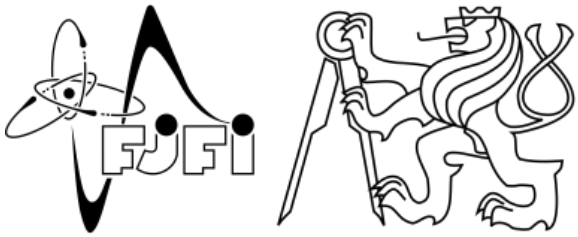
CZECH TECHNICAL UNIVERSITY IN PRAGUE
Faculty of Nuclear Sciences and Physical Engineering



Master's Thesis

Information flows and their role in complex systems

Author: Zlata Tabachová

Supervisor: Ing. Petr Jizba, PhD.

Prague 2022

Acknowledgment

I would like to thank to Ing. Petr Jizba, PhD. for supervising this work. Writing this thesis was accompanied by many immersive conversations and I'm looking forward to more of them. I'm very grateful for all new knowledge I have gained during my work on the thesis. Special thanks goes to Ing. Hynek Lavicka, PhD. for providing me with useful advice on data processing and providing me with several graphs I use in this work.

I would like to express my gratitude to all respectful colleagues from the department of physics at our faculty, whose high-quality lectures led me to realize the beauty of mathematical physics and helped me to master the skill of scientific thinking.

This work is completely dedicated to my beloved mother for constantly making impossible things possible, thus being an infinite source of inspiration. At the same time, my thanks also go to my caring siblings.

Název práce: Informační toky a jejich role v komplexních systémech
Autor: Zlata Tabachová
Obor: Matematická fyzika
Druh práce: Diplomová práce
Vedoucí práce: Ing. Petr Jizba, PhD. , Katedra fyziky, Fakulta jaderná a fyzikálně inženýrská, České vysoké učení technické v Praze

Abstrakt: Pro pochopení komplexích systémů je důležité stanovit jejich strukturu, obzvlášť příčinně následných vztahů mezi subsystémy, jinak také informační toky. Transférové entropie, určující směrově závislou míru chaosu, se osvědčily jako dobrý nástroj pro měření informačních toků mezi, obecně nelineárně provázanými systémy. V dané práci nejprve provádíme axiomatizaci zobecněných Rényiovských entropií, které vyústí do definice Rényiho transférových entropií (RTE). RTE umožňují volbou parametru vyzdvihovat určité části pravdepodobnostních distribucí. Toto je velice žádaná vlastnost v kontextu sytemů, ve kterých nás zajímají ocasní části, tedy málo pravděpodobné jevy, například ceny akcií. V praktické části nejdřív aplikujeme RTE na uměle vygenerovaná data parametricky provázaných Rösslerových systemů, a poté - na realná data z akciových trhů. Výsledky indikují, že RTE opravdu mohou kvantitativně meřit informační toky mezi závislými systémy.

Klíčová slova: Transférové entropie, informační toky, Rényiho entropie, finanční trhy, stochastické systémy, kauzalita.

Title: Information flows and their role in complex systems

Abstract: Fundamental aspect in understanding complex systems is determining its structure, especially causality dynamics between different parts of a system - *information flows*. *Transfer entropies*, directional measure of uncertainty, appeared to be a useful tool for measuring linear and non-linear information flows between or within the systems. In the following work we present an intuitive derivation of the family of α-entropies resulting with a concept of so called *Rényi transfer α-entropies*. The latter is able to selectively emphasize certain parts of probability distributions. That is a favourable property in analyzing processes, where marginal parts are the main source of the relevant information, i.e. stock prices. First, we use the derived methods to study model systems as coupled *Rössler systems*, and then we show the application on the real data from financial markets. Results show, that the Rényi transfer entropies can detect information flows between processes.

Key words: Transfer entropy, information flow, Rényi entropy, financial markets, stochastic systems, causality.

Contents

Chapter 0

Introduction

Without a doubt, we live in an era of information technologies. Obtaining data and its processing has become the core policy in many spheres of economic influence. Technologies are causing changes on governmental level, as well as in private sector and households. The rate at which these influences spread through the system is of exponential character, and, together with the globalization of human activities, systems became entwined on such a level, that it is hard to locate and identify causes and consequences. It can be scary to realize that at this rate we might not be able to predict the unwanted consequences in time. That's why it is desirable to develop sophisticated tools that will mediate hidden inner information from the *complex systems*.

Even though we can formulate many of complex systems' characteristics, such as non-linearity, emergence, adaptation, self-organization, spontaneous order among others, it is hard to formulate exact definition of it. The listed features appear in many different fields of human activities and natural phenomena, that's why complex systems' approach can be used in many diverse disciplines including statistical physics, information theory, anthropology, computer science, meteorology, sociology, biology, psychology, economics and others.

Complex dynamical system is a system composed of several subsystems (with given properties) interacting between each other and surrounding environment. These interactions bring to the whole system new properties, absent in composing parts. Very important question, that arises in this topic, is why by an "adding" process we obtain something completely different from what we started with. If we want to answer this question, we must turn our attention to the "adding" part, i.e. the way how subsystems interact within each other, how are they connected and what are the relationships between them. Therefore, a fundamental aspect in understanding a complex system

is determining its structure, especially causality dynamics between different parts of a system - *information flows*. It is desirable to know what mechanisms generate information, where is that information stored, and how is it transmitted within a system. Information flows might be possible between structures on a one scale, or from the micro- to the macro-scales. *Transfer entropies*, directional measure of ignorance, appeared to be a useful tool for measuring information flows between or within time-evolving systems.

Transfer entropy is an information-theoretic functional, originally introduced by means of Shannon entropy, and later generalized in terms of *one-parameter family of Rényi entropies*. Due to its many appealing properties, as equivalence to generalized dimensions, additivity and ability to emphasize or suppress particular parts of probability density functions. Rényi entropy and its derivative *Rényi transfer entropy* (RTE) is a good candidate to quantitatively characterize complex dynamical systems that are multi-scale and/or non-Gaussian, typically with heavy-tailed distributions, and, thus, detect information flows inside or between them.

This thesis is entitled *"Information flows and their role in data analysis"*. The aim of this work is to introduce the concept of transfer entropies in connection with all relevant fields, i.e. probability spaces, information-theoretic functionals, multifractal measures and stochastic systems, such that we can discuss and demonstrate how it can fit as a quantifier of causal interactions in complex systems. This work is also written with *financial systems* in mind, that is why we will omit commenting on other complex systems. Choice to use the Rényi transfer entropy to study financial systems is very natural due to the zooming property, that allows to concentrate on particular parts of probability distributions - especially tail parts. That might be a very appealing property from the point of view of a risk management, where low-probability (unexpected) events are monitored. That is why we, after theoretical discussion of the Rényi transfer entropy, turn our attention to the market risk management approaches, and in the end, estimate RTE for the real-market data.

This work is organized as follows:

Theoretical preliminaries This chapter serves as a building ground to the concepts discussed throughout the text. To introduce objects from probability, measure and information theories ideas of A.N.Kolmogorov, B.V.Gnedenko and A.Rényi are followed.

Transfer entropies Transfer entropy is the core of this work. The aim of this chapter is to "massage" the Rényi information measure in order to

get an intuition about it as well as about the Rényi transfer entropy, which is its derivative.

Causation Transfer entropy is a directional measure of information, which makes it a good candidate to be causality detector. The intention is to support this idea by discussion of the probabilistic nature of order.

Stochastic nature of markets Financial time series look chaotic and unpredictable. The truth is that the stock time series carry a large amount of information. Algorithmic and stochastic approach to the financial data is discussed in order to recognize and extract relevant information related to the risk management.

Estimation of the Rényi entropy We investigate financial data processing and numerical estimators of the Rényi entropy. The second half of the chapter is dedicated to the quantification of the Rényi transfer entropy of dummy data from Rössler system, GARCH model and pink noise. In the end, we also apply RTE on real-world data from financial markets.

Conclusion Final remarks and open questions.

(Ω, σ, P)	probability space
$(\Omega, \sigma, \varsigma, P(A\|B))$	conditional probability space
$\Omega, \omega \in \Omega$	set of elementary events
σ	σ-algebra
$\mathcal{B}$	Borel algebra
$A, B \in \sigma$	random events
$P(A) = P(\omega\|\omega \in A)$	probability measure
ξ, η	random variable
$\mathcal{F}_n$	filtration
$\mu(A)$	measure of set A
$H^s(A)$	s-dimensional Hausdorff measure
dim_H	Hausdorff dimension
dim_B	box-counting dimension
D_α	generalized dimension
$F(x)$	distribution function
$p(x)$	probability density function
$\mathcal{P}, \mathcal{Q}$	probability distribution
$\rho_\alpha(x)$	escort (zoom) distribution
$G(k)$	characteristic function
$\mathcal{F}[p(x)]$	Fourier transform
X, Y	stochastic process
$H(X) = H(\xi)$	Shannon entropy of a random variable
$H[\mathcal{P}]$	Shannon entropy of a probability distribution
$H_\alpha(X) = H_\alpha(\xi)$	Rényi entropy of a random variable
$H_\alpha[\mathcal{P}]$	Rényi entropy of probability distribution
$D(\mathcal{P}\|\|\mathcal{Q})$	Kullback-Leibler divergence
$I(X:Y) = I(\xi, \eta)$	mutual information
$T_{Y\to X}(k, l)$	Shannon transfer entropy
$T^R_{\alpha, Y\to X}(k, l)$	Rényi transfer entropy

Chapter 1

Theoretical preliminaries

1.1 Probability theory

1.1.1 The concept of probability

We begin by recalling basics of the probability theory by means of traditional approach of B.V.Gnedenko [9], which despite of being formulated in simple language doesn't lack mathematical rigor.

Originally, man was confronted with a probability in gambling. Attempts to solve the problem whether or not to bet even money on the occurrence of at least one "double six" during 24 throws led to an exchange of letters between Blaise Pascal and Pierre de Fermat [38] in which the fundamental principles of probability theory were formulated for the first time.

Advent of the probability theory was accompanied with an assumption of mutually exclusive elementary events with equal probabilities. For instance, throwing a fair dice has equiprobable outcomes of mutually exclusive elementary events from set $\{1, 2, 3, 4, 5, 6\}$. Thus, the probability of a random variable would be the number of elementary events satisfying that random variable divided by the number of all elementary events. Therefore, the probability of an odd number outcome is

$$P(\{1, 3, 5\}) = \frac{\#\{1, 3, 5\}}{\#\{1, 2, 3, 4, 5, 6\}} = \frac{3}{6} = \frac{1}{2}.$$

This is the most intuitive approach to the concept of a probability.

On the other hand, in terms of statistics, probability is equal to the frequency in which events do occur. It can be calculated from an empirical evidence, provided that an experiment is being repeated many times. Therefore, after N independent repetitions of the same experiment one will have

$n \leq N$ occurrences of a desired event A, and the probability of it would be $P(A) = \frac{n}{N}$. It is evident, that only for $N \to +\infty$ probability $P(A)$ converges to the theoretical value.

The latter two definitions of probability are sufficient for satisfying results in a broad range of scientific research, however, probability theory, as well as other mathematical disciplines, requires axiomatic grounds. Approach suggested by A.N.Kolmogorov includes classical and statistical definitions of probability as special cases. We will see that his formulation of axioms puts the probability theory on the domain of the set theory, and probability itself as a non-negative normalized additive function on measurable sets.

1.1.2 Probability space, random variable and its moments

Definition 1.1.1. Let Ω be a set with elements ω that we call *elementary events*. A set σ is called the *algebra of sets* if the following requirements are fulfilled

- $\Omega \in \sigma$ and $\emptyset \in \sigma$,
- if $A \in \sigma$, then $\overline{A} \in \sigma$,
- if $A \in \sigma$ and $B \in \sigma$, then $A \cup B \in \sigma$, $A \cap B \in \sigma$.

Adding one more requirement

- if $A_n \in \sigma, n = 1, 2, ..,$ then $\bigcup_n A_n \in \sigma$, $\bigcap_n A_n \in \sigma$

define σ as the $\sigma-$*algebra*. Elements of σ are called *random events*. Thus, operating with random events is equal to dealing with sets.

Note: When $\Omega = \mathbb{R}^n$, then $\sigma-$algebra is called the *Borel* $\sigma-$*algebra* that we denote by $\mathcal{B}$. *Borel set* is then a set from Borel $\sigma-$algebra.

Now we can formulate *Kolmogorovs' axioms of probability*:

- A1 Each element $A \in \sigma$ is assigned with a non-negative real number $P(A)$ called the *probability of the event* A.
- A2 $P(\Omega) = 1$.
- A3 If $A_1, A_2, ..., A_n \in \sigma$ is a finite sequence of a pairwise disjoint sets, then $P(A_1 + A_2 + ...A_n) = P(A_1) + P(A_2) + ... + P(A_n)$.

If we want to work with infinitely many elementary events, axiom 3 has to be extended by $n \to \infty$,

- A3' If $A_1, A_2, ... \in \sigma$ is an innumerably infinite sequence of pairwise disjoint sets and if event A is fulfilled by, at least, one of the event from the sequence, then $P(A) = P(A_1) + P(A_2) + ...$

Symbol	Set Theory	Probability Theory
Ω	Set, space	Set of elementary events
ω	Element of the set	Elementary event
A, B	Subsets	Random events
$A + B = A \cup B$	Union of sets A, B	Sum of random events A, B
$AB = A \cap B$	Intersection of events A, B	Product of random events A,B
$\bar{A}$	Set complementary to A	Event complementary to A
$A \setminus B$	Set difference of A and B	Difference of events A and B
$\emptyset$	Empty set	Impossible event
$AB = A \cap B = \emptyset$	Sets A and B have no intersection	Event AB is impossible
$A = B$	Set equality	Random events are the same
$A \subset B$	A is a subset of B	Occurrence of A is followed by the occurrence of B

Table 1.1: Probability theory terminology is analogous to the terminology from set theory.

Theorem 1.1.2. With every algebra of events an algebra of sets isomorphic to it can be associated.

We skip the proof of the latter theorem, that can be found in [33]. However, to illustrate its usefulness, we present a *dictionary* Table 1.1 [9] between terminology from the set theory and the probability theory. The main conclusion to be made is that operating with random variables is analogous to the operations on sets obtaining those variables.

Definition 1.1.3. *Probability space* is given by (Ω, σ, P), where Ω is a set of *elementary events*, σ is a $\sigma-$*algebra* with elements called *random events*, and $P(A)$ is a *probability* assigned to every element of σ. Every σ-measurable function $\xi : \Omega \to \mathbb{R}$ is called a *random variable*.

Definition 1.1.4. Let ξ be a random variable, the *cumulative probability distribution* is defined

$$F(x) := P(\xi(\omega) \leq x) = P(\omega \in \Omega | \xi(\omega) \leq x). \tag{1.1}$$

Definition 1.1.5. We call $\mathbb{E}[\xi]$ *expectation function of a random variable* ξ and define it as

$$\mathbb{E}[\xi] := \int_{\Omega} \xi(\omega) P(d\omega). \tag{1.2}$$

It is more convenient to transform probability space (Ω, σ, P) by the random variable ξ into the following structure - $(\mathbb{R}, \mathcal{B}, p(x)dx)$, where the function $p(x)$, defined by the relation $dF(x) = p(x)dx$, is a *probability density function (pdf)*. Hence, the expectation value of a random variable can be rewritten as

$$\mathbb{E}[\xi] = \int_{\mathbb{R}} xp(x)dx. \tag{1.3}$$

One can also think of expectation value as of a weighted average or as a *mean* that is also *the first moment* of a random variable. We can define the n'th moment of ξ as

$$\mathbb{E}[\xi^n] = \int_{\mathbb{R}} x^n p(x)dx. \tag{1.4}$$

$\mathbb{E}[\xi^2], \mathbb{E}[\xi^3], \mathbb{E}[\xi^4]$ are called *variance*, *skewness* and *kurtosis* respectively. Probability density function of a random variable can be described and analyzed in terms of its moments. Especially, it is convenient when there is no analytical description of the pdf. This brings us to the classification of distribution functions.

Extension of Kolmogorov probability space by means of Rényi

In the theory of Kolmogorov probability is a bounded measure. It is the second axiom that requires this condition, i.e. $P(\Omega) = 1$. However, A.Kolmogorov himself had mentioned (in a private conversation with B.V. Gnedenko, who passed this information to A.Rényi) that it would be desirable to extend his axioms such that unbounded measures could be included. Unbounded measures occur in statistics, quantum mechanics, or in the theory of Markov processes. As we will see in the following text, unbounded measures fit into the classical probability theory by means of conditional probabilities. Thus, Kolmogorovs' probability (not exceeding 1) is obtained as fraction of two unbounded measures of two sets, such that one of them is contained in the second one. A.Rényi introduced his own axiomatization of the probability space in [35], as a combination of conditional probability with classical Kolmogorovs' approach.

Let us have Ω, σ: set of elementary events and $\sigma-$algebra respectively. Let ς be a *non empty* subset of σ. For random events $A \in sigma$ and $B \in \varsigma$ a set function $P(A|B)$ is called *conditional probability of the event A with respect to the event B*, and it is defined if and only if B belongs to ς. $P(A|B)$ has to satisfy the following conditions:

- $P(A|B) \geq 0$ and $P(B|B) = 1$;
- for any $B \in \varsigma$ fixed, $P(A|B)$ is a measure, i.e. a countably additive set function of A, thus if $A_n \in \sigma$ $(n = 1, 2, ...)$ and $A_j A_k = \emptyset$ for $j \neq k$, $j, k \in n$ we have

$$P\left(\sum_{n=1}^{+\infty} A_n|B\right) = \sum_{n=1}^{+\infty} P(A_n|B);$$

- for $A \in \sigma$, $B, C, BC \in \varsigma$ we have

$$P(A|BC)P(B|C) = P(AB|C).$$

If all latter mentioned is satisfied then $(\Omega, \sigma, \varsigma, P(A|B))$ will be called a *conditional probability space.*

Rényis' conditional probability space can be easily linked with the Kolmogorovs' definition, which is a special case. Thus, if we choose σ' as a set of sets $B \in A$ for which $P(B) > 0$ and put $P(A|B) = \frac{P(AB)}{P(B)}$ for $A \in \sigma$ and $B \in \sigma'$, then $(\Omega, \sigma, \sigma', P(A|B))$ is a conditional probability space, or the *conditional probability space generated by the probability space* $(\Omega, \sigma, P(A))$.

Again, if $(\Omega, \sigma, \varsigma, P(A|B))$ is a conditional probability space, we can define $P_C := P(A|C)$ for an arbitrary $C \in \varsigma$, and then $(\Omega, \sigma, P_C(A))$ will be a probability space in the sense of Kolmogorov.

1.1.3 Stable distributions and infinite divisibility

Definition 1.1.6. Let $F_1(x)$ and $F_2(x)$ be distribution functions, we define a operation of *convolution* by

$$F_1(x) * F_2(x) := \int_{\mathbb{R}} F_1(x - y)F_2(y)dy. \tag{1.5}$$

Clearly, $F(x) = F_1(x) * F_2(x)$ is also a distribution function. The operation of convolution is commutative and associative. Therefore, the family of all distribution functions form a semigroup with respect to convolution operation defined by (1.5). Unity element of the semigroup is a *Dirac δ-function* defined by

$$\delta(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}. \tag{1.6}$$

By the *algebra of distributions* we mean the algebra of semigroup of distributions, as is introduced in [34].

Algebra of distributions is very useful in probability theory, since the convolution is related to the addition of random variables. Let ξ_1 and ξ_2 be independent random variables with distribution functions F_1 and F_2 respectively. The sum $\xi_1 + \xi_2$ has a distribution $F = F_1 * F_2$. Now we know how to solve a question what distribution is obtained by observing random variable $\xi = \sum \xi_i$, we call it a *composition problem.* On the other hand, one might be interested in the opposite - the distribution function of a sum of random variables is known, and distributions of components are to be determined - *factorization problem.* One of the thoroughly investigated concepts of the factorization problem concerns factorization of distributions into an arbitrary great number of equal distributions.

Definition 1.1.7. A distribution function K is called *infinitely divisible* if for any $n \geq 2$ there can be found a distribution function F such that

$$F^{(1)}(x) * F^{(2)}(x) * ... * F^{(n)}(x) = K(x). \tag{1.7}$$

Family of infinitely divisible functions is a subalgebra of the algebra of distributions.

The central limit theorem states, that the sum of N ($N \to \infty$) standardized random variables with finite first two moments is normally distributed. Restrictions of the latter statement are not strong, however, there is a big class of distributions with infinite second, or even first moments. The question is, whether these distributions belong to the domain of attraction of a specific group of distributions. The answer is given by the following theorem.

Theorem 1.1.8. A probability density can only be a limiting distribution of the sum of independent and randomly distributed random variables if it is stable.

Definition 1.1.9. A probability density is called *stable* if it is invariant under convolution, i.e., if there are constants $a > 0$ and b such that

$$p(a_1 l + b_1) * p(a_2 l + b_2) = \int_{\mathbb{R}} p(a_1(z-l) + b_1) p(a_2 + b_2) dl = p(az + b) \tag{1.8}$$

for all real constants $a_1 > 0$, b_1, $a_2 > 0$, b_2.

To test whether a distribution satisfies former relation the Fourier transform and its following property might be useful

$$\mathcal{F}[p_1(x) * p_2(x)] = \mathcal{F}[p_1(x)]\mathcal{F}[p_2(x)]. \tag{1.9}$$

Definition 1.1.10. For the probability distribution of a random variable Fourier transform is defined as

$$G(k) := \mathcal{F}[p(x)] = \int_{\mathbb{R}} e^{ikx} p(x) dx, \tag{1.10}$$

called *the characteristic function.* Substitution of the Taylor expansion gives

$$G(k) = \int_{\mathbb{R}} dx p(x) \sum_{n=0}^{+\infty} \frac{(ikx)^n}{n!} = \sum_{n=0}^{+\infty} \frac{(ik)^n}{n!} \mathbb{E}[\xi^n]. \tag{1.11}$$

It means that characteristic function of the random variable is generated by its moments, assuming they all exist.

French and Soviet mathematicians Paul Lévy and Aleksandr Khinchin completely specified a group of all possible stable distributions. They showed that the most general form of a characteristic function of a stable process is

$$\ln G(\xi) = i\mu\xi - \gamma|\xi|^\alpha \left[1 - i\beta \frac{\xi}{|\xi|} tg(\frac{\pi}{2}\alpha)\right]$$

for $\alpha \neq 1$,

$$\ln G(\xi) = i\mu\xi - \gamma|\xi| \left[1 + i\beta \frac{\xi}{|\xi|} \frac{2}{\pi} ln|\xi|\right]$$

for $\alpha = 1$, where $0 < \alpha \leq 2$, γ is a positive scale factor, also called the tail index or characteristic exponent. $\mu \in \mathbb{R}$ is location parameter, $\beta \in [-1, 1]$ is an asymmetry or skewness parameter. The tail index α defines the rate at which the tails of distributions are decreasing. For $\alpha = 2$ we obtain Gauss distribution, for $\alpha = 1$ and $\beta = 0$ we have Cauchy distribution and Lévy distribution for $\alpha = 1/2$ and $\beta = 1$. The latter three distributions are only three stable distributions that can be expressed in analytic form Fig.(1.1).

