European markets (see Figure 56). Their values are, typically, around 0.4 and 0.6 except for PSI-20 (that have Hurst exponents around 0.5 and 0.7 meaning that there is some persistence in this market behaviour).

The correlation coefficient $r(t)$ is also plotted for each point revealing the quality of the fit where the H exponent is evaluated; in all graphics the correlation coefficient is near 1. All correlation coefficients, $r(t)$, may be seen to fall in the range $0.95 - 1$, giving us confidence in the power law behaviour of $\langle F(t) \rangle$.

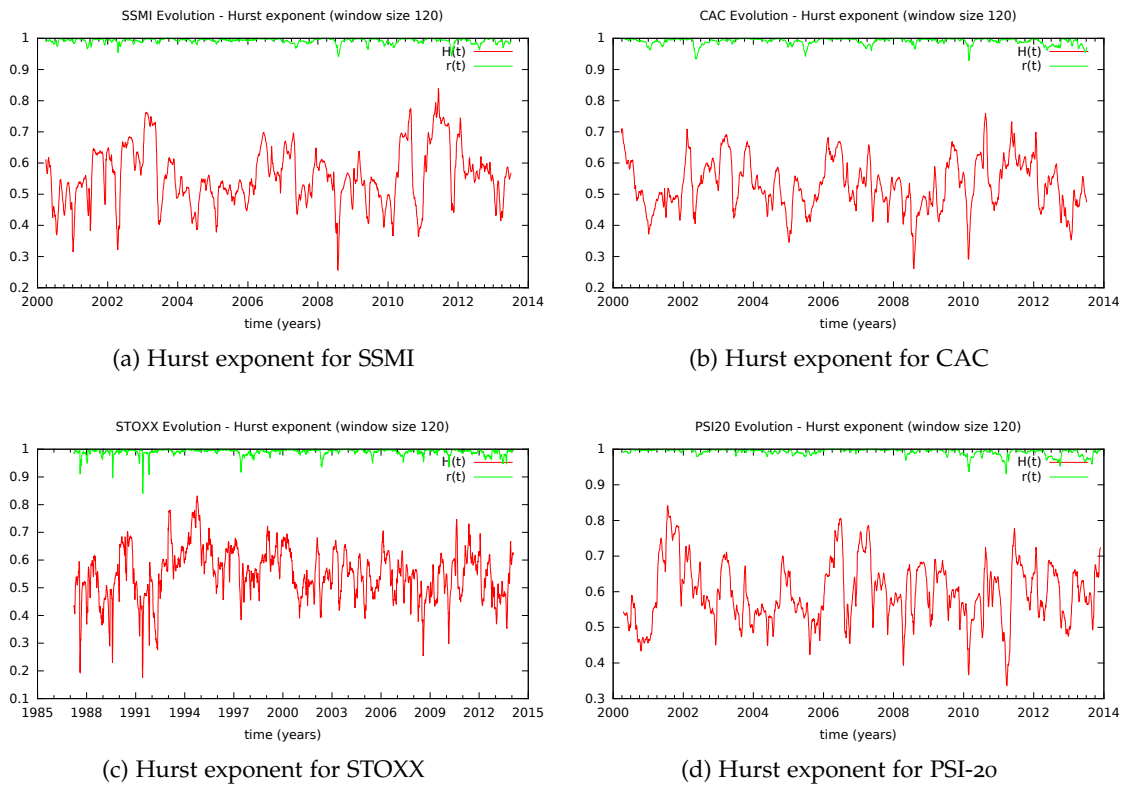


Figure 56: Hurst exponent for European markets

It should be noted, in what concerns PSI-20 (see Table 12), that despite having a Hurst exponent of 0.535 this market is having a very interesting evolution. In fact, a similar study in 2006 by Matos [2006], and using the same DFA method, estimated $H = 0.59$. It is clear that PSI-20 is going through a maturation process, that is, having less persistent behaviour and following less trends.

For a global Hurst exponent for the world markets we can view Table 12. It is noticeable that only 6 out of 23 markets have Hurst exponents, H , under 0.5, meaning that these 6 (CAC, DJI, FTSE, IBEX, SPY and SSMI) have anti-persistent behaviour and can be considered as mature markets. Looking at their geographical distribution we can count 4

European and the two North-American, which is not a surprise. Around $H = 0.5$ we find another 6 markets (AEX, ASX, DAX, MIB, NIK and STOXX), that is, 4 European more, the Japanese and the Australian. These markets can be also considered mature and random. Finally, all the others have $H > 0.5$.

Index	H	R	σ_H
^AEX	0.507	0.999	0.003
^ASX	0.509	1	0.002
^ATX	0.559	0.995	0.007
^BSESN	0.538	0.999	0.003
^BVSP	0.527	0.998	0.004
^CAC	0.46	0.999	0.003
^DAX	0.5	0.999	0.003
^DJI	0.462	0.999	0.003
^FTSE	0.452	0.999	0.003
^HSI	0.519	0.999	0.002
^IBEX	0.484	0.999	0.002
^IXIC	0.558	1	0.001
^JKSE	0.555	0.999	0.00302
^KOSPI	0.512	0.975	0.0121
^MERVAL	0.556	0.999	0.00309
^MIB	0.502	0.999	0.00341
^MXX	0.53	0.998	0.0053
^NIK	0.508	0.991	0.00839
^PSI20	0.535	0.997	0.00454
^SPY	0.476	0.992	0.00727
^SSMI	0.48	0.998	0.004
^STOXX	0.503	0.999	0.002
^STRAITS	0.526	0.998	0.005

Table 12: Hurst exponent for world markets

It should be noted, in what concerns PSI-20, that despite having a Hurst exponent of 0.535 this market is having a very interesting evolution, as we can see from a similar study in 2006 by Matos [2006], and using the same DFA method, estimated $H = 0.59$. It is clear that PSI-20 is going through a maturation process, that is, having less persistent behaviour and following less trends.

5.3 CONCLUDING REMARKS

In this chapter we have applied several Econophysics tools to the study of the World Markets set. First of all, some results found in literature are confirmed, namely the ones from random matrix theory and the ones for Hurst exponent. In this case, and based in previous results, we can go further and propose that all the world markets are becoming more mature, that is to say that they are becoming more transparent. It is noticeable when comparing with the results obtained eight years ago [Matos, 2006].

For Mutual Information or Kullback-Leibler Divergence the results are very sharp and a event related comparison was applied to find out the coincidences. This analysis has shown that we can match the more interesting values calculated with real events. Indeed, there are certain events that are clearly reflected in all markets, as expected since most events are due to external causes, and thus independent of the specific market.

The results from energy statistics are not so well defined as with PSI-20 stocks in Chapter 4. Despite that, we can find strong regional correlation for most of the markets and some, but a few, more global influence markets. There is, also, a strong connection between the North-American markets and most of the European ones. Also, it is possible to suggest that the Distance Correlation values tend to diminish after the most important event take place.

As a general conclusion we can say with enough confidence that the Distance Correlation has become higher since 2007, clearly showing that the world markets are in the way to act as one.

Distance Correlation results are not complemented here with Approximate Entropy like it was in Chapter 4. This measure, $ApEn$, peaks in periods of crisis, becoming agitated and with higher variations.

In general, a trend common to most markets is the progressive correlation over time for most of the studied markets. One possible reason to this is the progressive globalisation of markets, where the arbitrage opportunities are reduced thus producing more efficient markets.

CONCLUSIONS AND FUTURE WORK

"Prediction is very difficult, especially about the future" - Niels Bohr

"It's too early to tell", Zhou Enlai, Chinese premiere in the 1960s, about the impact of the French revolution

In this chapter all the results obtained in Chapter 4 and in Chapter 5 are merged and put into perspective in order to compose a coherent line of conclusions.

6.1 CONCLUSIONS

In this work we have addressed the analysis of financial time series from an econophysical point of view.

Financial data presents complex behaviour which needs to be decomposed effectively, that is, the breakdown of financial signals into component elements, in order to determine the nature of the fluctuations observed. This was done using a number of techniques:

- random matrix theory like the Correlation matrix;
- component analysis like the Forecastable Component Analysis;
- entropy measures like the Mutual Information, the Kullback-Leibler divergence and the Approximate entropy;
- energy statistics like the Distance Correlation;
- fractional Brownian motion like the Hurst exponent.

These techniques are twofold: measures of “disorder”/complexity and measures of coherence. We found that these techniques are in a sense complementary, that is, each provides a different view over the financial data studied, but they can be placed under the umbrella of Econophysics measures.

If entropy is disorder, implying lack of a common trading strategy, then coherence implies cooperative, or at least common tendencies in behaviour. We use the Correlation matrix as a measure of coherence among a closely related set of stocks or markets. Coherence can be either observed between each financial time series, like in Forecastable Component Analysis, Approximate entropy or Hurst exponent, or between different financial time series like in Mutual Information, Kullback-Leibler divergence, Distance Correlation or Correlation matrix.

Also, there were studied and used “sliding windows” of different sizes. The motivation and importance of this kind of analysis is the well known multi-fractal behaviour that financial data exhibits (see [Lux \[2004\]](#)). This was reflected in the output for 20, 60 and 120 trading days windows, that is, sensibly 1, 3 and 6 trading days (in months). A natural extension of this analysis is to consider other window sizes.

The first application of the techniques was to a set of 12 stocks from the PSI-20, the Portuguese index of the 20 most liquid assets of the Portuguese Stock market. PSI-20 index main characteristics are described in Appendix A. The Portuguese case is chosen both for: a) regional relevance; b) relatively little previous study and c) its relevance as a showcase both as an emerging young/mature market and its relevance to discuss features on the techniques presented.

The global results are presented in Chapter 4 and Chapter 5. We started by confirming some results found in literature, namely the ones from random matrix theory and the ones for the Hurst exponent. In this case, and based in previous results, we can go further and propose that the PSI-20 is becoming more mature. Indeed, it is noticeable when comparing the results for three and eight years ago (Matos et al. [2004], Matos et al. [2006] and Gomes [2012]).

It is safe to propose that an increasing number of markets achieving or mimicking mature behaviour relatively rapidly, irrespectively of their trading capability, which suggests that windows of opportunity are narrowing for investors since the arbitrage opportunities are reduced due to more efficient markets.

To our knowledge, it is the first time that energy statistics is applied to the PSI-20 data. It is interesting to note that this measure, and this is corroborated by Approximate entropy results, proposes two well defined behaviour for the PSI-20 stocks. One period, from 2000 to 2007, relatively calm, with low variation of Distance Correlation between stocks, and another period, from 2007 till now, much more agitated in what concerns this measure.

In Chapter 5 we have applied the above Econophysics tools to the study of the World Markets set. In this Chapter, we confirm some results found in literature, namely the ones from random matrix theory and the ones for Hurst exponent. In this case, and based in previous results, we can go further and propose that all the world markets are becoming more mature, that is to say that they are becoming more transparent. Indeed, it is noticeable when comparing with the results obtained in a previous study [Matos, 2006].

For Mutual Information or Kullback-Leibler Divergence the conclusions are similar to the ones obtained from PSI-20 stocks analysis. Indeed, there are certain events that are clearly reflected in all markets, as expected since most events are due to external causes, and thus independent of the specific market.

One event where this is clearly seen is the 9/11 (September 11th, 2001) attack against the World Trade Centre towers in Manhattan, NY, corresponding to the first XXI century recession. In all the markets this is clearly seen, both in markets present here and in Appendix B, where the same type of analysis reveals the same dominant stripe appearing around September 2001 and around 2008 when the second recession of XXI century happened.

It is, also, interesting to note that the results from energy statistics are not so well defined as with PSI-20 stocks. Despite that, we can find strong regional correlation for most of the markets and some, but a few, more global influence markets. There is, also, a strong connection between the North-American markets and most of the European ones. That correlation became higher since 2007.

Distance Correlation proposal is not complemented here with Approximate Entropy like it was for the PSI-20 stocks, which is somewhat disappointing because the pattern for stocks was very well defined.

In general, a trend common to most markets is the progressive correlation over time for most of the studied markets. One possible reason to this is the progressive globalisation of markets, where the arbitrage opportunities are reduced due to more efficient markets. Also, the information we got from Hurst exponent was vital to confirm that stocks and markets are getting more and more mature, that is, less autocorrelated. Would Bachelier liked this?

A good overall conclusion must include the understanding that we can not discard none of these methods. All of them show merits and the complementarity between them is an objective to pursue. Distance correlation have shown to be a good complement to entropy measures like Mutual Information or Kullback-Leibler divergence. Approximate entropy, as a stand alone method, have shown potential complementarity with Distance correlation.

The recession periods and in a comparative view, the chosen non-recession periods, have shown that these Econophysics tools behave quite differently in recession and non-recession times. This is a quite hopeful sign for the times to come.

6.2 FUTURE WORK

This work opened some new “windows” in the horizon, namely, to other variants of the techniques presented in this work that were not fully explored but have shown potential for further studies. These new “windows” are discriminated next.

1. The scale dependency can be further extended into comparing the detail levels. Instead of the whole time series, we must use the time dependent covariance matrix.
2. When studying the covariance matrix and its most significant eigenvalues, we could study the evolution of eigenvectors. This type of analysis should be useful to pick sudden jumps when the main eigenvectors changes suddenly, instead of smooth time dependency.
3. New libraries are needed for Mutual Information or Kullback-Leibler divergence calculation. Two good starting points are the R libraries “infotheo” and “FNN”.
4. Forecastable Component Analysis deserves a more profound study, that was not possible in this work.
5. Approximate entropy peaks in periods of crisis, becoming agitated and with higher variations. For the World markets set a closer look is a work in progress.
6. Finally, we have studied and used “sliding windows” of different sizes. The motivation and importance of this kind of analysis is the well known multi-fractal behaviour that financial data exhibits [Calvet and Fisher, 2002]. A natural extension to this question is to consider other window and step sizes.

DATA

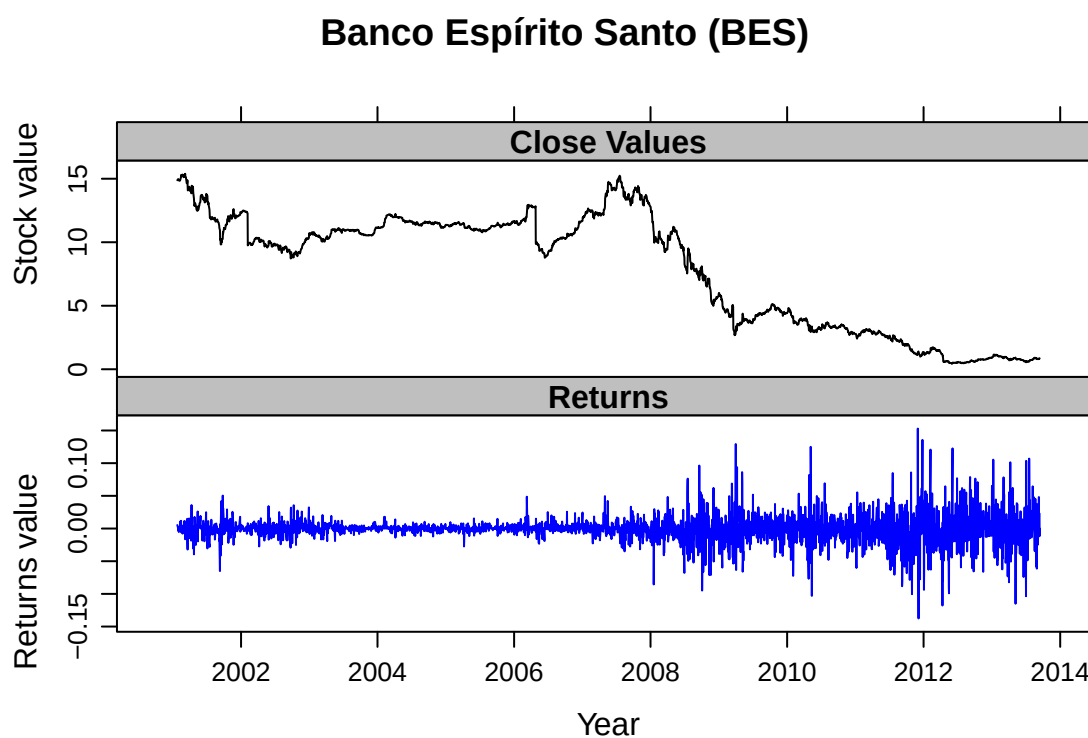
In this Appendix we visualise and present for each stock or market studied:

- Country and name of the index
- Historical index values.
- Historical return values.
- Statistical information: Observations, Minimum and Maximum, measures of central tendency like Arithmetic Mean, Geometric Mean, Median and Quartiles, Confidence Interval (95%), dispersion measures like variance and Standard Deviation, and Skewness and Kurtosis.

As previously described, all analyses deal with returns, as e.g. prices can be problematic due to currency exchanges. For each stock or market, therefore, we illustrate the original time series and the returns. The same scale is used for all plots to place comparisons in a context where they can be understood.

A.1 PSI-20 STOCKS

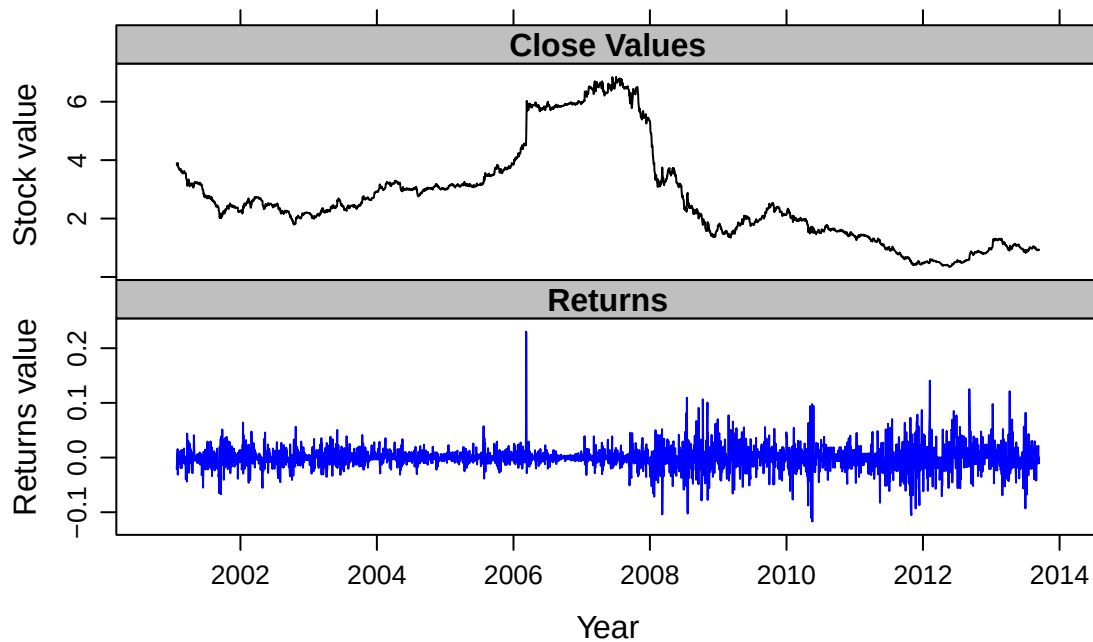
BES



```
table.Stats(BES returns, ci=0.95, digits=8)
```

	NA
Observations	3218.00000000
NAs	0.00000000
Minimum	-0.55961579
Quartile 1	-0.00659163
Median	0.00000000
Arithmetic Mean	-0.00093816
Geometric Mean	-0.00129075
Quartile 3	0.00548269
Maximum	0.15290767
SE Mean	0.00043587
LCL Mean (0.95)	-0.00179277
UCL Mean (0.95)	-0.00008355
Variance	0.00061136
Stdev	0.02472571
Skewness	-5.52336083
Kurtosis	115.05597353

BPI

Banco Português de Investimento (BPI)

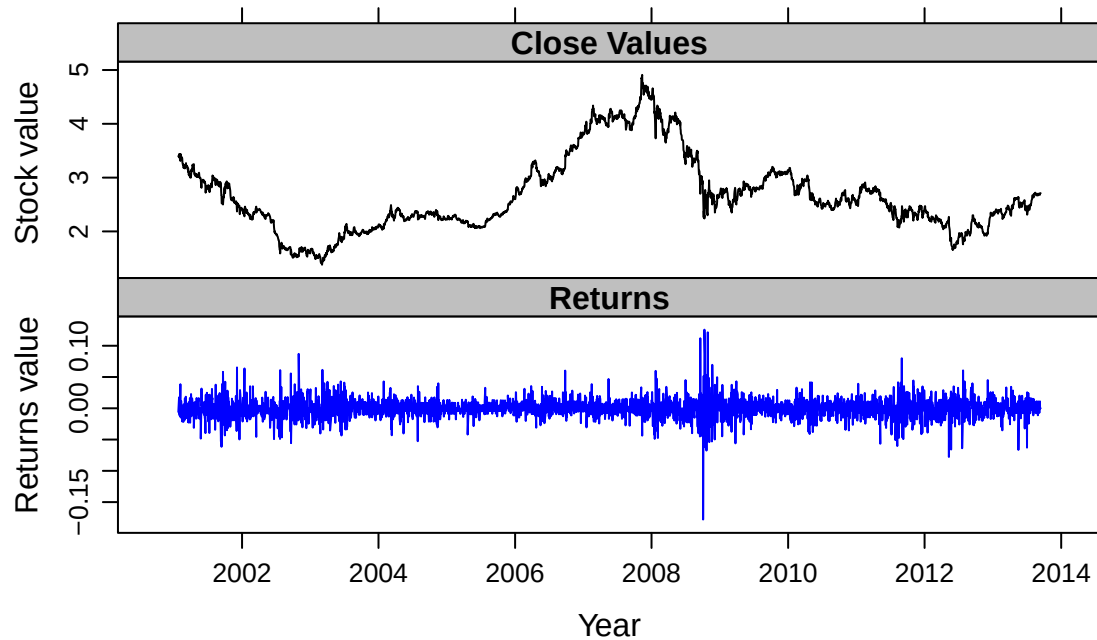
```
table.Stats(BPI returns, ci=0.95, digits=8)
```

```

NA
Observations    3218.00000000
NAs              0.00000000
Minimum         -0.11705656
Quartile 1      -0.00972062
Median          0.00000000
Arithmetic Mean -0.00044468
Geometric Mean  -0.00067470
Quartile 3      0.00840047
Maximum         0.23021660
SE Mean         0.00037934
LCL Mean (0.95) -0.00118844
UCL Mean (0.95) 0.00029908
Variance        0.00046306
Stdev           0.02151874
Skewness        0.63221621
Kurtosis        8.68241189
```

EDP

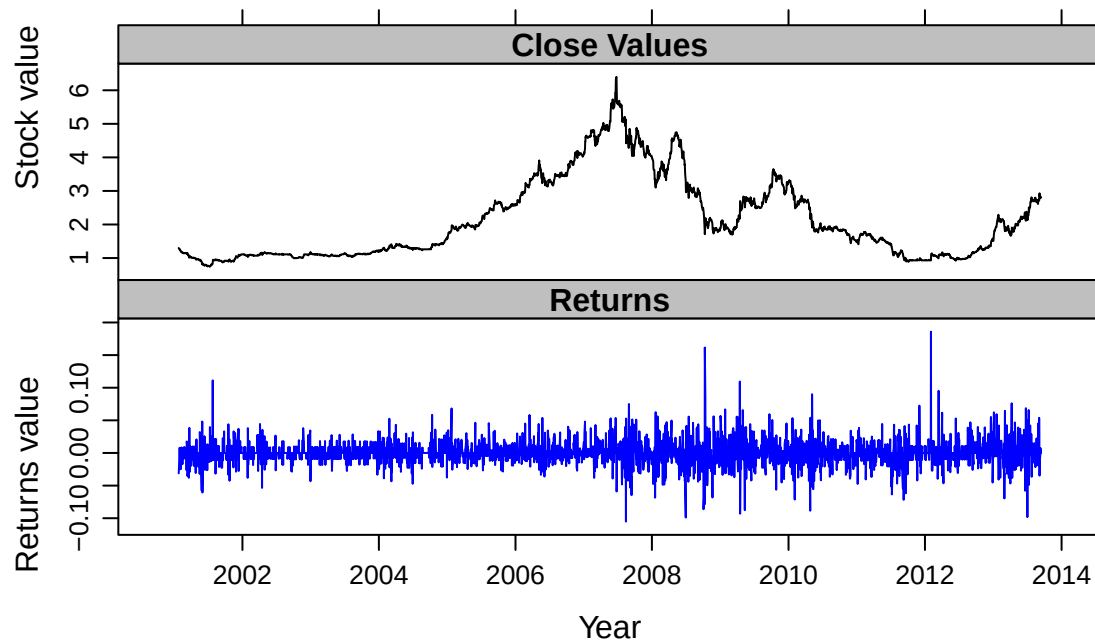
Energias de Portugal (EDP)



```
table.Stats(EDP returns, ci=0.95, digits=8)
```

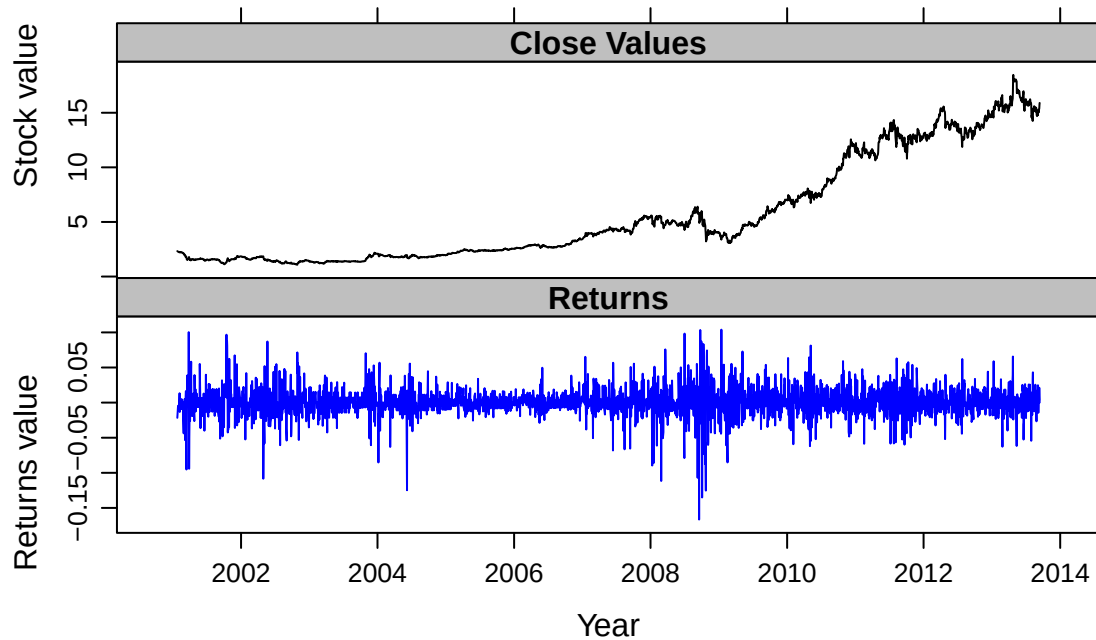
	NA
Observations	3218.00000000
NAs	0.00000000
Minimum	-0.17788696
Quartile 1	-0.00840047
Median	0.00000000
Arithmetic Mean	-0.00007049
Geometric Mean	-0.00020413
Quartile 3	0.00841225
Maximum	0.12568822
SE Mean	0.00028786
LCL Mean (0.95)	-0.00063490
UCL Mean (0.95)	0.00049393
Variance	0.00026666
Stdev	0.01632977
Skewness	-0.09063438
Kurtosis	8.95731757

EGL

Mota Engil (EGL)

```
table.Stats(EGL returns, ci=0.95, digits=8)
```

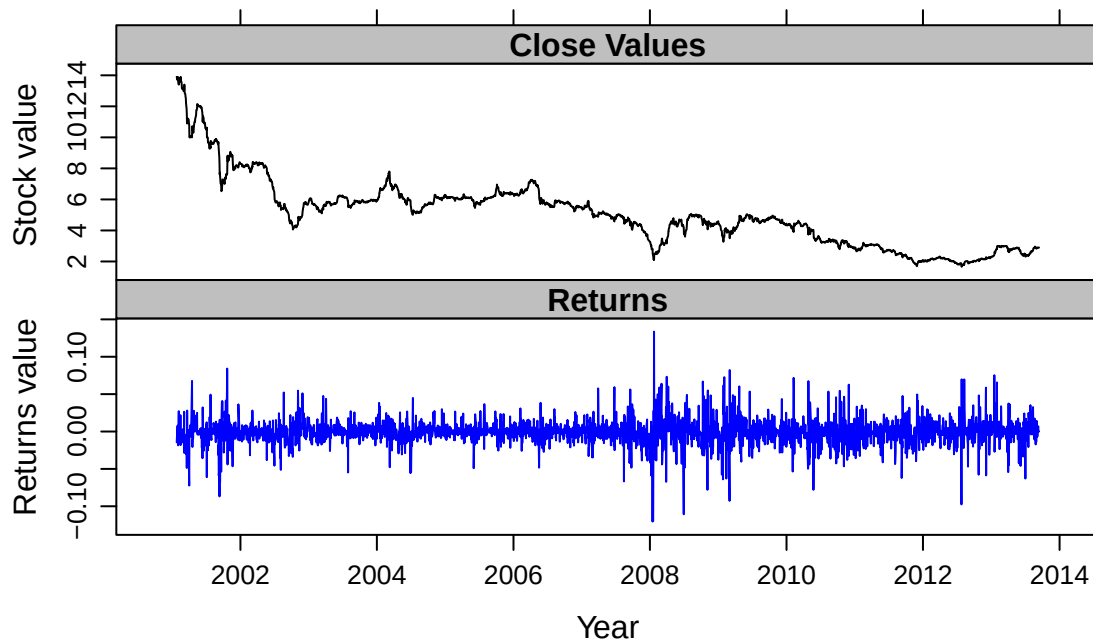
	NA
Observations	3218.00000000
NAs	0.00000000
Minimum	-0.10500331
Quartile 1	-0.00828173
Median	0.00000000
Arithmetic Mean	0.00016655
Geometric Mean	-0.00002486
Quartile 3	0.00843887
Maximum	0.18392284
SE Mean	0.00034573
LCL Mean (0.95)	-0.00051131
UCL Mean (0.95)	0.00084442
Variance	0.00038464
Stdev	0.01961214
Skewness	0.46309715
Kurtosis	7.34153549

*JMT***Jerónimo Martins (JMT)**

```
table.Stats(JMT returns, ci=0.95, digits=8)
```

	NA
Observations	3218.00000000
NAs	0.00000000
Minimum	-0.16658398
Quartile 1	-0.00816331
Median	0.00000000
Arithmetic Mean	0.00059678
Geometric Mean	0.00039113
Quartile 3	0.00904984
Maximum	0.10388013
SE Mean	0.00035638
LCL Mean (0.95)	-0.00010197
UCL Mean (0.95)	0.00129554
Variance	0.00040870
Stdev	0.02021644
Skewness	-0.39569875
Kurtosis	6.57536026

NBA

Novabase (NBA)

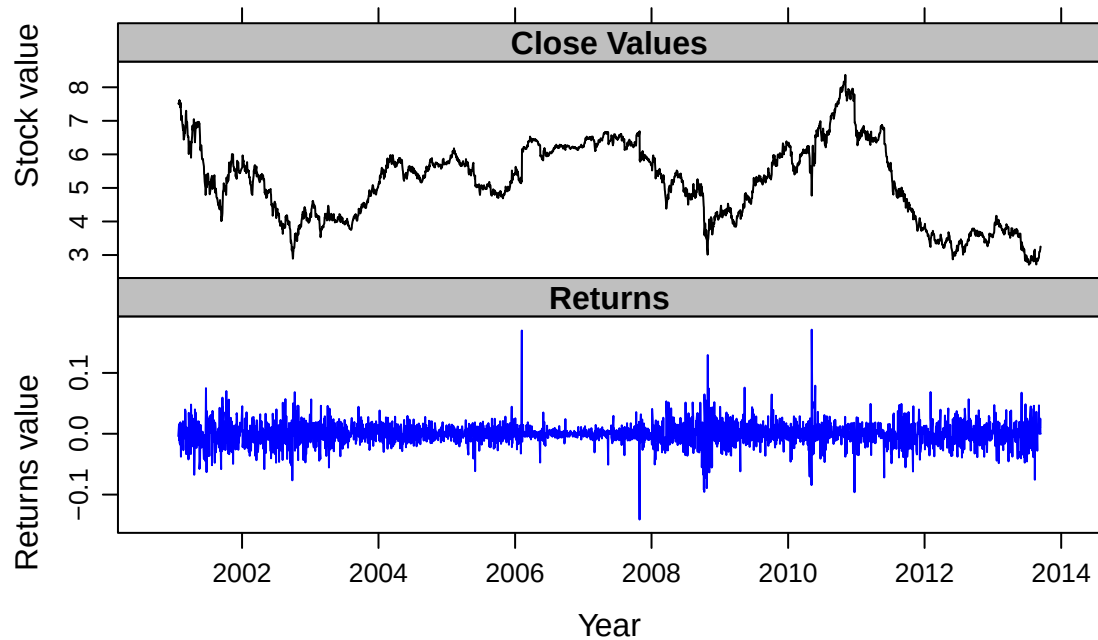
```
table.Stats(NBA returns, ci=0.95, digits=8)
```

```

NA
Observations    3218.00000000
NAs              0.00000000
Minimum         -0.12044615
Quartile 1      -0.00702991
Median          0.00000000
Arithmetic Mean -0.00048700
Geometric Mean  -0.00062420
Quartile 3      0.00613030
Maximum         0.13353139
SE Mean         0.00029160
LCL Mean (0.95) -0.00105874
UCL Mean (0.95) 0.00008473
Variance        0.00027363
Stdev           0.01654163
Skewness        -0.11074895
Kurtosis        7.54054925
```

PTC

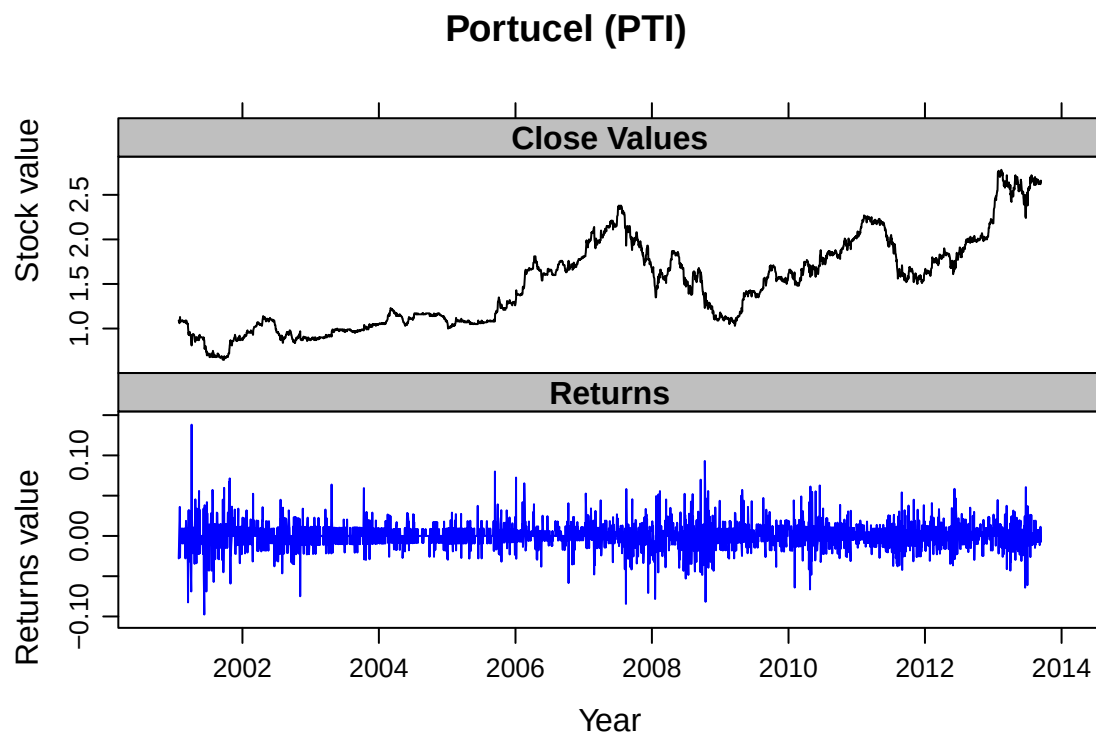
Portugal Telecom (PTC)



```
table.Stats(PTC returns, ci=0.95, digits=8)
```

	NA
Observations	3218.00000000
NAs	0.00000000
Minimum	-0.14047445
Quartile 1	-0.00900231
Median	0.00000000
Arithmetic Mean	-0.00040201
Geometric Mean	-0.00057878
Quartile 3	0.00860485
Maximum	0.17120027
SE Mean	0.00033095
LCL Mean (0.95)	-0.00105091
UCL Mean (0.95)	0.00024689
Variance	0.00035247
Stdev	0.01877419
Skewness	-0.06548821
Kurtosis	9.74735535

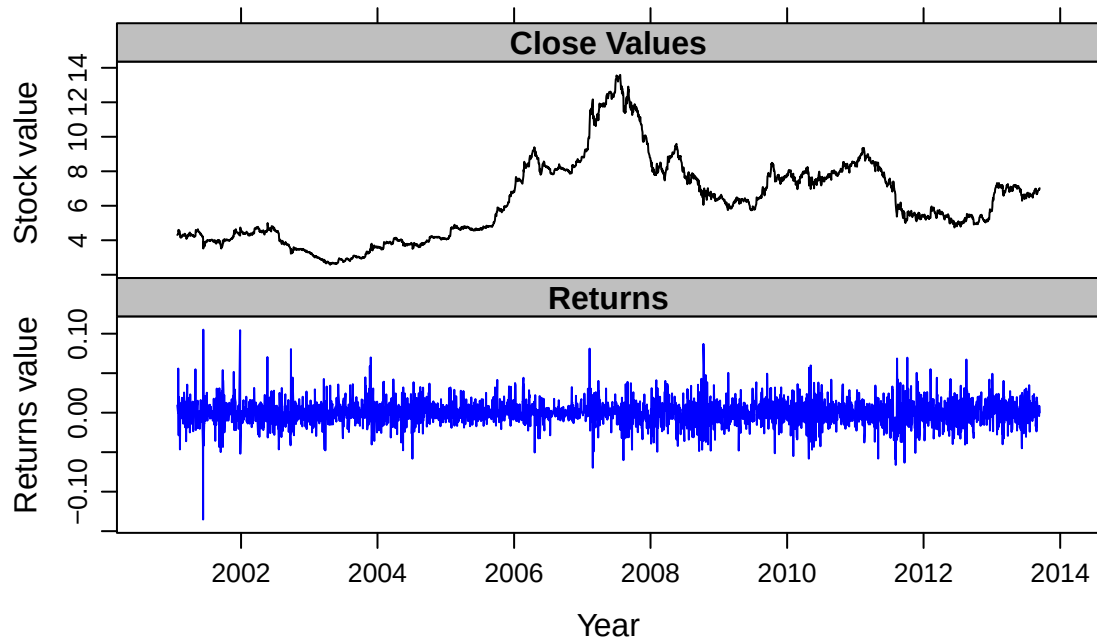
PTI



```
table.Stats(PTI returns, ci=0.95, digits=8)
```

	NA
Observations	3218.00000000
NAs	0.00000000
Minimum	-0.09389609
Quartile 1	-0.00734624
Median	0.00000000
Arithmetic Mean	0.00018621
Geometric Mean	0.00005986
Quartile 3	0.00751883
Maximum	0.13005313
SE Mean	0.00028024
LCL Mean (0.95)	-0.00036326
UCL Mean (0.95)	0.00073567
Variance	0.00025272
Stdev	0.01589727
Skewness	0.06148675
Kurtosis	5.58028443

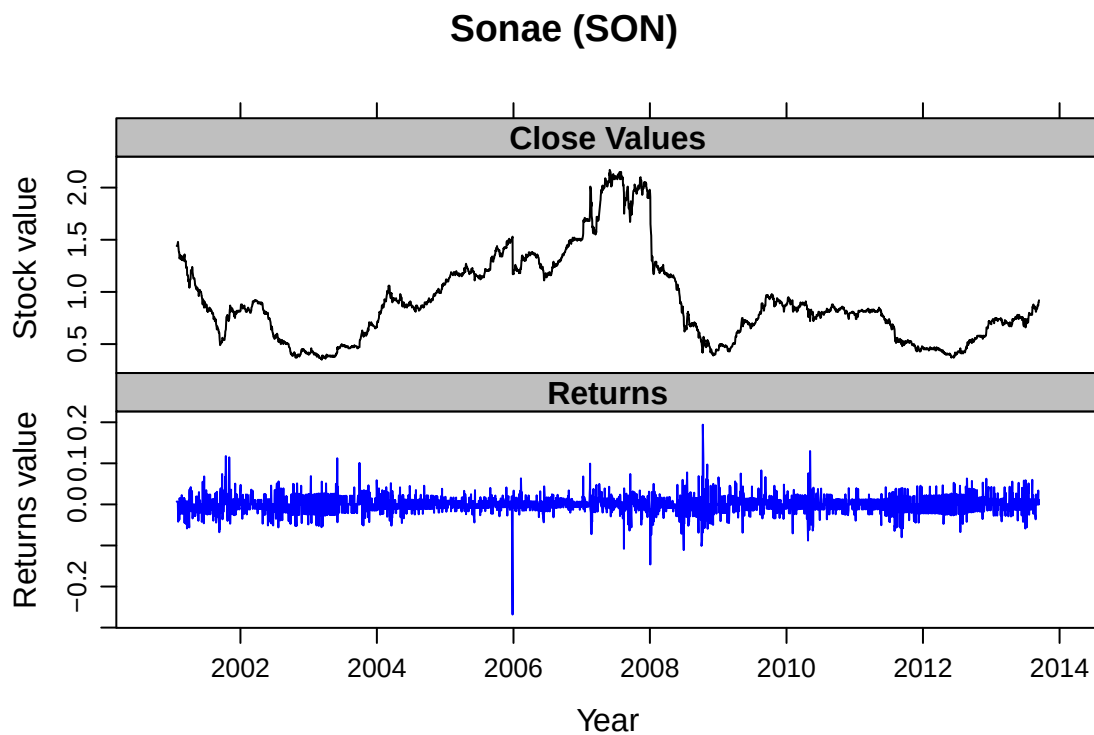
SEM

Semapa (SEM)

```
table.Stats(SEM returns, ci=0.95, digits=8)
```

	NA
Observations	3218.00000000
NAs	0.00000000
Minimum	-0.13530539
Quartile 1	-0.00804186
Median	0.00000000
Arithmetic Mean	0.00015159
Geometric Mean	0.00002307
Quartile 3	0.00814590
Maximum	0.10507638
SE Mean	0.00028277
LCL Mean (0.95)	-0.00040283
UCL Mean (0.95)	0.00070602
Variance	0.00025730
Stdev	0.01604068
Skewness	0.14014506
Kurtosis	4.69520935

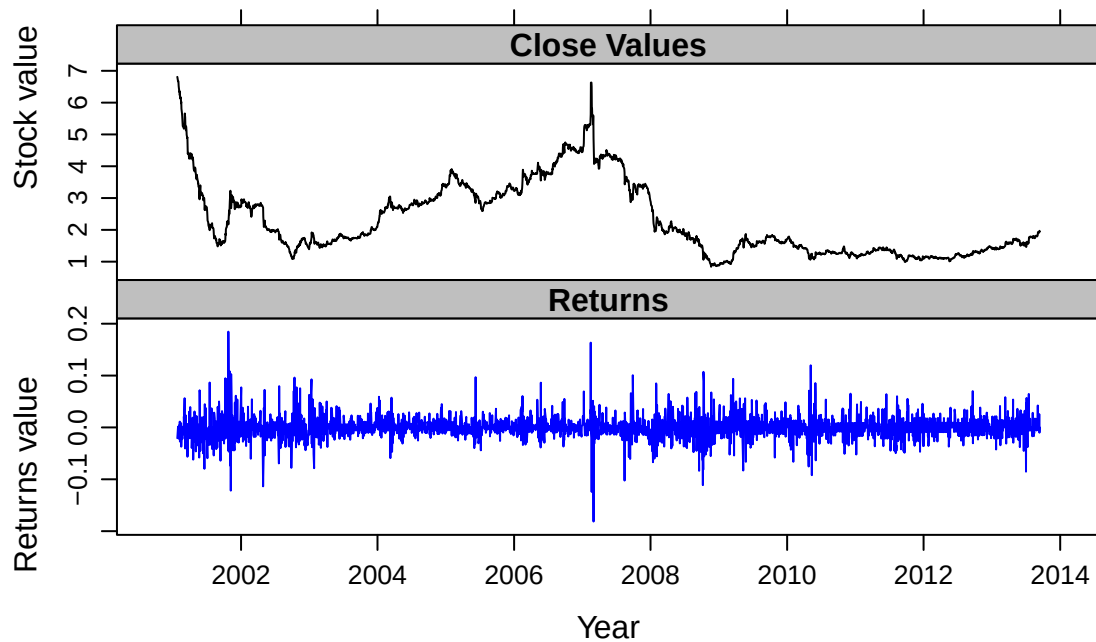
SON



```
table.Stats(SON returns, ci=0.95, digits=8)
```

	NA
Observations	3218.00000000
NAs	0.00000000
Minimum	-0.26826399
Quartile 1	-0.01169604
Median	0.00000000
Arithmetic Mean	-0.00013922
Geometric Mean	-0.00038495
Quartile 3	0.01156082
Maximum	0.19415601
SE Mean	0.00038936
LCL Mean (0.95)	-0.00090264
UCL Mean (0.95)	0.00062419
Variance	0.00048785
Stdev	0.02208731
Skewness	-0.25428937
Kurtosis	11.57492392

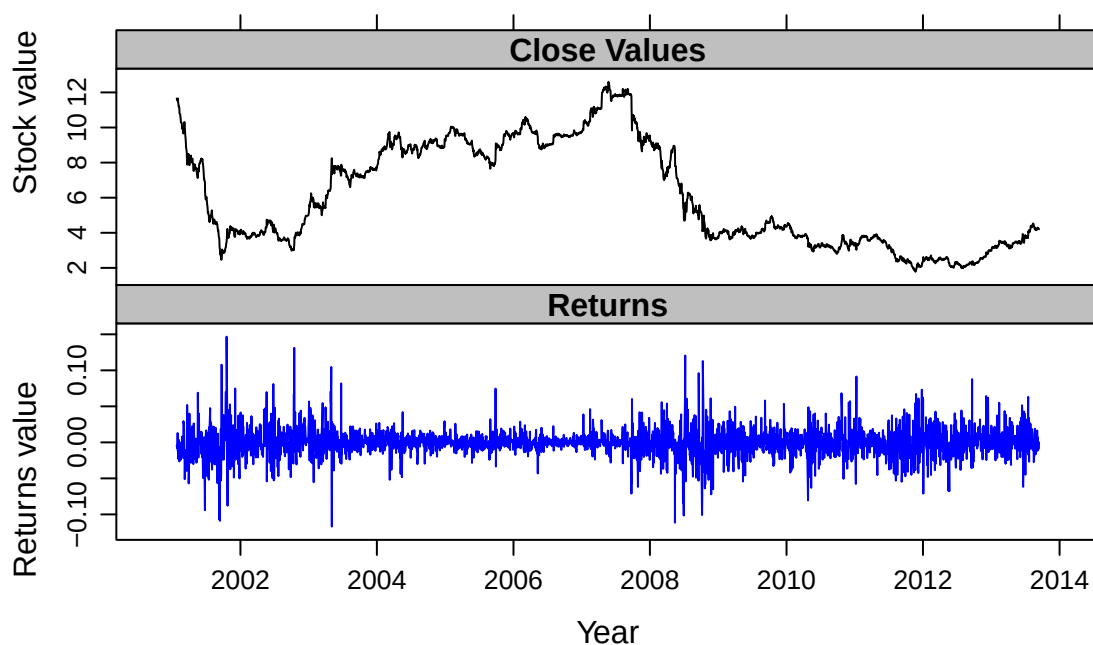
SONC

Sonae Com (SONC)

```
table.Stats(SONC returns, ci=0.95, digits=8)
```

	NA
Observations	3218.00000000
NAs	0.00000000
Minimum	-0.18015000
Quartile 1	-0.01010110
Median	0.00000000
Arithmetic Mean	-0.00042678
Geometric Mean	-0.00067183
Quartile 3	0.00816331
Maximum	0.18571715
SE Mean	0.00039073
LCL Mean (0.95)	-0.00119289
UCL Mean (0.95)	0.00033933
Variance	0.00049130
Stdev	0.02216523
Skewness	0.34516349
Kurtosis	7.86558480

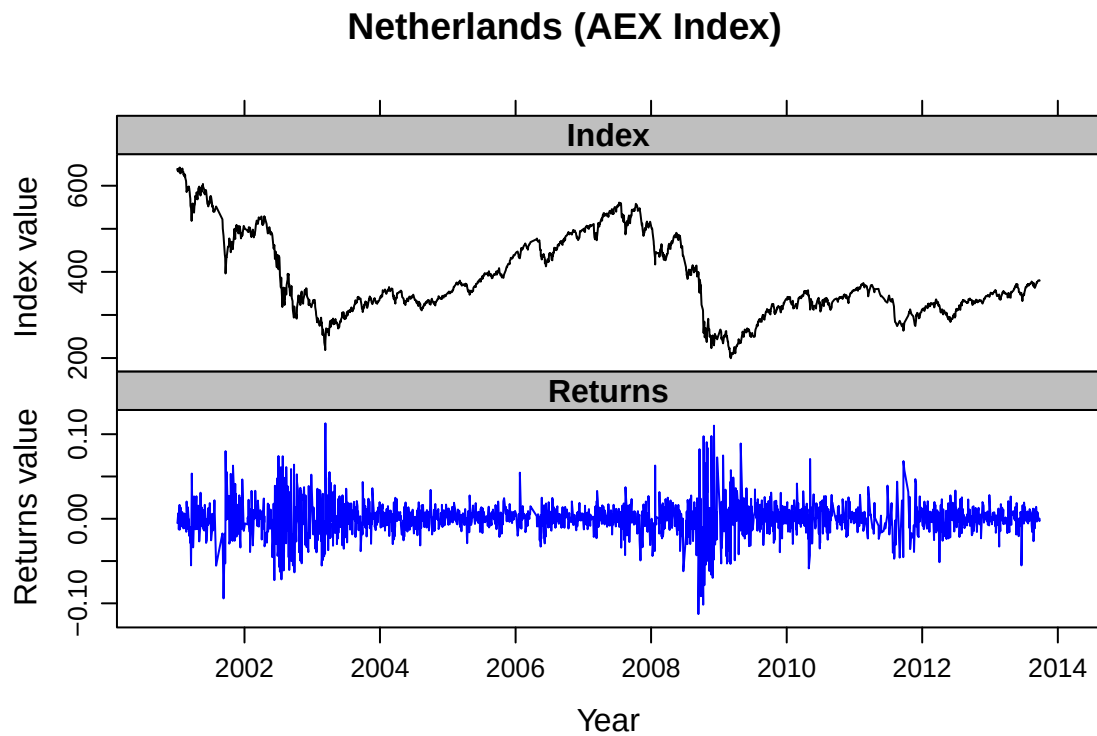
ZON

Zon Multimédia (ZON)

```
table.Stats(ZON returns, ci=0.95, digits=8)
```

	NA
Observations	3218.00000000
NAs	0.00000000
Minimum	-0.11687436
Quartile 1	-0.00847463
Median	0.00000000
Arithmetic Mean	-0.00031704
Geometric Mean	-0.00051066
Quartile 3	0.00809721
Maximum	0.14673408
SE Mean	0.00034725
LCL Mean (0.95)	-0.00099789
UCL Mean (0.95)	0.00036382
Variance	0.00038804
Stdev	0.01969870
Skewness	0.28515419
Kurtosis	6.49151035