Example 1.1.11. *Gauss probability density function* has the famous form

$$p_G(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{1.12}$$

where μ and σ^2 are two first moments. With the help of (1.9) we can check that the distribution is stable. The Fourier transform of the pdf is

$$\mathcal{F}[p_G(x)](k) = \int_{\mathbb{R}} e^{ixk} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx = e^{-\frac{\sigma^2 k^2}{2}} \equiv e^{-\gamma k^2}.$$

We have that

$$\mathcal{F}[p_{G_1}(x)](k)\mathcal{F}[p_{G_2}(x)](k) = e^{-k^2\sigma^2} \equiv P_G(k).$$

And inverse Fourier transform gives us

$$\mathcal{F}^{-1}[p_G(k)](x) = \frac{1}{2\pi} \int_R e^{-ixk} e^{-k^2\sigma^2} dk = \frac{1}{\sqrt{4\pi\sigma^2}} e^{-\frac{x^2}{4\sigma^2}}.$$

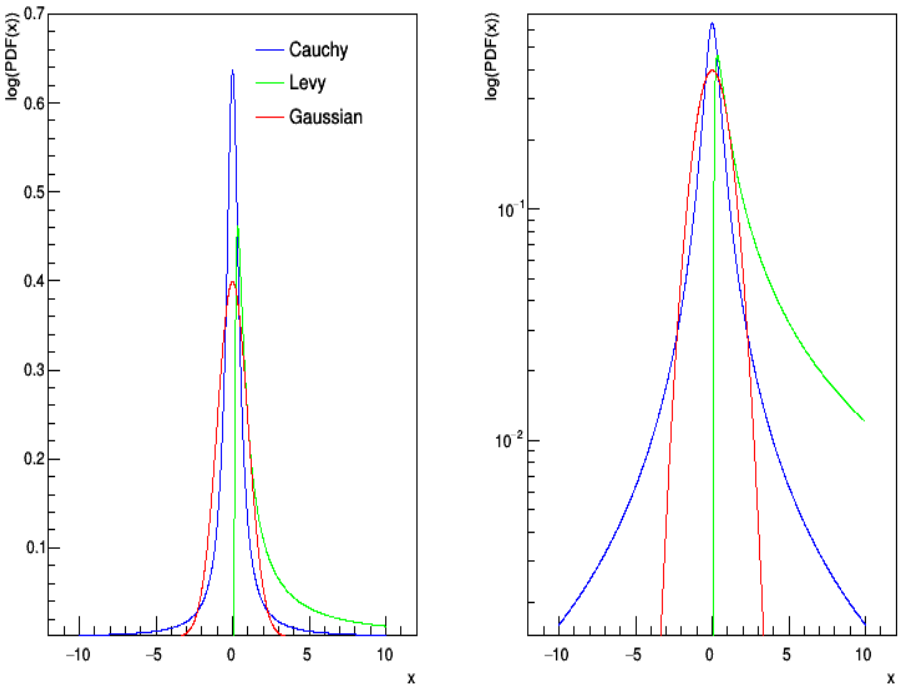


Figure 1.1: Gauss, Cauchy and Lévy distributions are only three stable probability functions that can be described analytically.

Example 1.1.12. *Cauchy probability density function* is

$$p_C(x) = \frac{1}{c\pi}\frac{c^2}{(x-\mu)^2 + c^2}, \tag{1.13}$$

where $c > 0$ and μ are real parameters of scale and position respectively. The mean and the variance, as well as higher moments of Cauchy distribution are undefined. In the same way as it is in the latter example, it can be proved that Cauchy pdf is stable.

Example 1.1.13. *Lévy probability density function* has the following analytic form

$$p_L(x) = \sqrt{\frac{c}{2\pi}}\frac{1}{(x-\mu)^{\frac{3}{2}}}\exp\left(-\frac{c}{2(x-\mu)}\right), \tag{1.14}$$

where $c > 0$ and μ are real valued parameters and $x > \mu$. All moments are undefined as well.

Lévy and Cauchy pdfs are both fat-tailed, i.e they exhibit polynomially decreasing tails and sharp peaks. Such distributions are very useful when we encounter systems with extreme events. Another two widely used heavy-tailed distributions are Tsallis and Student-t distributions.

Example 1.1.14. *Tsallis probability density function* is

$$p_T(x) = \frac{1}{Z}[1 - \gamma(1-q)x]^{\frac{1}{1-q}}, \tag{1.15}$$

where Z is a normalization constant, $q \in \mathbb{R}^+$ and γ is a scale parameter. For $q \to 1$ Tsallis distribution converges to the Normal distribution.

Example 1.1.15. *Student's-t pdf* is formulated in the following way

$$p_S(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}(a + \frac{x^2}{\nu})^{-\frac{\nu+1}{2}}, \tag{1.16}$$

where Γ is the Gamma function, $\nu \in \mathbb{R}^+$ is a parameter of the number of degrees of freedom. For $\nu \to \infty$ Student's-t distribution is Normal. For $\nu = 1$ it becomes Cauchy distribution.

1.2 Stochastic processes

Stochastic process is basically a family of random variables that follow a given rule. We will see that the concept of memory or independence in time is essential and is mostly described in terms of Markov stochastic processes.

Definition 1.2.1. Let (Ω, σ, P) be a probability space and X a random variable. *Stochastic process* is a collection of random variables, that can be written as

$$X = \{X(t, \omega) | t \in \mathbb{R}, \omega \in \Omega\}. \tag{1.17}$$

X is a function of two variables, where t (discrete or continuous) usually has a meaning of time. A realization $X(\omega)$ for given $\omega \in \Omega$ is called *trajectory* or a *sample path*. It can be discrete as well as continuous, with constant or co-evolving sample space.

Definition 1.2.2. A stochastic process X is called *stationary* if and only if for all n we have

$$p(X_{t_1+\Delta t}, X_{t_2+\Delta t}, ..., X_{t_n+\Delta t}) = p(X_{t_1}, X_{t_2}, ..., X_{t_n}), \tag{1.18}$$

where Δt is an arbitrary but fixed time shift.

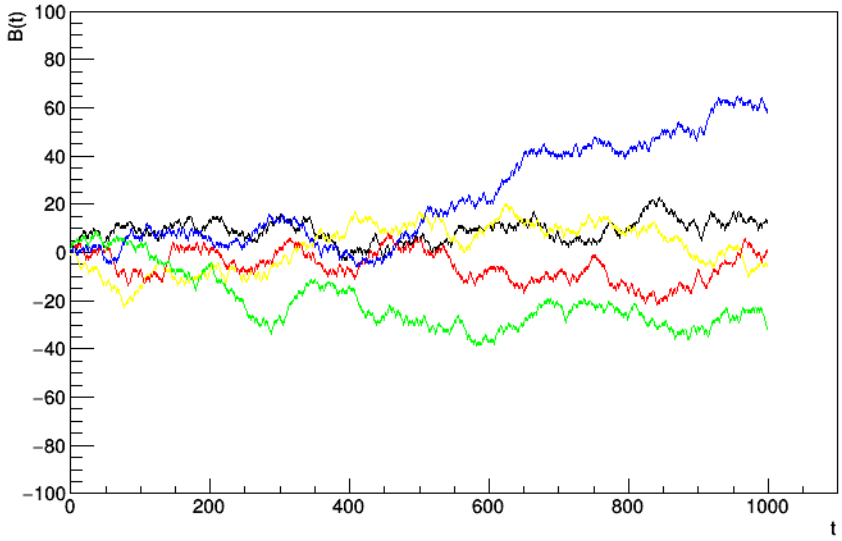


Figure 1.2: 5 different realizations of Wiener process.

Definition 1.2.3. A stochastic process X is said to be a *Lévy process* if it satisfies the following properties

- $X_0 = 0$ almost surely;
- increments are independent, i.e. for any $0 \leq t_1 < t_2 < ... < t_n < \infty$, $X_{t_2} - X_{t_1}, X_{t_3} - X_{t_2}, ...$ are independent random variables;
- stationary increments, i.e. for any $s < t$ $X_t - X_s$ is equal in distribution to X_{t-s};
- continuity in probability, i.e. for any $\epsilon > 0$ and $t \geq 0$ $\lim_{h \to 0}$

$$P(|X_{t+h} - X_t| > \epsilon) = 0.$$

Wiener process is the Lévy process with normally distributed increments Fig.(1.1), it is also known as the *Brownian motion* or *random walk* Fig.(1.2).

Definition 1.2.4. Let $\mathcal{F}_n$ with $n \geq 0$ be a $\sigma-$algebra on a sample space Ω. If for all $0 \leq m \leq n$

$$\mathcal{F}_m \subset \mathcal{F}_n, \tag{1.19}$$

is satisfied, then a collection of $\sigma-$algebras $(\mathcal{F}_m, m \geq 0)$ is called a *filtration*.

Theorem 1.2.5. A stochastic process X_n is said to be *adapted to the filtration* if

$$\sigma(X_n) \ \subset \ \mathcal{F}_n \,, \tag{1.20}$$

for all $t \geq 0$. A stochastic process X is always adapted to the *natural filtration (σ-field) generated by X*:

$$\mathcal{F}_n \ = \ \sigma(X_m, m \leq n) \,. \tag{1.21}$$

1.2.1 Markov processes

Definition 1.2.6. A sequence of random variables $x_1, x_2, ..., x_n$ is a discrete time *k-order Markov chain* with the *Markov property of order k* ($k \in \mathbb{N}$) if

$$P(x_{n+1}|x_n, ..., x_1) = P(x_{n+1}|x_n, ..., x_{n-k+1}) =: P(x_{n+1}|x_n^{(k)}), \tag{1.22}$$

Where

$$P(x|y) := \frac{P(\{x, y\})}{P(\{y\})} \tag{1.23}$$

is *conditional probability function*. Therefore, a probability of the next state depends only on the previous k-states. A process with the Markov property is called the *Markov process*. Another definition of a Markov process can be formulated by means of the filtration.

Definition 1.2.7. A stochastic process X is a *Markov process* with respect to a filtration $\mathcal{F}_t$ when X_t is adapted to the filtration, and, for any $s > t$, X_s is independent of $\mathcal{F}_t$ given X_t.

Typically, however, Markov property is defined as the *Markov property of order 1*, and thus can be interpreted as an independence of the future from the past, given the present, which is exactly what definition *(1.2.7)* implies. Markov used this approach in an attempt to weaken the conditions for the law of large numbers, i.e. exchange an assumption of strict statistical independence of variables by conditional independence. Indeed, limit theorems of probability are sustained for Markov processes as well. Showing it is the same task as proving limit theorems on Rényis' conditional probability spaces [35].

Given definitions of Markov processes are in a wide use, however, we consider important to present another, more detailed in terms of measurability, definition of the Markov process given by E. B. Dynkin in [7], which includes concept of a *transition probability function*.

Definition 1.2.8. Let

- $\zeta(\omega)$ be a non-negative function on some space Ω;
- let $x(t,\omega) = x_t(\omega)$ for $\omega \in \Omega$ and $t \in [0, \zeta(\omega))$ be a function with values in a measurable space $(E, \mathcal{B})$, where $\mathcal{B}$ is $\sigma-$algebra on one-point set;
- for every $0 \leq s \leq t$ we have $\sigma-$algebra $\mathcal{M}_t^s$ on space $\Omega_t = \{\omega | \zeta(\omega) > t\}$;
- for every $s \geq 0$, $x \in E$ $\mathbb{P}_{s,x}(A)$ be a function on $\sigma-$algebra $\mathcal{M}^s$ on space Ω, such that $\mathcal{M}_t^s \subset \mathcal{M}^s$.

Then we say, that $(x_t, \zeta, \mathcal{M}_t^s, \mathbb{P}_{s,x})$ defines a *Markov process* X if the following conditions are satisfied:

1. if $s \leq t \leq u$ and $A \in \mathcal{M}_t^s$, then $\{A, \zeta > u\} \in \mathcal{M}_u^s$;
2. $\{x_t \in \Gamma\} \in \mathcal{M}_t^s$ for any $0 \leq s \leq t$ and $\Gamma \in \mathcal{B}$;
3. $\mathbb{P}_{s,x}$ is the probability measure on $\sigma-$algebra $\mathcal{M}^s$;
4. for any $0 \leq s \leq t$, and $\Gamma \in \mathcal{B}$
$$P(s,x;t,\Gamma) = \mathbb{P}_{s,x}\{x_t \in \Gamma\}$$
is a $\mathcal{B}-$measurable function of x;
5. $P(s,x;s,E \setminus x) = 0$;
6. if $0 \leq s \leq t \leq u$, $x \in E$, $\Gamma \in \mathcal{B}$, then
$$\mathbb{P}_{s,x}\{x_u \in \Gamma | \mathcal{M}_t^s\} = P(t, x_t; u, \Gamma)$$
almost surely.

Ω is a space of elementary events. Measurable space $(E, \mathcal{B})$ is called a *phase space*, ζ is a *life-time*, $P(s,x;t,\Gamma)$ is a *transition probability function* of the process X. $\sigma-$algebra $\mathcal{M}_t^s$ can be thought of as a set of events obtained during the time interval $[s,t]$. $\mathbb{P}_{s,x}(A)$, $(A \in \mathcal{M}^s)$ can be interpreted as a probability of an event A, given that at the moment s we obtained x.

Condition (6) can be exchanged by the following statement

- if $0 \leq s \leq t \leq u$, $x \in E$, $\Gamma \in \mathcal{B}$, then
$$P(s,x;u,\Gamma) = \int_E P(s,x;t,dy)P(t,y;u,\Gamma). \tag{1.24}$$

The latter equation is also known as the *Chapman-Kolmogorov* equation.

Theorem 1.2.9. Every Markov process is completely and uniquely defined by its transition probability function

$$P(s,x;t,\Gamma) = \mathbb{P}_{s,x}\{x_t \in \Gamma\}. \tag{1.25}$$

Definition 1.2.10. Let $(E,\mathcal{B})$ be a measurable space. $P(s,x;t,\Gamma)$ $(0 \leq s \leq t, x \in E, \Gamma \in \mathcal{B})$ is a *transition probability function* if the following statements are satisfied:

1. $P(s,x;t,\Gamma)$ is a measure on Γ;
2. $P(s,x;t,\Gamma)$ is $\mathcal{B}-$measurable function of x;
3. $P(s,x;t,\Gamma) \leq 1$;
4. $P(s,x;u,\Gamma) = \int_E P(s,x;t,dy)P(t,y;u,\Gamma)$, for $0 \leq s \leq t \leq u$.

Example 1.2.11. Let $E \in \mathbb{R}^n$ and $\mathcal{B}$ $\sigma-$algebra. For every $x \in E$ and $\Gamma \in \mathcal{B}$ we define

$$P(s,x;t,\Gamma) = \begin{cases} [2\pi(t-s)]^{-\frac{n}{2}} \int_\Gamma \exp\left[-\frac{(y-x)^2}{2(t-s)}\right] dy & \text{for } 0 \leq s < t \\ \chi_\Gamma(x) & \text{for } 0 \leq s = t \end{cases}. \tag{1.26}$$

This is the *transition probability function of the Wiener process.* Integration is with respect to Lebesgue measure.

1.3 Measure

Definition 1.3.1. We call μ a *measure* on R^n if μ assigns a non-negative number, possibly ∞, to each subset of R^n such that

1. $\mu(\emptyset) = 0$,
2. $\mu(A) \leq \mu(B)$ if $A \subset B$,
3. if $A_1, A_2, ...$ is countable (or finite) sequence of sets then

$$\mu(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu(A_i)$$

with equality if A_i are disjoint Borel sets.

1.3.1 Measure on fractals

Let us cover any non-empty subset F of n-dimensional Euclidean space R^n with collection of sets U_i. Diameter is defined $|U| = \sup\{|x-y|; x,y \in U\}$. Let $0 < |U_i| < \delta$ for each i. We say that U_i is a $\delta - cover$ of F.

Definition 1.3.2. We define

$$H^s_\delta(F) = \inf\{\sum_{i=1}^{\infty} |U_i|^s\}. \tag{1.27}$$

We call $H^s(F)$ the s-dimensional *Hausdorff measure* of F, where

$$H^s(F) = \lim_{\delta \to 0} H^s_\delta(F). \tag{1.28}$$

Hausdorff measure satisfies the Definition 4.1.1. of a measure. Also it satisfies the scaling properties, that are fundamental to the theory of fractals.

Theorem 1.3.3. If $F \subset R^n$ and $\lambda > 0$ then

$$H^s(\lambda F) = \lambda^s H^s(F), \tag{1.29}$$

where $\lambda F = \{\lambda x; x \in F\}$, i.e. the set F scaled by a factor λ. (Proof can be found in [8].)

Definition 1.3.4. We define *Hausdorff dimension* (also called *Hausdorff–Besicovitch* dimension) as

$$\dim_H F = \inf\{s; H^s(F) = 0\} = \sup\{s; H^s(F) = \infty\}. \tag{1.30}$$

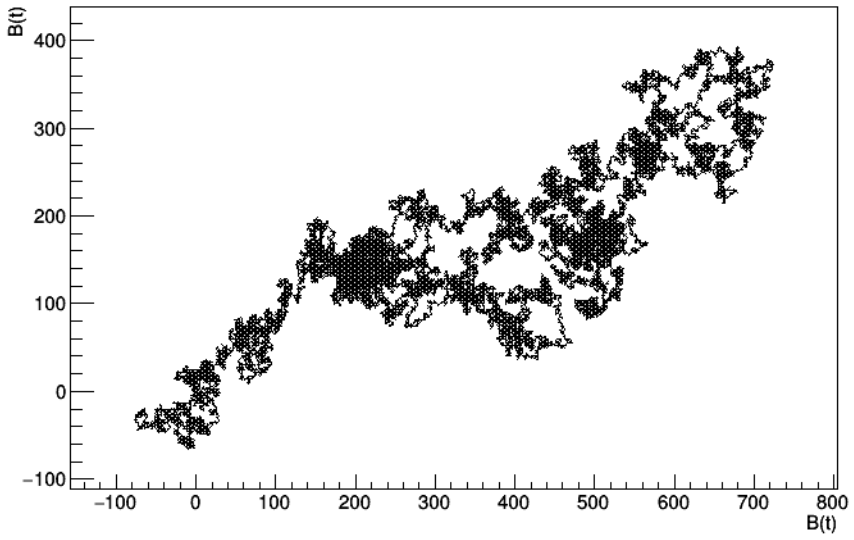


Figure 1.3: Random walk on 2-dimensional lattice.

For shapes of classical geometry mathematical formalism provides that the Hausdorff dimension of a single point is zero, of a line is 1, of a square is 2, and of a cube is 3. It is important to emphasize that $\dim_H$ can be a non-integer number, typically, fractals have fractional dimensions. For example, Brownian motion has Hausdorff dimension $\dim_H = 2$. Figure 1.3 shows two dimensional BM, one can notice, that plane is almost whole covered with paths. On the other hand BM motion shown on Figure 1.2 has fractal dimension $\frac{3}{2}$.

Former definition of Hausdorff dimension is fundamental, however, other definitions are in widespread use. Quite popular and intuitive is *box-counting dimension*. Hausdorff and box dimensions are not equal in general, although we get equality for many reasonably regular sets, which do appear often.

Definition 1.3.5. Let F be any non-empty bounded subset of R^n and let $N_\delta(F)$ be the smallest number of sets of diameter at most δ which cover F. We call *box-counting dimension* or box dimension of F

$$\dim_B(F) = \lim_{\delta \to 0} \frac{\ln N_\delta(F)}{-\ln \delta}. \tag{1.31}$$

This definition can be derived intuitively by consideration that volume V occupied by the set F is

$$V \simeq \delta^D N_\delta, \tag{1.32}$$

where N_δ is a number of a D-dimensional δ-cubes covering F. Taking logarithms

$$\ln V \simeq D \ln \delta + \ln N_\delta. \tag{1.33}$$

Then for $\delta \to 0$

$$D = \lim_{\delta \to 0} \frac{\ln V - \ln N_\delta}{\ln \delta} = \lim_{\delta \to 0} \frac{-\ln N_\delta}{\ln \delta} \equiv \dim_B . \tag{1.34}$$

1.3.2 Measure on multifractals

Let us cover the plane with δ mesh. The probability of occurrence in some given region i would be $p_i \sim \delta^{\beta_i}$. Thus, we have a spectrum of scaling exponents β_i for regions with different masses. Mass at each og these regions is distributed with $\varrho(\delta, \beta) \sim \delta^{f(\beta)}$, where $f(\beta)$ is called $multifractal\ spectrum$. It is a value of fractal dimension of subset with scaling exponent β.

Let us investigate partition function $Z(\delta, q) \sim \delta^{\tau(q)}$, where $\tau(q), q \in \mathbb{R}$ is called *scaling function* and

$$Z(\delta, q) = \sum_i p_i^q = \sum_i \delta^{q\beta_i} N_\delta(\beta_i) = \sum_i \delta^{q\beta_i} \delta^{-f(\beta_i)}, \tag{1.35}$$

thus

$$\tau(q) = f(\beta) - q\beta. \tag{1.36}$$

We have two pairs of functions and variables: $\{f(\beta), \beta\}$, where $\beta \in \mathbb{R}^+$, denotes the multifractal spectrum and $\{\tau(q), q\}$ is a mass function. Switching between these two pairs can be provided by the *Legendre transform*, thus minimizing (1.36) gives $\frac{d}{d\beta}f(\beta(q)) = q$. Moreover minimizing δ yields

$$\tau(q) = \lim_{\delta \to 0} \frac{\log Z(\delta, q)}{-\log \delta}, \tag{1.37}$$

that can be used to find the multifractal spectrum $f(\beta)$.

1.4 Information theory

1.4.1 Amount of information

In general, the information (in bits) necessary to characterize elements of a set E_N with N elements is given by *Hartley formula*

$$I(E_N) = \log_2 N. \tag{1.38}$$

The best way to get some knowledge about a particular object, let us denote it with A, is by asking questions about it. At the beginning of this process we have an uncertainty $h(A)$ about A, but each answer to relevant question brings an amount of information $i_k(A)$ $(k = 1, 2, ...)$ that reduces the uncertainty $h(A)$. It appears, that asking questions which can be answered by "yes" or "no" is the most efficient strategy. Moreover the probability of obtaining "yes"-answer should be equal to the probability of "no"-answer, therefore $p(yes) = p(no) = \frac{1}{2}$. An amount of information obtained by answering a question with only two possible and equally probable answers is *1 bit of information*.

To illustrate the latter we give an example. Let us have a list with 16 names on it and the task is to find one particular from it. According to the aforementioned strategy the first question asked should be whether the name is between the first half set of the names on the list. In that case $p(yes) = p(no) = \frac{1}{2}$ and any answer will lead us to the correct set of the names. We repeat the same strategy 3 more times to find the name we were searching for. Summing it up, one has to ask 4 questions to obtain the full information to be able to find the correct name. Also it is easy to see, that each answered question reduces uncertainty until it is diminished by the information at all. Hence, we can write $h(A) = -i(A)$, uncertainty can be seen as a negative information. Now it is clear that 4 bits of information

is necessary to encode the name. The problem could be solved using the Hartley's formula (1.38). For $N = 16$ we have

$$I(16) = \log_2 16 = \log_2(2{\cdot}2{\cdot}2{\cdot}2) = \log_2 2{+}\log_2 2{+}\log_2 2{+}\log_2 2 = 1{+}1{+}1{+}1 = 4.$$

1.4.2 Uncertainty and unexpectedness

Let us have a random variable ξ with the possible values $x_1, x_2, ..., x_N$ and probabilities $p_1, p_2, ..., p_N$ respectively, A_k denotes the event that x_i obtains a value from $\{x_k | k = 1, 2, ..., N\}$. Then we associate with every event a number $V(A_k)$ which represents the *unexpectedness of event* A_k. $V(A_k)$ has the following properties:

1. V is a monotone decreasing function of probability, i.e. $V = V(p(A_k))$ (the more improbable an event is, the bigger is its unexpectedness);

2. For two independent events A_i and A_j $i \neq j$ the unexpectedness of their simultaneous occurrence is equal to the sum of their individual unexpectedness
$$V(A_i A_j) = V(A_i) + V(A_j);$$

3. The unexpectedness of an event having probability $\frac{1}{2}$ is 1
$$V(p(A_i)) = \frac{1}{2} = 1.$$

These 3 assumptions hold if

$$V(A_i) = \log_2 \frac{1}{p(A_i)}. \tag{1.39}$$

The expectation (first moment) of the unexpectedness of the the random variable ξ is

$$H(\xi) = \sum_{i=1}^{N} p_i \log_2 \frac{1}{p_i}. \tag{1.40}$$

The latter is also known as the *entropy or uncertainty of the random variable* ξ. For equiprobable process $p_1 = p_2 = ... = p_N = \frac{1}{N}$ entropy reduces to the Hartley's formula (1.38)

$$H(\xi) = \sum_{i=1}^{N} \frac{1}{N} \log_2 N = \frac{N}{N} \log_2 N = \log_2 N. \tag{1.41}$$

Based on this on our derivation, entropy can be interpreted as a weighted average of unexpectedness of the random variable.

If random variable ξ is conditioned on an arbitrary event with positive probability, then $p(A_k|B) = \frac{p(A_k,B)}{p(B)}$. We can define *conditional entropy of random variable* ξ given condition B as

$$H(\xi|B) = \sum_{i=1}^{N} p(A_i|B) \log_2 \frac{1}{p(A_i|B)}. \tag{1.42}$$

1.4.3 Mutual information

Let ξ and η be two random variables with possible values $x_1, x_2, ..., x_N$ and $y_1, y_2, ..., y_M$ respectively. Let A_k be an event that $\xi = x_k$ and B_j be an event that $\eta = y_j$. If we observe B_j the unexpectedness of event A_k will change. In other words, the probability of A_k has changed to the conditional probability $p(A_k|B_j) = \frac{p(A_k,B_j)}{p(B_j)}$ and its unexpectedness will be $\log_2 \frac{1}{p(A_k|B_j)}$. Let $V(A_k, B_j)$ denote the change in the unexpectedness of A_k resulting from observation of B_j, then

$$V(A_k, B_j) = \log_2 \frac{1}{p(A_k)} - \log_2 \frac{1}{p(A_k|B_j)} = \log_2 \frac{p(A_k, B_j)}{p(A_k)p(B_j)}. \tag{1.43}$$

For $p(A_k, B_j) = p(A_k)p(B_j)$ we have $V(A_k, B_j) = 0$, which implies independence of A_k and B_j.

If we calculate the expectation of $V(A_k, B_j)$, we will see how much, on average, an unexpectedness of ξ will change, given that the η is also observed. We have

$$\sum_{k=1}^{N}\sum_{j=1}^{M} p(A_k, B_j)V(A_k, B_j) = \sum_{k=1}^{N}\sum_{j=1}^{M} p(A_k, B_j) \log_2 \frac{p(A_k, B_j)}{p(A_k), p(B_j)} =: I(\xi, \eta). \tag{1.44}$$

Quantity $I(\xi, \eta)$ is called *mutual information of random variables* ξ an η. It is easy to see that $I(\xi, \eta) = I(\eta, \xi)$ and $I(\xi, \eta) \geq 0$, equality implicates independence of ξ and η. Moreover, mutual information can be used to define the entropy of joint distribution of random variables as

$$H(\xi, \eta) = H(\xi) + H(\eta) - I(\xi, \eta), \tag{1.45}$$

that is a generalization of the law of additivity of information.

1.4.4 Gain of information

Let $A_1, A_2, ..., A_N$ be the mutually exclusive possible outcomes of random variable ξ with distribution $\mathcal{Q} = \{q_1, q_2, ..., q_N\}$, where $q_k = prob(A_k)$ $(k = 1, 2, ..., N)$ and $\sum q_i = 1$. Let us suppose that experimental circumstances have changed and random variable ξ is observed with distribution $\mathcal{P} =$

$\{p_1, p_2, ..., p_N\}$, where $\sum p_i = 1$. The change in unexpectedness of A_k is then

$$\log_2 \frac{1}{q_k} - \log_2 \frac{1}{p_k} = \log_2 \frac{p_k}{q_k}. \tag{1.46}$$

Analogically with previous calculations the expected value of this change can be calculated using the probabilities corresponding to the changed conditions. We call it *the gain of information*

$$D(\mathcal{P}||\mathcal{Q}) = \sum_{k=1}^{N} p_k \log_2 \frac{p_k}{q_k}. \tag{1.47}$$

$D(\mathcal{P}||\mathcal{Q})$ is also called *the information theoretical distance of two distributions* or *Kullback-Leibler divergence*. Because $\log_2 \frac{1}{x}$ is convex, $D(\mathcal{P}||\mathcal{Q})$ is always non-negative and equal to zero only if the $\mathcal{P}$ and $\mathcal{Q}$ distributions are the same. In special case, when $\mathcal{Q} = \frac{1}{N}, ..., \frac{1}{N}$ is the uniform distribution, we can write

$$D(\mathcal{P}||\mathcal{Q}) = \sum_{k=1}^{N} p_k \log_2 N p_k = \log_2 N - \sum_{k=1}^{N} p_k \log_2 p_k = \log_2 N - H(\xi). \tag{1.48}$$

1.4.5 Generalization of entropy

Let (Ω, σ, P) be a probability space, where Ω is an arbitrary non-empty set of elementary events, σ is a σ-algebra of subsets Ω, including Ω itself, P is a probability measure defined on σ, $P(\Omega) = 1$. Let $f = f(\omega)$ is defined for $\omega \in \Omega$, where $\Omega \in \sigma$ and $P(\Omega) > 0$, and is measurable with respect to σ. Function f is called *generalized random variable* with *generalized probability distribution* $\mathcal{P} = (P(f(x_1)), P(f(x_2)), ..., P(f(x_n))) \equiv (p_1, p_2, ..., p_n)$ and

$$W(\mathcal{P}) = \sum_{k=1}^{n} p_k \tag{1.49}$$

is *the weight of the distribution* $\mathcal{P}$. If $W(\mathcal{P}) < 1$ the distribution $\mathcal{P}$ is called *incomplete distribution*, if $W(\mathcal{P}) = 1$ it is called *complete distribution*.

The entropy $H[\mathcal{P}]$ of a finite discrete generalized probability distribution $\mathcal{P}$ is characterized by the following 5 postulates:

1. $H[\{p_1, ..., p_n\}] = H[\{p_{k(1)}, ..., p_{k(n)}\}]$, *i.e. is a symmetric function*
2. $H[\mathcal{P}]$ *is continuous function of* p *in the interval* $0 < p \leq 1$;
3. $H[\{\frac{1}{2}\}]$=1;

4. *If P and Q are finite discrete generalized probability distributions of two independent random variables, then $H[\mathcal{P} \times \mathcal{Q}] = H[\mathcal{P}] + H[\mathcal{Q}]$.*

5. *For $\mathcal{P} = (p_1, ..., p_n)$ and $\mathcal{Q} = (q_1, ..., q_m)$ from the set of the finite discrete generalized distributions, that $W(\mathcal{P}) + W(\mathcal{Q}) \leq 1$ holds*

$$H[\mathcal{P} \cup \mathcal{Q}] = \frac{W(\mathcal{P})H[\mathcal{P}] + W(\mathcal{Q})H[\mathcal{Q}]}{W(\mathcal{P}) + W(\mathcal{Q})},$$

where $(\mathcal{P} \cup \mathcal{Q}) = (p_1, ..., p_n, q_1, ..., q_m)$. Therefore, entropy of the union of two distributions is their weighted average.

Let $\{p\}$ be a generalized probability distribution of a single probability p, we put $h(p) = H[\{p\}]$. From the postulate (4) $h(p, q) = h(p) + h(q) = h(q) + h(p) = h(q, p)$ for $0 < p \leq 1$ and $0 < q \leq 1$. Therefore postulates (1)-(4) are fulfilled if $h(p) = \log_2 \frac{1}{p}$.

We see, that P can be written as a union of incomplete distributions p_i as $P = (p_1 \cup p_2 \cup ... \cup p_n)$, where $W(p_i) \leq 1$. Then by postulate (5)

$$H[\mathcal{P}] = \frac{W(p_1)H[\{p_1\}] + ... + W(p_n)H[\{p_n\}]}{W(p_1) + ... + W(p_n)} = \frac{p_1 \log_2 \frac{1}{p_1} + ... + p_n \log_2 \frac{1}{p_n}}{p_1 + ... + p_n}. \tag{1.50}$$

Therefore, postulates (1)-(5) are satisfied by the function

$$H[\mathcal{P}] = \frac{\sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i}}{\sum_{i=1}^{n} p_i}. \tag{1.51}$$

Which, for a complete distribution is the *Shannon entropy*

$$H[P] = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i}. \tag{1.52}$$

An interesting question to ask is what other quantity is obtained by replacing the arithmetic mean by other mean value. Let $x_1, ..., x_k$ be a set of numbers with weights $w_1, ..., w_k$, $\sum w_i = 1$. The general form of a mean value of this set is usually written in the following form

$$g^{-1}[w_i g(x_i)], \tag{1.53}$$

where $g(x)$ is called the *Kolmogorov-Nagumo function* and is defined as an arbitrary strictly monotonic continuous invertible function. The postulate (5) can be replaced by postulate

5' *There exist strictly monotonic and continuous invertible function g, that for discrete finite generalized probability distributions P, Q with $W(P) + W(Q) \leq 1$ holds*

$$H[\mathcal{P} \cup \mathcal{Q}] = g^{-1}\left[\frac{W(\mathcal{P})g(H[\mathcal{P}]) + W(\mathcal{Q})g(H[\mathcal{Q}])}{W(\mathcal{P}) + W(\mathcal{Q})}\right].$$

If we choose $g(x) = g_\alpha(x)$ for $\alpha > 0$, $\alpha \neq 1$ and $g_\alpha(x) = 2^{(1-\alpha)x}$ we have

$$\begin{aligned} H_\alpha[\mathcal{P} \cup \mathcal{Q}] &= \frac{1}{1-\alpha}\log_2\left[\frac{p2^{(\alpha-1)\log_2 p} + q2^{(\alpha-1)\log_2 q}}{p+q}\right] \\ &= \frac{1}{1-\alpha}\log_2\left[\frac{p^\alpha + q^\alpha}{p+q}\right]. \end{aligned}$$

Functional $H_\alpha[\mathcal{P}]$ defined by axioms (1)-(4) and (5') is called the *Rényi α-entropy*. It is discussed in detail in the Chapter 2.

Let us now show why is the choice of exponential function $g(x) = 2^{(1-\alpha)x}$ the only admissible. We consider two independent random variables ξ and η with probability density functions $\mathcal{P} = (p_1, ..., p_n)$ and $\mathcal{Q} = (q_1, ..., q_m)$ respectively. Let the information obtained by observing an event $\xi(p_k)$ is I_k bits, and J_i bits for observing an event $\eta(q_i)$. Thus, after this one experiment we receive $I_k + J_i$ bits of information with probability $p_k q_i$. If experiment is repeated many times, on average, the amount of information obtained from the sum is equal to the sum of means of information from the two random variables separately. Hence, (1.53) yields that

$$g^{-1}\left(\sum_{k=1}^{n}\sum_{i=1}^{m} p_k q_i g(I_k + J_i)\right) = g^{-1}\left(\sum_{k=1}^{n} p_k g(I_k)\right) + g^{-1}\left(\sum_{i=1}^{m} q_i g(J_i)\right). \tag{1.54}$$

If we put $J_i = J$, then

$$g^{-1}\left(\sum_{k=1}^{n}\sum_{i=1}^{m} p_k q_i g(I_k + J)\right) = g^{-1}\left(\sum_{k=1}^{n} p_k g(I_k)\right) + J. \tag{1.55}$$

From the theory of means equality (1.55) holds only for linear and exponential functions. Thus, if there exist real-valued constants $a(y) \neq 0$ and $b(y)$ for x and y we can write

$$g(x+y) = a(y)g(x) + b(y) \overset{b(y)=g(y)}{=} a(y)g(x) + g(y). \tag{1.56}$$

Equality $b(y) = g(y)$ we get from an assumption that $g(0) = 0$. Since this equation holds for every x and y we can interchange roles of x and y, hence

$g(x+y) = a(y)g(x)+b(y)$. By comparing this equation and (1.56) we obtain

$$\frac{a(y)-1}{g(y)} = \frac{a(x)-1}{g(x)} \tag{1.57}$$

if $x \neq 0$ and $y \neq 0$. Thus, there must exist constant k, such that

$$a(x) - 1 = kg(x). \tag{1.58}$$

For $k = 0$ substitution of (1.58) into (1.56) leads to

$$g(x+y) = g(x) + g(y), \tag{1.59}$$

that is $g(x) = cx$ for $c \neq 0$ is a constant. This choice of $g(x)$ transforms postulate (5') back to postulate (5), where $H = H_0$.

Let now be $k \neq 0$, then the substitution yields

$$a(x+y) = a(x)a(y). \tag{1.60}$$

As $g(x)$ is a monotonic function, from (1.58) $a(x)$ is also monotonic, hence $a(x)$ is an exponential function. If we choose $a(x) = 2^{(1-\alpha)x} + 1$ and $k = 1$, then from by substitution into (1.58) we have

$$g(x) = 2^{(1-\alpha)x}. \tag{1.61}$$

1.5 Summary

Chapter 1 is an overview of theoretical concepts used throughout this work. It covers conceptions from four separate, vast on its own, fields, which mostly intersect on the domain of probability theory.

To conclude this chapter, let us summarize with an overview of main takeaways:

- probability theory is built on the domain of the set theory; probability itself is a non-negative normalized additive function on measurable sets; operating with random variables is analogous to operations with sets obtaining those variables;
- Rényi conditional probability space with unbounded probability measure is a natural extension of the Kolmogorov probability space;
- stochastic process is a collection of random variables; Markov property is related to the idea of memory or independence in time;
- multifractal dimension is a variable that can be used to quantitatively characterize sets with unevenly distributed elements;
- generalized family of one-parameter entropies is derived as an exponentially weighted mean of unexpectedness functional $(-\log p)$.

Chapter 2

Transfer entropies

2.1 Rényi information measures

2.1.1 Rényi entropy

In the previous chapter we derived Rényi entropy as the quasi-linear mean of unexpectedness function. This derivation was presented by Alfred Rényi himself. Rényi $\alpha-$entropy is the most general class of information measures preserving additivity for independent systems and compatible with Kolmogorov's probability axioms.

We now present another set of axioms from [16] that are equivalent to the set of axioms (1)-(4),(5') from the previous chapter.

Definition 2.1.1. Let $\alpha > 0$, ξ_n a random variable with probability distribution $\mathcal{P} = (p_1, ..., p_n)$ ($\sum_i^n p_i = 1$) and $H_\alpha[\mathcal{P}]$ is such that

1. for a given $n \in \mathbb{N}$ $H_\alpha[\mathcal{P}(n)]$ is a continuous function;
2. for a given ($n \in \mathbb{N}$) function $H_\alpha[\mathcal{P}]$ has its largest value for $p_1 = p_2 = ... = p_n = 1/n$ with the normalization $H_1[\{\frac{1}{2}, \frac{1}{2}\}] = 1$;
3. for a given α, $H_\alpha[\mathcal{P} \cap \mathcal{Q}] = H_\alpha[\mathcal{P}] + H_\alpha[\mathcal{Q}|\mathcal{P}]$, where
$$H_\alpha[\mathcal{Q}|\mathcal{P}] = g^{-1}(\sum_i \rho_\alpha(q_i) g(H_\alpha[\mathcal{Q}|p_i]))$$
and
$$\rho_\alpha(q_i) = \frac{q_i^\alpha}{\sum_j q_j^\alpha};$$
4. g is invertible and positive on $[0, +\infty)$;
5. $H_\alpha[\mathcal{P}, \{0\}] = H_\alpha[\mathcal{P}]$, therefore, adding impossible event doesn't change the amount of information.

Then all these axioms are satisfied if

$$H_\alpha[\mathcal{P}] := \frac{1}{1-\alpha}\log_2 \sum_{i=1}^{n} p_i^\alpha. \tag{2.1}$$

$H_\alpha[\mathcal{P}]$ is called the *Rényi entropy* of a probability distribution $\mathcal{P}$.

Note: The function ρ_α from the axiom (3) is known as the *zooming* or *escort distribution*. Its distinguishing property is that the small values of probabilities p_i are emphasized for $\alpha < 1$ and, on the contrary, higher probabilities are being emphasized for $\alpha > 1$ as can be seen on Fig.(2.1).

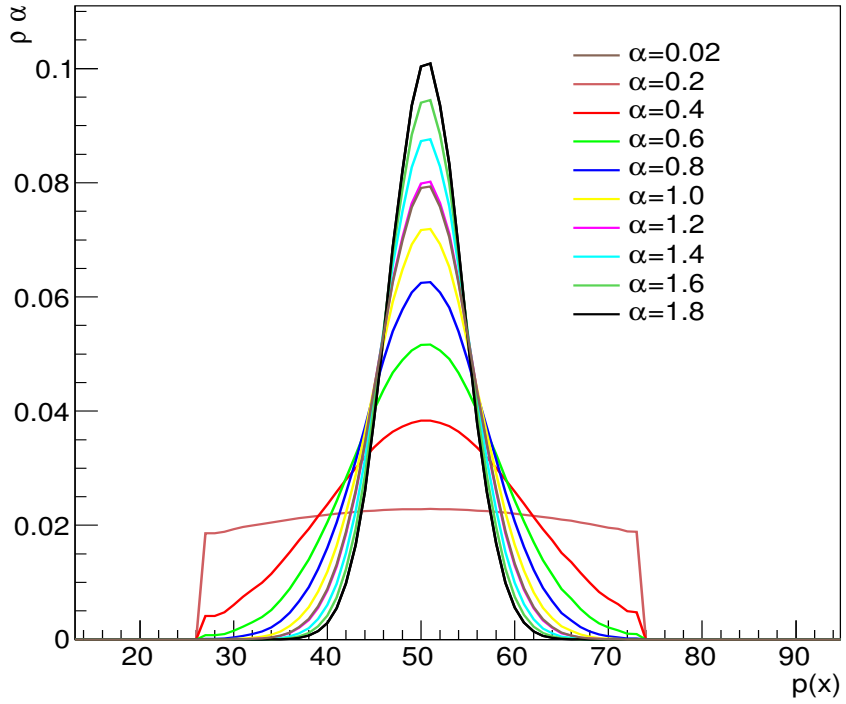


Figure 2.1: Escort distribution of Normal distribution for different values of parameter α.

Rényi entropy (RE) has some interesting properties as:

- RE is symmetric, i.e. $H_\alpha[\{p_1, .., p_n\}] = H_\alpha[\{p_{\pi(1)}, ..., p_{\pi(n)}\}]$;
- RE is non-negative, i.e. $H_\alpha \geq 0$;
- $\lim_{\alpha\to 1} H_\alpha = H_1$, where $H_1 = H$ is the Shannon entropy;

- $H_0 = \log_2 n$ is the Hartley entropy (1.38) and $H_2 = -\log_2 \sum_{i=1}^{n} p_i^2$ is the Collision entropy;
- $0 \leq H_\alpha[\mathcal{P}] \leq \log_2 n$;
- H_α is a positive, decreasing function of $\alpha > 0$.
  proof: It can be proved that for $x_k > 0, w_k > 0, \sum w_k = 1$ the average $(\sum w_k x_k^\beta)^{\frac{1}{\beta}}$ is a monotone increasing function of β. In the same way we can write

$$H_\alpha[\mathcal{P}] = \log_2 \left[\sum_{i=1}^{n} p_k \left(\frac{1}{p_k} \right)^{1-\alpha} \right]^{\frac{1}{1-\alpha}}. \tag{2.2}$$

  Thus H_α is an increasing function of $1-\alpha$ and decreasing function of α.

Another appealing property of the Rényi entropy is its connection to multifractals, in particular to the multifractal dimension. Using (1.37) and (2.45) we can introduce generalized dimensions as in [16]

$$D_\alpha = \lim_{\delta \to 0} \left(\frac{1}{1-\alpha} \frac{\log Z(\delta)}{\log \delta} \right) = -\lim_{\delta \to 0} \frac{H_\alpha(\delta)}{\log \delta}. \tag{2.3}$$

2.1.2 Gain of information of order α

Definition 2.1.2. Let $\alpha > 0$, ξ and η are random variables with probability distributions $\mathcal{P} = (p_1, ..., p_n)$ $(\sum_i^n p_i \leq 1)$ and $\mathcal{Q} = (q_1, ..., q_n)$ $(\sum_i^n q_i \leq 1)$ respectively. We define quantity $I_\alpha(\mathcal{P}||\mathcal{Q})$ called the *measure of order α of the gain of information* [33] as

$$I_\alpha(\mathcal{P}||\mathcal{Q}) := \frac{1}{\alpha - 1} \log_2 \left(\frac{1}{\sum_{k=1}^n p_k} \sum_{k=1}^{n} \frac{p_k^\alpha}{q_k^{\alpha-1}} \right), \tag{2.4}$$

for limit $\alpha \to 1$ we obtain Kullback-Leibler divergence (1.47). I_α is an increasing function of α and for $\mathcal{P}, \mathcal{Q}$ complete distributions $I_\alpha(\mathcal{P}||\mathcal{Q}) \geq 0$ for $\alpha \geq 0$. If $\mathcal{Q} = \mathcal{U}_n$ is a uniform distribution with $q_k = \frac{1}{n}$ for all $k = 1, ..., n$, then equation (2.4) reads

$$I_\alpha(\mathcal{P}||\mathcal{U}_n) = \log_2 n - \frac{1}{1-\alpha} \log_2 \left(\frac{\sum_{k=1}^n p_k^\alpha}{\sum_{k=1}^n p_k} \right), \tag{2.5}$$

which in case of $\mathcal{P}$ complete distribution is

$$I_\alpha(\mathcal{P}||\mathcal{U}_n) = H_\alpha[\mathcal{U}_n] - H_\alpha[\mathcal{P}]. \tag{2.6}$$

For limit $\alpha \to 1$ we obtain (1.48).

One can first axiomatize $\alpha-$gain of information I_α (2.4) as is done in [33], and then define $\alpha-$entropy H_α with equation (2.6). It is especially useful when we want to discuss information-theoretic properties of H_α. From non-negativity of I_α for $\alpha \geq 0$ and (2.6) we see that H_α is maximal for uniform distribution. However, for $\alpha < 0$ and $p_i \to 0$ $H_\alpha[\mathcal{P}]$ tends to infinity, but for $\alpha = 0$, $H_0[\mathcal{P}] = \log_2 n$. Because of this discontinuity, it is more convenient to consider only H_α for positive α as a proper information measure.

Example 2.1.3. α-information measure can be used in proofs of probability theory limit theorems. Let $X = x_1, ..., x_n$ be a stationary Markov chain (discrete Markov process). Let p_{jk} denote transition probability from j to k in one step, and $p_{jk}^{(n)}$ in n steps. All transition probabilities are positive. We recall Chapman-Kolmogorov equation (1.24) for Markov chain

$$\sum_{j=1}^{n} p_j p_{jk} = p_k, \tag{2.7}$$

where

$$\sum_{k=1}^{n} p_k = 1. \tag{2.8}$$

The well-known result is that

$$\lim_{n \to +\infty} p_{jk}^{(n)} = p_k. \tag{2.9}$$

In [32] (2.9) is proved with the help of $I_1(\mathcal{P}||\mathcal{Q})$, we give another, but technically similar, proof using $I_\alpha(\mathcal{P}||\mathcal{Q})$. First, we take for granted that the system of equations has a solution $\mathcal{P} = (p_1, ..., p_m)$. We denote $\mathcal{P}_j^{(n)} = (p_{j1}^{(n)}, ..., p_{jm}^{(n)})$ and write

$$I_\alpha(\mathcal{P}^{(n)}||\mathcal{P}) = \frac{1}{1-\alpha} \log_2 \sum_{k=1}^{m} p_{jk}^{(n)} \left(\frac{p_{jk}^{(n)}}{p_k} \right)^{\alpha-1}. \tag{2.10}$$

We denote $w_{lk} = \frac{p_l p_{lk}}{p_k}$, it is a so-called *backward transition probability* such that $\sum_{l=1}^{m} w_{lk} = 1$ and from the Chapman-Kolmogorov equation (1.24) we recall that $p_{jk}^{(n+1)} = \sum p_{jl}^{(n)} p_{lk}$. Substitution into (2.10) gives

$$I_\alpha(\mathcal{P}^{(n+1)}||\mathcal{P}) = \sum_{k=1}^{m} p_{jk}^{(n+1)} \left(\frac{p_{jk}^{(n+1)}}{p_k} \right)^{\alpha-1} \tag{2.11}$$

$$= \sum_{k=1}^{m} p_k \left[\sum_{l=1}^{m} w_{lk} \frac{p_{jl}^{(n)}}{p_l} \left(\sum_{l=1}^{m} w_{lk} \frac{p_{jl}^{(n)}}{p_l} \right)^{\alpha-1} \right] \tag{2.12}$$

$$\equiv \sum p_k x f(x), \tag{2.13}$$

where $xf(x) = xx^{\alpha-1}$ is a convex function. In the following we will need the *Jensen's inequality* which for $g(x)$ convex and w_k, such that $\sum w_k = 1$ reads

$$g(\sum_{j=1}^{n} w_j x_j) \leq \sum_{j=1}^{n} w_j g(x_j). \tag{2.14}$$

We substitute into the rhs of (2.14) w_{lk}, $g(x) = xf(x)$ and sum over p_k:

$$\sum_{k=1}^{m} p_k \sum_{l=1}^{m} w_j g(x_j) = \sum_{k=1}^{m} p_k \sum_{l=1}^{m} w_{lk} \frac{p_{jl}^{(n)}}{p_l} \left(\frac{p_{jl}^{(n)}}{p_l} \right)^{\alpha-1} \tag{2.15}$$

$$= \sum_{l=1}^{m} p_{jl}^{(n)} \left(\frac{p_{jl}^{(n)}}{p_l} \right)^{\alpha-1} \tag{2.16}$$

$$= I_\alpha(\mathcal{P}^{(n)}||\mathcal{P}), \tag{2.17}$$

where we used that $\sum p_k w_{lk} = p_l$. From Jensen's inequality we have

$$I_\alpha(\mathcal{P}^{(n+1)}||\mathcal{P}) \leq I_\alpha(\mathcal{P}^{(n)}||\mathcal{P}), \tag{2.18}$$

moreover we know that $I_\alpha(\mathcal{P}^{(n)}||\mathcal{P}) \geq 0$ and thus the limit

$$\lim_{n \to +\infty} I_\alpha(\mathcal{P}^{(n)}||\mathcal{P}) \tag{2.19}$$

exists, and it is straightforward to show that it is equal to zero. Therefore we have proved that (2.9) holds.