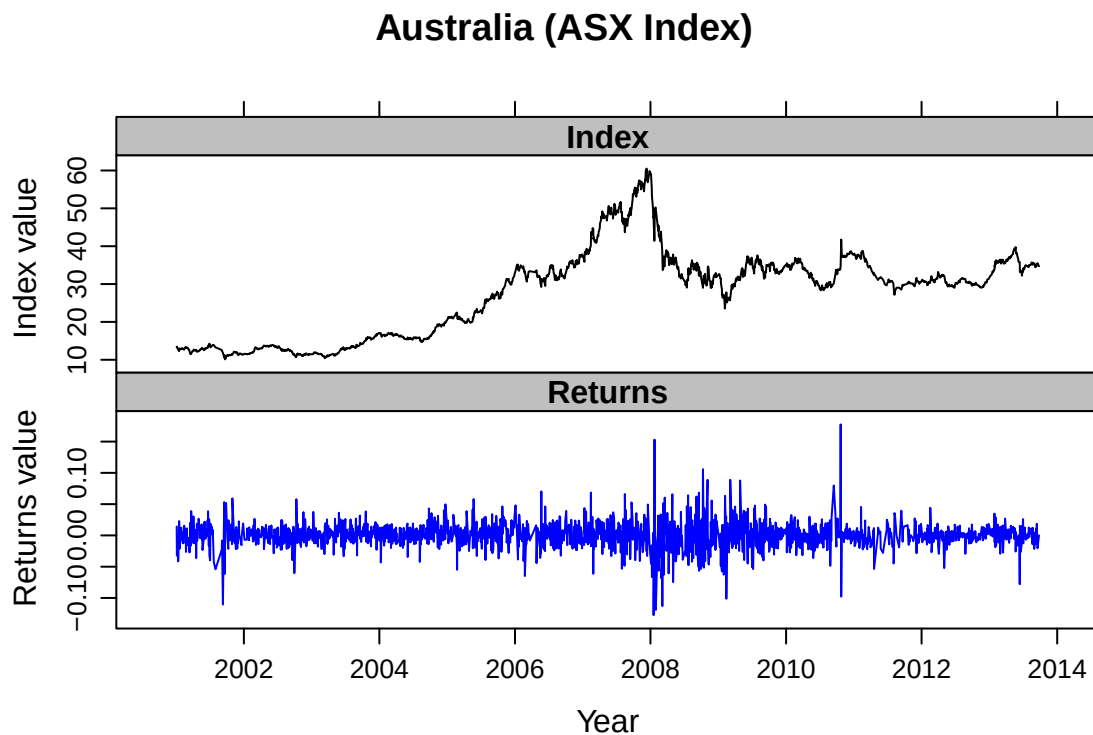
A.2 MARKETS

AEX

```
> table.Stats(AEX returns, ci=0.95, digits=4)
```

	NA
Observations	2024.0000
NAs	0.0000
Minimum	-0.1127
Quartile 1	-0.0086
Median	0.0002
Arithmetic Mean	-0.0003
Geometric Mean	-0.0004
Quartile 3	0.0084
Maximum	0.1129
SE Mean	0.0004
LCL Mean (0.95)	-0.0011
UCL Mean (0.95)	0.0006
Variance	0.0004
Stdev	0.0196
Skewness	0.1986
Kurtosis	5.5145

ASX

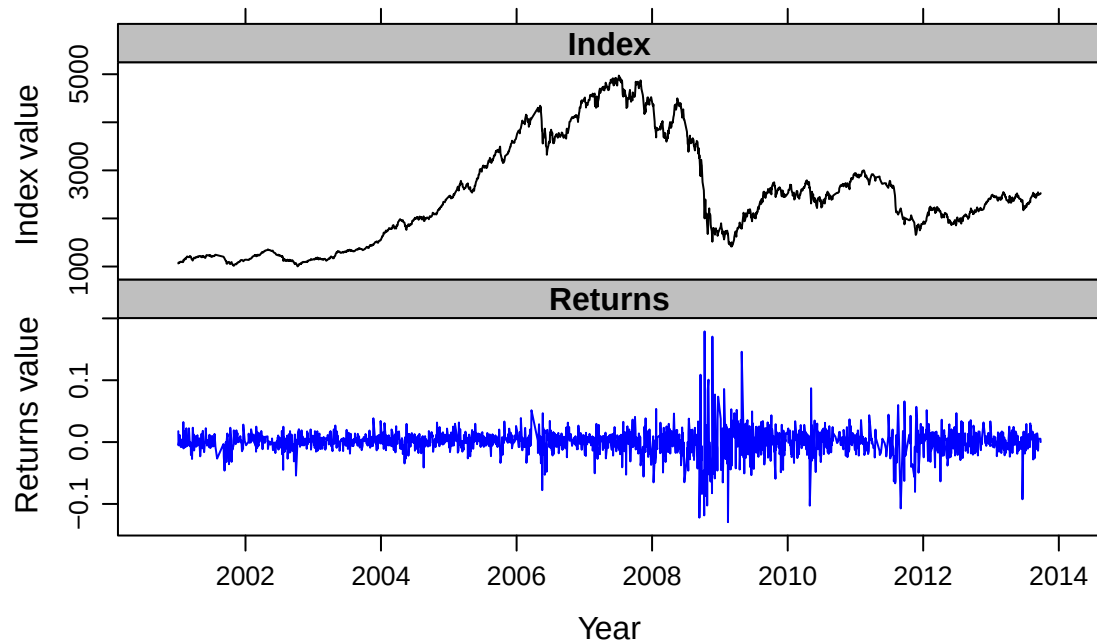


```
> table.Stats(ASX returns, ci=0.95, digits=4)
```

	NA
Observations	2024.0000
NAs	0.0000
Minimum	-0.1275
Quartile 1	-0.0086
Median	0.0003
Arithmetic Mean	0.0005
Geometric Mean	0.0003
Quartile 3	0.0099
Maximum	0.1775
SE Mean	0.0004
LCL Mean (0.95)	-0.0004
UCL Mean (0.95)	0.0013
Variance	0.0004
Stdev	0.0200
Skewness	0.0843
Kurtosis	9.3220

ATX

Austria (ATX Index)

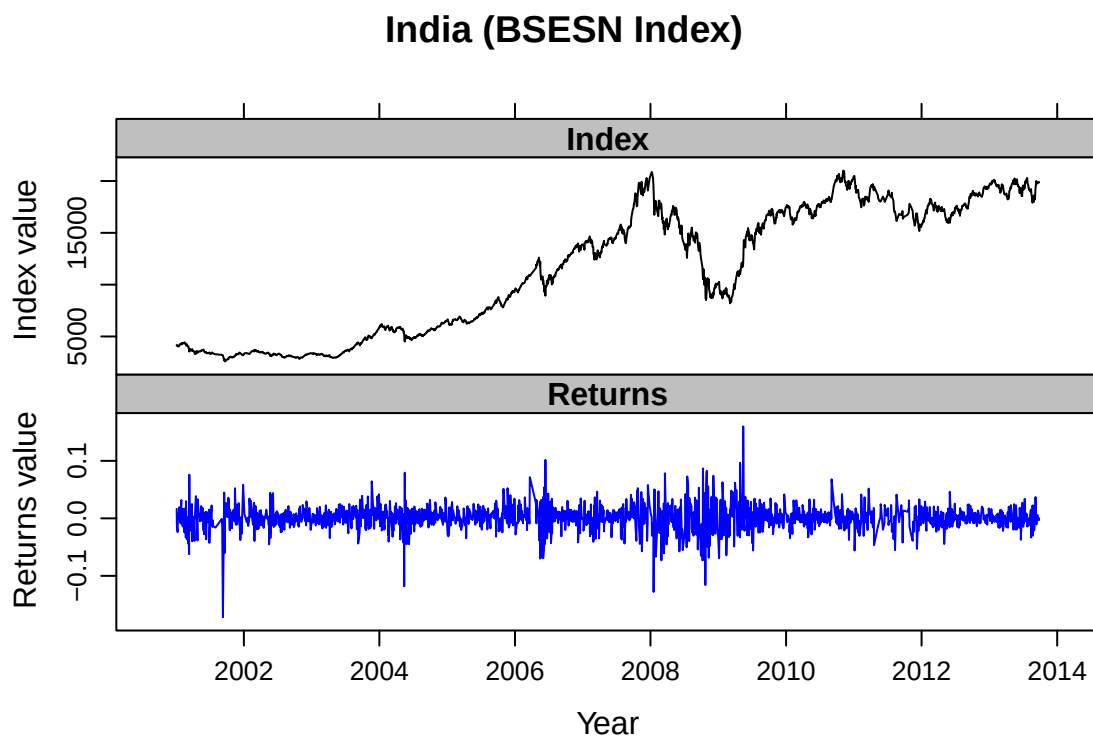


```
> table.Stats(ATX returns, ci=0.95, digits=4)
```

```

NA
Observations    2024.0000
NAs              0.0000
Minimum         -0.1294
Quartile 1      -0.0072
Median          0.0011
Arithmetic Mean  0.0004
Geometric Mean  0.0002
Quartile 3      0.0092
Maximum         0.1789
SE Mean         0.0004
LCL Mean (0.95) -0.0004
UCL Mean (0.95)  0.0013
Variance        0.0004
Stdev           0.0198
Skewness         0.2031
Kurtosis        13.1087
```

BSESN

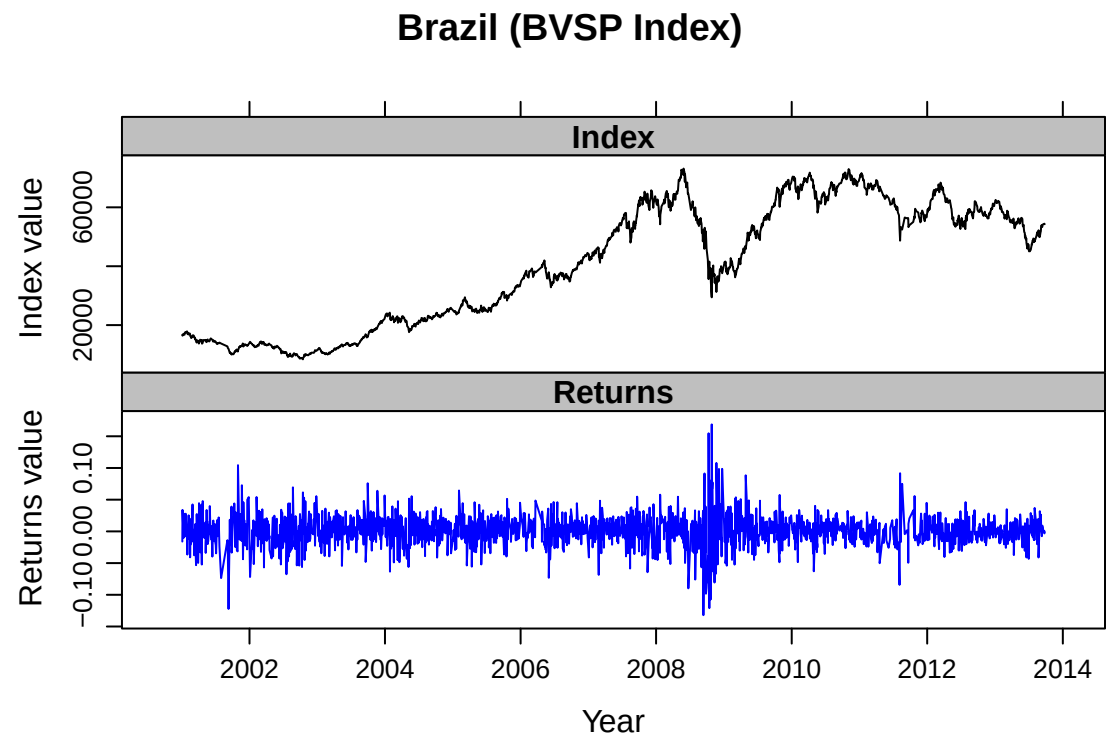


```
table.Stats(BSESN returns, ci=0.95, digits=4)
```

NA

Observations	2024.0000
NAs	0.0000
Minimum	-0.1718
Quartile 1	-0.0084
Median	0.0012
Arithmetic Mean	0.0008
Geometric Mean	0.0006
Quartile 3	0.0103
Maximum	0.1599
SE Mean	0.0005
LCL Mean (0.95)	-0.0001
UCL Mean (0.95)	0.0017
Variance	0.0004
Stdev	0.0203
Skewness	-0.2492
Kurtosis	8.0857

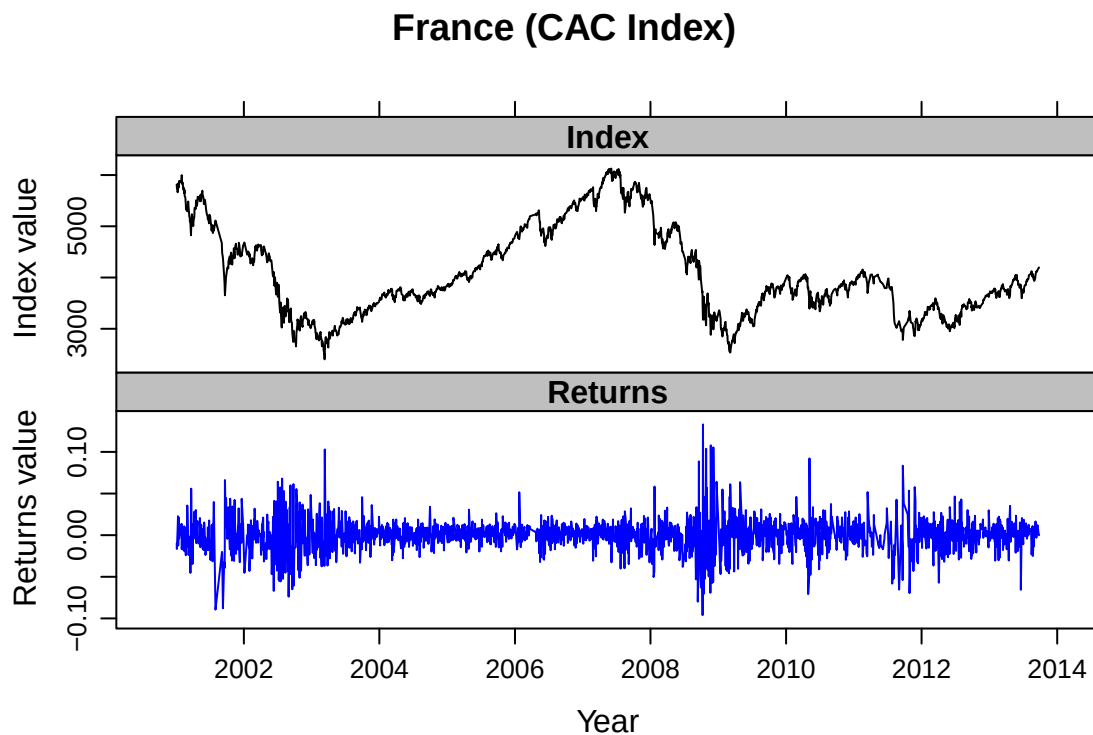
BVSP



```
table.Stats(BVSP returns, ci=0.95, digits=4)
```

	NA
Observations	2024.0000
NAs	0.0000
Minimum	-0.1321
Quartile 1	-0.0110
Median	0.0006
Arithmetic Mean	0.0006
Geometric Mean	0.0003
Quartile 3	0.0129
Maximum	0.1687
SE Mean	0.0005
LCL Mean (0.95)	-0.0004
UCL Mean (0.95)	0.0016
Variance	0.0005
Stdev	0.0233
Skewness	0.1234
Kurtosis	5.4069

CAC

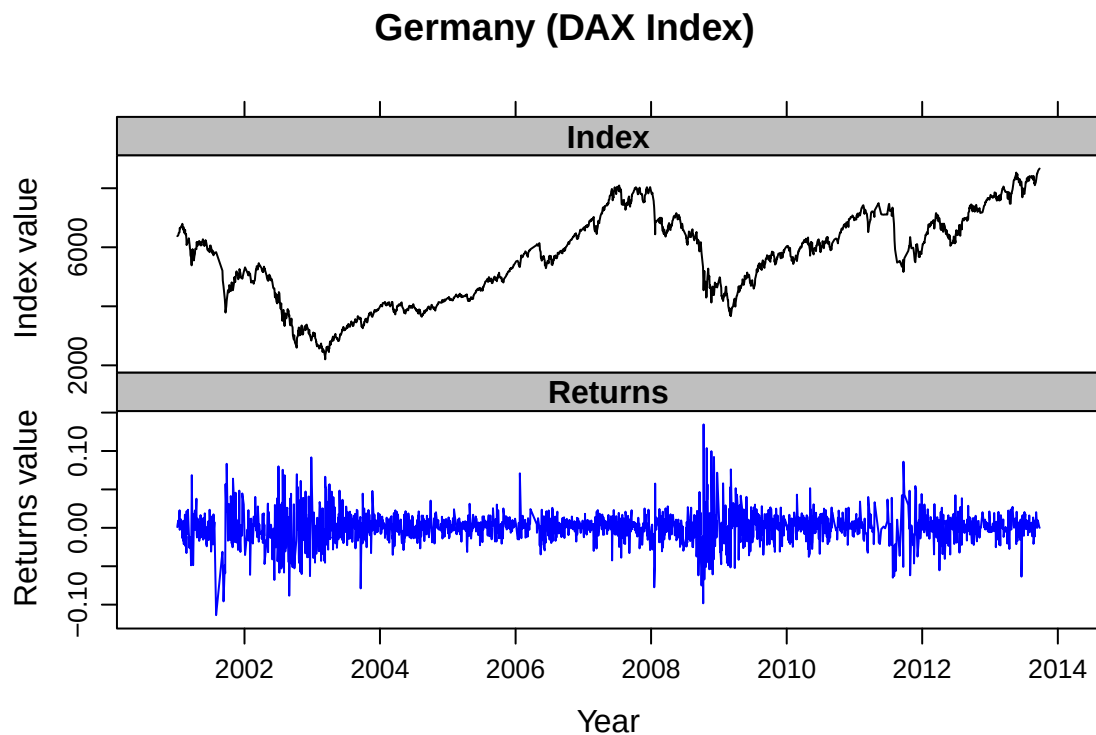


```
> table.Stats(CAC returns, ci=0.95, digits=4)
```

```

NA
Observations    2024.0000
NAs              0.0000
Minimum         -0.0961
Quartile 1      -0.0087
Median          0.0003
Arithmetic Mean -0.0002
Geometric Mean  -0.0003
Quartile 3       0.0090
Maximum          0.1330
SE Mean         0.0004
LCL Mean (0.95) -0.0010
UCL Mean (0.95)  0.0007
Variance        0.0004
Stdev           0.0193
Skewness         0.2561
Kurtosis         5.3707
```

DAX

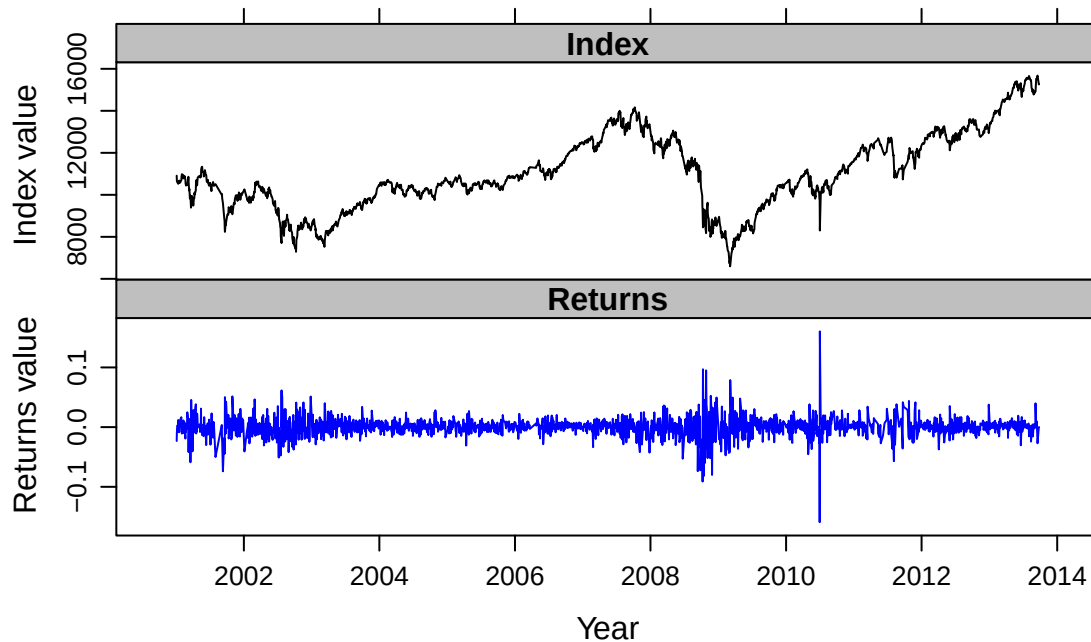


```
> table.Stats(DAX returns, ci=0.95, digits=4)
```

```

NA
Observations    2024.0000
NAs              0.0000
Minimum         -0.1137
Quartile 1      -0.0091
Median          0.0009
Arithmetic Mean  0.0002
Geometric Mean  0.0000
Quartile 3      0.0094
Maximum         0.1346
SE Mean         0.0004
LCL Mean (0.95) -0.0007
UCL Mean (0.95)  0.0010
Variance        0.0004
Stdev           0.0200
Skewness        0.0526
Kurtosis        4.7335
```

DJI

United States (DJI Index)

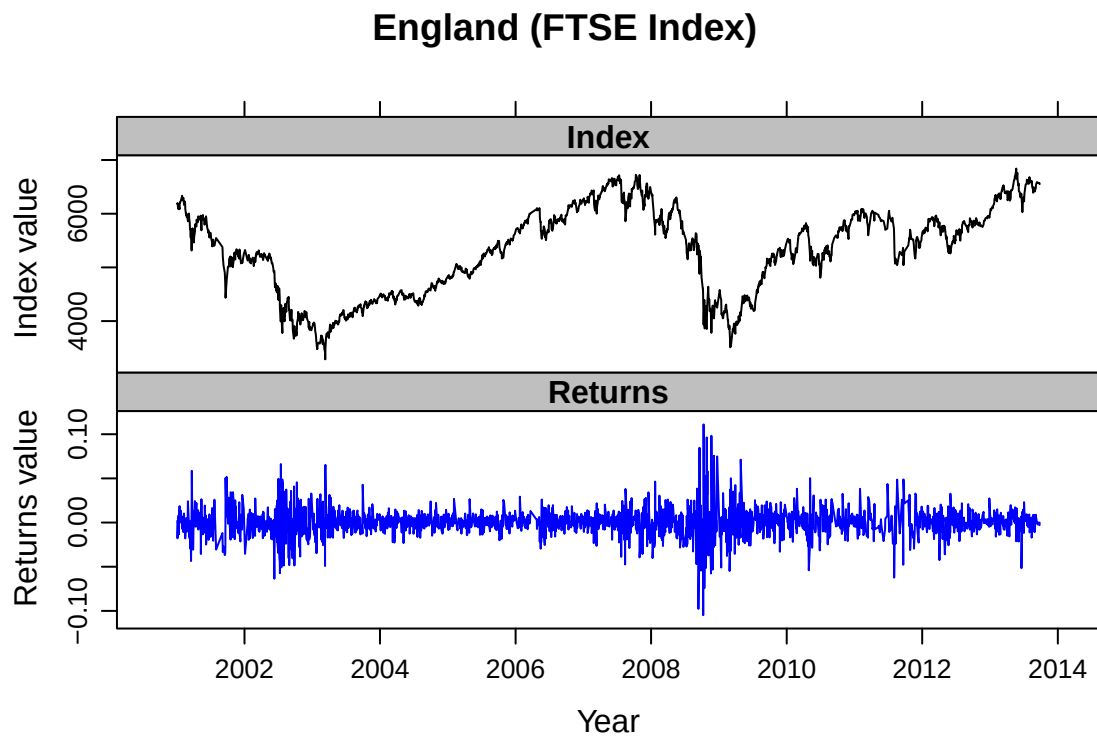
```
table.Stats(DJI returns, ci=0.95, digits=4)
```

```

NA
Observations    2024.0000
NAs              0.0000
Minimum         -0.1592
Quartile 1      -0.0065
Median          0.0005
Arithmetic Mean  0.0002
Geometric Mean  0.0000
Quartile 3      0.0066
Maximum         0.1604
SE Mean         0.0003
LCL Mean (0.95) -0.0005
UCL Mean (0.95)  0.0008
Variance        0.0002
Stdev           0.0157
Skewness        -0.0279
Kurtosis        15.0527