2.1.3 Differential Rényi entropy

So far we have defined Rényi information measure of discrete probability distributions. When the distribution is continuous, the information tends to infinity as $n \to \infty$. However, in many cases the limit

$$d_\alpha(\xi) = \lim_{n \to +\infty} \frac{H_\alpha(\xi_n)}{\log_2 n}$$

exists. The quantity $d_\alpha(\xi)$ is called *dimension of order α of ξ*. If also the limit

$$\lim_{n \to +\infty} (H_\alpha(\xi_n) - d_\alpha(\xi) \log_2 n) = H_{\alpha,d}(\xi)$$

exists, the quantity $H_{\alpha,d}(\xi)$ is called the *d-dimensional information of order α contained in the value of the random variable ξ*. In the important case when the distribution of ξ is absolutely continuous, we have the following.

Theorem 2.1.4. Let ξ be a random variable having an absolutely continuous distribution with density function $p(x)$. If we put $\xi_n = \frac{[n\xi]}{n}$ $(n \in \mathbb{N})$ and if we suppose that ξ_1 is finite and that for $\alpha > 0$

$$\lim_{n\to+\infty} \frac{H_\alpha(\xi_n)}{\log_2 n} = 1, \tag{2.20}$$

i.e. the dimension of order α of ξ is equal to 1; if the the integral for $\alpha \neq 1$

$$\int_{\mathbb{R}} p^\alpha(x)dx \tag{2.21}$$

exists, then

$$\lim_{n\to\infty} (H_\alpha(\xi_n) - \log_2 n) = H_{\alpha,1}(\xi) = \frac{1}{1-\alpha} \log_2 \int_{\mathbb{R}} p^\alpha(x)dx. \tag{2.22}$$

For $\alpha = 1$ if

$$-\int_{\mathbb{R}} p(x) \log_2 p(x)dx \tag{2.23}$$

exists, then

$$\lim_{n\to\infty} (H_1(\xi_n) - \log_2 n) = H_{1,1}(\xi) = -\int_{\mathbb{R}} p(x) \log_2 p(x)dx. \tag{2.24}$$

Therefore, the two latter limits define generalized Rényi information measure of continuous distributions. Definitions also hold for higher dimensions $d > 1$ of the random variable ξ.

Example 2.1.5. At this point we can derive RE of the N-dimensional multivariate Gaussian distribution:

$$\begin{aligned}
\frac{1}{1-\alpha} \log_2 \int_{\mathbb{R}} d^N x p_G^\alpha(x) &= \frac{1}{1-\alpha} \log_2 \int_{\mathbb{R}} d^N x \det(2\pi\Sigma)^{-\frac{\alpha}{2}} \exp(-\frac{\alpha}{2} x^T \Sigma^{-1} x) \\
&= \frac{1}{1-\alpha} \log_2 \left[\det(2\pi\Sigma)^{-\frac{\alpha}{2}} (\det(\Sigma^{-1}))^{\frac{1}{2}} \left(\frac{2\pi}{\alpha}\right)^{\frac{N}{2}} \right] \\
&= \frac{1}{1-\alpha} \log_2 \left[(2\pi)^{\frac{N}{2}(1-\alpha)} \det(\Sigma)^{\frac{1}{2}(1-\alpha)} \alpha^{-\frac{N}{2}} \right] \\
&= \log_2 \left[(2\pi)^{\frac{N}{2}} \det(\Sigma)^{\frac{1}{2}} \right] + \frac{1}{1-\alpha} \log_2 \left[\alpha^{-\frac{N}{2}} \right],
\end{aligned}$$

where Σ is $N \times N$ covariance matrix. RE of 1-dimensional Gaussian distribution is depicted on the Figure 2.2.

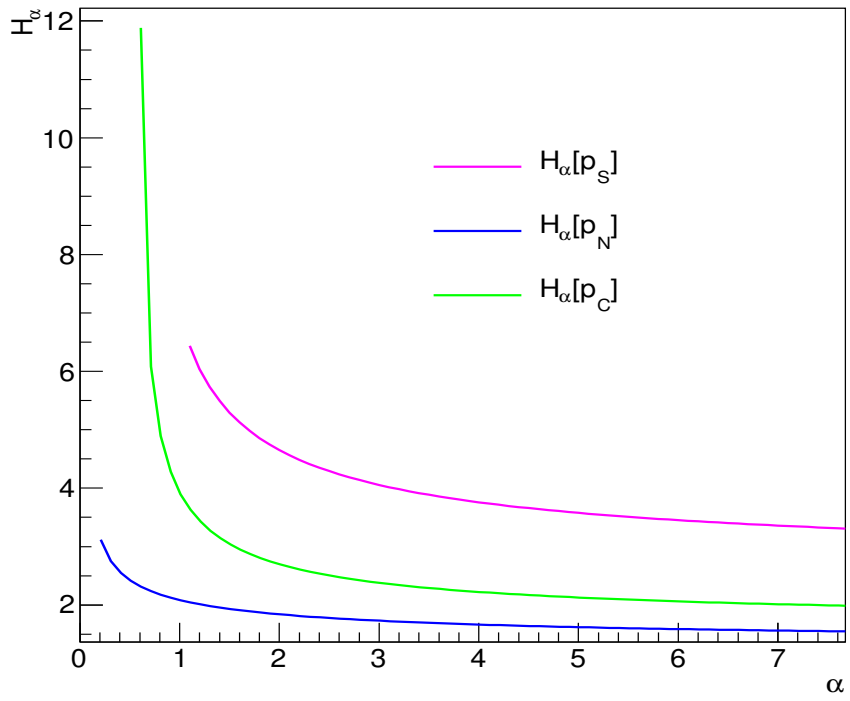


Figure 2.2: RE of probability distributions: Normal, Cauchy and 2-dimensional Student's t-distribution with $\nu = 1$.

Example 2.1.6. RE of the Cauchy probability density function (1.13) can be quantified in a following way (we put $\mu = 0$)

$$\frac{1}{1-\alpha}\log_2\int_{\mathbb{R}} dx p_C^{\alpha}(x) = \frac{1}{1-\alpha}\log_2\int_{\mathbb{R}} dx \left(\frac{1}{\pi}\frac{1}{(x/c)^2+1}\right)^{\alpha}.$$

To calculate the integral we will use *Schwinger trick*, widely used in quantum field theory. It stem from integral definition of Gamma function, that can be rewritten as

$$\frac{1}{A(z)^n} = \frac{1}{\Gamma(n)}\int_0^{+\infty} du\ u^{n-1}\exp(-uA(z)). \tag{2.25}$$

Therefore, we can write

$$\begin{aligned}\int_{\mathbb{R}} dy \left(\frac{1}{\pi}\frac{1}{y^2+1}\right)^{\alpha} &= \frac{1}{\pi^{\alpha}\Gamma(\alpha)}\int_0^{+\infty} du\ u^{\alpha-1}\int_{\mathbb{R}} dy \exp(-u(1+y^2)) \\ &= \frac{1}{\pi^{\alpha}\Gamma(\alpha)}\int_0^{+\infty} du\ u^{\alpha-1}\sqrt{\frac{\pi}{u}}e^{-u} \\ &= \frac{\sqrt{\pi}}{\pi^{\alpha}\Gamma(\alpha)}\int_0^{+\infty} du\ u^{\alpha-1-\frac{1}{2}}e^{-u} \\ &= \frac{\sqrt{\pi}}{\pi^{\alpha}\Gamma(\alpha)}\Gamma(\alpha-\frac{1}{2}),\end{aligned}$$

where $\alpha > \frac{1}{2}$. Finally we can write RE of the Cauchy distribution Fig.(2.2)

$$H_{\alpha}[p_C] = \frac{1}{1-\alpha}\log_2\left[\frac{1}{\pi^{\alpha-\frac{1}{2}}}\frac{\Gamma(\alpha-\frac{1}{2})}{\Gamma(\alpha)}\right]. \tag{2.26}$$

Example 2.1.7. One more analytical result can be obtained for Student's t-distribution. We generalize (1.16) as n-dimensional multivariate Student's t-distribution defined as

$$p_S(\mathbf{x}) = \frac{\Gamma\left(\frac{\nu+n}{2}\right)}{\Gamma(\frac{\nu}{2})(\nu\pi)^{\frac{n}{2}}\sqrt{\det(\Sigma)}}\left[1+\frac{1}{\nu}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]^{-\frac{\nu+n}{2}}. \tag{2.27}$$

Σ is a PD $n\times n$ matrix and $\boldsymbol{x}, \boldsymbol{\mu}$ are real n-dimensional vectors. We put $\boldsymbol{\mu} = \mathbf{0}$, $a \equiv \frac{\nu+n}{2}$ and calculate (using the Schwinger trick (2.25)) following integral

$$\begin{aligned}\int_{\mathbb{R}} \frac{d^n x}{\left(1+\frac{1}{\nu}\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}\right)^{\alpha a}} &= \frac{1}{\Gamma(\alpha a)}\int_{\mathbb{R}} d^n x \int_{\mathbb{R}^+} du\ u^{\alpha a-1}\exp(-u[1+\nu^{-1}\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}]) \\ &= \frac{1}{\Gamma(\alpha a)}\int_{\mathbb{R}^+} du\ u^{\alpha a-1}e^{-u}\left(\frac{\nu\pi}{u}\right)^{\frac{n}{2}}\sqrt{\det(\Sigma)} \\ &= \frac{\sqrt{\det(\Sigma)}}{\Gamma(\alpha a)}(\nu\pi)^{\frac{n}{2}}\int_{\mathbb{R}^+} du\ u^{\alpha a-1-\frac{n}{2}}e^{-u} \\ &= \frac{\sqrt{\det(\Sigma)}}{\Gamma(\alpha a)}(\nu\pi)^{\frac{n}{2}}\Gamma(\alpha a-\frac{n}{2}).\end{aligned}$$

Finally, we can write RE of multivariate Student's t-distribution Fig.(2.2)

$$H_{\alpha}[p_S] = \frac{1}{1-\alpha}\log_2\left[\frac{\Gamma^{\alpha}(\frac{\nu+n}{2})\Gamma(\alpha\frac{\nu+n}{2}-\frac{n}{2})}{\Gamma^{\alpha}(\frac{\nu}{2})\Gamma(\alpha\frac{\nu+n}{2})}|\Sigma|^{\frac{1}{2}(1-\alpha)}(\nu\pi)^{\frac{1}{2}n(1-\alpha)}\right], \tag{2.28}$$

which hold for ν and α such that $Re(\alpha\frac{\nu+n}{2}-\frac{n}{2}) > 0$. For $\nu = 1$ Student's t-distribution is equal to the Cauchy distribution. One can check that substitution of $\nu = 1$ and $n = 1$ into (2.28), using that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, gives indeed (2.26).

2.2 Transfer entropy

2.2.1 Shannon transfer entropy

Let $X \equiv x_1, x_2, ..., x_N$ be a discrete stochastic process with complete probability distribution P_X, then the Shannon entropy of this process is

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x). \tag{2.29}$$

Let $Y \equiv y_1, y_2, ..., y_N$ be another stochastic process, then mutual information between X and Y is

$$I(X:Y) = \sum_{x \in X y \in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} = H(X) - H(X|Y). \tag{2.30}$$

Mutual information measures an average information shared between two processes X and Y. It is symmetric $I(X:Y) = I(Y:X)$, therefore can not be considered as directional information flow. Additionally, $I(X:X) = H(X)$, i.e. amount of information about X contained in X is its entropy.

The averaged mutual information between two processes X and Y conditioned on the third process Z is called *conditional mutual information* and is defined as

$$I(X:Y|Z) = H(X|Z) - H(X|Y,Z) = I(X:(Y,Z)) - I(X:Y), \tag{2.31}$$

where quantity $H(X|Y)$ is the *conditional entropy*, that is averaged entropy of X under assumption that Y is known and is defined as

$$H(X|Y) = -\sum_{x \in X, y \in Y} p(x,y) \log_2 p(x|y). \tag{2.32}$$

We assume X and Y to be a Markov processes of order k and l respectively. Let us try to replace the process Z from (2.31) with the values of from the process X denoting them $x_n^{(k)} \equiv (x_n, x_{n-1}, ...x_{n-k+1})$. We want to know, what is the average information shared between the future value of the process X and known historical values of the process Y provided that previous values of X are also known, i.e. we use equation (2.31) and obtain

$$
\begin{aligned}
I(x_{n+1} : y_n^{(l)} | x_n^{(k)}) &= H(x_{n+1}|x_n^{(k)}) - H(x_{n+1}|y_n^{(l)}, x_n^{(k)}) \\
&= -\sum_{x\in X} p(x_{n+1}, x_n^{(k)}) \log_2 \left(\frac{p(x_{n+1}, x_n^{(k)})}{p(x_n^{(k)})} \right) \\
&+ \sum_{x\in X y\in Y} p(x_{n+1}, x_n^{(k)}, y_n^{(l)}) \log_2 \left(\frac{p(x_{n+1}, x_n^{(k)}, y_n^{(l)})}{p(x_n^{(k)}, y_n^{(l)})} \right) \\
&= \sum_{x\in X y\in Y} p(x_{n+1}, x_n^{(k)}, y_n^{(l)}) \log_2 \left(\frac{p(x_n^{(k)})}{p(x_{n+1}, x_n^{(k)})} \frac{p(x_{n+1}, x_n^{(k)}, y_n^{(l)})}{p(x_n^{(k)}, y_n^{(l)})} \right) \\
&= \sum_{x\in X y\in Y} p(x_{n+1}, x_n^{(k)}, y_n^{(l)}) \log_2 \left(\frac{p(x_{n+1}|x_n^{(k)}, y_n^{(l)})}{p(x_{n+1}|x_n^{(k)})} \right).
\end{aligned}
$$

The quantity

$$
\left(\frac{p(x_{n+1}|x_n^{(k)}, y_n^{(l)})}{p(x_{n+1}|x_n^{(k)})} \right) \tag{2.33}
$$

has the meaning of the deviation from the Markov property. If $p(x_{n+1}|x_n^{(k)}, y_n^{(l)}) = p(x_{n+1}|x_n^{(k)})$ for all $x \in X$ and $y \in Y$ the logarithm function will return 0 and the gain in information is equal to 0. The separate values of logarithm in the sum can be negative as well as positive, but the sum is always non-negative.

Conditional mutual information is also known as *Shannon Transfer Entropy* or, in our case, *gain in information about x_{n+1} caused by the history of Y up to $y_n^{(l)}$ under the assumption that the history of X up to $x^{(k)}$ is known*. As a directional measure of information transfer, transfer entropy was introduced by T. Schreiber in [39], and we will denote it as

$$
T_{Y\to X}(k, l) = \sum_{x\in X y\in Y} p(x_{n+1}, x_n^{(k)}, y_n^{(l)}) \log_2 \left(\frac{p(x_{n+1}|x_n^{(k)}, y_n^{(l)})}{p(x_{n+1}|x_n^{(k)})} \right). \tag{2.34}
$$

For an independent processes transfer entropy is equal to zero. For a non-zero cases transfer entropy measures the deviation from the independence of the two processes. An important property of the transfer entropy is that it is directional, i.e. in general $T_{Y\to X} \neq T_{X\to Y}$.

2.2.2 Rényi transfer entropy

Following the previous ideas one can now obtain generalizations using Rényi's information measures with parameter $\alpha > 0$ defined for a stochastic

process X as

$$H_\alpha(X) := \frac{1}{1-\alpha}\log_2 \sum_{x\in X} p^\alpha(x). \tag{2.35}$$

In the same manner with (2.31) we can derive Rényi transfer entropy first introduced in [14] as

$$T^R_{\alpha,Y\to X}(k,l) = H_\alpha(x_{n+1}|x_n^{(k)}) - H_\alpha(x_{n+1}|x_n^{(k)}, y_n^{(l)}) = I_\alpha(x_{n+1} : y_n^{(l)}|x_n^{(k)}), \tag{2.36}$$

where $H_\alpha(X|Y)$ is the *conditional entropy of order* α and $I_\alpha(X:Y)$ is the *mutual information of order* α defined in [14] as

$$H_\alpha(X|Y) = \frac{1}{1-\alpha}\log_2 \frac{\sum_{x\in X, y\in Y} p^\alpha(x,y)}{\sum_{y\in Y} p^\alpha(y)}, \tag{2.37}$$

$$I_\alpha(X:Y) = \frac{1}{1-\alpha}\log_2 \frac{\sum_{x\in X, y\in Y} p^\alpha(x)p^\alpha(y)}{\sum_{x\in X, y\in Y} p^\alpha(x,y)}. \tag{2.38}$$

Rényi's transfer α-entropy also includes the case of Shannon entropy, i.e.

$$\lim_{\alpha\to 1} T^R_{\alpha,Y\to X} = T_{Y\to X}. \tag{2.39}$$

Rényi TE can also end up with a negative result, that is not true for the Shannon TE. It implicates that uncertainty of the process X becomes bigger knowing the past of Y, i.e. $H_\alpha(x_{n+1}|x_n^{(k)}) \le H_\alpha(x_{n+1}|x_n^{(k)}, y_n^{(l)})$. If X and Y are independent, then $T^R_{\alpha,Y\to X} = 0$. Unlike in Shannon's case $T^R_{\alpha,Y\to X} = 0$ does not necessarily implicate independence of the processes.

2.2.3 Escort distribution

The Rényi transfer entropy is able to locate an information transmissions between certain parts of underlying distributions. For $0 < \alpha < 1$ information flows accentuates for marginal events. For $\alpha > 1$ more probable events are emphasized. Therefore one can "zoom" different parts of probability density functions involved by choosing different values of α. This is particularly very helpful in describing systems, where marginal events are of the interest.

In order to understand a "zooming" property of RTE we rewrite (2.36) explicitly in the form

$$T^R_{\alpha,Y\to X}(k,l) = \frac{1}{1-\alpha}\log_2 \left(\frac{\sum \frac{p^\alpha(x_n^{(k)})}{\sum p^\alpha(x_n^{(k)})} p^\alpha(x_{n+1}|x_n^{(k)})}{\sum \frac{p^\alpha(x_n^{(k)}, y_n^{(l)})}{\sum p^\alpha(x_n^{(k)}, y_n^{(l)})} p^\alpha(x_{n+1}|x_n^{(k)}, y_n^{(l)})} \right). \tag{2.40}$$

The latter form presents how the underlying distribution is changed with the change of parameter α. Numerator and denominator inside the log-function contain the *escort distributions* ρ_α

$$\rho_\alpha \equiv \frac{p^\alpha(x)}{\sum p^\alpha(x)}, \tag{2.41}$$

which is emphasizing less probable events when $0 < \alpha < 1$ and more probable events when $\alpha > 1$, see Fig.(2.1). One can see, that numerator in (2.40) presents the average probability weighted with $\rho_\alpha(x_n^{(k)})$ and the denominator is average with respect to the $\rho_\alpha(x_n^{(k)}, y_n^{(l)})$.

For $0 < \alpha < 1$ factor before log is positive, therefore RTE is negative, i. e. $I_\alpha(x_{n+1} : y_n^{(l)}|x_n^{(k)}) < 0$, if the rate in which the marginal parts of X obtaining the past of Y are enhanced is bigger than the rate of suppression for large probability events.

Analogically, for $\alpha > 1$ RTE is negative when, by learning Y, the large probability events are suppressed more than marginal events are enhanced.

2.2.4 Transfer entropy by means of filtration

Let $\mathcal{X}$ be a natural filtration of the process X and $\mathcal{Y}$ - natural filtration of the process Y, then we can rewrite conditional RE and thus RTE in the following manner

$$H_\alpha(x_{n+1}|x_n^{(k)}) = H_\alpha(X|\mathcal{X}), \tag{2.42}$$

$$T^R_{\alpha,Y\to X} = H_\alpha(X|\mathcal{X}) - H_\alpha(X|\mathcal{X}\cap\mathcal{Y}), \tag{2.43}$$

where we consider $k \to +\infty$ and $l \to +\infty$. Conditioning with the respect to filtration is a generalization of the discussed case of Markov process to any stochastic process. Natural filtration contains information about all possible realizations of trajectories up to the moment n. Moreover, scaling by the parameter $\alpha \geq 1$ produces filtrations which are adapted to the natural filtration.

In case one wants to control an information concealed in a process he or she can choose to work with subsequences, i.e. conditioning on some chosen subset $X^{\{k\}}$ of past values from the processes X. The subset would be a coarser version of original time sequence with a natural filtration $\mathcal{F}_n(X^{\{k\}}) \subset \mathcal{F}_n(X)$. Therefore the subprocess $X^{\{k\}}$ is adapted to the natural filtration of the process X and they bear some common information. This approach would be especially desirable for proper numerical processing

of the data (Chapter 5), because high dimensions negatively influence accuracy of estimation. For further convenience we introduce a new notation of the RTE Eq.(2.36) by means of conditioning on subsequences as

$$\begin{aligned} T^{R}_{\alpha,Y\to X}(\{k\},\{m\},\{l\}) &= H_\alpha(x_{n+m}|x_n^{\{k\}}) \;-\; H_\alpha(x_{n+m}|x_n^{\{k\}},y_n^{\{l\}}) \\ &= I_\alpha(x_{n+m}:y_n^{\{l\}}|x_n^{\{k\}})\,, \end{aligned} \tag{2.44}$$

where $x_n^{\{k\}}$ is a subset of past values of X up to the time n with number of elements equal to k, such that $\{k\}=\{\kappa_1,...,\kappa_k\}$ is a set of indices and $x_n^{\{k\}}=\{x_{n-\kappa_1},x_{n-\kappa_2},...,x_{n-\kappa_k}\}$ is a subsequence (the same logic applies to $y_n^{\{l\}}$). We also added a third parameter m, so called future step. Such parametrization is used in literature as well, for instance [42].

2.3 Summary

Main takeaways of this chapter are:

- Rényi $\alpha-$entropy

$$H_\alpha[\mathcal{P}] := \frac{1}{1-\alpha}\log_2\sum_{i=1}^{n} p_i^\alpha \tag{2.45}$$

  is the most general class of information measures preserving additivity for independent systems and compatible with Kolmogorov's probability axioms;

- Shannon transfer entropy between two processes is defined as a Kullback-Leibler divergence or average deviation from a Markov property;

- concept of transfer entropy can be extended by means of Rényi entropy, resulting in a more general definition of the TE with an "$\alpha-$zooming" property;

- zooming property is given by the presence of escort distributions, which can emphasize or suppress different parts (for instance tails or peaks) of probability distributions.

Chapter 3

Causation

3.1 Order

Cause and *effect* are intrinsic to the concept of *causality*. Intuitively, cause must happen before effect and causality is something between them. By "something" we may mean some physical measurable quantity or it might be an abstract term arising from the existence of cause and effect itself. Causality is also usually falsely understood as correlations, however, correlations does not necessarily exclude causation. Moreover, the concept of cause and effect is very similar to *induction* principle, which is core to the logical thinking in exact sciences.