```

FTSE



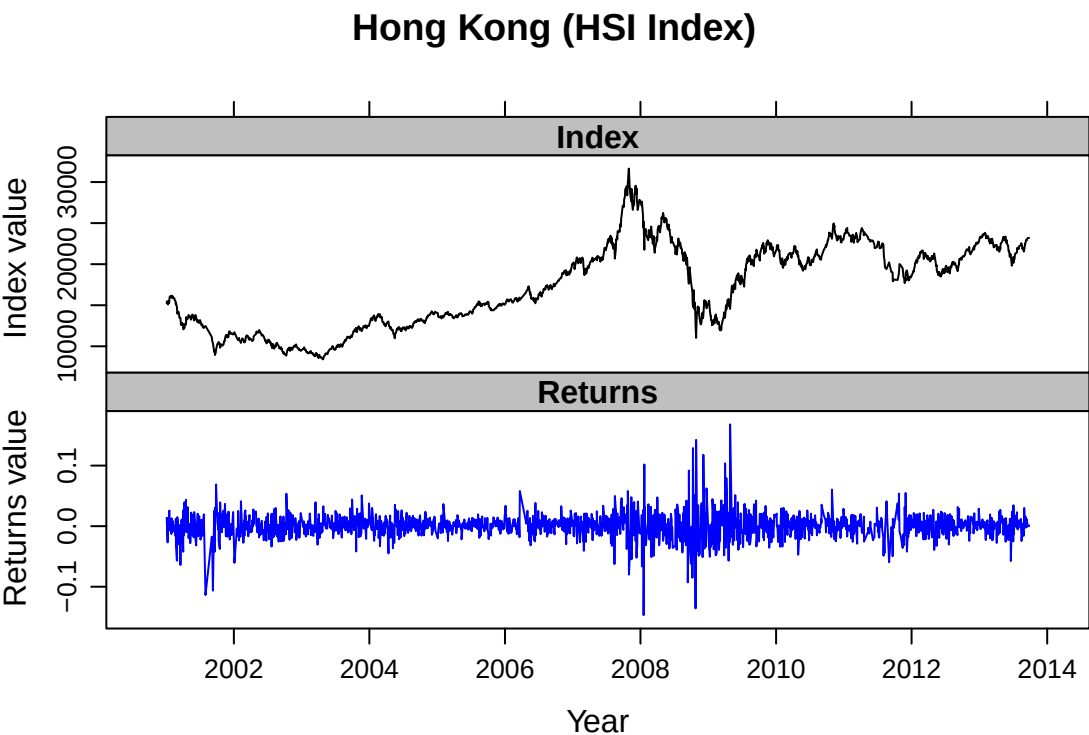
```
table.Stats(FTSE returns, ci=0.95, digits=4)
```

```

      NA
Observations    2024.0000
NAs              0.0000
Minimum         -0.1048
Quartile 1      -0.0067
Median          0.0004
Arithmetic Mean  0.0000
Geometric Mean  -0.0001
Quartile 3       0.0070
Maximum         0.1127
SE Mean         0.0004
LCL Mean (0.95) -0.0007
UCL Mean (0.95)  0.0007
Variance        0.0002
Stdev           0.0158
Skewness         0.3454
Kurtosis         8.1444

```

HSI

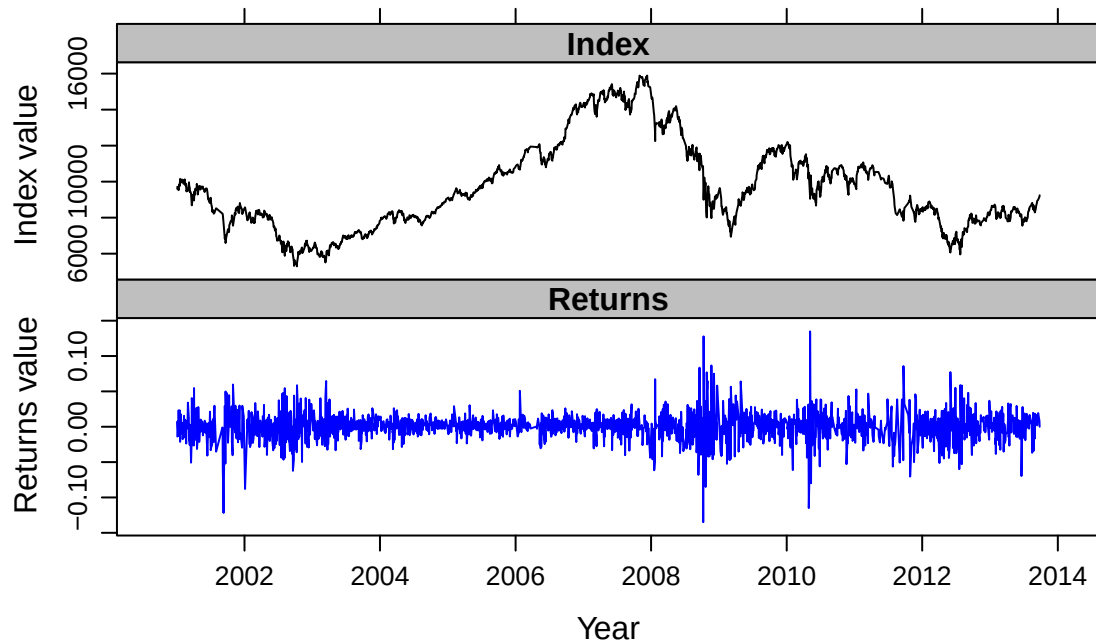


```
table.Stats(HSI returns, ci=0.95, digits=4)
```

	NA
Observations	2024.0000
NAs	0.0000
Minimum	-0.1470
Quartile 1	-0.0075
Median	0.0005
Arithmetic Mean	0.0002
Geometric Mean	0.0000
Quartile 3	0.0086
Maximum	0.1680
SE Mean	0.0004
LCL Mean (0.95)	-0.0006
UCL Mean (0.95)	0.0010
Variance	0.0004
Stdev	0.0191
Skewness	0.1709
Kurtosis	12.1247

IBEX

Spain (IBEX Index)



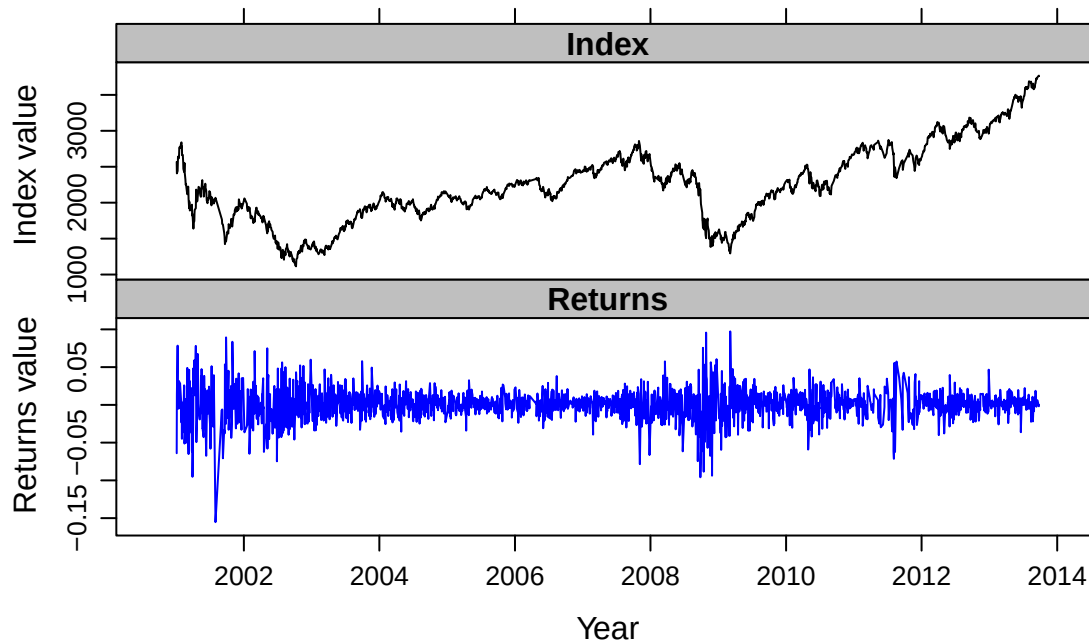
```
> table.Stats(IBEX returns, ci=0.95, digits=4)
```

```

NA
Observations    2024.0000
NAs              0.0000
Minimum         -0.1520
Quartile 1      -0.0086
Median          0.0005
Arithmetic Mean  0.0000
Geometric Mean  -0.0002
Quartile 3      0.0092
Maximum         0.1348
SE Mean         0.0004
LCL Mean (0.95) -0.0009
UCL Mean (0.95)  0.0008
Variance        0.0004
Stdev           0.0194
Skewness        -0.1921
Kurtosis        7.5566
```

IXIC

United States (IXIC Index)

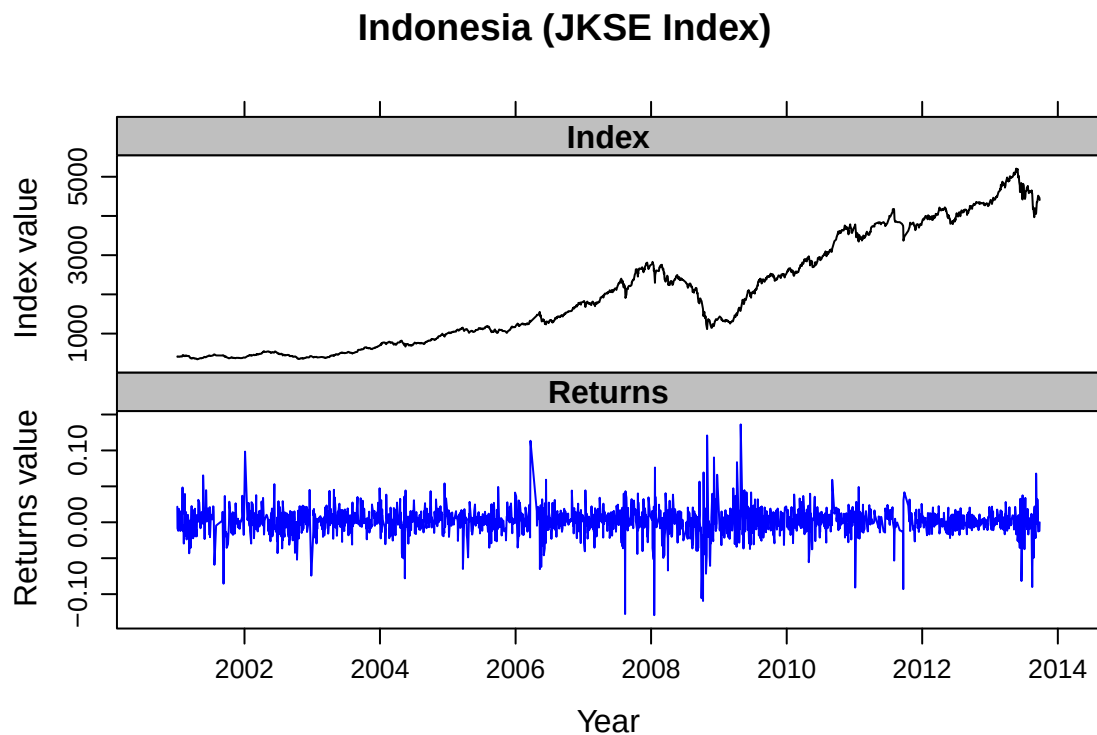


```
> table.Stats(IXIC returns, ci=0.95, digits=4)
```

```

NA
Observations    2024.0000
NAs              0.0000
Minimum         -0.1553
Quartile 1      -0.0088
Median           0.0005
Arithmetic Mean  0.0002
Geometric Mean  0.0000
Quartile 3       0.0095
Maximum          0.0973
SE Mean          0.0004
LCL Mean (0.95) -0.0007
UCL Mean (0.95)  0.0010
Variance         0.0004
Stdev            0.0197
Skewness         -0.3322
Kurtosis         4.9143
```

JKSE



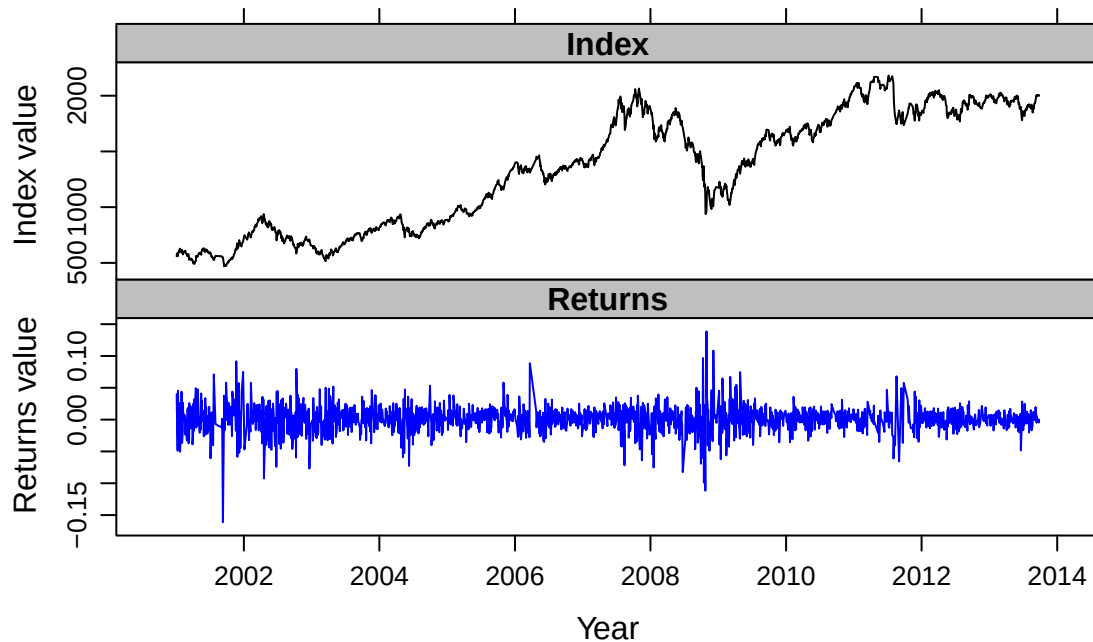
```
> table.Stats(JKSE returns, ci=0.95, digits=4)
```

```

NA
Observations    2024.0000
NAs              0.0000
Minimum         -0.1293
Quartile 1      -0.0071
Median          0.0014
Arithmetic Mean 0.0012
Geometric Mean  0.0010
Quartile 3      0.0102
Maximum         0.1362
SE Mean         0.0004
LCL Mean (0.95) 0.0004
UCL Mean (0.95) 0.0020
Variance        0.0004
Stdev           0.0187
Skewness        -0.2494
Kurtosis         8.8300
```

KOSPI

South Korea (KOSPI Index)



```
table.Stats(KOSPI returns, ci=0.95, digits=4)
```

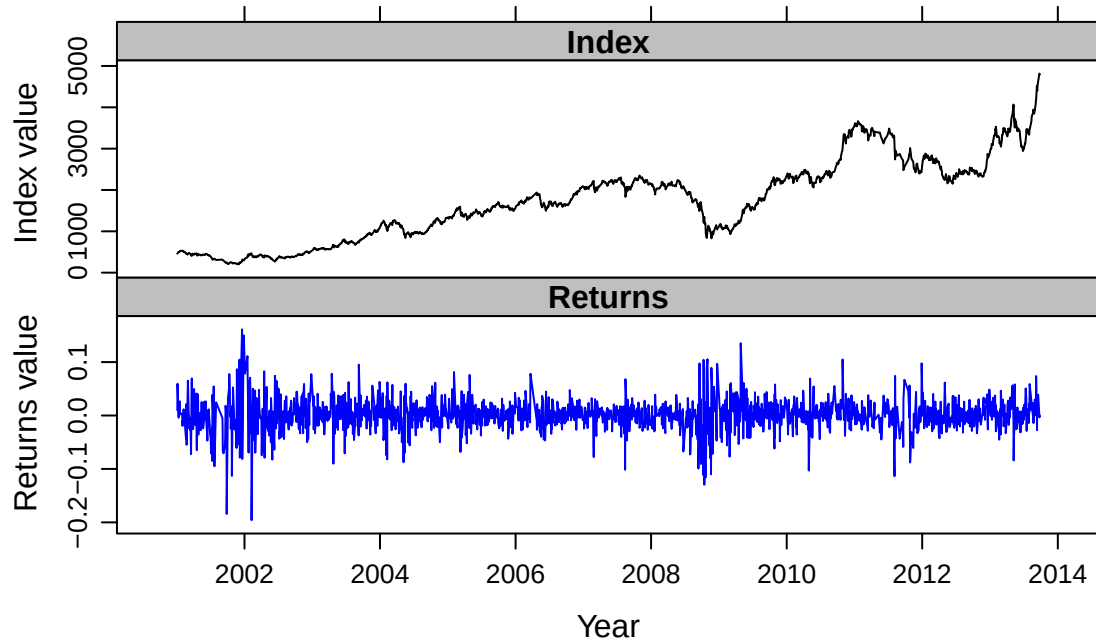
```

NA
Observations    2024.0000
NAs              0.0000
Minimum         -0.1612
Quartile 1      -0.0079
Median          0.0011
Arithmetic Mean  0.0006
Geometric Mean  0.0004
Quartile 3      0.0100
Maximum         0.1386
SE Mean         0.0004
LCL Mean (0.95) -0.0002
UCL Mean (0.95)  0.0015
Variance        0.0004
Stdev           0.0195
Skewness        -0.2725
Kurtosis        6.9870

```

MERVAL

Argentina (MERVAL Index)

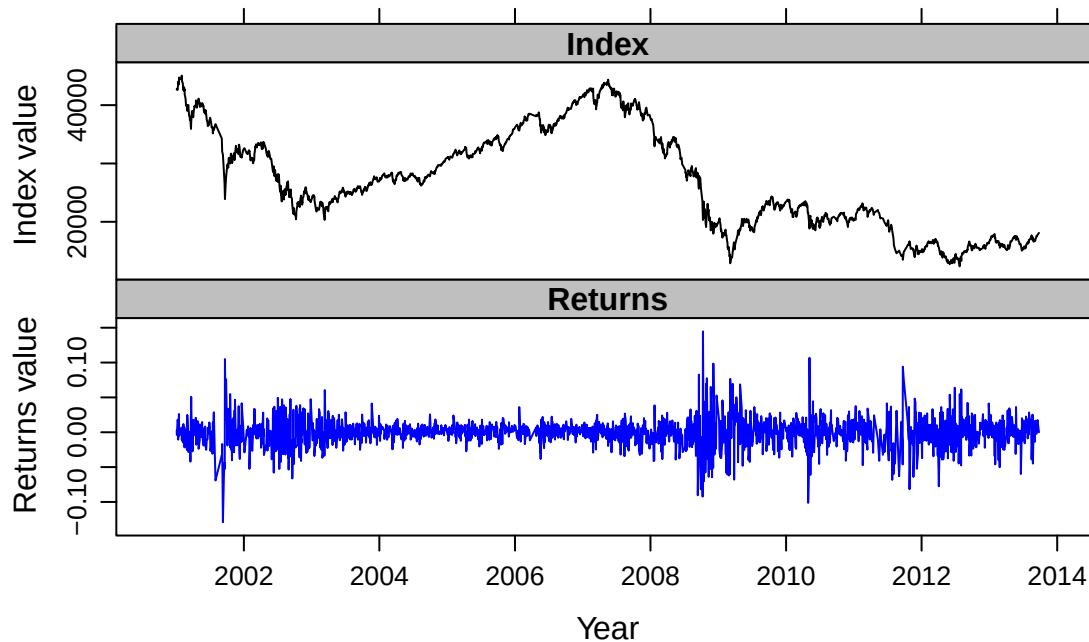


```
> table.Stats(MERVAL returns, ci=0.95, digits=4)
```

	NA
Observations	2024.0000
NAs	0.0000
Minimum	-0.1959
Quartile 1	-0.0110
Median	0.0010
Arithmetic Mean	0.0012
Geometric Mean	0.0008
Quartile 3	0.0133
Maximum	0.2310
SE Mean	0.0006
LCL Mean (0.95)	-0.0001
UCL Mean (0.95)	0.0024
Variance	0.0008
Stdev	0.0278
Skewness	0.0518
Kurtosis	7.3188

MIB

Italia (MIB Index)

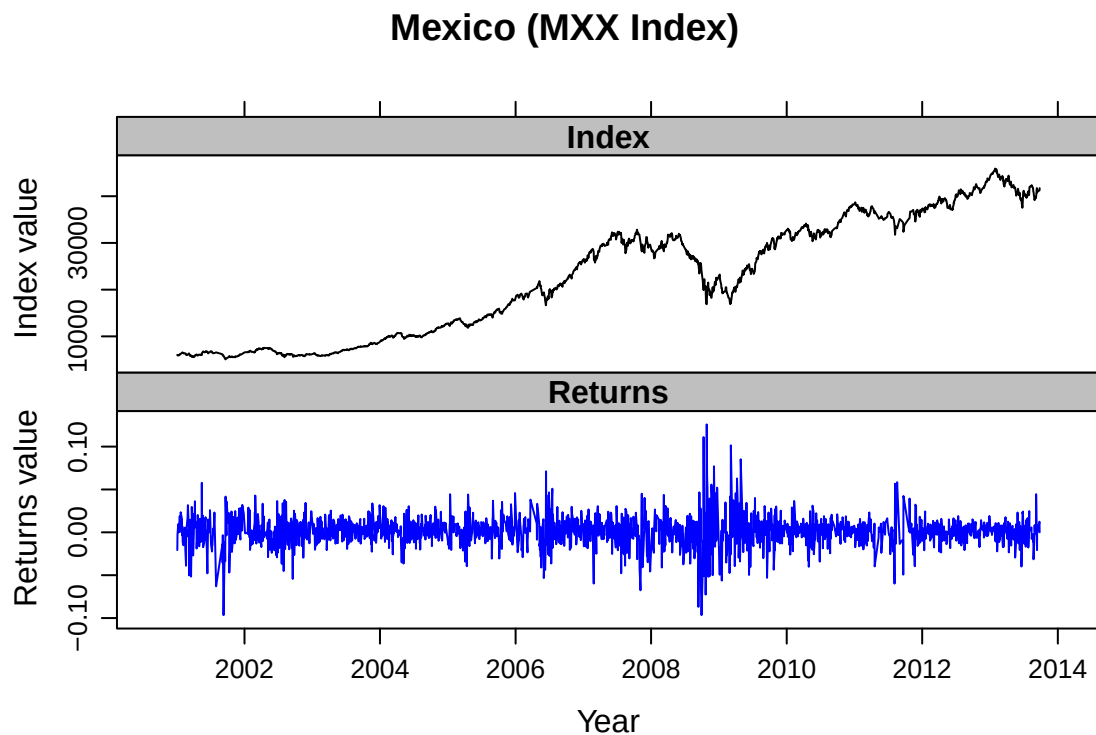


```
> table.Stats(MIB returns, ci=0.95, digits=4)
```

NA

Observations	2024.0000
NAs	0.0000
Minimum	-0.1291
Quartile 1	-0.0088
Median	0.0006
Arithmetic Mean	-0.0004
Geometric Mean	-0.0006
Quartile 3	0.0085
Maximum	0.1447
SE Mean	0.0004
LCL Mean (0.95)	-0.0013
UCL Mean (0.95)	0.0004
Variance	0.0004
Stdev	0.0197
Skewness	-0.0899
Kurtosis	6.3869

MXX

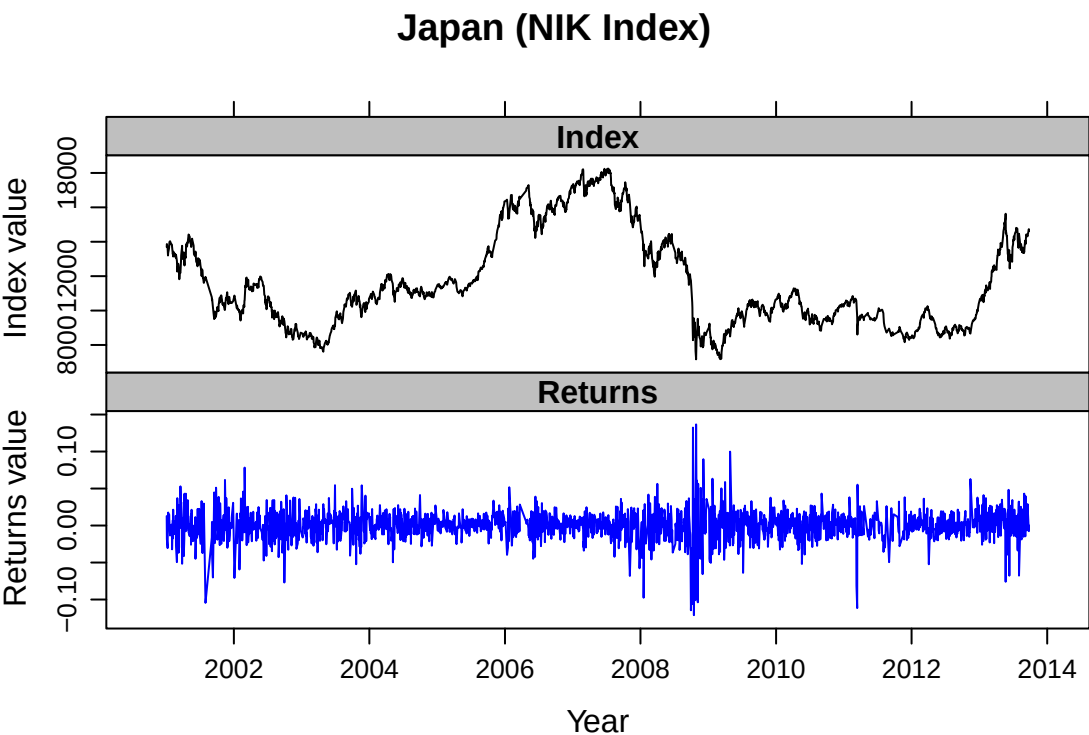


```
> table.Stats(MXX returns, ci=0.95, digits=4)
```

```

NA
Observations    2024.0000
NAs              0.0000
Minimum         -0.0966
Quartile 1      -0.0067
Median           0.0014
Arithmetic Mean  0.0010
Geometric Mean  0.0008
Quartile 3       0.0087
Maximum          0.1259
SE Mean          0.0004
LCL Mean (0.95)  0.0002
UCL Mean (0.95)  0.0017
Variance         0.0003
Stdev            0.0167
Skewness         0.1871
Kurtosis         6.9050
```

NIK

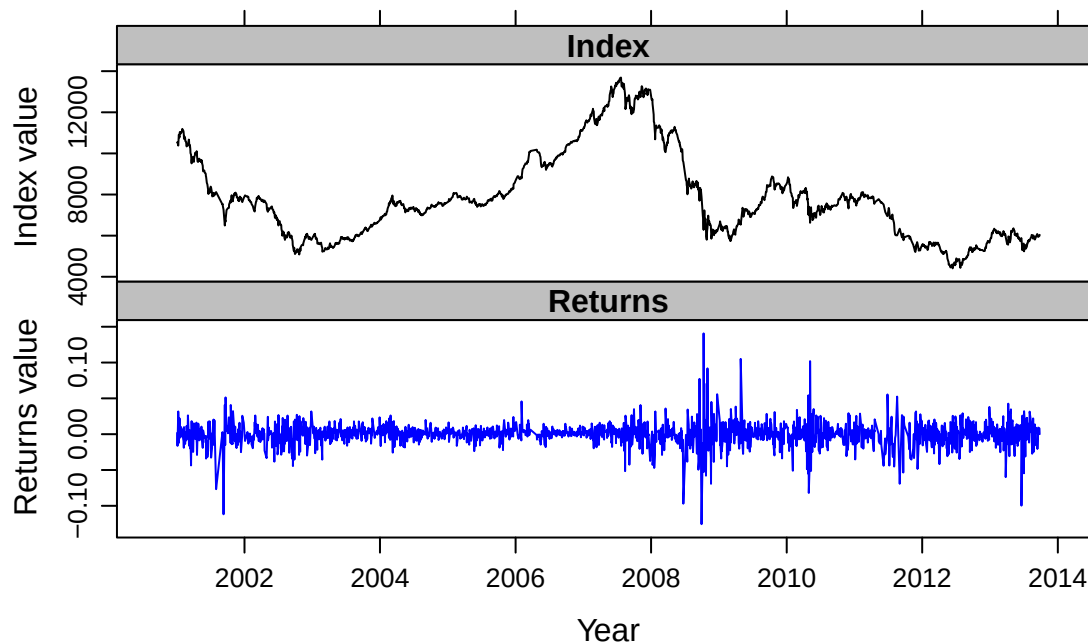


```
table.Stats(NIK returns, ci=0.95, digits=4)
NA
```

Observations	2024.0000
NAs	0.0000
Minimum	-0.1211
Quartile 1	-0.0092
Median	0.0005
Arithmetic Mean	0.0000
Geometric Mean	-0.0002
Quartile 3	0.0100
Maximum	0.1367
SE Mean	0.0004
LCL Mean (0.95)	-0.0008
UCL Mean (0.95)	0.0009
Variance	0.0004
Stdev	0.0197
Skewness	-0.4147
Kurtosis	7.0111

PSI

Portugal (PSI Index)

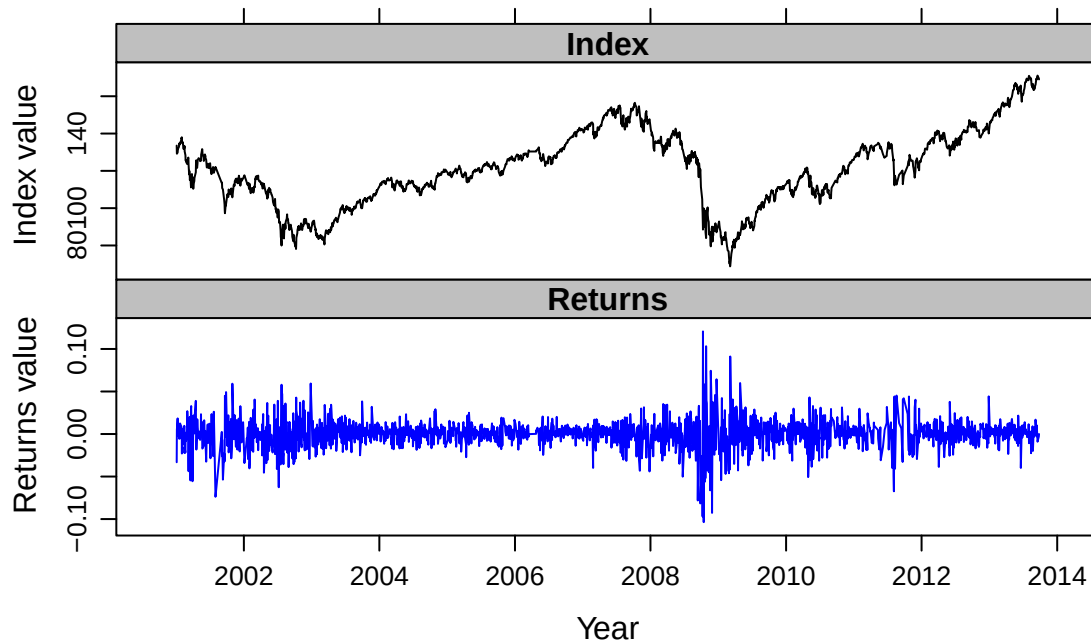


```
> table.Stats(PSI returns, ci=0.95, digits=4)
```

	NA
Observations	2024.0000
NAs	0.0000
Minimum	-0.1378
Quartile 1	-0.0063
Median	0.0007
Arithmetic Mean	-0.0003
Geometric Mean	-0.0004
Quartile 3	0.0063
Maximum	0.1407
SE Mean	0.0003
LCL Mean (0.95)	-0.0010
UCL Mean (0.95)	0.0004
Variance	0.0002
Stdev	0.0156
Skewness	-0.3625
Kurtosis	15.4539

SPY

United States (SPY Index)

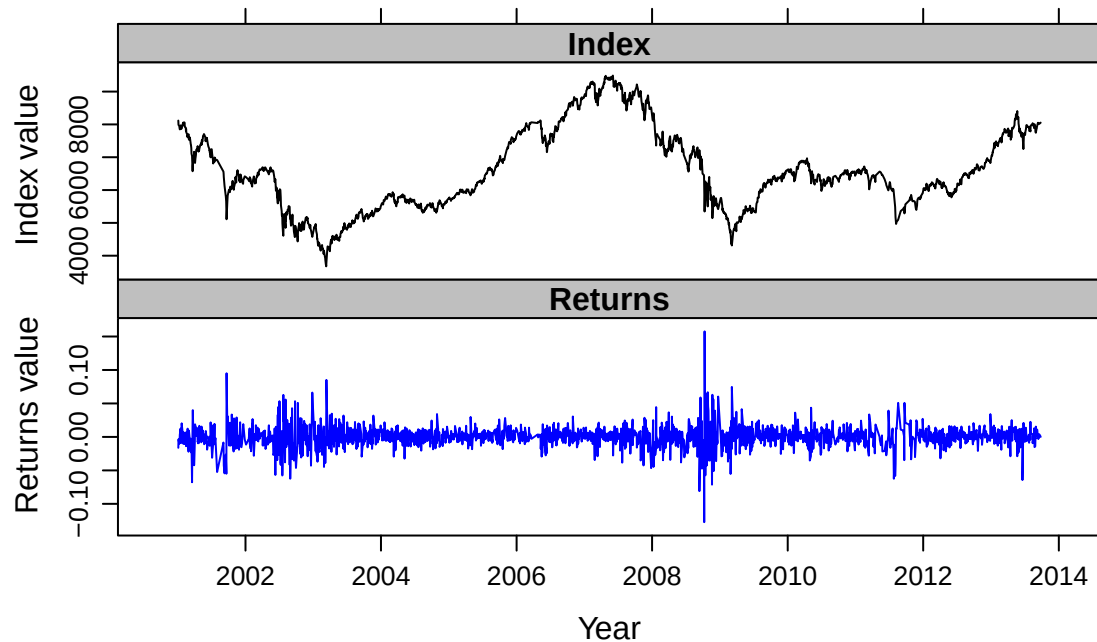


```
> table.Stats(SPY returns, ci=0.95, digits=4)
```

	NA
Observations	2024.0000
NAs	0.0000
Minimum	-0.1036
Quartile 1	-0.0064
Median	0.0006
Arithmetic Mean	0.0001
Geometric Mean	0.0000
Quartile 3	0.0072
Maximum	0.1207
SE Mean	0.0004
LCL Mean (0.95)	-0.0006
UCL Mean (0.95)	0.0008
Variance	0.0003
Stdev	0.0160
Skewness	-0.1062
Kurtosis	7.3934

SSMI

Switzerland (SSMI Index)

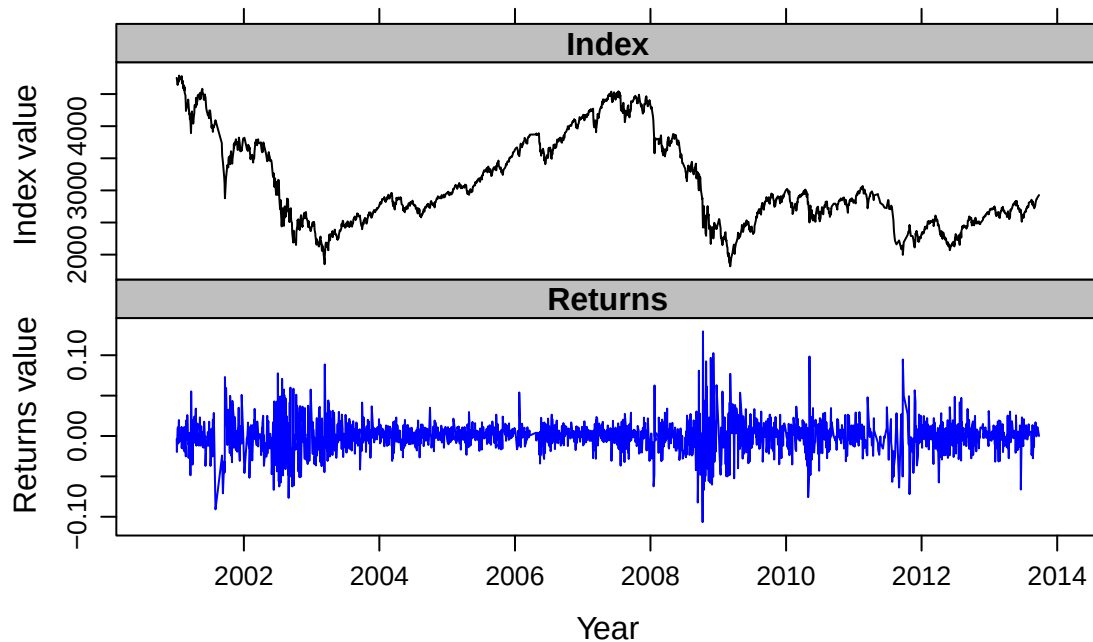


```
> table.Stats(SSMI returns, ci=0.95, digits=4)
```

```

NA
Observations    2024.0000
NAs              0.0000
Minimum         -0.1274
Quartile 1      -0.0069
Median          0.0004
Arithmetic Mean  0.0000
Geometric Mean  -0.0001
Quartile 3      0.0075
Maximum         0.1576
SE Mean         0.0004
LCL Mean (0.95) -0.0007
UCL Mean (0.95)  0.0007
Variance        0.0003
Stdev           0.0159
Skewness        0.2232
Kurtosis        10.4162
```

STOXX

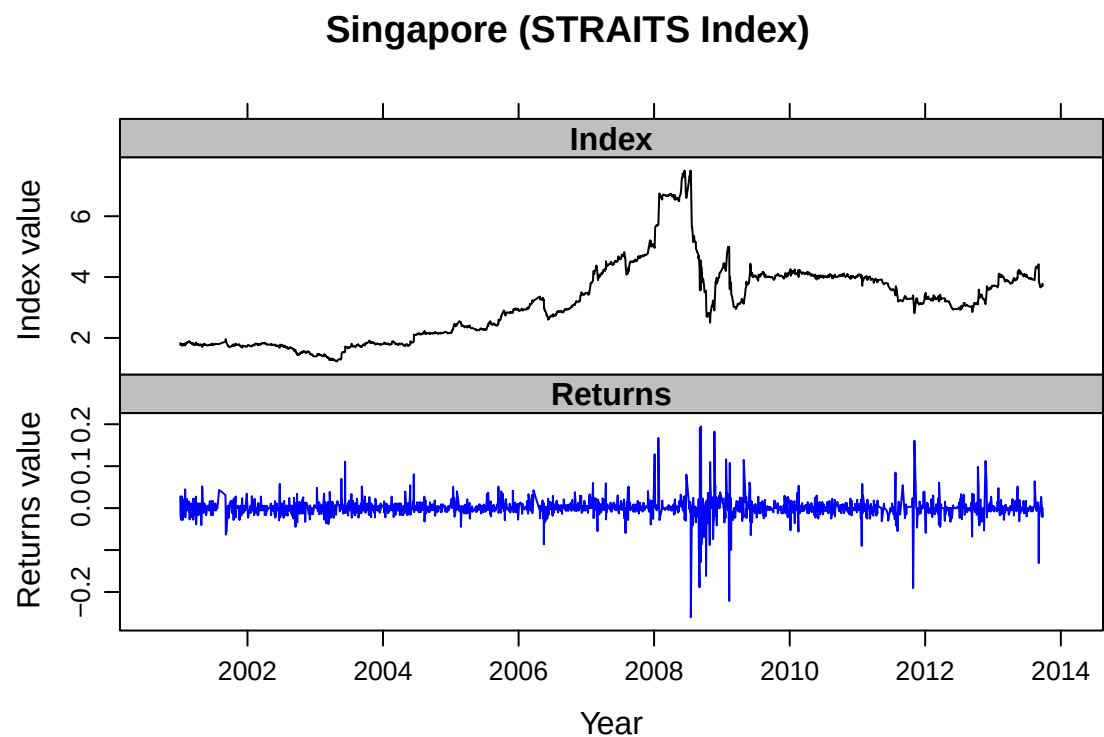
Europe (STOXX Index)

```
table.Stats(STOXX returns, ci=0.95, digits=4)
```

NA

Observations	2024.0000
NAs	0.0000
Minimum	-0.1067
Quartile 1	-0.0088
Median	0.0000
Arithmetic Mean	-0.0002
Geometric Mean	-0.0004
Quartile 3	0.0089
Maximum	0.1295
SE Mean	0.0004
LCL Mean (0.95)	-0.0011
UCL Mean (0.95)	0.0006
Variance	0.0004
Stdev	0.0194
Skewness	0.1935
Kurtosis	4.9081

STRAITS



```
table.Stats(STRAITS returns, ci=0.95, digits=4)
```

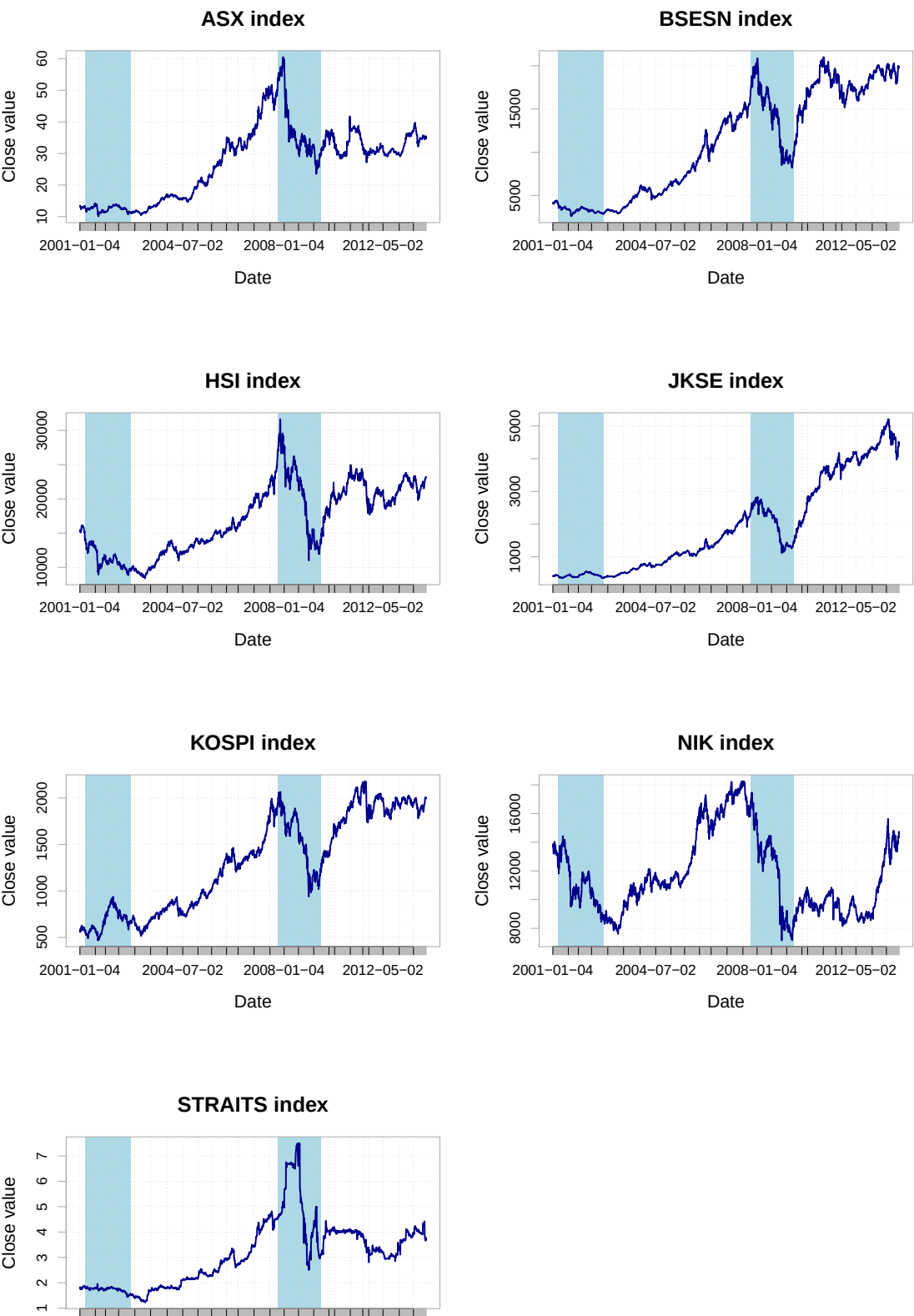
	NA
Observations	2024.0000
NAs	0.0000
Minimum	-0.2600
Quartile 1	-0.0058
Median	0.0000
Arithmetic Mean	0.0004
Geometric Mean	0.0001
Quartile 3	0.0060
Maximum	0.1948
SE Mean	0.0005
LCL Mean (0.95)	-0.0006
UCL Mean (0.95)	0.0014
Variance	0.0005
Stdev	0.0229
Skewness	-0.6769
Kurtosis	31.5261

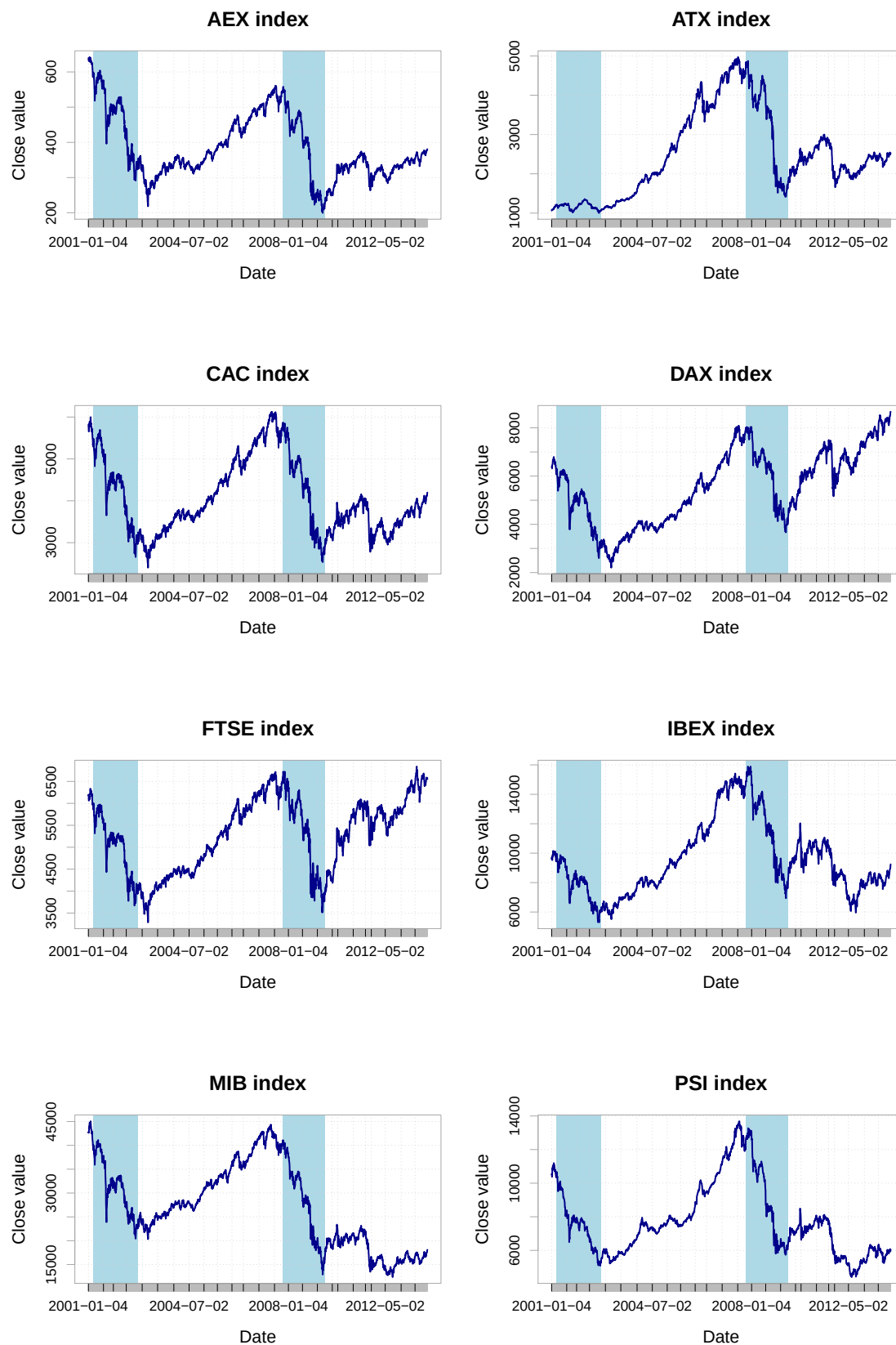
B

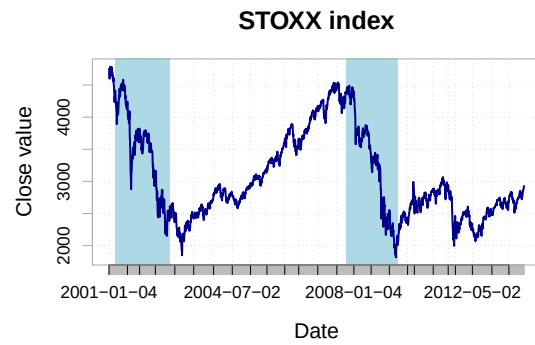
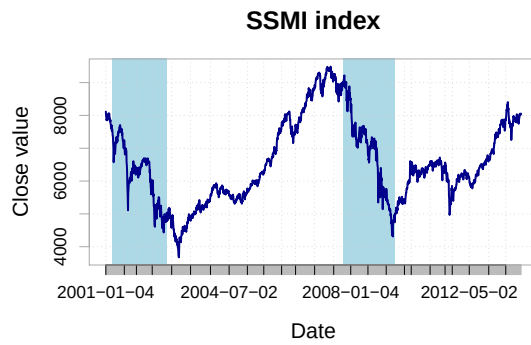
CATALOGUE OF RESULTS