In classical physics, causation is associated with deterministic laws or arrow of the time. Modern physics, however, changed the status quo by revealing a probabilistic nature of the quantum mechanics (Heisenberg uncertainty principle) or backward causation in the quantum field theory (electron-positron annihilation). These observations remind us about uncompromising nature of physical laws, that lead us to the necessity of *methodological scepticism* advocated by *René Descartes*. Indeed, in one of the rules on method proposed in [6] he states *"...and even supposing an order among those things that do not naturally precede one another"*. Further in the text, giving comments on the rules proposed by him, he writes *"...and that one always maintains the order to be followed in deducing the ones from the other"*.

In this chapter we want to study connection of the Rényi transfer entropy with causality. The former mentioned suggests that causality, despite of its simplicity, is an abstract concept, which can be to a certain extent associated with *order*.

3.2 Predictability

With order comes predictability. Let us investigate what predictability is in terms of classical mechanics. We consider a simple system of the falling object associated with a mass m. Differential equation capturing dynamics of this particle is Newton's equation of motion

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = F(x), \tag{3.1}$$

where t is time, $x(t)$ is trajectory and $F = mg$ is gravitational force. Knowing initial or boundary conditions, one can solve the equation and get

$$x(t) = x_0 + v_0t + \frac{1}{2}gt^2. \tag{3.2}$$

Now, prediction of a position $x(t)$ in any instance is possible. Hence, predictability in this case means finding a solution of differential equation.

However, predictions become complicated when system of interacting particles is under consideration. In statistical mechanics a probabilistic framework is used to study macroscopic behaviour arising from the microscopic components of a system. Trajectories of composing particles do not follow any specific trajectory with probability equal to 1. On the opposite, here is a number of different trajectories occurring with different probabilities less than 1. For example, a gas particle is observed in a state associated with energy ϵ_i with a probability given by the *Boltzmann-Gibbs distribution*

$$p_i \propto e^{-\frac{\epsilon_i}{kT}}, \tag{3.3}$$

where k is the *Boltzmann constant* and T is thermodynamic temperature.

Probabilistic approach is, in fact, more general, since the classical deterministic trajectories (3.2) can be associated with the Dirac $\delta-$distribution (1.6) by

$$prob(x) \propto \prod_i \delta(x - x(t_i)). \tag{3.4}$$

Thus, predictability stems from probability functions and even for deterministic processes it is more realistic to say that the present state determines the *distribution* of the future sates (which is precisely the *Markov property*). We already know that information given by a probability p can be decoded by unexpectedness functional (1.39)

$$-\log p$$

or its average - entropy. Entropies naturally appear in thermodynamics as an "order measure", or can be used to find a probability distribution from given conditions by means of *max-entropy principle*. Nevertheless, in this work we don't go further into these concepts and proceed with information-theoretic entropies.

3.3 Granger causality

In terms of time series analysis, causality was introduced by C. Granger, the 2003 Nobel prize winner in economy. He succeeded N. Wiener who formulated the problem of causality between two processes in the following way [44]: "*For two simultaneously measured signals, if we can predict the first signal better by using the past information from the second one than by using the information without it, then we call the second signal causal to the first one.*" Granger identified two conditions of causal dependence as [11]

- (i) the cause occurs before the effect;
- (ii) the cause contains information about the effect that is unique, and is in no other variable.

It suggests that one system can be used to predict another one. Let $\{X_t\}_{t\in N}$ and $\{Y_t\}_{t\in N}$ be two stationary processes, we say that the process X_t *Granger causes* another process Y_t if future values of Y_t can be better predicted using the past values of X_t and Y_t rather than only past values of Y_t. Thus from [10]

$$Y_t = a_0 + \sum_{i=1}^{l} a_i Y_{t-i} + \sum_{j=1}^{k} b_j X_{t-j} + e_t, \tag{3.5}$$

where e_t is a random variable with zero mean and variance σ^2, a_0, a_i, b_j are constants and l and k are memory indices. If X_t does not Granger cause Y_t then $b_j = 0\ \forall j \in \{1, 2, ..., k\}$.

The equation (3.5) can be reformulated in terms of probability theory. Let $P(y_{t+1}|x_t^{(k)}, y_t^{(l)})$ denote the probability distribution function of the process Y conditional on the joint (k, l)-history of both processes. Let $P(y_{t+1}|y_t^{(l)})$ denote the probability distribution function of Y conditional on just its own l-history. Then variable X is said to Granger-cause variable Y (with lags k, l) iff

$$P(y_{t+1}|x_t^{(k)}, y_t^{(l)}) \neq P(y_{t+1}|y_t^{(k)}). \tag{3.6}$$

This statement is similar to the term (2.33) contained in the transfer entropy. Therefore, transfer entropy can be used as a statistic test for Granger

causality. On the other hand, Granger causality can only detect linear dependencies between variables, thus transfer entropy holds as its generalization. However, for Gaussian variables RTE and Granger causality are equivalent.

3.4 Information flow

We begin by clarifying concepts of information transfer and information flow. To do so, we adopt approach presented in [1] and further discussed in [31]: *information transfer* is viewed as the amount of information that a source adds to the next state of a destination, while *information flow* refers to the extent to which a source has a direct influence on the next state of a destination variable. It is, indeed, a flow present and transmitted through a system with direction and dynamics.

Let us denote

$$p(x|y) = p(x) \tag{3.7}$$

as a conditional independence, i.e. *correlation* is absent, and

$$p(x|\hat{y}) = p(x) \tag{3.8}$$

as an interventional (causal) independence, i.e. absence of *information flow*. For instance, if x is correlated with z and y is also correlated with z, then $p(x|y) \neq p(x)$ (correlation is detected), on the other hand, $p(x|\hat{y}) = p(x)$ still holds, since the correlation caused by z doesn't imply causal dependence between x and y. So far, $p(x|y)$ can be viewed as a variable of observational character and $p(x|\hat{y})$ is rather based on inside mechanisms.

To be able to measure causal flow, one needs to interfere with hidden mechanism of information flow transmission. *Intervention* is a path towards causality detection, otherwise, one might be left measuring correlations.

To detect information flow with RTE, one should vary number of past values of process X, i.e. parameter k, in RTE to see if the flow changes, at some instance RTE should remain constant. This is equal to imposing intervention source to be sure that we measure causality, not correlations. Moreover, by varying α we consider new trajectories Fig.(3.1), thus new sources of information.

3.5 Summary

In this chapter we aimed to discuss ideas linking entropy with a measure of causality. Our conception can be summarized as follows:

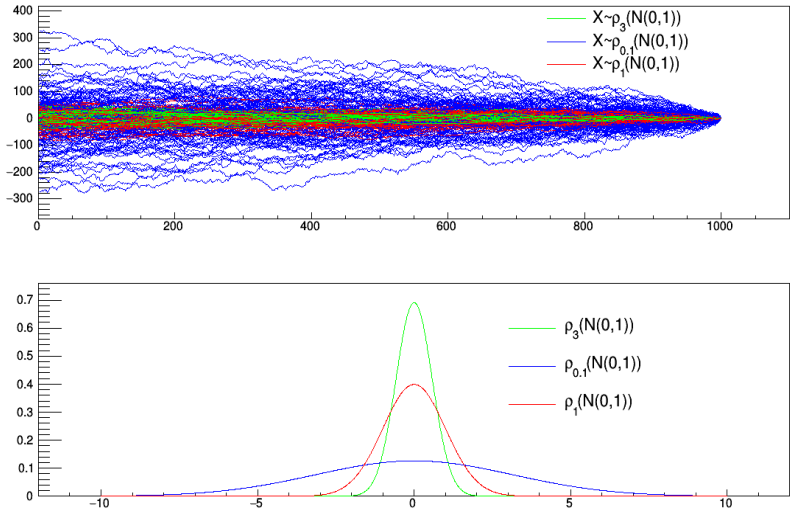


Figure 3.1: Escort distribution (2.41) for three different values of α and trajectories generated with respect to these distributions.

- future steps of deterministic systems with given initial conditions are predictable with a probability equal to 1, however, systems with some extent of randomness are predictable with probability $\in (0, 1)$;
- entropy measures ignorance we have about a system, for deterministic systems it is equal to 0;
- transfer entropy can be used as a statistic test for Granger causality;
- entropic temporal differences or temporal rates are useful for the description of the dynamics of a system, thus can be thought of as a good candidate for causality measure;
- intervention is a path towards causality detection, otherwise, one might be left measuring correlations;
- parameter k present in the definition of RTE can be used as an intervention source.

Chapter 4

Stochastic nature of markets

4.1 Complex systems

It is a well known fact that physics is an exact science, guided by logic and objectiveness, followed by experiments, in order to deal with matter and its interactions. Notwithstanding the fact how successful Newtonian approach of differential equations was in the last five centuries, we are beginning to recognize a new phenomenon for which purely deterministic laws are not giving satisfactory results. We call it *complex systems*.

Complex systems arise in physics, biology, chemistry or social sciences. One can think of a complex system as of a system composed of number of subsystems interacting with each other in generally non-linear way. Dynamics of complex systems, resulting from such behaviour, typically perform self-similarity, scaling, power laws, self-organization, phase transitions, emergence, evolution or chaos among others. These phenomena are not mutually exclusive, on the other hand, they usually appear simultaneously or dependently. This is an important feature of complex system, because it can not be studied in terms of decomposition. Sometimes, it is possible to obtain separated subsystems, however, the dynamics of the whole system would be violated. In other words, we can not control experimentally intrinsic dynamics inside the system. This is quite in contrast with traditional approach of physical science - to separate system into elementary, well describable, components, in order to apply deterministic laws on it, and make predictions, which can be verified experimentally. The traditional approach stems also from the fact that four fundamental forces act on different scales, thus, when dealing with any physical phenomenon, one can neglect all forces but one. That is not a case in complex systems. This suggests that phenomenon of complex systems might be a new scientific discipline. With its own understanding of composing elementary "particles" (which are not limited to the matter, as it is in physics) and the laws of interactions (beyond the four

fundamental forces). More about complex systems as a new scientific field can be found in [41].

4.2 Financial markets

Financial markets are typical complex systems. Time series from stocks perform self-similarity, power laws, phase transitions or turbulence among others. It is, thus, tempting to take well-known tools from physics and apply them on market data. A branch of scientific discipline studying markets or economic activity from the physical point of view is called *econophysics*.

Brownian motion was the first approximation of stock prices behaviour in terms of physics. French mathematician *Louis Bachelier* is recorded to be the first, who, at the beginning of the $20th$ century, investigated that price returns are normally distributed, i.e. in the same way as step increments of random walker or Brownian particle. A half century later, *Benôit Mandelbrot* recognized a self-similar behaviour of stock prices. A number of books or articles exploring and generalizing stochasticity and scaling invariance of markets have appeared ever since. However, vast bulk of this literature didn't lead to real use of the obtained results.

In this context, one should be aware that financial data, or any data obtained from economic activity, differ, in terms of information they carry, from the data obtained via physical experiment. Basically, when information gained from stock activity is used, it alters the stock activity. It is what happens during the trading. Agents are trying to understand behaviour on the markets, "decode" the information emitted by some hidden source, in order to know what and when to buy or sell. At the same time, however, another agent can buy stocks in such amount that will signal growing demand, hence, growing price leading to change of market behaviour. The latter can be associated with the change in the hidden, information emitting, mechanism. This is something known as a self-organization. Inside forces of a system move it forward, making it to evolve over time. Understanding that stock behaves in such manner, and ability to prove it on data, would be satisfying result for physicist whose job is to objectively describe phenomenon. However, this particular knowledge does not provide any relevant information to the trader who wants to decide whether to buy or sell. Moreover, a number of problems in physics with well defined equation of motion can not be solved analytically. Interactions in complex systems are evolving over time, meaning that the internal structure is changed as the system operate. This would be very challenging to describe analytically, however, one can use an algorithmic approach. *Algorithm* is basically a self-updating list of

rules, enabling to describe evolution of the states as well as interactions between them. Algorithmic analogue to the classical equation of motion (3.1) can be written as a recurrence relation

$$x_{n+1} = f(x_n), \tag{4.1}$$

$$x_{n+2} = f(f(x_n)). \tag{4.2}$$

Well known example is the *logistic map*

$$f_r(x_n) = rx_n(1 - x_n), \tag{4.3}$$

or the Rössler chaotic system that will be discussed in the Chapter 5.

Information-theoretic approach to stock markets provide a new view on the problems arising from the stochastic description of markets. The random nature of stochastic process can be successfully dealt with by means of probability theory and statistics. Probability density functions can be defined by its moments (1.11), but it is usually hard to obtain them from real data that are non-stationary. Thus, with stochastic approach we must always be aware of limitations arising from the domain on which our theory is valid. In the following sections we give more details about stochastic nature of price changes on stocks, and present Black and Scholes theory, which is, so far, the greatest contribution of econophysics to the real-world markets. Nevertheless, as will be presented, it suffers from a number of limitations stemming from Brownian assumption among others. That is why, at the end of the chapter, we discuss an information-theoretic approach that doesn't load real data with any special requirements.

4.3 Financial risks

Modern financial institutions are exposed to many risks including reputation, legal or climate risks. Nevertheless, *financial risk* is the biggest threat to the daily functioning of these companies. Financial risk can be categorized into three different kinds of risk - credit risk, liquidity risk, and market risk. *Credit risk* is the risk of a borrower defaulting on a loan, or related financial obligation. In the context of traded markets, *liquidity risk* is the risk of being unable to buy or sell assets in a given size over a given period without adversely affecting the price of the asset. *Market risk* is the risk of losses on financial investments caused by adverse price movements. In this work we are interested in the market risk.

Uncertain periods, for instance the so called *trade war* between China and the USA in 2019, are usually followed by *volatility* on markets. If an

asset is volatile - it is unpredictable, thus risky. In terms of thermodynamics, volatility can be associated with temperature. If temperature is low, particles have low kinetic energy and remain around a still position. On the other hand, with rising temperature particles start to move and interact, resulting in random behaviour. The direction and size of the increments in time are, thus, hard to predict. Let us now clarify how the price increments are defined.

Price increment has no unique definition. It is a stochastic variable mimicking the price dynamics of the underlying asset. The first, very straightforward definition of a price increment is called the *price change*

$$\delta x_\tau(t) := x(t+\tau) - x(t). \tag{4.4}$$

It is an intuitive approach, however it might be misinterpreted when different scales are under the study. Another definition is called the *discounted price change*

$$Z_\tau(t) := \delta x_\tau(t) D(t), \tag{4.5}$$

where $D(t)$ is a discounting factor. The latter definition accounts for real-time money value, however $D(t)$ is time dependent and is unpredictable in long term. Third definition is called the *price return*

$$R_\tau(t) := \frac{\delta x_\tau(t)}{x(t)}. \tag{4.6}$$

The merit is that returns provide information about relative change. The problem is that they are sensitive to scale changes in long time horizons. Another choice of definition is called the *price log-return*

$$S_\tau(t) := \log\left(\frac{x(t+\tau)}{x(t)}\right). \tag{4.7}$$

The appealing property of this definition is that the average correction of scale changes is incorporated without a need of including discounting factor, as in (4.5). However, non-linearity strongly affects the statistical properties of the underlying process.

Once one decides for the definition of the stochastic variable, he or she can proceed to study it by means of statistical analysis. This usually involves moments of random variable (1.4). We recall that first, second, third and fourth moments are called *mean*, *variance*, *skewness* and *kurtosis* respectively. Variance and kurtosis can provide us with information relevant to the risk.

Variance is the square of the standard deviation, thus tells how far from the mean value price can most probably get. Kurtosis describes tail parts of the underlying pdf. Tails can decay exponentially, as it is with the Gauss distribution (1.12), or decay polynomially - Cauchy (1.13) distribution Fig(1.1). Polynomial tails inform us about higher probability of events that are afar from the mean value. Probability distributions with sharp peaks and slowly decaying tails are typical for the distributions of the price increments Fig.(5.1). It is gaining more attention among researchers, however normal pdf remains popular in the real-world use for its analytical simplicity. Gaussian approximation can give good results during calm and certain periods of economic activity, nevertheless, it fails to describe turnovers characterized by the fat tails.

4.4 Stochastic processes with memory

4.4.1 Short-memory processes

In order to obtain a stochastic process similar in behaviour to the market processes we present simple processes first. They are also used to study and predict real processes, however in a limited way. Nevertheless, these models can be combined, which results in more sophisticated ones. At the same time, refinements of the basic models attempt to bring the characteristics of the time series closer to actual data. All models presented further are recurrence relations (4.1) with stochastic term, i.e.

$$x_{n+1} \equiv f(x_n, x_{n-1}, ...x_0) + \xi(n, n-1, n-2, ..., 0). \tag{4.8}$$

where ξ is a random variable.

Autoregressive process

The first basic model is the autoregressive process. The autoregressive process is one in which the change in a variable at a point in time is linearly correlated with the previous change. In general, the correlation declines exponentially with time and is gone in a relatively short period. A general form follows

$$C_t = e_t + aC_{t-1} + bC_{t-2}, \tag{4.9}$$

where C_t is a change at time t, a and b are constants such that $\mid a \mid \leq 1$, $\mid b \mid \leq 1$, and e_t is a white noise series with zero mean and unit variance. Equation (4.9) is the autoregressive process of order 2, or AR(2). One can extend it to the AR(n), where C_t is dependent on the previous n steps.

Moving average process

In a moving average (MA) process, the time series is the result of the moving average of an unobserved time series

$$C_t = \mu + ce_{t-1} + e_t, \tag{4.10}$$

where e is a random variable with μ mean and unit variance, c is a constant such that $\mid c \mid < 1$. The restriction on the moving average parameter c ensures that the process is invertible. $c > 1$ would imply that future events affect the present, which would be unrealistic. The observed time series C, is the result of the moving average of an unobserved random time series e. Because of the moving average process, there is a linear dependence on the past and a short-term memory effect.

ARMA process

In this type of model, we have both an autoregressive and a moving average terms

$$C_t = aC_{t-1} + \mu + be_{t-1} + e_t. \tag{4.11}$$

Models of this type are called mixed models and are typically denoted as ARMA(n,m), where n is the number of autoregressive terms, and m represents the number of moving average terms. On Fig.(4.1) we can see random paths of all three processes. It can be seen that ARMA is a sum of the other two. Despite of having information from both processes, ARMA is still not similar to real process (Fig.(5.1)) of some asset over a long period of time, nevertheless it can be used as a good approximation for short periods.

ARCH process

Models that exhibit autoregressive conditional heteroskedasticity (ARCH) have become popular for a number of reasons. They are a family of non-linear stochastic processes, as opposed to the linear-dependent AR and MA processes. Their frequency distribution is a high-peaked and fat-tailed. Empirical studies have shown that financial time series exhibit statistically significant ARCH.

The basic ARCH model was developed by R. F. Engle in 1982. Engle considered time series that were defined by normal probability distributions but time-dependent variances. The expected variance of a process is conditional on what it was previously. The basic ARCH(1) model is defined by

$$x_t \equiv \sigma_t e_t, \tag{4.12}$$

where

$$\sigma_t^2 = \beta_0 + \beta_{t-1} x_{t-1}^2, \tag{4.13}$$

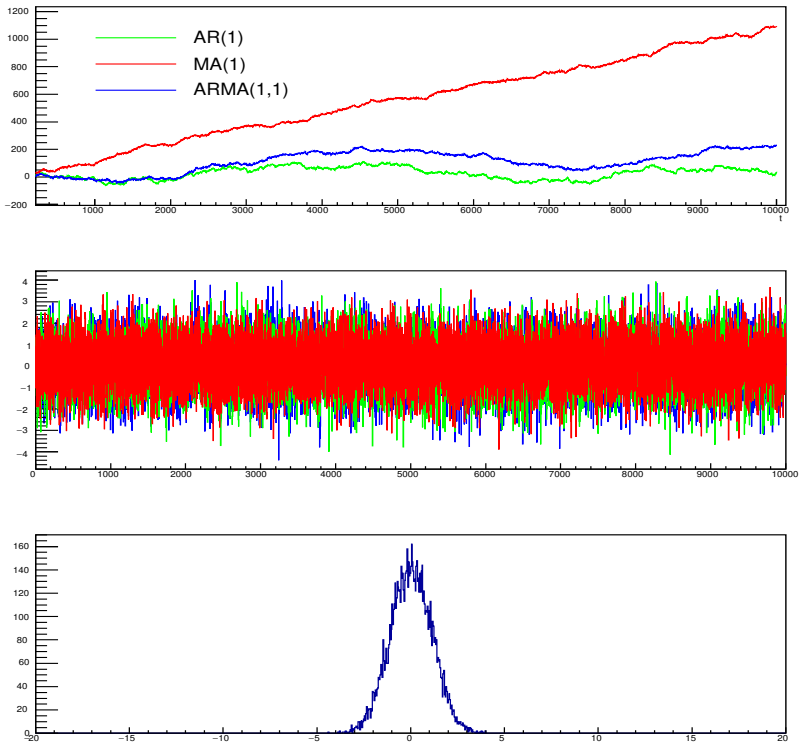


Figure 4.1: Simulation of AR(1), MA(1) and ARMA(1,1) processes. Increments of ARMA are normally distributed, thus no big jumps are visible in the noise diagram.

e_t is a random variable with zero mean and unit variance, and $\beta_i (i \in 0, 1, ..., t-1)$ are constants. By (4.12) we mean that x_t is a random variable with zero mean and variance σ_t^2. ARCH process is nonlinear. Small changes will likely be followed by other small changes, and large changes by other large changes, but the sign will be unpredictable. Also, because ARCH is nonlinear, large changes will amplify and small changes will contract. This results in the fat-tailed, high-peaked distribution.

GARCH process

The ARCH model was modified to make the σ variable dependent on the past of the process as well. T. Bollerslev proposed in 1986 a generalized ARCH or GARCH(1,1) model in the following manner

$$x_t \equiv \sigma_t e_t, \tag{4.14}$$

$$\sigma_t^2 = \beta_0 + \beta_{t-1} x_{t-1}^2 + \gamma_{t-1} \sigma_{t-1}^2. \tag{4.15}$$

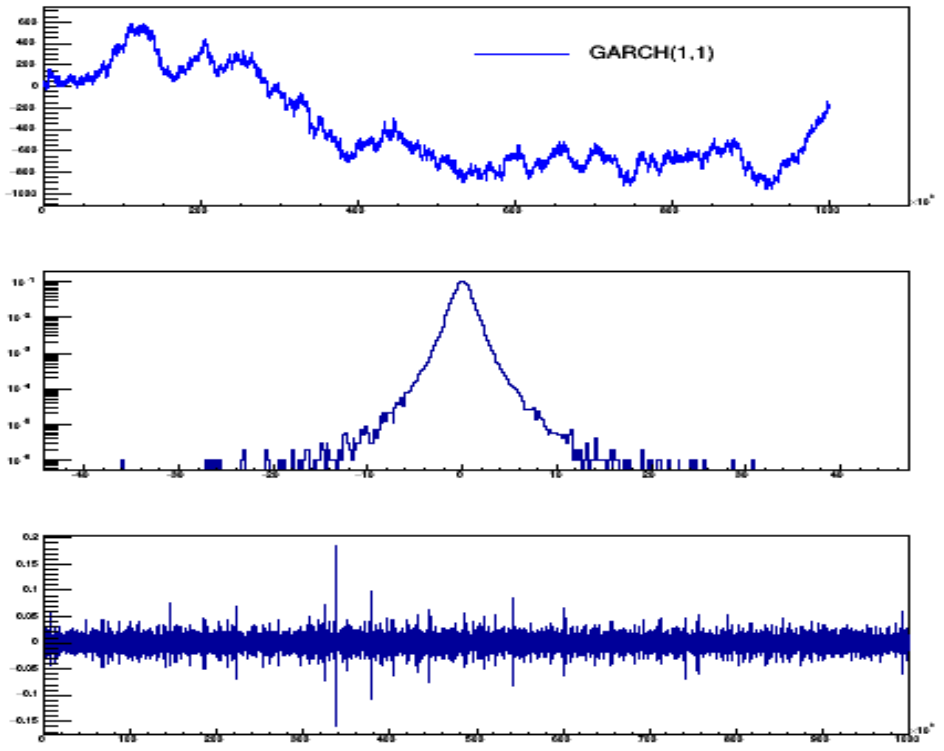


Figure 4.2: GARCH(1,1) process has fat-tailed distribution of increments. Noise histogram confirms that the process exhibits non-Gaussian behaviour.