B.1 MARKETS INDEX VERSUS CRISIS DATES

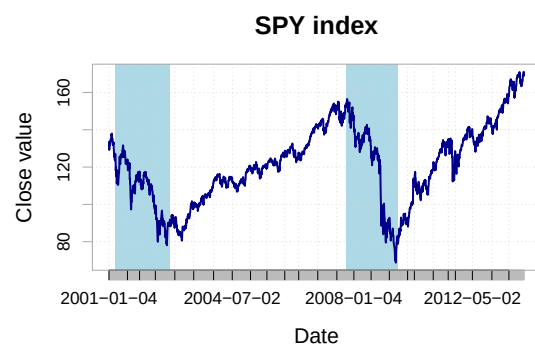
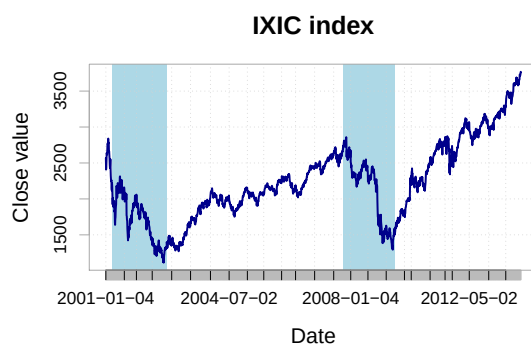
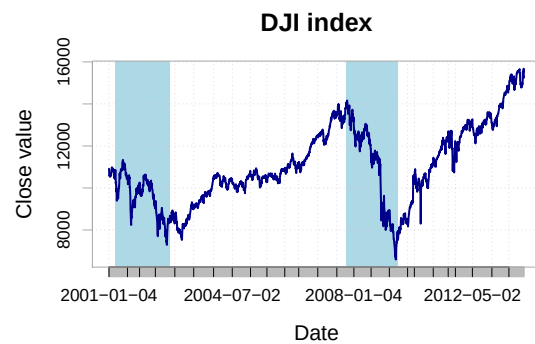
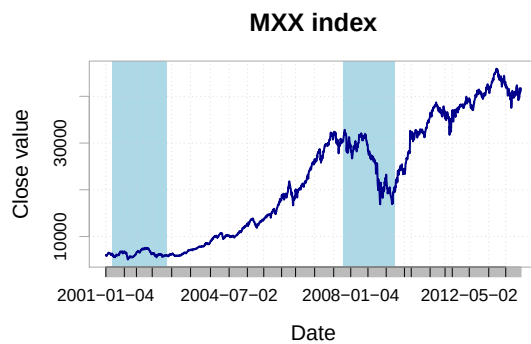
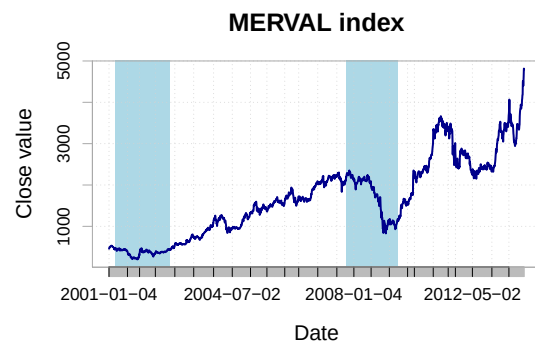
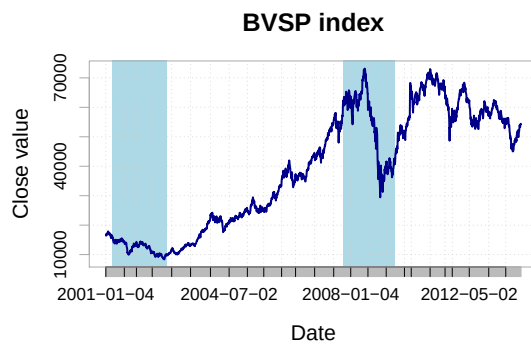
Asia-Pacific markets



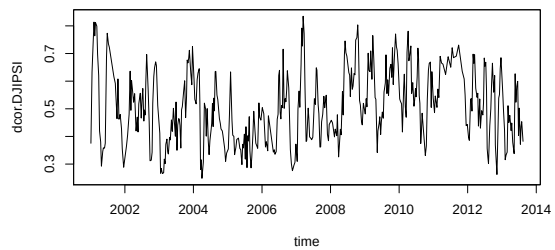
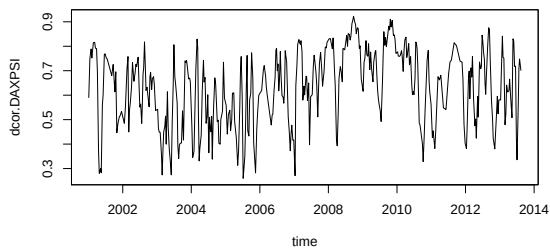
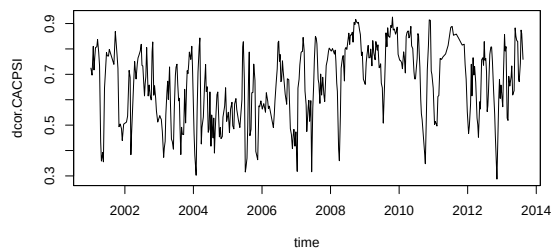
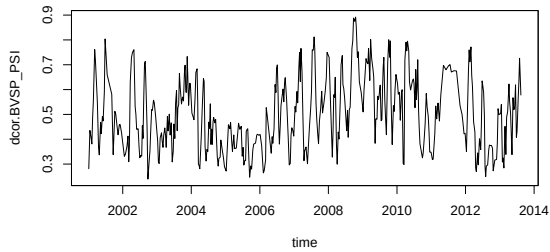
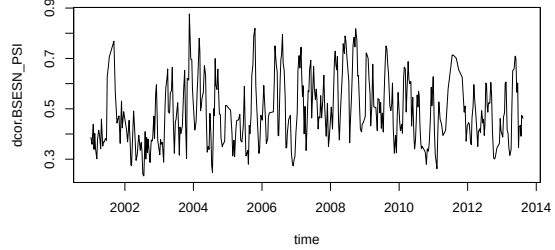
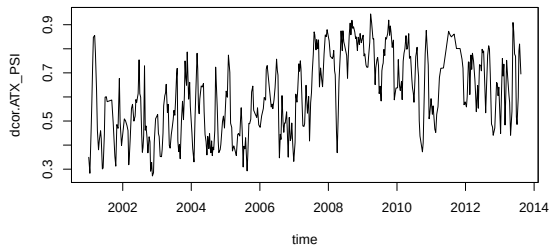
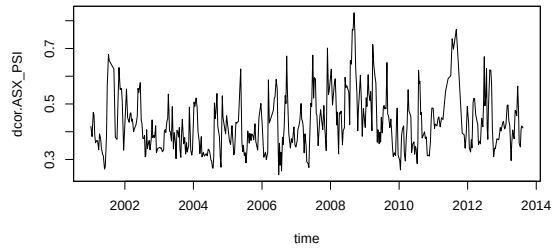
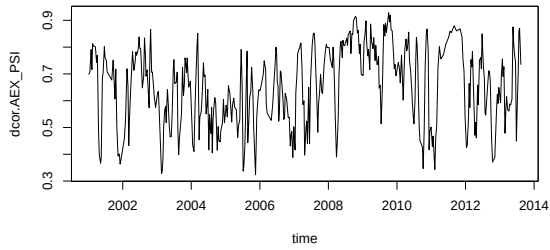
European markets

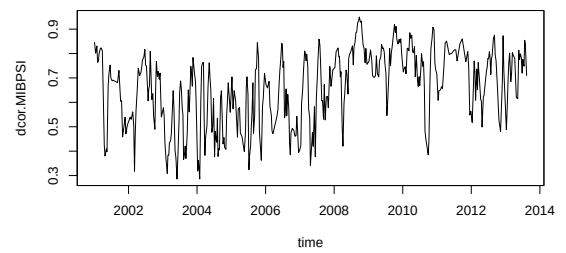
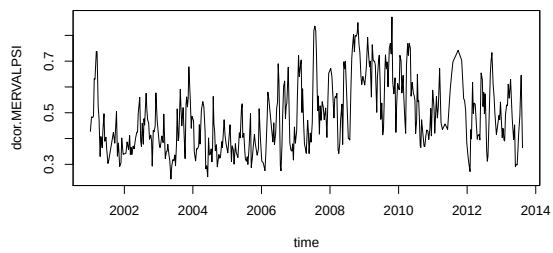
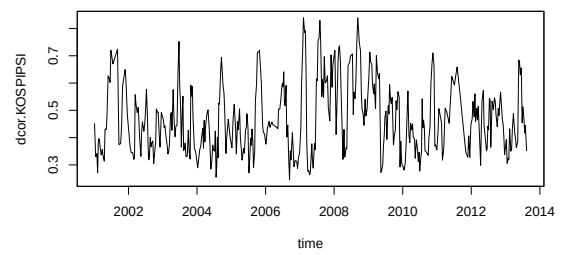
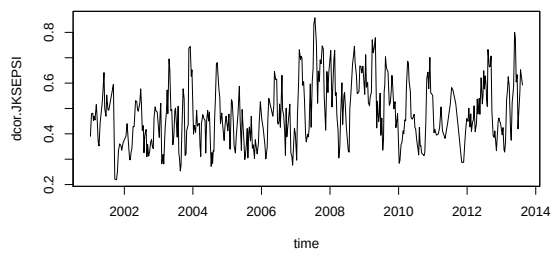
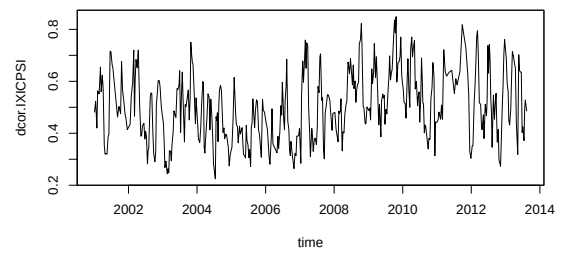
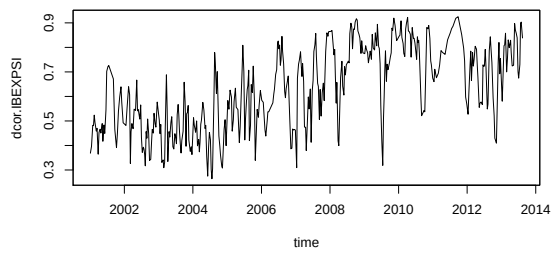
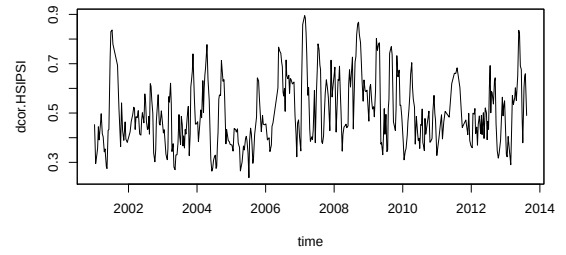
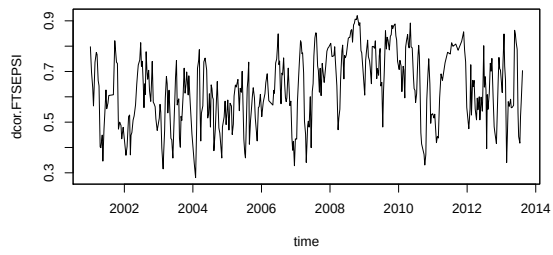


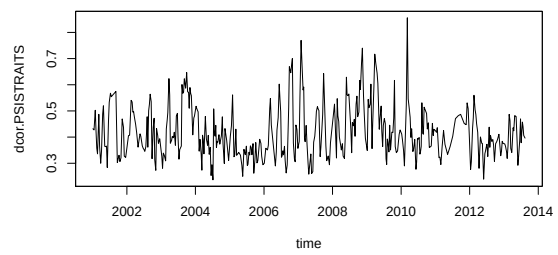
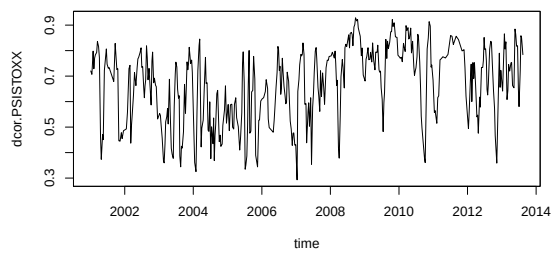
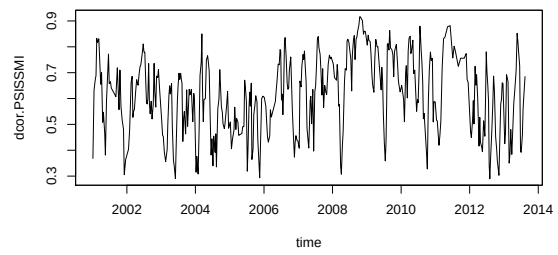
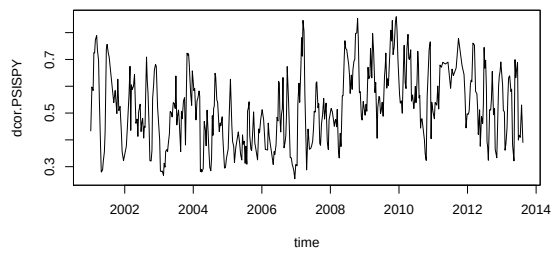
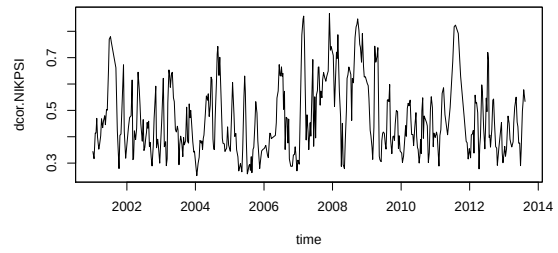
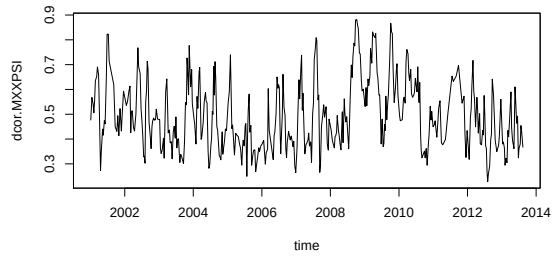
American markets



B.2 DISTANCE CORRELATION FOR PSI-20

Distance Correlation for pairs with PSI-20





PACKAGE DESCRIPTION

All the packages listed in this appendix can be found at cran.r-project.org/web/packages/

C.1 HASH

- Details

PACKAGE: hash

AUTHOR: Christopher Brown

TITLE: Full feature implementation of hash/associated arrays/dictionaries

DATE: 2013-02-20

DESCRIPTION: This package implements a data structure similar to hashes in Perl and dictionaries in Python but with a purposefully R flavor. For objects of appreciable size, access using hashes outperforms native named lists and vectors.

VERSION: 2.2.6

DEPENDS: R ($\geq 2.12.0$), methods, utils

SUGGESTS: testthat

LICENSE: GPL (≥ 2)

C.2 PERFORMANCEANALYTICS

- Details

PACKAGE: performanceAnalytics

AUTHORS: Brian G. Peterson [cre, aut, cph], Peter Carl [aut, cph], Kris Boudt [ctb, cph], Ross Bennett [ctb], Joshua Ulrich [ctb], Eric Zivot [ctb], Matthieu Lestel [ctb], Kyle Balkissoon [ctb], Diethelm Wuertz [ctb]

TITLE: Econometric tools for performance and risk analysis

DATE: 2014-09-15

DESCRIPTION: Collection of econometric functions for performance and risk analysis. This package aims to aid practitioners and researchers in utilizing the latest research in analysis of non-normal return streams. In general, it is most tested on return (rather than price) data on a regular scale, but most functions will work with irregular return data as well, and increasing numbers of functions will work with P&L or price data where possible.

VERSION: 1.4.3541

IMPORTS: zoo

DEPENDS: R ($\geq 3.0.0$), xts (≥ 0.9)

SUGGESTS: Hmisc, MASS, quantmod, gamlss, gamlss.dist, robustbase, quantreg, gplots

LICENSE: GPL-2 | GPL-3

URL: <http://r-forge.r-project.org/projects/returnanalytics/>

C.3 zoo

- Details

PACKAGE: zoo

AUTHORS: Achim Zeileis [aut, cre], Gabor Grothendieck [aut], Jeffrey A. Ryan [aut], Felix Andrews [ctb]

TITLE: S3 Infrastructure for Regular and Irregular Time Series (Z's ordered observations)

DATE: 2014-02-27

DESCRIPTION: An S3 class with methods for totally ordered indexed observations. It is particularly aimed at irregular time series of numeric vectors/matrices and factors. zoo's key design goals are independence of a particular index/date/time class and consistency with ts and base R by providing methods to extend standard generics.

VERSION: 1.7-11

DEPENDS: R ($\geq 2.10.0$), stats

SUGGESTS: coda, chron, DAAG, fts, its, ggplot2, mondate, scales, strucchange, timeDate, time-Series, tis, tseries, xts Imports utils, graphics, grDevices, lattice ($\geq 0.20-27$)

LICENSE: GPL-2 | GPL-3

URL: <http://zoo.R-Forge.R-project.org/>

C.4 PRACMA

- Details

PACKAGE: pracma

AUTHORS: Hans Werner Borchers

TITLE: Practical Numerical Math Functions

DATE: 2014-11-01

DESCRIPTION: Functions from numerical analysis and linear algebra, numerical optimization, differential equations, plus some special functions. Uses Matlab function names where appropriate to simplify porting.

VERSION: 1.7.7
 DEPENDS: R ($\geq 2.11.1$)
 LICENSE: GPL (≥ 3)

C.5 ENERGY

- Details

PACKAGE: energy
 AUTHORS: Maria L. Rizzo and Gabor J. Szekely
 TITLE: E-statistics (energy statistics)
 DATE: 2014-10-27
 DESCRIPTION: E-statistics (energy) tests and statistics for comparing distributions: multivariate normality, multivariate distance components and k- sample test for equal distributions, hierarchical clustering by e-distances, multivariate independence tests, distance correlation, goodness-of-fit tests. Energy- statistics concept based on a generalization of Newton's potential energy is due to Gabor J. Szekely.
 VERSION: 1.6.2
 IMPORTS: boot
 LICENSE: GPL (≥ 2)

C.6 LATTICE

- Details

PACKAGE: lattice
 AUTHORS: Deepayan Sarkar
 TITLE: Lattice Graphics
 DATE: 2014/04/01
 DESCRIPTION: Lattice is a powerful and elegant high-level data visualization system, with an emphasis on multivariate data, that is sufficient for typical graphics needs, and is also flexible enough to handle most non standard requirements.
 VERSION: 0.20-29
 DEPENDS: R ($\geq 2.15.1$)
 SUGGESTS: KernSmooth, MASS Imports grid, grDevices, graphics, stats, utils
 LICENSE: GPL (≥ 2)
 URL: <http://lattice.r-forge.r-project.org/>

C.7 xts

- Details

PACKAGE: xts

AUTHORS: Jeffrey A. Ryan, Joshua M. Ulrich

TITLE: eXtensible Time Series

DATE: 2013-06-26

DESCRIPTION: Provide for uniform handling of R's different time-based data classes by extending zoo, maximizing native format information preservation and allowing for user level customization and extension, while simplifying cross-class interoperability.

VERSION: 0.9-7

DEPENDS: zoo ($\geq$ 1.7-10)

SUGGESTS: timeSeries, timeDate, tseries, its, chron, fts, tis

LICENSE: GPL ($\geq$ 2)

URL: <http://r-forge.r-project.org/projects/xts/>

C.8 XTSEXTRA

- Details

PACKAGE: xtsExtra

AUTHORS: Michael Weylandt

TITLE: xtsExtra

DATE: 2012

DESCRIPTION: For the community who makes the most heavy use of xts, xtsExtra introduces a new set of plotting functions for xts objects available as part of Google Summer of Code 2012. This work represents a major overhaul of previously existing plot.xts and should provide you with the most comprehensive and flexible time series plotting available

VERSION: 0.0-1

URL: <https://stat.ethz.ch/pipermail/r-sig-finance/2012q3/010652.html>

C.9 ENTROPY

- Details

PACKAGE: entropy

AUTHORS: Jean Hausser and Korbinian Strimmer

TITLE: Estimation of Entropy, Mutual Information and Related Quantities

DATE: 2013-07-16

DESCRIPTION: This package implements various estimators of entropy, such as the shrinkage estimator by Hausser and Strimmer, the maximum likelihood and the Millow-Madow estimator, various Bayesian estimators, and the Chao-Shen estimator. It also offers an R interface to the NSB estimator. Furthermore, it provides functions for estimating Kullback-Leibler divergence, chi2-squared, mutual information, and chi2-squared statistic of independence. In addition there are functions for discretizing continuous random variables.

VERSION: 1.2.0

DEPENDS: R ($\geq 2.15.1$)

LICENSE: GPL (≥ 3)

URL: <http://strimmerlab.org/software/entropy/>

C.10 FORECA

- Details

PACKAGE: ForeCA

AUTHORS: Georg M. Goerg

TITLE: ForeCA - Forecastable Component Analysis

DATE: 2014-03-01

DESCRIPTION: Forecastable Component Analysis (ForeCA) is a novel dimension reduction (DR) technique for temporally dependent signals. Contrary to other popular DR methods, such as PCA or ICA, ForeCA explicitly searches for the most "forecastable" signal. The measure of forecastability is based on negative Shannon entropy of the spectral density of the transformed signal. This R package provides the main algorithms and auxiliary function(summary, plotting, etc) to apply ForeCA to multivariate data (time series).

VERSION: 0.1

IMPORTS: R.utils, sapa, mgcv, astsa

DEPENDS: R ($\geq 2.15.0$), ifultools ($\geq 2.0-0$), splus2R ($\geq 1.2-0$), nlme ($\geq 3.1-64$)

LICENSE: GPL-2

URL: <http://www.gmge.org>

SOFTWARE

In this thesis there were developed several R scripts for analysing and calculating the needed measures over the stocks and markets chosen, as follows.

For simplicity, it is only shown the code with respect to markets calculation. Similar programs were applied to PSI-20 stocks.

D.1 MARKETS MATRIX CODE

```

1  # Copyright (C) 2013-2014 José Miguel Salgado <portosalgado@gmail.com>
  #
  # This program is free software; you can redistribute it and/or modify it under the terms of
  #the GNU General Public License as published by the Free Software Foundation; either version
5  #2 of the License or (at your option) any later version.
  #
  # This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY;
  #without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
  #See the GNU General Public License for more details.
10 #
  # You should have received a copy of the GNU General Public License along with this program;
  #if not, write to the Free Software Foundation, Inc., 59 Temple Place - Suite 330, Boston,
  #MA 02111-1307, USA.
  #
15 #####
  #Real Markets#
  #market.names.europe=list("aex","atx","cac","dax","ibex","ftse","mib","psi20","ssmi","stoxx
    ")
  #market.names.eua=list("dji","iixc","spy")
  #market.names.latinamerica=list("bvsp","merval","mxx")
20 #market.names.asia=list("bsebn","hsi","kospi","jkse","nik","strait")
  #market.names.oceania=list("asx")
  market.names=list("AEX","ASX","ATX","BSEBN","BVSP","CAC","DAX","DJI","FTSE",
    "HSI","IBEX","IXIC","KOSPI","JKSE","MERVAL","MIB","MX",
    "NIK","PSI20","SPY","SSMI","STOXX","STRAITS")
25
  #markets complete data
  markets=list()
  for (m in 1:length(market.names)){
    markets[[m]]=read.csv(paste(market.names[[m]], "csv", sep="."), header=TRUE)
30 }

  library(hash)

  #markets data and close value
35 markets.hash=list()
  for (m in 1:length(market.names)){
    markets.hash[[m]]=hash(markets[[m]]$Date,markets[[m]]$Close)
  }

```

```

40 #markets dates
   dates=keys(markets.hash[[1]])
   for (m in 2:length(market.names)){
       dates=dates[has.key(dates,markets.hash[[m]])]
   }
45 #same days markets close values
   markets.common=list()
   for (m in 1:length(market.names)){
       markets.common[[m]]=values(markets.hash[[m]],dates)
50 }

   markets.matrix=matrix(unlist(markets.common),length(dates),length(market.names))

```

Listing 1: Markets Matrix calculation code

D.2 RETURNS CODE

```

1  # Copyright (C) 2013-2014 José Miguel Salgado <portosalgado@gmail.com>
   #
   # This program is free software; you can redistribute it and/or modify it under the terms of
   #the GNU General Public License as published by the Free Software Foundation; either version
5  #2 of the License or (at your option) any later version.
   #
   # This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY;
   #without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
   #See the GNU General Public License for more details.
10  #
   # You should have received a copy of the GNU General Public License along with this program;
   #if not, write to the Free Software Foundation, Inc., 59 Temple Place - Suite 330, Boston,
   #MA 02111-1307, USA.
   #
15  #####
   ##### 1.markets returns calculation
   ret.matrix=matrix(0,length(dates)-1,length(market.names))

   for (k in 1:length(market.names)){
20     ret.matrix[,k]=diff(log(markets.matrix[,k]))
   }

   ##### 2.stocks returns calculation
25  ret.stocks.matrix=matrix(0,length(stocks.dates)-1,length(stock.names))

   for (l in 1:length(stock.names)){
       ret.stocks.matrix[,l]=diff(log(stocks.matrix[,l]))
   }
30

   ##### 3.statistics returns
   library(PerformanceAnalytics)
   table.Stats(ret.stocks.matrix[,1], ci=0.95, digits=8)

```

Listing 2: Returns calculation code

D.3 EIGENVALUES CODE

```

1  # Copyright (C) 2013-2014 José Miguel Salgado <portosalgado@gmail.com>
  #
  # This program is free software; you can redistribute it and/or modify it under the terms of
  # the GNU General Public License as published by the Free Software Foundation; either version
5  #2 of the License or (at your option) any later version.
  #
  # This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY;
  # without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
  # See the GNU General Public License for more details.
10 #
  # You should have received a copy of the GNU General Public License along with this program;
  # if not, write to the Free Software Foundation, Inc., 59 Temple Place - Suite 330, Boston,
  # MA 02111-1307, USA.
  #
15 #####
  due.dates.idx = seq(1, length(dates)-20, by=5)
  dt=1:length(due.dates.idx)

  #eigenvalues para cov.matrix
20 total.eig=list()
  idx=1

  for (k in due.dates.idx){
    cov.matrix=matrix(0,length(market.names),length(market.names))
25    cov.matrix=cov(diff(log(markets.matrix[k:(k+20),])),)
    cor.matrix=cov2cor(cov.matrix)
    total.eig[[idx]]=eigen(cor.matrix)$values
    idx=idx+1
  }
30
  max.eig12=vector("double",length(dates)/20-1)
  max.eig13=vector("double",length(dates)/20-1)

  for (k in dt){
35    max.eig12[k]=total.eig[[k]][1]/total.eig[[k]][2]
    max.eig13[k]=total.eig[[k]][1]/total.eig[[k]][3]
  }

  #eigenvalues para cov.weighted.matrix
40 R=0.9
  weight.vector=R^(20-1:20)
  total.weighted.eig = list()
  idx=1

45 for (k in due.dates.idx){
  cov.weighted.matrix=matrix(0,length(market.names),length(market.names))
    cov.weighted.matrix=cov.wt(diff(log(markets.matrix[k:(k+20),])),weight.vector)
    cor.weighted.matrix=cov2cor(cov.weighted.matrix$cov)
    total.weighted.eig[[idx]]=eigen(cor.weighted.matrix)$values
50    idx=idx+1
  }

  max.weighted.eig12=vector("double",length(dt)/20-1)
  max.weighted.eig13=vector("double",length(dt)/20-1)

```

```

55   for (k in dt){
      max.weighted.eig12[k]=total.weighted.eig[[k]][1]/total.weighted.eig[[k]][2]
      max.weighted.eig13[k]=total.weighted.eig[[k]][1]/total.weighted.eig[[k]][3]
    }
60   #eigenvalues para cov.random.matrix
   markets.random.common=list()
   markets.returns.random=list()

65   for (m in 1:length(market.names)){
      rmarket=diff(log(markets.common[[m]][dates]))
      markets.returns.random[[m]]=c(0,sample(rmarket))
      markets.random.common[[m]]=markets.common[[m]][dates[1]]*exp(cumsum(
        markets.returns.random[[m]]))
    }
70   markets.random.matrix=matrix(unlist(markets.random.common),length(dates),length(market.names
    ))

   total.random.eig=list()
   idx=1
75   for (k in due.dates.idx){
      cov.random.matrix=matrix(0,length(market.names),length(market.names))
      cov.random.matrix=cov(diff(log(markets.random.matrix[k:(k+20)],)))
      cor.random.matrix=cov2cor(cov.random.matrix)
80     total.eig[[idx]]=eigen(cor.random.matrix)$values
      idx=idx+1
    }

   max.random.eig12=vector("double",length(dates)/20-1)
85   max.random.eig13=vector("double",length(dates)/20-1)

   for (k in dt){
      max.random.eig12[k]=total.eig[[k]][1]/total.eig[[k]][3]
      max.random.eig13[k]=total.eig[[k]][1]/total.eig[[k]][2]
90   }

   #####plots
   library(zoo)
95   time.max.eig12 = zoo(max.eig12, order.by = as.Date(dates[due.dates.idx]))
   time.max.eig13 = zoo(max.eig13, order.by = as.Date(dates[due.dates.idx]))
   time.max.weighted.eig12 = zoo(max.weighted.eig12, order.by = as.Date(dates[due.dates.idx]))
   time.max.weighted.eig13 = zoo(max.weighted.eig13, order.by = as.Date(dates[due.dates.idx]))
   time.max.random.eig12 = zoo(max.random.eig12, order.by = as.Date(dates[due.dates.idx]))
100  time.max.random.eig13 = zoo(max.random.eig13, order.by = as.Date(dates[due.dates.idx]))

   ###plots
   ##plot max.eig12 vs max.weighted.eig12
   pdf(file="eig12vsweighted eig12.pdf", paper="special", width=7, height=4)
105  plot(time.max.eig12, xlab="time", ylab="max.eig12 vs max.weighted.eig12(red)",type="l",ylim=
      range(max.eig12,max.weighted.eig12))
   points(time.max.weighted.eig12, type="l", col='red')
   dev.off()

```

```

110 ##plot max.eig13 vs max.weighted.eig13
pdf(file="eig13vsweighted.eig13.pdf", paper="special", width=7, height=4)
plot(time.max.eig13, xlab="time", ylab="max.eig13 vs max.weighted.eig13(red)", type="l", ylim=
      range(max.eig13, max.weighted.eig13))
points(time.max.weighted.eig13, type="l", col='red')
dev.off()

115 ##plot max.eig13 vs max.eig12
pdf(file="eig13vseig12.pdf", paper="special", width=7, height=4)
plot(time.max.eig13, xlab="time", ylab="max.eig13 vs max.eig12(red)", type="l", ylim=range(
      max.eig13, max.eig12))
points(time.max.eig12, type="l", col='red')
dev.off()

120 ##plot max.eig12 vs max.random.eig12
pdf(file="eig12vsrandomeig12.pdf", paper="special", width=7, height=4)
plot(time.max.eig12, xlab="time", ylab="max.eig12 vs max.random.eig12(red)", type="l", ylim=
      range(max.eig12, max.random.eig12))
points(time.max.random.eig12, type="l", col='red')
125 dev.off()

##plot max.eig13 vs max.random.eig13
pdf(file="eig13vsrandomeig13.pdf", paper="special", width=7, height=4)
plot(time.max.eig13, xlab="time", ylab="max.eig13 vs max.random.eig13(red)", type="l", ylim=
      range(max.eig13, max.random.eig13))
130 points(time.max.random.eig13, type="l", col='red')
dev.off()

```

Listing 3: Eigenvalues calculation code

D.4 APPROXIMATE ENTROPY CODE

```

1 # Copyright (C) 2013-2014 José Miguel Salgado <portosalgado@gmail.com>
#
# This program is free software; you can redistribute it and/or modify it under the terms of
# the GNU General Public License as published by the Free Software Foundation; either version
5 #2 of the License or (at your option) any later version.
#
# This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY;
# without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
# See the GNU General Public License for more details.
10 #
# You should have received a copy of the GNU General Public License along with this program;
# if not, write to the Free Software Foundation, Inc., 59 Temple Place - Suite 330, Boston,
# MA 02111-1307, USA.
#
15 #####
##### apen.total calculation
apen.total=vector()
library(pracma)

20 for(i in 1:length(market.names)){
      apen.total[i]=approx_entropy(diff(log(markets.matrix[,i])), edim=2,
      r=0.2*sd(markets.matrix[,i]), elag=1)
}

```

```

}
25 #####
##### apen.slidewind calculation
##sliding window
due.dates.idx = seq (1, length(dates)-120, by=5)
dt=1:length(due.dates.idx)
30
##calculate ApEn for markets.matrix

library(pracma)
markets.matrix.apen=matrix(0,length(due.dates.idx),length(market.names))
35
idx=1
for (k in due.dates.idx){
  window.matrix=(diff(log(marks.matrix[k:(k+120),])))
  for(i in 1:length(market.names)){
40     marks.matrix.apen[idx,i]=approx_entropy(window.matrix[,i], edim=2,
                                             r=0.2*sd(window.matrix[,i]), elag=1)
  }
  idx=idx+1
}
45
##### plots

library(zoo)

50 for(i in 1:length(market.names)){
  time=zoo(marks.matrix.apen[,i], order.by = as.Date(dates[due.dates.idx]))
  pdf(file=paste("ApEn_",market.names[i],".pdf",sep=""), paper="special", width=7, height=4)
  plot(time, xlab="time", ylab=paste("ApEn_",market.names[i],sep=""), type="l")
  dev.off()
55 }

```

Listing 4: Approximate Entropy calculation code

D.5 DISTANCE CORRELATION CODE

```

1 # Copyright (C) 2013-2014 José Miguel Salgado <portosalgado@gmail.com>
#
# This program is free software; you can redistribute it and/or modify it under the terms of
#the GNU General Public License as published by the Free Software Foundation; either version
5 #2 of the License or (at your option) any later version.
#
# This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY;
#without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
#See the GNU General Public License for more details.
10 #
# You should have received a copy of the GNU General Public License along with this program;
#if not, write to the Free Software Foundation, Inc., 59 Temple Place - Suite 330, Boston,
#MA 02111-1307, USA.
#
15 #####
##sliding window

```

```

due.dates.idx = seq (1, length(dates)-20, by=5)
dt=1:length(due.dates.idx)

20 ##calculate dcor for markets.matrix

total.dcor=list()
total.dcor.obj=list()

25 markets.matrix.dcor=matrix(0,length(market.names),length(market.names))

library(energy)

idx=1

30 for (k in due.dates.idx){
  window.matrix=(diff(log(markets.matrix[k:(k+20),])))
  for(i in 1:length(market.names)){
    markets.matrix.dcor[i,i]=1
35     for (j in min(i+1,length(market.names)-1):length(market.names)){
       markets.matrix.dcor[i,j]=dcor(window.matrix[,i],window.matrix[,j])
       markets.matrix.dcor[j,i]=markets.matrix.dcor[i,j]
     }
  }
40 total.dcor[[idx]]=markets.matrix.dcor
total.dcor.obj[[idx]]=markets.matrix.dcor[22,23]
idx=idx+1
}

45 #####plots
z=unlist(total.dcor.obj)
library(zoo)
time = zoo(z, order.by = as.Date(dates[due.dates.idx]))

50 ##plot total.dcor
pdf(file="totaldcor.STOXXSTRAITS_20.pdf", paper="special", width=7, height=4)
plot(time, xlab="time", ylab="dcor.STOXXSTRAITS",type="l")
dev.off()

```

Listing 5: Distance Correlation calculation code

D.6 PLOTS CODE

```

1 # Copyright (C) 2013-2014 José Miguel Salgado <portosalgado@gmail.com>
#
# This program is free software; you can redistribute it and/or modify it under the terms of
# the GNU General Public License as published by the Free Software Foundation; either version
5 #2 of the License or (at your option) any later version.
#
# This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY;
# without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
# See the GNU General Public License for more details.
10 #
# You should have received a copy of the GNU General Public License along with this program;
# if not, write to the Free Software Foundation, Inc., 59 Temple Place - Suite 330, Boston,

```

```

#MA 02111-1307, USA.
#
#####
15 ##### 1.plot markets and returns
library(zoo)
time.markets.common = zoo(markets.common[[19]][dates], order.by = as.Date(dates))

20 pdf(file="psi-20.pdf", paper="special", width=7, height=4)
plot(time.markets.common, xlab="time",
      ylab="index values",type="l",ylim=range(markets.matrix[,19]))
points(ret.matrix[,19], type="l", col='red')
dev.off()

25

##### 2.plot markets
vect=numeric(length(market.names))
ret.total.matrix=rbind(ret.matrix, vect)

30 library(lattice)
time.markets = zoo(markets.common[[1]][dates], order.by = as.Date(dates))
z=zoo(cbind(time.markets,ret.total.matrix[,1]))
xyplot(z,xlab="Year",col=list(1,4),las=1,
35       ylab=("Returns value      Index value"),
       main="Netherlands (AEX Index)",
       strip=strip.custom(bg="gray75",factor.levels=c("Index","Returns"),
                          par.strip.text=list(font=2)))

40

##### 3.plot returns
library(zoo)
time.markets.common = zoo(markets.common[[19]][dates], order.by = as.Date(dates))

45 pdf(file="psi20returns.pdf", paper="special", width=7, height=4)
plot(ret.matrix[19], xlab="time",
      ylab="psi-20 returns",type="l",ylim=range(ret.matrix[,19]))
dev.off()

50

##### 4.plot stocks
stock.vector=numeric(length(stock.names))
ret.total.stocks.matrix=rbind(ret.stocks.matrix, stock.vector)

55 library(lattice)
time.stocks = zoo(stocks.common[[12]][stocks.dates], order.by = as.Date(stocks.dates))
z=zoo(cbind(time.stocks,ret.total.stocks.matrix[,12]))

xyplot(z,xlab="Year",col=list(1,4),las=1,
60       ylab=("Returns value      Stock value"),
       main="Zon Multimédia (ZON)",
       strip=strip.custom(bg="gray75",factor.levels=c("Close Values","Returns"),
                          par.strip.text=list(font=2)))

65

##### 5.plot markets, cycles and events
## http://www.nber.org-cycles.html
cycles.dates<-c("1857-06/1858-12",
                "1860-10/1861-06",

```

```

70         "1865-04/1867-12",
          "1869-06/1870-12",
          "1873-10/1879-03",
          "1882-03/1885-05",
          "1887-03/1888-04",
75         "1890-07/1891-05",
          "1893-01/1894-06",
          "1895-12/1897-06",
          "1899-06/1900-12",
          "1902-09/1904-08",
80         "1907-05/1908-06",
          "1910-01/1912-01",
          "1913-01/1914-12",
          "1918-08/1919-03",
          "1920-01/1921-07",
85         "1923-05/1924-07",
          "1926-10/1927-11",
          "1929-08/1933-03",
          "1937-05/1938-06",
          "1945-02/1945-10",
90         "1948-11/1949-10",
          "1953-07/1954-05",
          "1957-08/1958-04",
          "1960-04/1961-02",
          "1969-12/1970-11",
95         "1973-11/1975-03",
          "1980-01/1980-07",
          "1981-07/1982-11",
          "1990-07/1991-03",
          "2001-03/2001-11",
100        "2007-12/2009-06"
          #           "2001-03/2002-10",
          #           "2007-10/2009-03"
    )

105 # Events list
    #risk.dates=c("2000-03-11", "2001-09-11", "2007-10-31")
    #risk.labels=c("dotcom", "terror", "credit")
    risk.dates=c("2005-09-11", "2007-10-31")
    risk.labels=c("terror", "credit")
110 #risk.dates=c("2005-12-08", "2007-08-09", "2008-02-17", "2008-09-07", "2008-09-15", "2010-04-23",
    # "2010-11-21", "2011-04-06", "2012-06-27", "2012-06-27")
    #risk.labels=c("ECB first warning", "global liquidity shortage", "Northern Rock (UK) goes
        public",
        #"Fannie Mae and Freddie MacLB Bankruptcy", "Greece financial support", "Ireland financial
        support",
        #"Portugal financial support", "Spain financial support",
115 #"Cyprus financial support")
    # Markets
    market=list()
    library(xts)
    library(xtsExtra)
120 library(PerformanceAnalytics)
    library(zoo)
    for (m in 1:length(market.names)){
        time.markets.common = zoo(markets.common[[m]][dates], order.by = as.Date(dates))
        market[[m]]=time.markets.common
    }

```

```

125 }

chart.TimeSeries(market[[23]], main="STRAITS index",ylab="Close value", colorset="darkblue",
                period.areas=cycles.dates, period.color="lightblue")

```

Listing 6: Plots representation code

D.7 KULLBACK-LEIBLER DIVERGENCE CODE

```

1  # Copyright (C) 2013-2014 José Miguel Salgado <portosalgado@gmail.com>
  #
  # This program is free software; you can redistribute it and/or modify it under the terms of
  #the GNU General Public License as published by the Free Software Foundation; either version
5  #2 of the License or (at your option) any later version.
  #
  # This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY;
  #without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
  #See the GNU General Public License for more details.
10 #
  # You should have received a copy of the GNU General Public License along with this program;
  #if not, write to the Free Software Foundation, Inc., 59 Temple Place - Suite 330, Boston,
  #MA 02111-1307, USA.
  #
15 #####
  ##### estimates KL Divergence

  library(entropy)

20 ##sliding window
  due.dates.idx = seq (1, length(dates)-20, by=5)
  dt=1:length(due.dates.idx)

  ##calculate KL for markets.matrix
25
  total.KL=list()
  total.KL.obj=list()

  KL_matrix=matrix(0,length(market.names),length(market.names))
30

  idx=1

  for (k in due.dates.idx){
35   window.matrix=(diff(log(markets.matrix[k:(k+20),])))
   for(i in 1:length(market.names)){
     KL_matrix[i,i]=1
     for (j in min(i+1,length(market.names)-1):length(market.names)){
       KL_matrix[i,j]=KL.Dirichlet(window.matrix[,i], window.matrix[,j],
40         1/2, 1/2)
       KL_matrix[j,i]=KL_matrix[i,j]
     }
   }
   total.KL[[idx]]=KL_matrix
45   total.KL.obj[[idx]]=KL_matrix[6,7]

```

```

    idx=idx+1
  }

  z=unlist(total.KL.obj)
50 library(zoo)
   time = zoo(z, order.by = as.Date(dates[due.dates.idx]))

  ##plot total.KL
  pdf(file="KL.CACDAX_20.pdf", paper="special", width=7, height=4)
55 plot(time, main="CAC_DAX KL_Divergence", xlab="time", ylab="KL.AEXPSI", type="l")
  dev.off()

```

Listing 7: Kullback-Leibler Divergence calculation code

D.8 MUTUAL INFORMATION CODE

```

1  # Copyright (C) 2013-2014 José Miguel Salgado <portosalgado@gmail.com>
   #
   # This program is free software; you can redistribute it and/or modify it under the terms of
   # the GNU General Public License as published by the Free Software Foundation; either version
5  # 2 of the License or (at your option) any later version.
   #
   # This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY;
   # without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
   # See the GNU General Public License for more details.
10  #
   # You should have received a copy of the GNU General Public License along with this program;
   # if not, write to the Free Software Foundation, Inc., 59 Temple Place - Suite 330, Boston,
   # MA 02111-1307, USA.
   #
15  #####
   ##### estimates Mutual Information

   library(entropy)

20  ##sliding window
   due.dates.idx = seq(1, length(dates)-20, by=5)
   dt=1:length(due.dates.idx)

   ##calculate MI for markets.matrix
25  total.MI=list()
   total.MI.obj=list()

   MI_matrix=matrix(0,length(market.names),length(market.names))
30
   idx=1

   for (k in due.dates.idx){
35     window.matrix=(diff(log(markets.matrix[k:(k+20),])))
       for(i in 1:length(market.names)){
           MI_matrix[i,i]=1
           for (j in min(i+1,length(market.names)-1):length(market.names)){

```

```

    adj=rbind(window.matrix[,i], window.matrix[,j])
40    MI_matrix[i,j]=mi.Dirichlet(adj, 1/2)
    MI_matrix[j,i]=MI_matrix[i,j]
  }
}
total.MI[[idx]]=MI_matrix
45 total.MI.obj[[idx]]=MI_matrix[22,23]
    idx=idx+1
}

z=unlist(total.MI.obj)
50 library(zoo)
time = zoo(z, order.by = as.Date(dates[due.dates.idx]))

##plot total.MI
pdf(file="MI.STOXXSTRAITS_20.pdf", paper="special", width=7, height=4)
55 plot(time, main="STOXX_STRAITS Mutual Information", xlab="time", ylab="MI.STOXXSTRAITS",type
      ="l")
dev.off()

```

Listing 8: Mutual Information calculation code

D.9 FORECA CODE

```

1  # Copyright (C) 2013-2014 José Miguel Salgado <portosalgado@gmail.com>
  #
  # This program is free software; you can redistribute it and/or modify it under the terms of
  #the GNU General Public License as published by the Free Software Foundation; either version
5  #2 of the License or (at your option) any later version.
  #
  # This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY;
  #without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
  #See the GNU General Public License for more details.
10 #
  # You should have received a copy of the GNU General Public License along with this program;
  #if not, write to the Free Software Foundation, Inc., 59 Temple Place - Suite 330, Boston,
  #MA 02111-1307, USA.
  #
15 #####
  #####analyse ForeCA

  library(ForeCA)

20 YY= ts(diff(log(markets.matrix)))
  #plot(ts(YY))
  ff=foreca(YY, n.comp=2)
  plot(ff)

```

Listing 9: Forecastable Component Analysis calculation code

D.10 MARCHENKO-PASTUR CODE

```

1 ##### Markets Marchenko-Pastur
  due.dates.idx = seq (1, length(stocks.dates)-1000, by=100)
  dt=1:length(due.dates.idx)
  dtotal=2:length(stock.names)

5  #eigenvalues para cov.matrix
  total.eig=list()
  total.eig.norm=list()
  idx=1

10 max.eig12=vector("double",length(stocks.dates)/1000-1)
  max.eig13=vector("double",length(stocks.dates)/1000-1)
  eig=vector("double",length(stocks.dates)/1000-1)
  total=vector("double",length(stocks.dates)/1000-1)

15 for (k in due.dates.idx){
  cov.matrix=matrix(0,length(stock.names),length(stock.names))
  cov.matrix=cov(diff(log(stocks.matrix[k:(k+20),])))
  cor.matrix=cov2cor(cov.matrix)
20  total.eig[[idx]]=eigen(cor.matrix)$values
  # total.eig[[idx]]=eigen(cov.matrix)$values
  idx=idx+1
}

25 for (k in dt){
  soma=0
  for (j in dtotal){
    soma=total.eig[[k]][j]+soma
  }
30 total[k]=soma
  total.eig.norm[[k]]=total.eig[[k]]/soma
}

all.eig.norm=unlist(total.eig.norm)
35 ###less.eig.norm=all.eig.norm(x<4)

###plot density
T=20
N=12
40 Q=T/N
q=1/Q
x=seq(0.24,6.2,0.001)

#calculate marcenko-pastur
45 #library(RMTstat)
#plot(x,dmp(x,ndf=N-1,pdim=(N-1)/Q))

#another way
#x=seq(0.0,6.5,0.001)
50 mp=function(x,q) {return(sqrt(4*x*q-(x+q-1)^2)/(2*pi*x*q))}

#calculate my eigenvalues
all.eig=unlist(total.eig)

55 h=hist(all.eig,plot=FALSE,nclass=100)
plot(x,mp(x,1/Q))

```

```
lines(h$mids,h$density)
```

Listing 10: Marchenko-Pastur calculation code

BIBLIOGRAPHY

- Rules for psi-20 weights. <http://www.euronext.pt/bvlp/files/pubs/calcpsien.pdf>, 2003. (Cited on page 57.)
- A. Abhyankar, L.S. Copeland, and W. Wong. Uncovering nonlinear structure in real-time stock-market indexes: The s&p 500, the dax, the nikkei 225, and the ftse-100. *Journal of Business & Economic Statistics, American Statistical Association*, 15(1):1–14, January 1997. (Cited on page 13.)
- S. Amari, A. Cichoki, and H.H. Yang. A new learning algorithm for blind signal separation. *Advances in Neural Information Processing Systems*, pages 757–763, 1996. (Cited on pages 31, 32, and 37.)
- P.A. Ammermann and D.M. Patterson. The cross-sectional and cross-temporal universality of nonlinear serial dependencies: Evidence from world stock indices and the taiwan stock exchange. *Pacific-Basin Finance Journal, Elsevier*, 11(2):175–195, April 2003. (Cited on page 13.)
- T. Araújo and F. Louçã. Complex behavior of stock markets: process of synchronization and desynchronization during crises. In *Perspectives on Econophysics*. Universidade de Évora - Portugal, 2006. (Cited on page 21.)
- M. Ausloos. Financial time series and statistical mechanics. *arXiv:cond-mat/0103068*, 2001. (Cited on page 6.)
- M. Ausloos. Econophysics of stock and foreign currency exchange markets. *arXiv:physics/0606012*, 2006. (Cited on page 47.)
- L. Bachelier. Théorie de la Spéculation. *Ann. Sci. Ecole Norm. S.*, III(17):21–86, 1900. (Cited on pages 3, 10, 11, and 19.)
- A.D. Back and A.S. Weigend. A first application of independent component analysis to extracting structure from stock returns. *International Journal of Neural Systems*, 8, 1997. (Cited on pages 29 and 31.)
- N. K. Bakirov, M. L. Rizzo, and Székely. A multivariate nonparametric test of independence. *J. Multivariate Anal.*, 93:1742–1756, 2006. (Cited on page 39.)
- P. Baldi and K. Hornik. Neural networks and principal component analysis: learning from examples without local minima. *Neural Networks*, 2:53–58, 1989. (Cited on page 30.)
- P. Ball. Culture Crash. *Nature*, 441:686–688, 2006. (Cited on page 14.)
- M. Bartolozzi, D.B. Leinweber, and A.W. Thomas. Scale-free avalanche dynamics in the stock market, 2006. URL <http://www.citebase.org/cgi-bin/citations?id=oai:arXiv.org:physics/0601171>. (Cited on page 47.)