The latter equation can be extended to GARCH(n, m), thus the variance σ_t^2 would be dependent on longer past. On Fig.(4.2) one can see GARCH(1,1) trajectory and high-peaked, fat-tailed distribution of the increments.

4.4.2 Long-memory processes

Pink noise B. Mandelbrot postulated the pink noise as a sum of a large number of parallel processes occurring over many different frequencies. These processes can be considered as several equilibrium states through which system is passing. In markets it could be seasonal trends, political conditions or the interest rate development. Therefore, the market as a whole would have many different parallel relaxation times in reaction to the information coming from the outside, and volatility would undergo many parallel shifts with different correlation or relaxation times.

Pink noise is also known as the *fractional noise* since it has fractional dimension (1.30). The formula for the pink noise X involves adding several

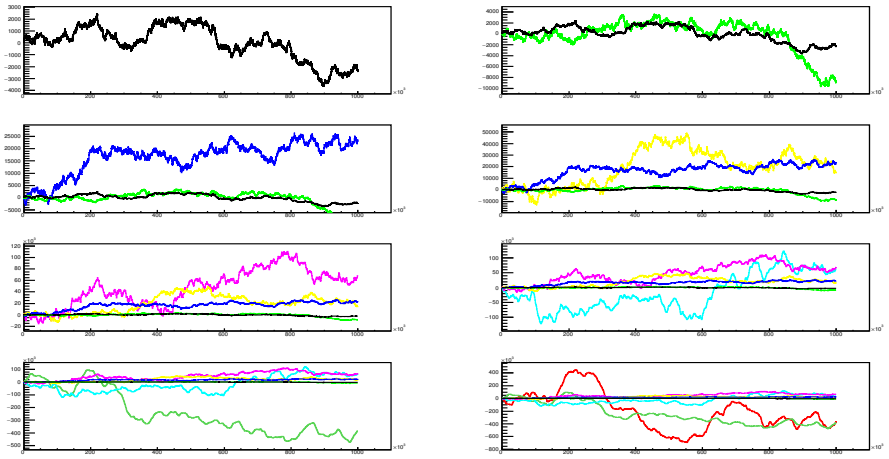


Figure 4.3: Pink noise sub-processes $x^{(t)}$ generated for $t \in T = \{0, 4, 32, 128, 512, 1024, 8192\}$, black trajectory is for $t = 0$, green for $t = 4$ and so on. The red trajectory (right bottom) is a pink process X^T generated as the sum of sub-processes.

parallel processes together, i.e.

$$X_n^T = x_n^{(1)} + x_n^{(2)} + ... + x_n^{(k)}, \tag{4.16}$$

where

$$x_n^{(t)} \equiv x_n(t) = c(t)x_{n-1} + u_{n-1}\sqrt{1 - c^2(t)} \tag{4.17}$$

is a single-frequency process with $x_0 = 0$, u_n is uniformly distributed random number and $c(t) = \exp(-1/t)$ is the frequency parameter. "Relaxation time" $t \in T$ of processes $\{x^{(t)}\}_{n \in N}$ should be evenly separated in log space, for example $T = \{1, 10, 100, ...\}$ or $T = \{1, 2, 4, 8, 16, 32, ...\}$. One can see, that for t small the process $\{x^{(t)}\}_{n \in N}$ will be determined by the random term u_n, however, with growing number of higher t process will switch to persistence, Fig.(4.3) and Fig.(4.4) fulfill these expectations. This is why the pink noise can mimic long-term dependence as well.

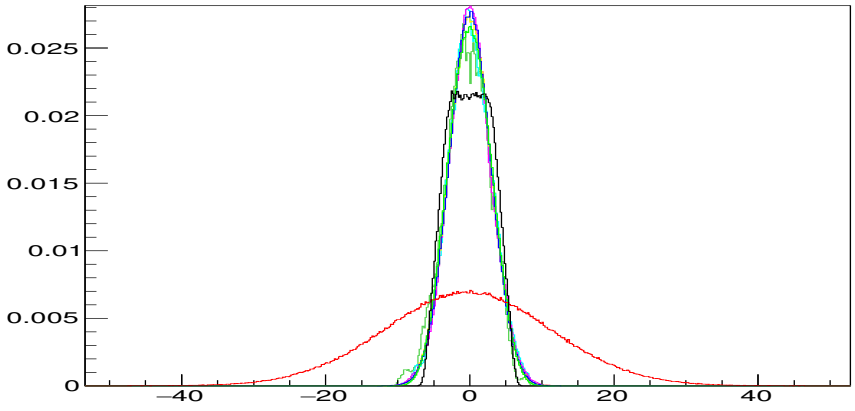


Figure 4.4: Distributions of increments of pink noise sub-processes from the Fig.(4.3). The red distribution represents the pink noise X^T

4.5 Hedging and portfolio optimization

4.5.1 Black & Scholes

Hedging is one of the most popular risk management strategies. It allows one to pay a given amount of money in advance, in order to be protected from unexpected market changes in the future. A financial instrument allowing this kind of contract is called an *option*. Price of an option depends on the underlying asset, therefore it is a so called *derivative*. More about the financial instruments can be found in [28]. The task is to find the fair-market price $C(Y,t)$ of the option, where $Y(t)$ is a stochastic price of the underlying asset.

The answer was given in 1973 by Black and Scholes. They used an assumption that prices follow *geometric Brownian motion*

$$dY = \mu Y dt + \sigma Y dW, \tag{4.18}$$

where μ and σ^2 are the first and second moments of the *Wiener process* W. Equation (4.18) is an example of *stochastic differential equation* which can be integrated by means of *Itô stochastic integral*. We recall that $Y(t)$ and $W(t)$ are stochastic variables, therefore YdW is also a stochastic variable. Since $Y(t)$ is supposed to be modeled by a geometric Brownian motion, any function of it, including $C(Y,t)$, must be a solution of the partial differential equation

$$dC = \left[\frac{\partial C}{\partial Y}\mu Y + \frac{\partial C}{\partial t} + \frac{1}{2}\frac{\partial^2 C}{\partial Y^2}\sigma^2 Y^2\right] dt + \frac{\partial C}{\partial Y}\sigma Y dW, \tag{4.19}$$

which is a special case of *Itô's lemma* for a geometric Brownian motion.

The change in the value of the portfolio ϕ over a time interval Δt is from [28] given by

$$\Delta\phi = -\Delta C + \frac{\partial C}{\partial Y}\Delta Y. \tag{4.20}$$

We also assume that the market interest rate r is constant, i.e.

$$\Delta\phi = r\phi\Delta t. \tag{4.21}$$

If we rewrite differentials d to Δ in (4.18) and (4.19) we can substitute them into (4.20). Finally, equating obtained equation with (4.21) results in

$$rC = \frac{\partial C}{\partial t} + rY\frac{\partial C}{\partial Y} + \frac{1}{2}\frac{\partial^2 C}{\partial Y^2}\sigma^2 Y^2. \tag{4.22}$$

This is known as the *Black & Scholes partial differential equation*. This equation is valid for both call and put European options and it can be solved after setting the boundary conditions. Since the owner of a call option has the right to buy the underlying asset at *maturity time* T for a given *strike price* K, it is desirable to have $(Y - K) > 0$. It means that the owner will pocket the difference. Therefore, boundary condition for a call option is $C = max\{Y - K, 0\}$ at $t = T$. Assuming further that

$$C(Y,t) \propto e^{r(t-T)} f(Y,t,\sigma,r,K,T) \tag{4.23}$$

Black and Scholes obtained the equation for the European call option pricing given by

$$C(Y,t) = YN(d_1) - Ke^{r(t-T)}N(d_2). \tag{4.24}$$

$N(x)$ is the Gaussian cumulative density function

$$N(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{z^2}{2}}dz \tag{4.25}$$

and

$$d_1 \equiv \frac{\ln(Y/K) + (r + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}, \tag{4.26}$$

$$d_2 \equiv d_1 - \sigma\sqrt{T-t}. \tag{4.27}$$

Hence, the estimation of the option price depends on five parameters: the asset stock price $Y(t)$, the asset volatility rate σ, the interest rate r, strike price K and the maturity time T. The maturity time and the strike price are given by the contract, Y and r are known from the market. Therefore the only parameter that needs to be quantified is the volatility rate. Even though σ is time-dependent, as well as r can be, and the rough approximation of Brownian motion is used, the Black and Scholes differential equation and its solutions led to improvement of market risk management and changed the the way how market trading is done.

4.5.2 Minimum Rényi entropy portfolio

Another risk management strategy is by the risk minimization and simultaneous return maximization. This approach is known as the *portfolio diversification* or *modern portfolio theory*. The foundation of it was built by the Nobel price laureate H. Markowitz. It is called the *Markowitz portfolio optimization model* and is based on expected return (mean) $E[R_p]$ and standard deviation (variance) of the portfolio σ_p^2. In general it can be written as

$$E[R_p] = \sum_{i=1}^{n} w_i E[R_i] \tag{4.28}$$

and

$$\sigma_p^2 = \sum_{i=1}^{n} w_i^2 \sigma_i^2 + \sum_{i=1}^{n} \sum_{i \neq j=1}^{n} w_i w_j \sigma_i \sigma_j \rho_{ij}, \tag{4.29}$$

where R_i is the return of asset i, w_i is the asset weight, σ_i is the standard deviation of returns of the underlying asset i and ρ_{ij} is the correlation coefficient between returns of assets i and j. Minimization of σ_p^2 and maximization of $E[R_p]$ boils down to the optimization problem which can be solved by means of Lagrange multipliers method.

In Markowitz model risk is measured by the variance, i.e the second moment of the stochastic process R_p. However, in [19] the *exponential Rényi entropy* is proposed as a measure of the risk. The exponential Rényi entropy is defined as

$$H_\alpha^{\exp}(X) := \exp(H_\alpha(X) = \left(\int_{\mathbb{R}} (p_X(x))^\alpha dx \right)^{\frac{1}{1-\alpha}}. \tag{4.30}$$

Its appealing property is that it incorporates higher order moments, thus including second and fourth ones. We show it by rewriting the latter equation by means of Kolmogorov-Nagumo function (1.61) and denoting $dF_X(x) = p_X(x)dx$, thus

$$H_\alpha^{\exp}(X) = \left(\int_{\mathbb{R}} e^{(1-\alpha)x} dF_X(x) \right)^{\frac{1}{1-\alpha}} \tag{4.31}$$

$$= \left(\int_{\mathbb{R}} dx \; p_X(x) \sum_{n=0}^{\infty} \frac{(1-\alpha)^n}{n!} x^n \right)^{\frac{1}{1-\alpha}} \tag{4.32}$$

$$= \left(\sum_{n=0}^{\infty} \frac{(1-\alpha)^n}{n!} \mathbb{E}_X[x^n] \right)^{\frac{1}{1-\alpha}}, \tag{4.33}$$

where we used the Taylor expansion of exponential function. Therefore H_α obtains uncertainty stemming from higher-order moments.

The minimum Rényi entropy portfolio for a given α is defined as

$$w_\alpha := \min_{w \in W} H_\alpha^{\exp}(R), \tag{4.34}$$

where W is a set of constraints on w and $R = (R_1, ..., R_n)$ is a random vector of asset returns. (4.34) can be solved via global optimization techniques. This approach, as is demonstrated in [19], yields to the better risk-return-turnover trade-off than by using the minimum-variance Markowitz model.

4.6 Summary

Main takeaways of this chapter are:

- financial and economic data can be approached by means of complex systems theory;
- in financial time series analysis one is mostly interested in moments of the underlying random processes, for example, second and fourth moments are especially relevant to market volatility and thus market risk;
- in order to manage market risk, past behaviour of the stocks is used to predict future behaviour, this is where stochastic financial models are widely used;
- another, more sophisticated, approach to the market risk management is via Black & Scholes differential equation;
- minimum Rényi entropy portfolio incorporates higher moments of the underlying process, thus yields to the better risk-return-turnover trade-off than classical minimum-variance Markowitz portfolio.

Chapter 5

Estimation of the Rényi entropy

5.1 Data processing

In data analysis one is always confronted with imperfections caused by numerical limitations or inaccuracies of records. For instance, theory we want to use suggest infinitely many data points, which is a priori unrealizable. Also, data we usually obtain contain number of operational errors, for example small precision they are measured with. As a result, one have to compromise between theory and existing methods of data analysis. Hopefully, there is a wide range of literature sources covering this topic, in particular we use [24], [26], [27].

In this work we are interested in financial data recorded by stocks in the form of time series. All methods we discuss in what follows are under this assumption. To get an idea of what market data are like, we present minute prices of *Johnson & Johnson* recorded between 1998 and 2014, Fig(5.1). We see that the stock price of *J & J* is following a positive trend during the whole period. On the second half of the trajectory one can notice a downturn of the 2007-2008 crisis with a little stagnation afterwards. The first histogram of price increments (upper Fig.(5.1)) is quite indistinguishable, however it demonstrate limitations of financial records. We see that the zero-return (difference of prices in time) is causing a deformation of distribution. The second (bottom) histogram is a reproduction of the first one, zoomed around returns with lower probability. If the data were recorded with higher precision we would see a spread of counts from 0-bin to bins around zero, which are empty in our case. Also, the first histogram can give a wrong impression that price changes are very rare events, that is not true. In our case only around 30% of the records are obtained in the zero-bin and the rest is obtained in, at first sight insignificant, lower part of the histogram.

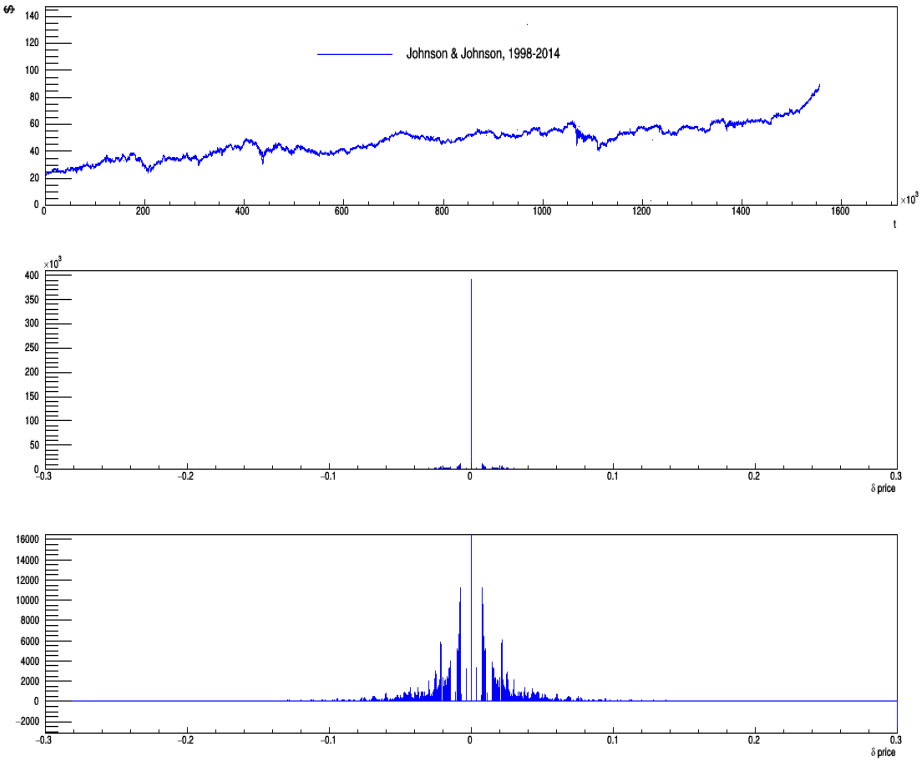


Figure 5.1: Minute prices of Johnson & Johnson company recorded from 1998 to 2014. Histogram of price increments and its zoomed version.

As can be expected market data are time dependent, i.e. non-stationary, which impede further use of methods validate for stationary processes. Typically, this problem is solved by describing market time series in terms of *price increments* (4.4), (4.5), (4.6), (4.7). Each approach is suitable for different situation and one must be aware of its limitations. However, it allow us to consider time series to be stationary, while preserving information they carry.

5.1.1 Partitioning

Financial data is discrete, however the resolution is still too high with respect to the amount of records. Partitioning the data is quite similar to filling a histogram. Partition is generated by dividing the domain of data set into disjoint intervals. Number of the intervals is denoted by S_A, and is usually referred to as a size of so called *alphabet*. Therefore, every data point is uniquely labeled by its interval. For example, $S_A = 2$ corresponds to the sign of return. For $S_A = 3$ data points are labeled as neutral, larger gain or larger loss. Of course, the latter is possible for approximately symmetric

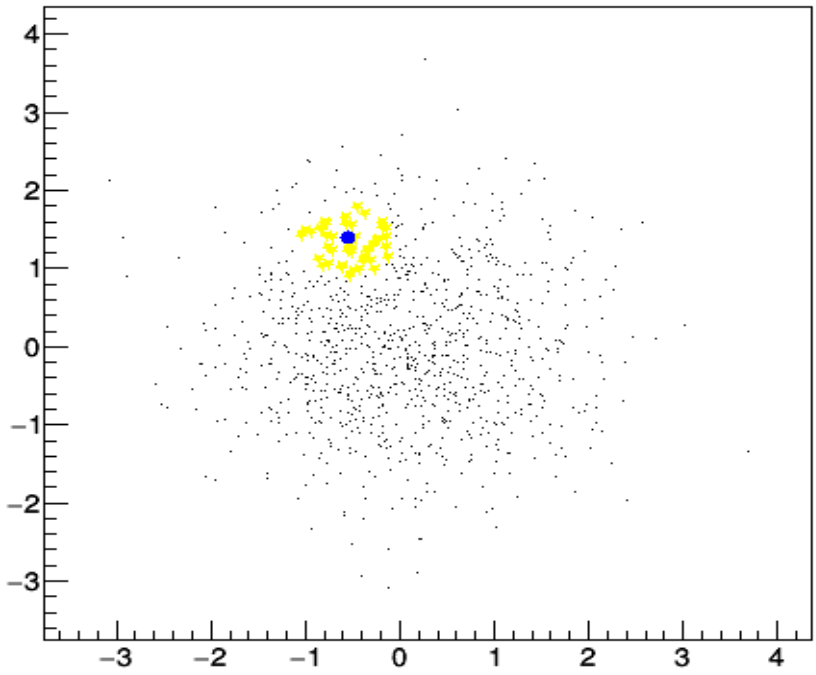


Figure 5.2: Normally distributed 2-dimensional points. By yellow color we denote neighbouring points within $\rho = 0.5$ distance from the blue point.

distributions, which holds for the market data.

Numerical results of Rényi entropy will be influenced by the choice of alphabet. It is a rough approximation and some of the information contained in the data will be lost. However, in some cases partitioning has proved to be useful. For instance, in [23], [14] alphabet of size $S_A = 3$ is applied on stock index time series. In absence of economic downturns they are typically not as much volatile as stock prices, therefore daily returns can be considered as noise with a drift. On the other side, marginal returns of indexes are usually a sign of major changes spreading across markets. Therefore, in order to study an impact of marginal events, it is reasonable to approximate stock index time series by three intervals representing larger loss or gain and neutral changes.

5.1.2 Covering

Covering is very similar to the process of box-counting we presented earlier in Chapter 1. Let $x_1, ..., x_N$ be a randomly distributed points in $\mathbb{R}^m$ and

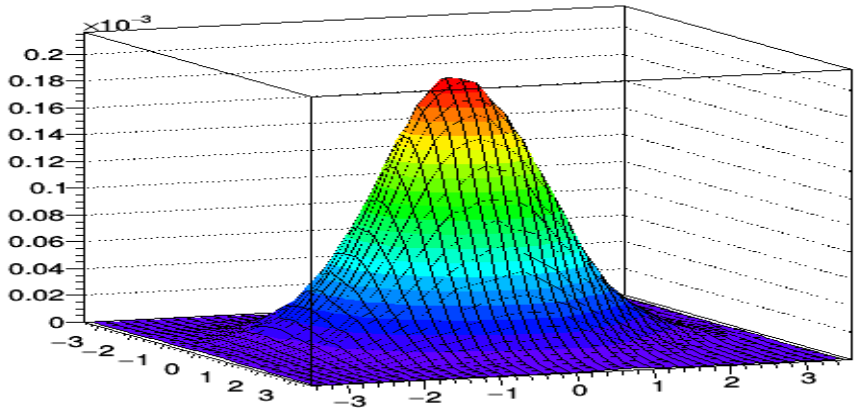


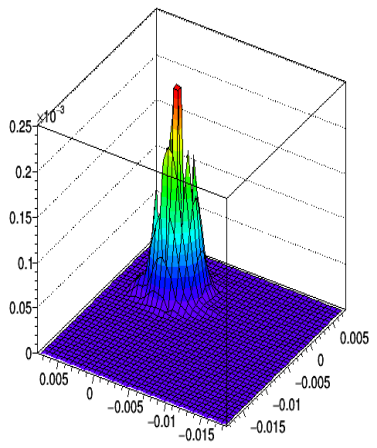
Figure 5.3: 2-dimensional Normal distribution obtained by applying the kth nearest neighbor method on the data set from Fig.(5.2).

$\rho(x, y)$ be the Euclidean distance between two points x and y of $\mathbb{R}^m$. For a given point x_i we form the ordered statistics $\rho^{(i)}_{1,N-1} \leq \rho^{(i)}_{1,N-1} \leq ... \leq \rho^{(i)}_{N-1,N-1}$, so that $\rho^{(i)}_{k,N-1}$ is the *kth nearest-neighbor distance* from x_i to some other x_j $(i \neq j)$.

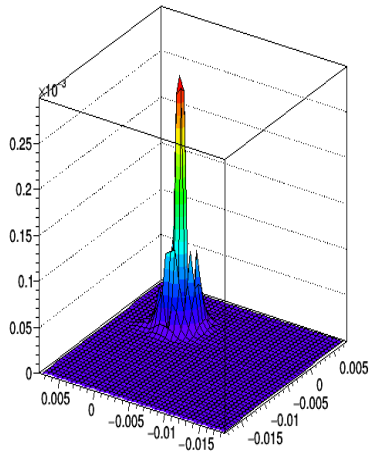
We can use this statistics to approximate a probability density distribution over the data set. For a point x_i and fixed distance ρ we count a number of nearest neighbors. This number is then proportional to the weight of the point x_i. Another approach suggests to fix the kth nearest neighbor and sum all the distances up to k, i.e. $\sum_{j \leq k} \rho^{(i)}_{j,N-1} =: K$. Weight of the point x_i is then proportional to $\frac{1}{K}$. Each approach is more or less suitable for different data sets.

Example 5.1.1 To demonstrate the kth nearest neighbor method we generate normally distributed 2-dimensional data set Fig.(5.2). We fix $\rho = 0.5$ and find all nearest points within this distance to every point separately, as is depicted on Fig.(5.2). The probability distribution obtained from this method is on Fig.(5.3). As expected, it is a 2-dimensional Normal distribution.

Example 5.1.2. Now we apply the kth nearest neighbor method on *Johnson & Johnson* price data from Fig.(5.1). In order to demonstrate a sensitivity to the distance parameter ρ, on Fig.(5.4a) distribution obtained with $\rho = 0.0025$ is presented, and on Fig.(5.4b) - with $\rho = 0.0015$. We see that the second case, when $\rho = 0.0015$ is more suitable for the correct approximation.



(a) Distribution obtained by applying the kth nearest neighbor method (with $\rho = 0.0025$) on *Johnson & Johnson* price log-returns from Fig.(5.1)



(b) Distribution obtained by applying the kth nearest neighbor method (with $\rho = 0.0015$) on *Johnson & Johnson* price log-returns from Fig.(5.1)

Figure 5.4

5.1.3 Estimators

Using the notion of kth nearest neighbor $\rho_k^{(i)}$ we now present estimator of the Rényi entropy. Let us recall Rényi entropy of absolutely continuous probability distribution p in $\mathbb{R}^m$ defined by

$$H_\alpha = \frac{1}{1-\alpha} \log \int_{\mathbb{R}^m} p^\alpha dx. \tag{5.1}$$

N. Leonenko, L. Prozanto and V. Savani introduced in [20] estimation of

$$\mathcal{I}_\alpha = \int_{\mathbb{R}^m} p^\alpha dx \tag{5.2}$$

for $\alpha \neq 1$ by

$$\hat{\mathcal{I}}_{N,k,\alpha} = \frac{1}{N} \sum_{i=1}^{N} (\zeta_{N,i,k})^{1-\alpha}, \tag{5.3}$$

where

$$\zeta_{N,i,k} = (N-1) C_k V_m (\rho_k^{(i)}, N-1)^m \tag{5.4}$$

with

$$V_m = \frac{\pi^{\frac{m}{2}}}{\Gamma(m/2+1)} \tag{5.5}$$

is the volume of a unit ball $\mathcal{B}(0,1)$ in $\mathbb{R}^m$ and

$$C_k = \left(\frac{\Gamma(k)}{\Gamma(k+1-\alpha)} \right)^{1/(1-\alpha)}. \tag{5.6}$$

Hence, (5.1) is from [20] estimated by

$$\hat{H}_{N,k,\alpha} = \frac{1}{1-\alpha} \log(\hat{\mathcal{I}}_{N,k,\alpha}). \tag{5.7}$$

For special case of the Shannon entropy estimation is of the following form

$$\hat{H}_{N,k,1} = \frac{1}{N} \sum_{i=1}^{N} \log \xi_{N,i,k}, \tag{5.8}$$

where

$$\xi_{N,i,k} = (N-1) \exp(-\psi(k)) V_m(\rho^{(i)}_{k,N-1})^m, \tag{5.9}$$

where $\psi(x)$ is the digamma function.

5.1.4 Effective transfer entropy

Entropies are often misestimated due to finite sample effects. To avoid these finite-size effects Marchinski and Kantz [23] defined *effective transfer entropy* as

$$T^{eff}_{Y\to X}(k,l) := T_{Y\to X}(k,l) - T_{Y_{schuffled}\to X}(k,l), \tag{5.10}$$

where the second term is evaluated for the *schuffled* Y series via *surrogate data* technique, destroying potential correlations with X. Thus, non-zero $T^{eff}_{Y\to X}(k,l)$ should indicate information flow free of correlations.

5.2 Transfer entropy in toy-model systems

To be able to understand and interpret Rényi transfer entropy applied on real data, we first apply RTE to the model systems, where we know the dynamics. We choose three toy-model systems: pink noise, Rössler chaotic system and GARCH process.

Pink noise is desirable since it is a system created by several parallel processes existing on different scales Fig.(4.3). It is a very good approach to the market behavior, where prices are not formed by inside dynamics solely, but are influenced by external seasonality as political decisions or central banks policy. Moreover, since the pink noise is generated by linear composition of sub-processes, we can easily decompose it, and study information flows from subsystems separately.

Rössler system is on the contrary nonlinear and exhibits chaotic behaviour. Nevertheless we show that the RTE can give reasonable results and detect information flows between two coupled Rössler subsystems. GARCH model is chosen as a typical representative of stock data. Each of those systems exhibit different types of couplings which are described further.

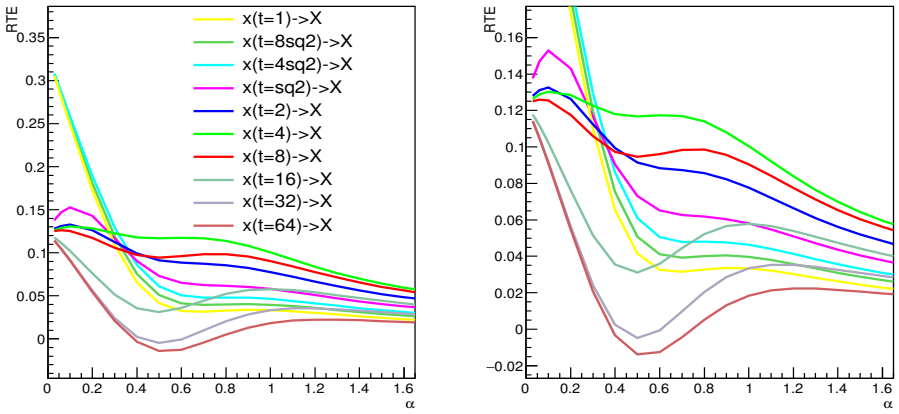


Figure 5.5: Rényi transfer entropy $T^R_{\alpha,x^{(t)}\to X^T}(1,1)$ for every $t \in T = \{1, 2^{1/8}, 2^{1/4}, 2^{1/2}, 2, 4, 8, 16, 32, 64\}$, with the right graph zoomed around $\alpha = 1$

5.2.1 Pink noise

In Chapter 4 we introduced the pink noise (4.16), which we recall here for convenience

$$X^T_n = x^{(1)}_n + x^{(2)}_n + ... + x^{(k)}_n, \tag{5.11}$$

where

$$x^{(t)}_n \equiv x_n(t) = c(t)x_{n-1} + u_{n-1}\sqrt{1 - c^2(t)}, \tag{5.12}$$

$c(t) = \exp(-1/t)$, u_n is uniformly distributed random variable, $k \in T$. T is such that its elements are evenly distributed on the log-scale, i.e. $T = \{1, \sqrt{2}, 2, 4, 8, ...\}$ or $T = \{, 10, 100, ...\}$.

We study information flows (RTE) between composing processes $x^{(t)}$ and the final process X^T obtained from the parallel summation of $K = \#T$ processes $x^{(t)}$. X^T is fully determined by all $x^{(t)}$, moreover we know that for small values of parameter $t \equiv t_s$ process $x^{(t_s)}$ is non-persistent, thus it contributes to the tail parts of the distribution underlying to X^T. On the contrary, $x^{(t_h)}$ for higher values of parameter $t \equiv t_h$ is persistent process

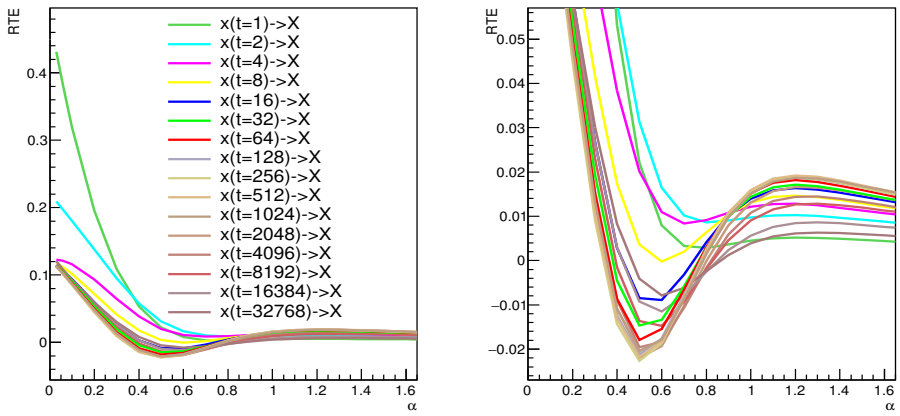


Figure 5.6: Rényi transfer entropy $T^R_{\alpha,x^{(t)}\to X^T}(1,1)$ for every $t \in T = \{1,2,4,8,16,32,64,128,256,512,1024,2048,4096,8192,16384,32768\}$, with the right graph zoomed around $\alpha = 1$

contributing to the central parts of the final distribution of the pink noise. Considering conclusions we made in Chapter 2 about escort distribution, we expect Rényi transfer entropy from rare events to be highlighted for $\alpha < 1$.

Numerical method

In this case we use *partitioning* described earlier in this chapter. Size of the alphabet is $S_A = 3$ and parameters from (2.36) are $(k,l) = (1,1)$. One can see that it is a rough approximation and we won't be able to estimate RTE precisely. Nevertheless it still allows us to get some reasonable results.

Results

We first generate pink noise for $T = \{1, 2^{1/8}, 2^{1/4}, 2^{1/2}, 2, 4, 8, 16, 32, 64\}$ and compute Rényi transfer entropy $T^R_{\alpha,x^{(t)}\to X^T}(1,1)$, results can be seen on Fig.(5.5). One can notice three types of behaviour. RTE for $t \in \{1, 2^{1/8}, 2^{1/4}\}$ is higher for $\alpha < 0.4$ in comparison with other flows, it plummets till $\alpha = 0.5$ and then continuous to decrease slowly with slight peak around $\alpha = 1$. Similar behaviour, but shifted, can be detected for $t \in \{16, 32, 64\}$. The third group with $t \in \{2^{1/2}, 2, 4, 8\}$ has distinguishable two peaks, while decreasing slowly. We can conclude that the process X^T is guided by the third group in the largest extent, since information flows are present in tail parts ($\alpha < 1$) as well as in central part ($\alpha > 1$) of the underlying distribution. As we expected, for the low values of t (the first group) information flow is highlighted for small α. The second group seems to be suppressed for all α.

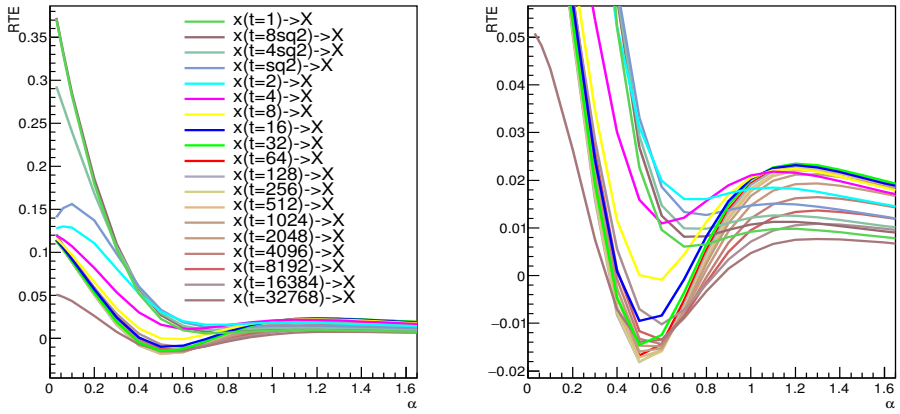


Figure 5.7: Rényi transfer entropy $T^R_{\alpha,x^{(t)}\to X^T}(1,1)$ for every $t \in T = \{1, 2^{1/8}, 2^{1/4}, 2^{1/2}, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384, 3276\}$, with the right graph zoomed around $\alpha = 1$

In the latter setting of the experiment different types of subsystems are evenly represented. We try to add more persistent subsystems and vary the number of random ones, results can be seen on Fig.(5.6) and Fig.(5.7). Even thought the persistent subsystems are numerically superior in these cases, we see that RTE can still detect flows for $\alpha < 1$ from the minority processes represented by small values of t. For $\alpha > 1$ all graphs tend to zero. This is most probably caused by the partitioning method which we use, therefore we refrain from interpreting those results.

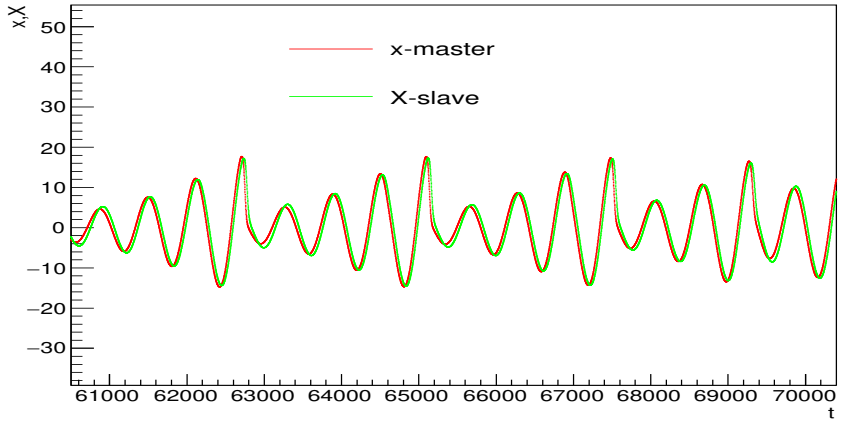


Figure 5.8: Synchronized Rossler systems, projections of x and X.

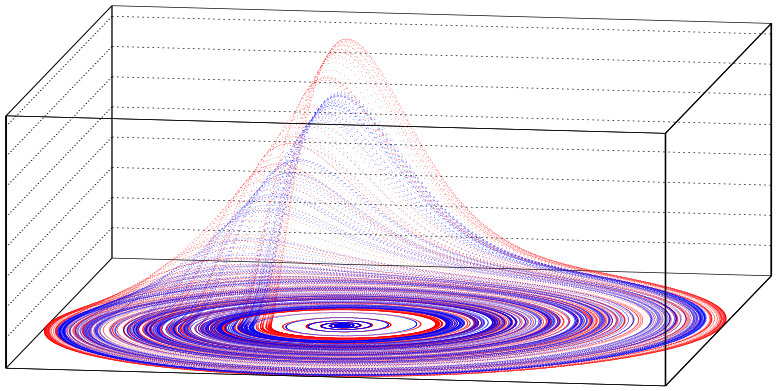


Figure 5.9: Attractors of the synchronized Rossler systems (master system is red and slave system is blue).

5.2.2 Rössler system

Rössler system is a system of ordinary differential equations

$$\begin{aligned}\dot{x} &= -\omega_1 y - z \\ \dot{y} &= \omega_1 x + a y \\ \dot{z} &= b + z(x - c)\end{aligned}$$

which are non-linearly coupled. It is also known as a *chaotic system* since the small perturbation of the initial conditions can alter solutions in significant way. We investigate RTE between two Rössler systems coupled in one direction by parameter ϵ. The latter system is than called a *master system* and a *slave system* is obtained as

$$\begin{aligned}\dot{X} &= -\omega_2 Y - Z + \epsilon(x - X) \\ \dot{Y} &= \omega_2 X + a Y \\ \dot{Z} &= b + Z(X - c)\end{aligned}$$

with parameters $a = 0.15$, $b = 0.2$, $c = 10.0$ and frequencies $\omega_1 = 1.015$ and $\omega_2 = 0.985$. We adopt this experiment from [25], where Shannon transfer entropy between x and X is studied.

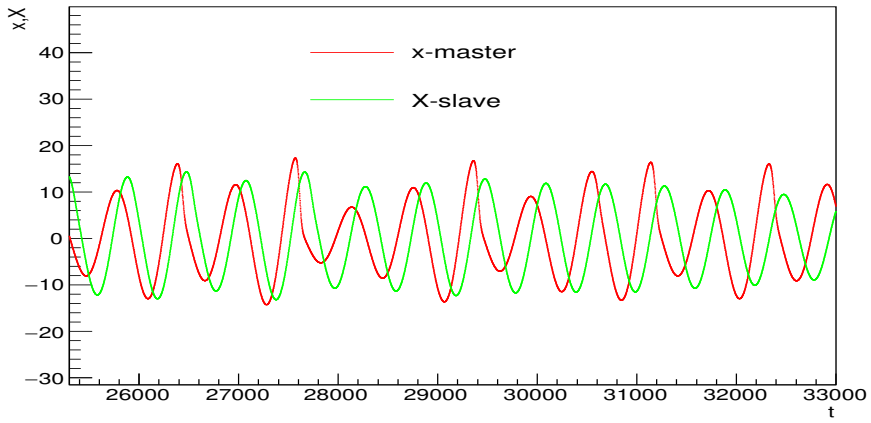


Figure 5.10: Coupled Rossler systems, projections of x and X.

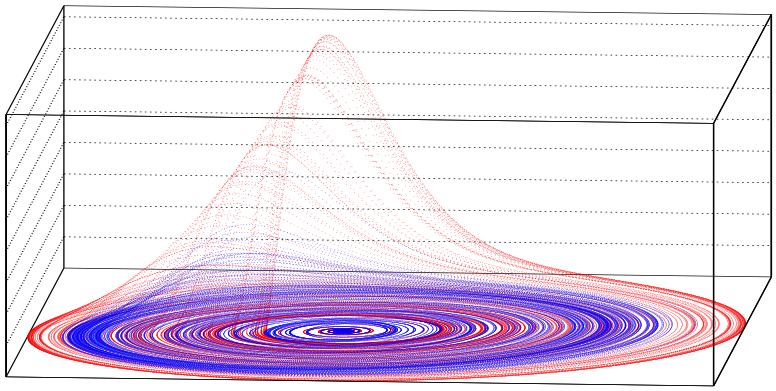


Figure 5.11: Attractors of the coupled Rossler systems (master system is red and slave system is blue).

Numerical method

In this case estimator (5.7) method is used and RTE is evaluated for different values of memory parameters (k, l). In this example we use notation given by (2.44). We also calculate the effective Rényi transfer entropy (5.10).

Results

We evaluate information transfer, dependent on coupling parameter ϵ, from the master system x to the slave system X. RTE with different memories is depicted on Fig.(5.13), Fig.(5.15) and Fig.(5.17). However, calculations are affected by the limited amount of data and we obtain biased results. It is thus more convenient to evaluate the effective Rényi transfer entropy (5.10) in this case: Fig.(5.12), Fig.(5.14) and Fig.(5.16). In [25] authors obtained synchronization of systems x and X for $\epsilon \approx 0.12$. It means that after this value there should be zero information flow, since there is no additional information from x. Nevertheless, projection X is also dependent on projections Y and Z, thus only after conditioning on Y and Z one would obtain a zero flow. In what follows we will interpret information flows for $\epsilon < 0.12$ only. Results for $\epsilon \geq 0.12$ do not represent coupling interaction we want to examine here.

Let us study how RTE changes with additional memory terms. All three settings perform increasing information flow with larger coupling. Adding more historical values lessens the information flow slightly. For $\alpha > 1$ information transfer is quantitatively almost similar in cases on Fig.(5.14) and Fig.(5.16), however in the latter case the synchronization threshold for $\epsilon \approx 0.12$ is more distinguishable. It suggest that using additional memory is crucial in order to detect information flow in detail. On the other hand, satisfactory results are obtained even for memory of two values.

The setting of this experiment accounts only for $l = 1$. One can also variate this parameter to obtain better results. However, we wanted to demonstrate conclusions from the Chapter 3, that causality can be detected through the variation of the parameter k.

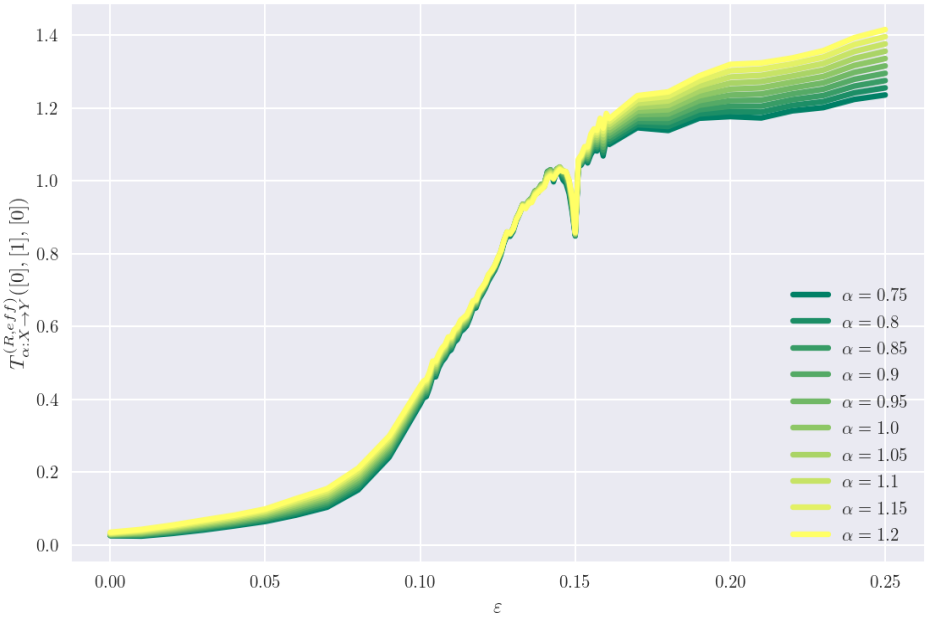


Figure 5.12: Rényi transfer entropy $T^{R,eff}_{\alpha,x\to X}(\{0\},\{1\},\{0\})$ between x-projection of master Rössler system to X-projection of slave system. (source [45])

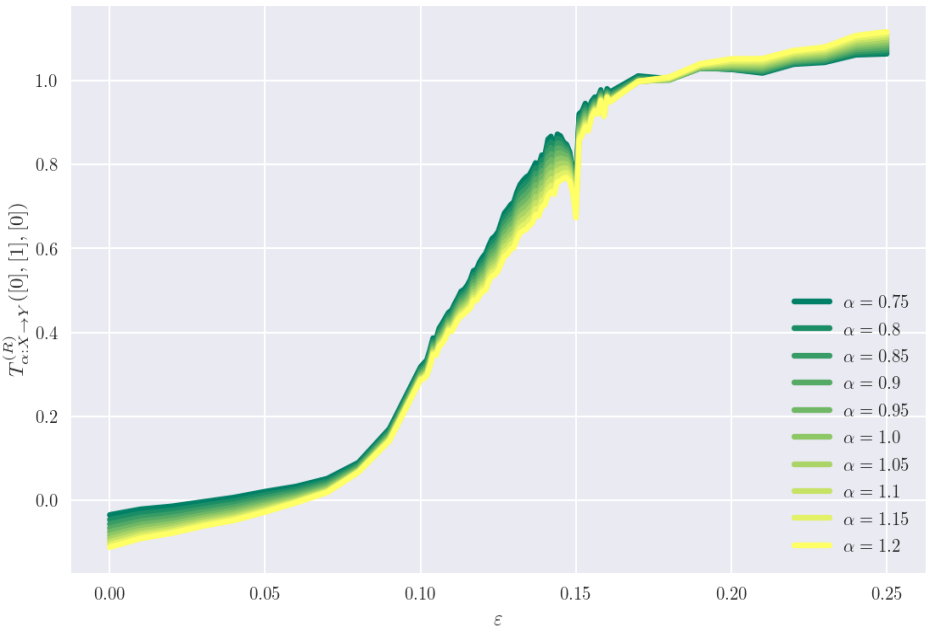


Figure 5.13: Rényi transfer entropy $T^{R}_{\alpha,x\to X}(\{0\},\{1\},\{0\})$ between x-projection of master Rössler system to X-projection of slave system.(source [45])

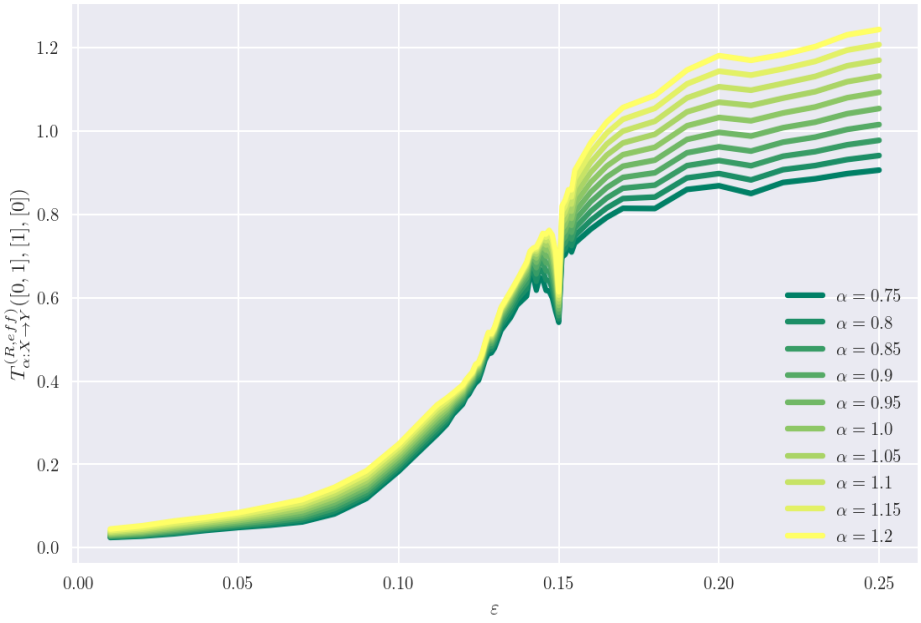


Figure 5.14: Rényi transfer entropy $T^{R,eff}_{\alpha,x\to X}(\{1\},\{1\},\{0\})$ between x-projection of master Rössler system to X-projection of slave system.(source [45])