- A. Beattie. Market crashes, 2013. URL www.investopedia.com. (Cited on page 16.)
- A.J. Bell and T.J. Sejnowski. An information maximisation approach to blind source separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995. (Cited on pages 31 and 32.)
- A. Belouchrani, K. Abed-Meraim, J.F. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing*, 45 (2):434–444, 1997. (Cited on page 32.)
- S.R. Bentes. Econophysics: a new discipline. *Science and Culture*, 76, 2010. (Cited on page 5.)
- F. Black and M. Scholes. The pricing of options and corporate liabilities. *J. Polit. Econ.*, 81:637–659, 1973. (Cited on pages 4, 5, and 12.)
- T. Bollerslev, R.F. Engle, and D.B. Nelson. Arch models. *Handbook of econometrics*, 4: 2959–3038, 1994. (Cited on page 12.)
- G. Bonanno, F. Lillo, and R.N. Mantegna. Levels of complexity in financial markets. *Physica A: Statistical Mechanics and its Applications*, 299 (1):16–27, 2001. (Cited on page 46.)
- J.P. Bouchaud and M. Potters. *Theory of Financial Risks: from Statistical Physics to Risk Management*. Cambridge University Press, Cambridge, 2003. (Cited on pages 13, 22, 24, and 26.)
- J.P. Bouchaud and M. Potters. Financial applications of random matrix theory: a short review. *The Oxford Handbook of Random Matrix Theory*, Oxford University Press, Part III, number 40, 2011. (Cited on pages 25, 27, 29, 47, and 63.)
- G.E.P. Box and G.C. Tiao. A canonical analysis of multiple time series. *Biometrika*, 64 (2): 355–365, 1977. (Cited on page 29.)
- L. Calvet and A. Fisher. Multifractality in asset returns: Theory and evidence. *The Review of Economics and Statistics*, 84(3):381–406, 2002. (Cited on pages 13 and 103.)
- J.F. Cardoso. Blind identification of independent components with higher-order statistics. *Proc. Workshop on Higher-Order Spect. Anal.*, pages 157–160, 1989. (Cited on page 32.)
- J.F. Cardoso and A. Souloumiac. An efficient technique for blind separation of complex sources. *Proc. IEEE SP Workshop on Higher-Order Stat.*, pages 275–279, 1993. (Cited on page 32.)
- A. Chakarborti, M. Patriarca, and M.S. Santhanam. Financial time series analysis: a brief overview. *Econophysics of Markets and Business Networks: Proceedings of the Econophysics-Kolkata III*, pages 51–68, 2007. (Cited on page 19.)
- A. Chakarborti, I.M. Toke, M. Patriarca, and F. Abergel. Econophysics review i: Empirical facts. *Quantitative Finance*, 11:991–1012, 2011. (Cited on page 3.)
- C. Chatfield. *The Analysis of Time Series: An Introduction*. Chapman & Hall, 6th edition, 2003. (Cited on page 10.)

- P. Common. Independent component analysis. a new concept? *Signal Processing*, 36: 287–314, 1994. (Cited on pages 30, 31, and 32.)
- R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1:223–236, 2001. (Cited on page 12.)
- R. Cont, M. Potters, and J.P. Bouchaud. Scaling in stock market data: stable laws and beyond. *arXiv: cond-mat/9705087*, 1997. (Cited on page 12.)
- T. Di Matteo, T. Aste, and Michel M. Dacorogna. Using the scaling analysis to characterize financial markets. *Journal of Banking & Finance*, 29:827–851, 2005. (Cited on page 12.)
- T. Di Matteo, F. Pozzi, and T. Aste. The use of dynamical networks to detect the hierarchical organization of the financial markets sectors. *Eur Phys J B*, 73(1):3–11, 2010. (Cited on page 24.)
- Z. Ding, C.W.J. Granger, and R. Engle. A long memory property of stock returns and a new model. *Journal of Empirical Finance*, 1:83–106, 1993. (Cited on page 44.)
- A. Dionisio, R. Menezes, and D.A. Mendes. An econophysics approach to analyse uncertainty in financial markets: an application to the portuguese stock market. *The European Physical Journal B - Condensed Matter and Complex Systems*, 50:161–164, 2006. (Cited on page 31.)
- P. Doukhan, G. Oppenheim, and M.S. Taqqu, editors. *Theory and Applications of Long-Range Dependence*. Birkhäuser, 2003. (Cited on page 44.)
- S. Drozd, J. Kwapien, and P. Oswiecimka. Empirics versus rmt in financial cross-correlations. *Acta Physica Polonica, B*, 58:4027–4039, 2007. (Cited on page 28.)
- J.P. Eckman and D. Ruelle. Ergodic theory of chaos and strange attractors. *Review of Modern Physics*, 57(3):617–656, 1985. (Cited on page 38.)
- A. Einstein. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Ann. Phys-Berlin*, 17:549–560, 1905. (Cited on pages 3 and 19.)
- P. Embrechts. Copulas: a personal view. *Journal of Risk and Insurance*, 76:639–650, 2009. (Cited on page 47.)
- P. Embrechts, A. McNeil, and D. Straumann. Correlation and dependence in risk management: properties and pitfalls. In M. Dempster, editor, *Risk Management: Value at Risk and Beyond*, pages 176–223. Cambridge University Press, 2002. (Cited on pages 24 and 39.)
- P. Erdős and A. Rényi. On Random Graphs I. *Publicationes Mathematicae*, 6:290–297, 1959. (Cited on page 46.)
- E.F. Fama. *J. Business*, 38, 1965. (Cited on page 3.)

- E.F. Fama. Efficient capital markets: A review of theory and empirical work. *J. Financ.*, 25:383–417, 1970. (Cited on pages 4 and 12.)
- W. Feller. *An Introduction to Probability Theory and its Applications*. John Wiley & Sons, Inc., third edition edition, 1968. (Cited on page 11.)
- D.J. Fenn, M.A. Porter, S. Williams, M. McDonald, N.F. Johnson, and N.S. Jones. Temporal evolution of financial-market correlations. *Physical Review E*, 84, 2011. (Cited on pages 24, 48, and 60.)
- K. Fergusson and E. Platen. On the distributional characterization of daily log-returns of a world stock index. *Applied Mathematical Finance*, 13:01:19–38, 2006. (Cited on page 59.)
- A. Feuerverger. A consistent test for bivariate dependence. *International Statistical Review*, 61 (2):419–433, 1993. (Cited on page 39.)
- G. Fraham and U. Jaekel. Random matrix theory and robust covariance matrix estimation for financial data. ??, ??, 2008. (Cited on page 23.)
- J.H. Friedman and J.W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, 23 (9):881–890, 1974. (Cited on page 37.)
- X. Gabaix, Gopikrishnan P., V. Plerou, and H. Stanley. A theory of power-law distributions in financial market fluctuations. *Nature*, 423:267–270, 2003. (Cited on page 13.)
- S. Galluccio, J.P. Bouchaud, and M. Potters. Rational decisions, random matrices and spin glasses. *Physica A*, 259:449–456, 1998. (Cited on page 27.)
- G.M. Goerg. Forecastable component analysis. *Journal of Machine Learning Research (JMLR) W&CP*, 28 (2):64–72, 2013. (Cited on pages 32, 33, 66, and 80.)
- L.M.P. Gomes. *Memória de Longo Prazo nos Retornos Acionistas dos Índices de Referência da Euronext, Implicações para a Hipótese de Mercados Eficientes e Contributo Fractal para Aperfeiçoamento do Capital Asset Pricing Model*. Universidade Portucalense, 2012. (Cited on pages 46 and 102.)
- P. Gopikrishnan, V. Plerou, L.A.N. Amaral, and H.E. Stanley. Scaling of the distribution of fluctuations of financial market indices. *Physical Review E*, 60:5305–5316, 1999. (Cited on pages 22 and 28.)
- A.C. Harvey. Long memory in stochastic volatility. Research Report 10, London School of Economics, 1993. (Cited on page 44.)
- J. Heraut and C. Jutten. Space or time adaptive signal preprocess by neural network models. *Neural Networks for Computing*, 151(1):206–211, 1986. (Cited on pages 30 and 32.)
- T. Higushi. Approach to an irregular time series on the basis of the fractal theory. *Physica D*, pages 277–283, 1988. (Cited on page 13.)

- K.K.L. Ho, G.B. Moody, C.K. Peng, J.E. Mietus, M.G. Larson, D. Levy, and A.L. Goldberger. Predicting survival in heart failure case and control subjects by use of fully automated method for deriving nonlinear and conventional indices of heart rate dynamics. *Circulation*, 96 (3):842–848, 1997. (Cited on page 38.)
- P.J. Huber. What is projection pursuit? *Journal of the Royal Statistical Society*, 13 (2): 435–475, 1985. (Cited on page 37.)
- H.E. Hurst. Long-term storage capacity of reservoirs. *Trans. Am. Soc. Civ. Eng.*, 116: 770–808, 1951. (Cited on page 45.)
- C. Jutten and J. Herault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991. (Cited on page 30.)
- N. Kaldor. A model of economic growth. *The Economic Journal*, 67 (268):591–624, 1957. (Cited on page 12.)
- J.W. Kantelhardt, E. Koscielny-Bunde, H.A. Rego, S. Havlin, and A. Bunde. Detecting long-range correlations with detrended fluctuation analysis. *Physica A*, 295:441–454, 2001. (Cited on page 46.)
- H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, second edition, 2004. (Cited on pages 34 and 35.)
- D.E. Knuth. *The TeXbook*. Addison-Wesley, 1984. (Cited on page 49.)
- A.N. Kolmogorov. A new invariant of transitive dynamical systems. *Dokl. Akad. Nauk. SSSR*, 119:861, 1958. (Cited on page 36.)
- I. Koponen. Analytical approach to the problem of convergence of truncated Lévy flights towards the Gaussian stochastic process. *Physical Review Letter E*, 52:1197, 1995. (Cited on page 3.)
- S. Kullback and R.A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, 1951. (Cited on pages 37 and 38.)
- J. Kwapień, P. Oświecimka, and S. Drożdż. The bulk of the stock market correlation matrix is not pure noise. *Physica A*, 359:589–606, 2005. (Cited on page 28.)
- L. Laloux, P. Cizeau, J.P. Bouchaud, and M. Potters. Noise Dressing of Financial Correlation Matrices. *Physical Review Letters*, 83(7):1467–1470, 1999. (Cited on page 28.)
- L. Laloux, P. Cizeau, and M. Potters. Random matrix theory and financial correlations. *International Journal of Theoretical and Applied Finance*, 3(3):391–397, 2000. (Cited on pages 6, 24, and 28.)
- L. Lamport. *LaTeX: A Document Preparation System*. Addison-Wesley, 1986. (Cited on page 49.)
- J. Lee and H.E. Stanley. Phase transition in the multifractal spectrum of diffusion-limited aggregation. *Physical Review Letters*, 61(26):2945–2948, Dec 1988. doi: 10.1103/PhysRevLett.61.2945. (Cited on page 47.)

- F. Lillo and R.N. Mantegna. Power-law relaxation in a complex system: Omori law after a financial market crash. *Physical Review E*, 68, 2003. (Cited on page 47.)
- J.K. Lindsey. *Statistical Analysis of Stochastic Processes in Time*. Number 14 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2004. (Cited on page 19.)
- R. Litterman and K. Winkelmann. *Estimating Covariance Matrices*. Goldman-Sachs Risk Management Series. Goldman, Sachs and Co., 1998. (Cited on page 24.)
- A. Lo. Long-Term memory in stock market prices. *Econometrica*, 59:1279–1313, 1991. (Cited on page 44.)
- T. Lux. Detecting Multi-Fractal Properties in Asset Returns: An Assessment of the ‘Scaling Estimator’. *International Journal of Modern Physics*, 15:481 – 491, 2004. (Cited on pages 13 and 101.)
- E. Maasoumi and J. Racine. Entropy and predictability of stock markets returns. *Journal of Econometrics*, 107:291–312, 2002. (Cited on page 34.)
- E. Majorana. *Scientia*, 36:58, 1942. (Cited on page 2.)
- B.B. Mandelbrot. The variation of certain speculative prices. *J. Bus.*, XXXVI(4):394–419, 1963. (Cited on pages 3, 11, and 13.)
- B.B. Mandelbrot. *Statistical Models and turbulence: Possible refinements of the lognormal hypothesis concerning the distribution of energy dissipation in intermittent turbulence*. Springer Verlag (New York), 1972. (Cited on page 47.)
- B.B. Mandelbrot. *Fractals: Form, Chance and Dimension*. W H Freeman and Co, 1977. (Cited on page 4.)
- B.B. Mandelbrot. *The Fractal Geometry of Nature*. W H Freeman and Co, 1982. (Cited on page 4.)
- B.B. Mandelbrot and J.W. Van Ness. Fractional brownian motion, fractional noises and applications. *SIAM Review*, 10:422, 1968. (Cited on page 45.)
- B.B. Mandelbrot, A.J. Fisher, and L.E. Calvet. A Multifractal Model of Asset Returns. Cowles Foundation Discussion Paper 1164, 1997. Available at SSRN: <http://ssrn.com/abstract=78588>. (Cited on page 47.)
- R.N. Mantegna. Presentation of the english translation of etto majorana’s paper: The value of statistical laws in physics and social sciences. *Quant*, 5:133–140, 2005. (Cited on page 2.)
- R.N. Mantegna. The tenth article of etto majorana. *Europhysics News*, 37:15–17, 2006. (Cited on page 2.)
- R.N. Mantegna and H.E. Stanley. Stochastic process with ultraslow convergence to a gaussian: the truncated lévy flight. *Physical Review Letter*, 73:2946, 1994. (Cited on page 3.)

- R.N. Mantegna and H.E. Stanley. Scaling behaviour in the dynamics of an economic index. *Nature*, 376:46 – 49, 1995. (Cited on page 12.)
- R.N. Mantegna and H.E. Stanley. Turbulence and financial markets. *Nature*, 383:587–588, 1996. (Cited on page 47.)
- R.N. Mantegna and H.E. Stanley. Stock market dynamics and turbulence: parallel analysis of fluctuation phenomena. *Physica A: Statistical Mechanics and its Applications*, 239: 255–266, 1997. (Cited on page 47.)
- R.N. Mantegna and H.E. Stanley. *An Introduction to Econophysics: Correlations and Complexity in Finance*. Cambridge University Press, Cambridge, 2000. (Cited on pages 2, 13, and 47.)
- V.A. Marchenko and L.A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mat. Sb.*, 72(114):507–536, 1967. (Cited on pages 21, 25, 63, and 77.)
- J.A.O. Matos. *Entropy Measures Applied to Financial Time Series - an Econophysics Approach*. Departamento de Matematica Aplicada, Universidade do Porto, 2006. (Cited on pages 97, 98, 99, and 102.)
- J.A.O. Matos, S.M.A. Gama, H.J. Ruskin, and J.A.M.S. Duarte. An econophysics approach to the portuguese stock index, psi-20. *Physica A*, 342(3-4):665–676, 2004. (Cited on page 102.)
- J.A.O. Matos, S.M.A. Gama, H.J. Ruskin, A. Sharkasi, and M. Crane. Correlation of worldwide markets entropies. *Proceedings of the Workshop: Perspectives on Econophysics*, 259:449–456, 2006. (Cited on pages 11, 24, and 102.)
- J. McCauley. Thermodynamics analogies in economics and finance: instabilities of markets. *Physica A*, 329:199–212, 2003. (Cited on page 34.)
- R.V. Mendes, T. Araújo, and F. Louçã. Reconstructing an economic space from a market metric. *Physica A*, 323:635–650, 2003. (Cited on page 21.)
- I Meric and G Meric. Co-movements of european markets before and after the 1987 crash. *Multinational Finance Journal*, 1:137–152, 1997. (Cited on page 30.)
- J. Moody and L. Wu. What is the "true price"? In Berlin Springer, editor, *State Space Models for High Frequency Financial Data. Progress in Neural Information Processing (ICONIP'96)*, pages 697–704, 1996. (Cited on page 31.)
- M.E.J. Newman. The structure and function of networks. *SIAM Review*, 45:167–256, 2003. (Cited on page 46.)
- J.P. Nolan. *Lévy Processes: Theory and Applications*, chapter Maximum likelihood estimation of stable parameters, pages 379–400. Boston: Birkhäuser, 2001. (Cited on page 3.)
- J.P. Nolan. *Stable Distributions - Models for Heavy Tailed Data*. Boston: Birkhäuser, 2006. (Cited on page 13.)

- E. Oja. Neural networks, principal components and subspaces. *International Journal of Neural Systems*, 1:61–68, 1989. (Cited on page 30.)
- J.P. Onnela, A. Chakraborti, K. Kaski, J. Kertész, and A. Kanto. Dynamics of market correlations: Taxonomy and portfolio analysis. *Phys. Rev. E*, 68, 2003. (Cited on page 48.)
- M.F.M. Osborne. Brownian motion in the stock market. *Oper. Res.*, 7:145–173, 1959. (Cited on page 12.)
- M.F.M. Osborne. *The Stock Market and Finance from a Physicist's Viewpoint*. Crossgar Press, 1977. (Cited on page 12.)
- A. Papoulis. *Probability, Random Variables and Stochastic Processes*. Mc Graw Hill, 1985. ISBN 0-07-048468-6. (Cited on pages 19 and 37.)
- V. Pareto. *Cours d'Économie Politique*. 1897. (Cited on page 3.)
- E. Parzen. *Stochastic Processes*. SIAM, 1999. (Cited on page 19.)
- D. Peña and G.E.P. Box. Identifying a simplifying structure in time series. *Journal of the American Statistical Association*, 82 (399):836–843, 1987. (Cited on page 29.)
- H.O. Peitgen, H. Jürgens, and D. Saupe. *Chaos and Fractals, New Frontiers of Science*. Springer-Verlag, 1992. (Cited on page 4.)
- C.K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, and A.L. Golderberger. On the mosaic organization of dna sequences. *Phys. Rev. E*, 49:1685–1689, 1994. (Cited on page 45.)
- J.P. Pereira and T. Cutelo. Tiny prices in a tiny market - evidence from portugal on optimal share prices. Available at SSRN: <http://ssrn.com/abstract=1728712>, 2010. (Cited on pages 53 and 54.)
- S.M. Pincus. Approximate entropy as a measure of system complexity. *Proc. Natl. Acad. Sci.*, 88:2297–2301, 1991. (Cited on page 38.)
- S.M. Pincus. Approximate entropy as an irregularity measure for financial data. *Econometric Reviews*, 27:4-6:329–362, 2008. (Cited on page 39.)
- S.M. Pincus and R.E. Kalman. Irregularity, volatility, risk, and financial market time series. *Proc. Natl. Acad. Sci.*, 101:13709–13714, 2004. (Cited on page 39.)
- V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, and H.E. Stanley. Universal and non-universal properties of cross correlations in financial time series. *Physical Review Letters*, 83(7):1471–1474, 1999. (Cited on pages 22, 27, 28, and 29.)
- V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, and H.E. Stanley. A random matrix theory approach to financial cross correlations. *Physica A*, 287:374–382, 2000. (Cited on pages 6 and 28.)
- V. Plerou, P. Gopikrishnan, and B. Rosenow. Collective behaviour of stock price movement: A random matrix approach. *Physica A*, 299:175–180, 2001. (Cited on page 28.)

- V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, and T. Guhr. Random matrix approach to cross correlations in financial time series. *Physical Review E*, 65, 2002. (Cited on pages 28 and 29.)
- S.R. Rege, J.C.A. Teixeira, and A.G. Menezes. The daily returns of the portuguese stock index: a didistribution characterization. *Journal of Risk Model Validation*, 7(4):53–70, 2013. (Cited on page 59.)
- Pierre Alain Reigneron, Romain Allez, and Je. Principal regression analysis and the index leverage effect. *Physica A*, 390:3026–3035, 2011. (Cited on page 14.)
- A. Rényi. On measures of information and entropy. In *4th Berkeley Symposium on Mathematics, Statistics and Probability*, pages 547–561, 1961. (Cited on pages 6, 34, and 35.)
- B.D. Ripley. *Pattern recognition and neural networks*. Cambridge University Press, 1996. (Cited on page 37.)
- B.M. Roehner. Patterns of speculation: a study in observational econophysics. *Journal of Economic Literature*, 42:838–840, 2004. (Cited on page 2.)
- B.M. Roehner. fifteen years of econophysics: worries, hopes and prospects. *Science and Culture*, 76, 2010. (Cited on page 2.)
- D. Ruelle. *Thermodynamic formalism. The Mathematical Structures of Equilibrium Statistical Mechanics*. Cambridge University Press, 2004. (Cited on page 34.)
- A.L. Rukhin. Approximate entropy for testing randomness. *J. Appl. Probab.*, 37:88–100, 2000. (Cited on page 39.)
- P.A. Samuelson. Mathematics of speculative prices. *SIAM Rev.*, 15:1–34, 1973. (Cited on page 12.)
- T. Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85:461, 2000. (Cited on page 6.)
- C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948. (Cited on pages 6 and 33.)
- S. Sharifi, M. Crane, A. Shamaie, and H.J. Ruskin. Random matrix theory for portfolio optimization: a stability approach. *Physica A*, 335(3-4):629–643, 2004. (Cited on page 28.)
- A. Sharkasi, M. Crane, H.J. Ruskin, and J.A.O. Matos. The reaction of stock markets to crashes and events: A comparison study between emerging and mature markets using wavelet transforms. *Physica A*, 368(2):511–521, 2006a. (Cited on pages 24 and 28.)
- A. Sharkasi, H.J. Ruskin, M. Crane, J.A.O. Matos, and S.M.A. Gama. A wavelet-based method to measure stages of stock market development. In preparation, 2006b. (Cited on page 47.)
- M. F. Shlesinger, U. Frisch, and G. Zaslavsky, editors. *Lévy Flights and Related Phenomena in Physics*. Springer, 1995. (Cited on page 3.)

- A. Silberschatz and A. Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE transactions on knowledge and data engineering*, 8:970–974, 1996. (Cited on page 37.)
- A.G. Sinai. On the concept of entropy of a dynamical system. *Dokl. Akad. Nauk. SSSR*, 124:768, 1959. (Cited on page 36.)
- D. Sornette. Predictability of catastrophic events: material rupture, earthquakes, turbulence, financial crashes and human birth. *Proc. Natl. Acad. Sci.*, 99:2522–2529, 2002. (Cited on page 47.)
- H.E. Stanley. name? *Physica A*, 224:302, 1996. (Cited on page 2.)
- H.E. Stanley. Econophysics: can physicists contribute to the science of economics? *Computing in Science & Engineering*, 1(1):74–77, 1999. (Cited on page 4.)
- J.H. Stock and M.W. Watson. Forecast using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97 (460):1167–1179, 2002. (Cited on page 29.)
- G.J. Székely and M.L. Rizzo. Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1236–1265, 2009. (Cited on pages 40 and 42.)
- G.J. Székely, M.L. Rizzo, and N.K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007. (Cited on pages 39, 42, and 44.)
- M. S. Taqqu, V. Teverovsky, and W. Willinger. Estimators for long-range dependence: An empirical study. *Fractals*, 3, No. 4:785–798, 1995. (Cited on page 44.)
- H. Theil. *Economics and Information Theory*. North- Holland Amsterdam, 1967. (Cited on page 6.)
- G. Tilak. Studies of the recurrence-time interval distribution in financial time-series data at low and high frequencies. Master’s thesis, Université Paris Dauphine, 2012. (Cited on page 51.)
- C. Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of Statistical Physics*, 52:479, 1988. (Cited on pages 6, 33, and 36.)
- C. Tsallis, C. Anteneodo, L. Borland, and R. Osorio. Nonextensive statistical mechanics and economics. *Physica A*, 324:89–100, 2003. (Cited on page 36.)
- R.S. Tsay. *Analysis of Financial Time Series*. Wiley Interscience, Hoboken, NJ, 2005. (Cited on page 10.)
- T.A. Vuorenmaa. *Proceedings of SPIE: Noise and Fluctuations in Econophysics and Finance*, Vol. 5848, chapter A Wavelet Analysis of Scaling Laws and Long-Memory in Stock Market Volatility, pages 39–54. 2005. (Cited on page 47.)
- E. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Ann. of Math.*, 62:548–564, 1955. (Cited on page 21.)

- E. Wigner. On the distribution of the roots of certain symmetric matrices. *Ann. of Math.*, 67:325–328, 1958. (Cited on page 21.)
- D. Wilcox and T. Gebbie. On the analysis of cross-correlations in South African market data. *Physica A*, 344(1-2):294–298, 2004. (Cited on page 28.)
- D. Würtz. *Rmetrics: an environment for teaching financial engineering and computational finance with R*. Rmetrics, ITP, ETH Zürich, Zürich, Switzerland, 2004. <http://www.rmetrics.org>. (Cited on page 50.)