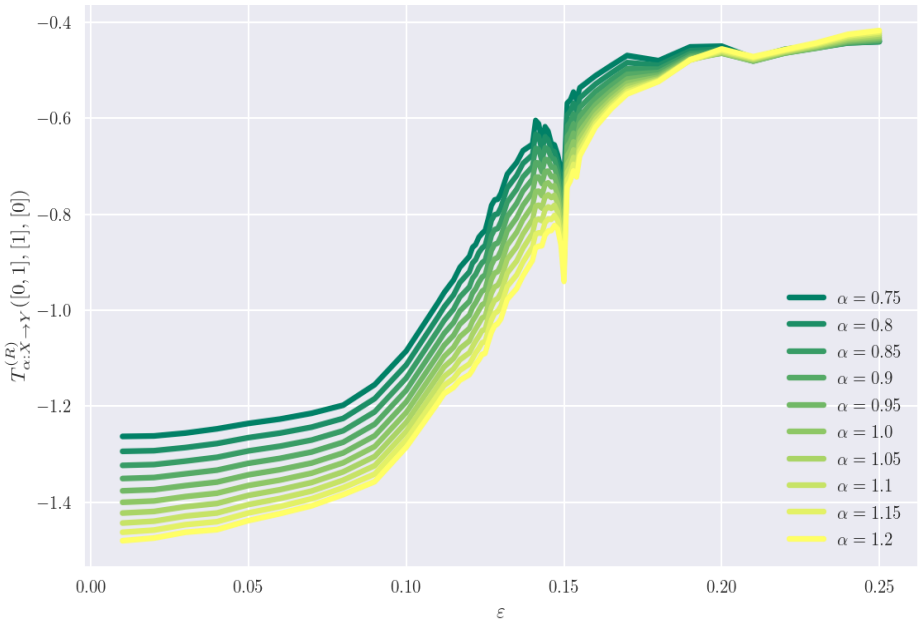


Figure 5.15: Rényi transfer entropy $T^{R}_{\alpha,x\to X}(\{1\},\{1\},\{0\})$ between x-projection of master Rössler system to X-projection of slave system.(source [45])

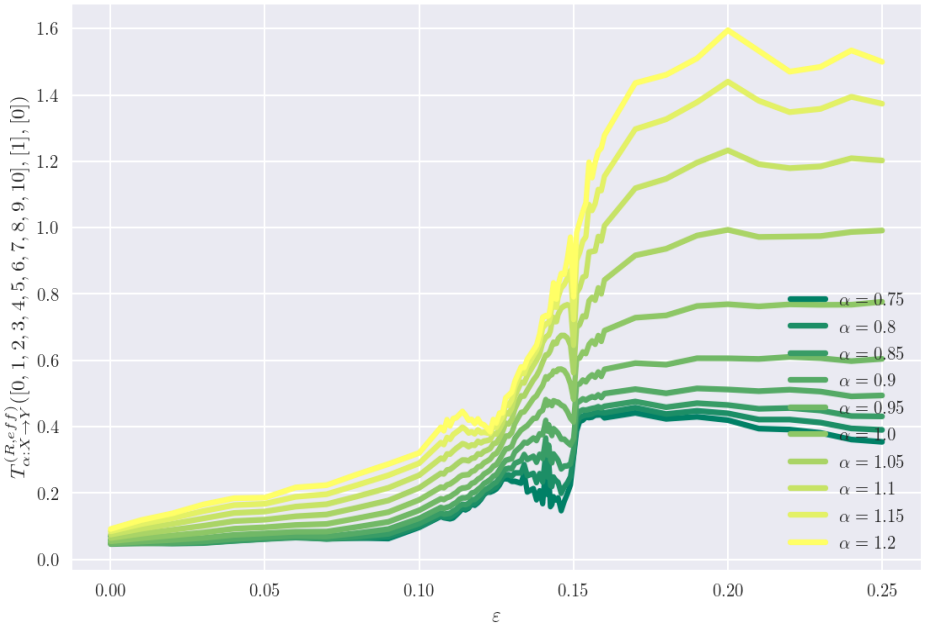


Figure 5.16: Rényi transfer entropy $T^{R,eff}_{\alpha,x\to X}(\{10\},\{1\},\{0\})$ between x-projection of master Rössler system to X-projection of slave system.(source [45])

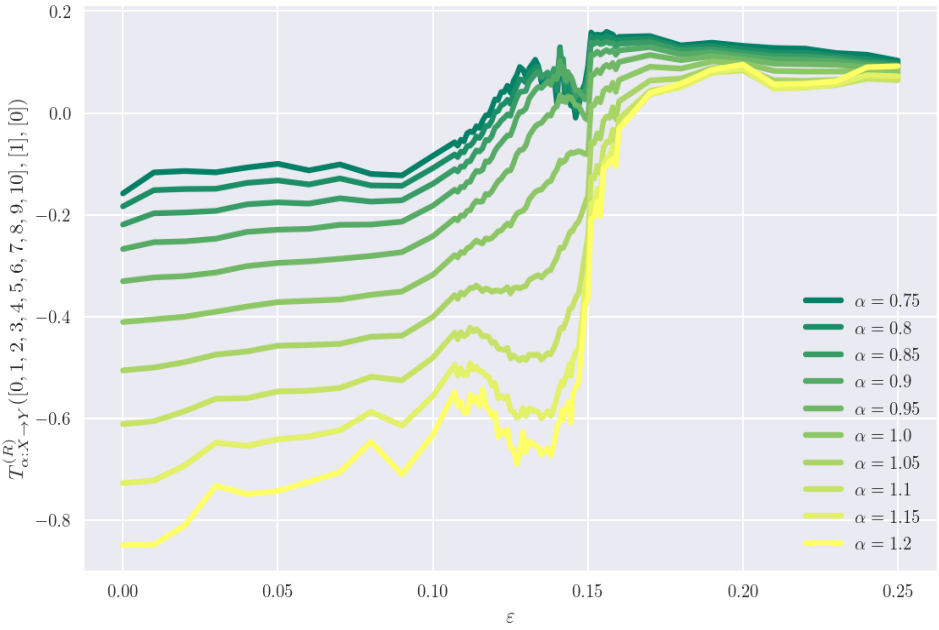


Figure 5.17: Rényi transfer entropy $T^{R}_{\alpha,x\to X}(\{10\},\{1\},\{0\})$ between x-projection of master Rössler system to X-projection of slave system.(source [45])

5.2.3 GARCH

Generalized ARCH process is a stochastic process with increments generated with respect to the time-varying volatility dependent on previous increments and variance. In this case we don't couple processes by any explicit parameter. The experiment is set such that one sample GARCH(1,1) process is generated, we denote it G_t, as

$$G_t \equiv \sigma_t e_t, \tag{5.13}$$

$$\sigma_t^2 = \beta_0 + \beta_{t-1} G_{t-1}^2 + \gamma_{t-1} \sigma_{t-1}^2. \tag{5.14}$$

The second process g_t is generated as

$$g_t \equiv \sigma_t e_t, \tag{5.15}$$

therefore it is a random process with zero mean and time-dependent variance $\sigma_t^2 = \sigma_t^2(G_{t-1})$. Increments of g_t are independent and are determined completely by G_t. Systems G_t and g_t are thus dependent in variance.

For comparison, we also investigate RTE for ARCH processes. We generate A_t as (4.12) with variance (4.13) dependent on previous increment A_{t-1}. The second process a_t is again obtained with respect to the variance of A_t.

Numerical method

We again use *partitioning* approach of data processing. Size of the alphabet is $S_A = 3$ and parameters from (2.36) are $(k, l) = (1, 1)$.

Results

Rényi transfer entropy between G/ARCH-like processes is depicted on Fig.(5.18). The information flow between GARCH-generated processes is much more significant than between ARCH processes. This suggests importance of variance dependency in markets. One can notice that $T^R_{\alpha, A\to a}$ is close to zero for $\alpha = 1$, thus processes are independent in terms of Granger causality or Shannon transfer entropy. However, RTE can detect correlations that would be suppressed by the linear averaging.

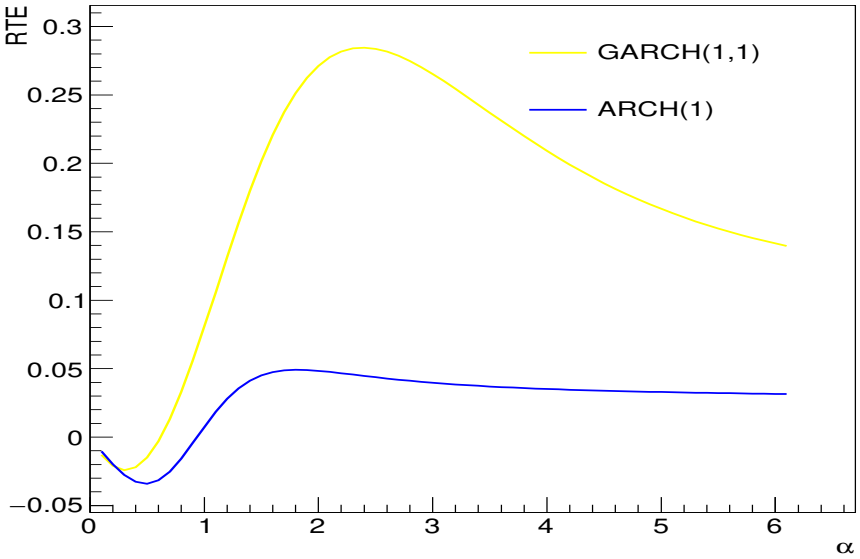


Figure 5.18: Rényi transfer entropy $T^R_{\alpha,G\to g}(1,1)$ (yellow) between GARCH processes dependent in volatility and $T^R_{\alpha,A\to a}(1,1)$ (blue) between ARCH processes dependent in volatility.

5.2.4 Interactions in complex systems

In general, inner interactions within complex systems are hard to detect. In previous toy-model examples we investigate three types of systems with different entanglements. Pink noise is a fractal-like system obtained by linear superposition of other subsystems. In this case correlations are linear, which we consider to be the best first testing ground.

Another kind of interaction is coupling. In this case, we are able to control at which extent systems are influencing one another through a coupling parameter. Even though we use coupled chaotic systems, we receive distinguishable information flows.

The last dummy system we study is GARCH and ARCH processes. Interconnection in this case is with respect to the same algebra of events. We can't control influences between these systems explicitly, as it is with other examples. However, in the stock markets one also can not do this and have to deal with a hidden web of relations. The setting of our third experiment can be thought of as a situation with one leading asset at the stock and other one influenced by its volatility.

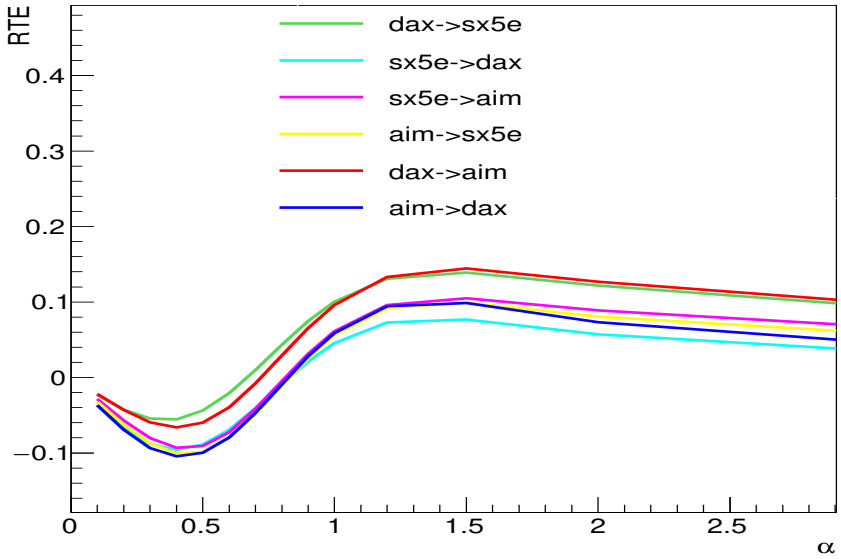


Figure 5.19: Rényi transfer entropy between European stock indices.

5.3 Real data

In the following section RTE is quantified for real market data. We study minute increments of European and American stock indices on the interval from 1.7.2012 to 1.10.2012.

Glossary of Indices

- **aim** The Alternative Investment Market is a sub-market of the London Stock Exchange
- **sx5e** EURO STOXX 50 is a stock index of Eurozone stocks
- **dax** Deutscher Aktienindex, index performance for Deutsche Boerse
- **dji** Dow Jones Industrial Average
- **nya** index covering all common stock listed on the New York Stock Exchange
- **ccmp** NASDAQ Composite

Numerical method

To mimic calculations done in [14], where RE is calculated for stock indices as well, we use *partitioning* approach of data processing. Size of the alphabet

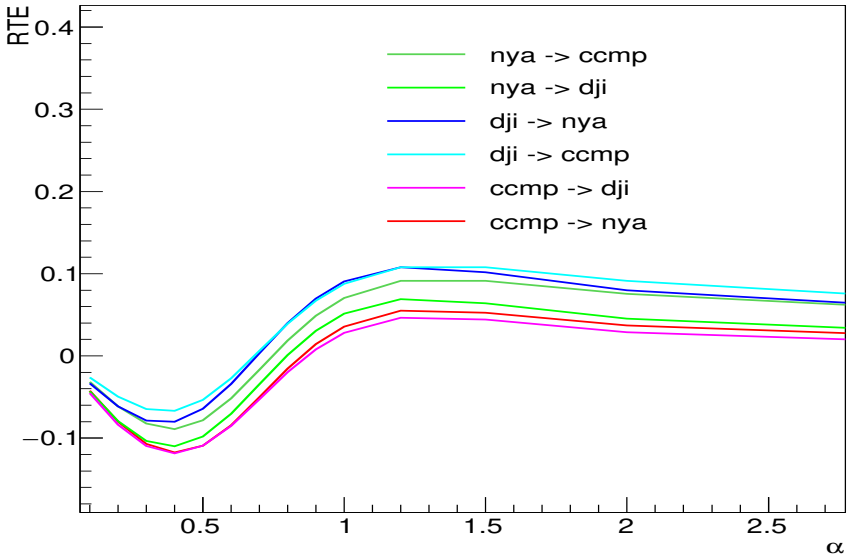


Figure 5.20: Rényi transfer entropy between American stock indices.

is $S_A = 3$ and parameters from (2.36) are $(k, l) = (1, 1)$.

Results

Information flows between European stocks are depicted on Fig.(5.19). We detect highest flows from *dax* into *sx5e* as well as *aim* indices. *sx5e* is composed of stocks listed on *Deutsche Boerse*, moreover most of the German stocks are with high weighting coefficient, thus one would expect additional information flow in this direction. Positive values of RTE for some values of $\alpha < 1$ and for $\alpha \geq 1$ suggest interdependence between all European indices on short and long time intervals.

Information transfer between American indices is on Fig.(5.20). *dji index* is derived from stocks listed on *NYSE* as well as on *NASDAQ*, however we detect more additional information transfer in the direction from *dji* to *nya* and *ccmp*. This suggest that agents on American markets derive their strategies from the *dji index*. We also notice, that $nya \to dji$ flow is larger than $ccmp \to dji$ which make sense, since there is less *NASDAQ* stocks in *dji* than from *NYSE*. It is important to note, that RTE between all indices cross zero and become positive for $\alpha < 1$, i.e. there is non-zero information flow between tail events. RTE is also positive for $\alpha \geq 1$, thus American indices are correlated on all scales.

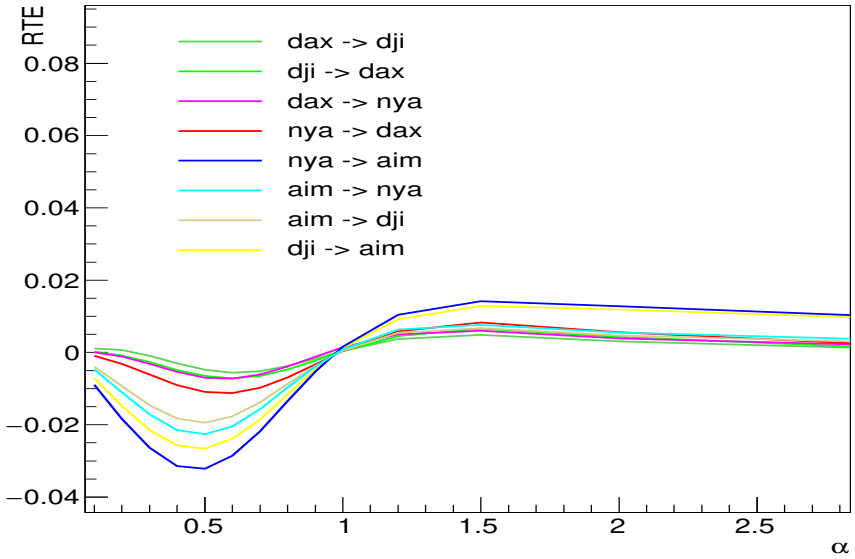


Figure 5.21: Rényi transfer entropy between European and American stock indices.

Atlantic information transfer is on Fig.(5.21). In this case Shannon transfer entropy for all indices is zero, thus they are Granger-independent. However RTE for $\alpha \neq 1$ detected non-zero information flow. Largest values for $\alpha > 1$ belong to $nya \rightarrow aim$ and $dji \rightarrow aim$, therefore London markets are influenced by American ones. Flow in the opposite direction is not so strong since *aim* represent only sub-market of *London Stock Exchange* and has small relative weight.

5.4 Summary

In this chapter we discussed some methods of data processing and estimated RTE for dummy and real data. Main takeaways are:

- with data analysis one must seek a trade-off between data processing and theory applied;
- *partitioning* and *covering* are examples of data processing, they can be used to estimate probability measure of sets;
- RTE can be estimated by means of *partitioning* or *covering*, the latter, however, gives better results for higher dimensions of data-sets;
- RTE measures directional information transfers between causally interconnected dummy data as well as real data from financial markets.

Chapter 6

Conclusion

We conclude by providing a final link between the all chapters. Rényi entropy is an information-theoretic concept. It, however, is closely related to multifractal measures and can be applied on stochastic processes. This wider view can be traced throughout the works of A.Rényi, whose scientific research touches fundamentals of probability theory and proceeds with Markov chains, information theory and graph theory. The latter is a building block for the network theory, that is especially used in connection with complex systems. In order to study complex systems, A.Rényi recognized necessity to extend Kolmogorov probability space to spaces with unbounded probability measure. Unbounded measures appear, for instance, in context of stochastic processes. In Chapter 1 we discuss all these concepts, such that Rényi entropy can be understood from a wider perspective and its application in the theory of complex systems is justified. We present definitions of Kolmogorov probability space, as well as conditional probability space - Rényi space, and show that the first mentioned is a special case of the latter. Another section is devoted to the stochastic processes, in particular to Markov processes. We present 3 equivalent definitions of Markov processes, two of them in terms of conditional probabilities and one by means of filtrations. Originally, TE was defined with conditional probabilities using Markov order parameter, however we aimed to discuss the possibility that it can be re-defined with the use of filtrations. Such approach would provide a link to the theory of stochastic processes and its properties, that will make RTE more easier to use for data analysis.

In Chapter 2 we exercise concept of RE by using it to prove a limit theorem for Markov chain transition probabilities, or quantifying explicit RE of Gauss, Cauchy and Student's-t distributions. The latter analytic results can be used to check and compare numerical estimators of RE. Moreover, we prove that RTE for normal distributions is $\alpha-$independent, thus can be used as a test of normality of a processes by comparing results for different

αs. In Chapter 3 we discuss that TE is a suitable tool for quantification of causal inference. Its equivalence with Granger causality for Gaussian processes is mentioned as well. Best numerical results showing, that RTE is able to detect causal dynamics is for the coupled Rössler systems. Results show that information flow increases with the growing coupling between the systems. Moreover, at the point where systems become synchronized, thus there is no new information that is transferred, RTE decreases.

In order to put RE and RTE into the context of real-world complex systems, Chapter 4 is devoted to financial systems and risk management. Financial institutions are directly or indirectly exposed to market risks from unexpected big (far from standard deviation) price changes. $\alpha-$zooming property of the RTE is thus suitable for these cases, where we want to concentrate on marginal events. We also quantified RTE for data from the stocks, in particular European and American stock indices. Results indicated, for instance, that the Dow Jones Industrial Average (DJI) is superior to both (NASDAQ and NYSE) New York stock markets. Such conclusion can be supported by the fact, that the value of the DJI is the weighted sum of the stocks traded on both aforementioned markets.

In this work we aimed to be mathematically rigorous on the one side, but accessible for real-world data analysis on the other side. This, however, revealed many gaps in the current formulation of the theory of transfer entropies, for instance, explicitly unclear connection of the RTE with deviation from the Markov property. Nevertheless, we would like to focus on it in the future research.

Future research: Next steps following this work would be in re-defining RTE by means of filtrations, connecting RTE with Pearl's causal model or improving computational part such that it can be applied on real-time financial data.

Bibliography

[1] N. Ay, D. Polani, "Information Flows in Causal Networks." Advances in Complex Systems, 11, 17–41. https://doi.org/10.1142/S0219525908001465 (2008)

[2] C. Beck, F. Schlögl, "Thermodynamics of Chaotic Systems", Cambridge Nonlinear Science Series (Book 4), (1995)

[3] S. Behrendt, T. Dimpfl, F. J. Peter, D. J. Zimmermann, "RTransferEntropy: Measuring information flow between time series with effective transfer entropy in R", will be published in Journal of Statistical Software

[4] J. F. Bercher,"Source Coding with Escort Distributions and Rényi Entropy Bounds" arXiv:1206.5127 (2011)

[5] T. Bossomaier, L. Barnett, M. Harré, J. T. Lizier, "An Introduction to Transfer Entropy, Information Flow in Complex Systems.", Springer (2016)

[6] R. Descartes, "Discourse on Method and Meditations on First Philosophy. ", Hackett Publishing Company (1980)

[7] E. B. Dynkin, "Osnovania teorii markovskih protsecov. ", Gosudarstvennoje Izdatelstvo Fiziko-Matematicheskoj literatury, Moskva (1959)

[8] K. Falconer, "Fractal geometry: Mathematical foundations and applications. ", John Wiley & Sons (1990)

[9] B. V. Gnedenko, "Kurs teorii veroiatnostei. ", Nauka, Moskva (1988)

[10] C. Granger, "Investigating causal relations by econometric and cross-spectral methods. ", Econometrica 37 424–438, (1969)

[11] C. Granger, "Time series analysis, cointegration, and applications. ", Nobel Lecture, December 8, 2003, in: Les Prix Nobel. The Nobel Prizes 2003, ed. Tore Frängsmyr, [Nobel Foundation] pp. 360–366, Stockholm (2004)

[12] V. M. Ilic, M. S Stankovic, "Generalized Shannon-Khinchin axioms and uniqueness theorem for pseudo-additive entropies.", arXiv:1311.0323v1 (2014)

[13] R. G. James, N. Barnett, James P. Crutchfield, "Information Flows? A Critique of Transfer Entropies.", arXiv:1512.06479 (2016) (2002)

[14] P. Jizba, H. Kleinert, M.Shefaat, "Rényi's information transfer between financial time series", Physica A, 391, 2971-2989 (2012)

[15] P. Jizba, T. Arimitsu, "The world according to Rényi: thermodynamics of multifractal systems", Annals of Physics 312 (2004) 17–59

[16] P.Jizba, "Information theory and generalized statistics", arXiv:cond-mat/0301343 (2003)

[17] J. Korbel, X. Jiang, B. Zheng "Transfer entropy between communities in complex networks", arXiv:1706.05543 (2017)

[18] S. Kullback , "Information Theory and Statistics", Dover Publications (1968)

[19] N. Lassance, F. Vrins, "Minimum Rényi Entropy Portfolios.", arXiv:1705.05666 (2019)

[20] N. Leonenko, L. Pronzato, and V. Savani "A class of Rényi information estimators for multidimensional densities." Annals of Statistics 36 (5), pp. 2153-2182. 10.1214/07-AOS539 (2008)

[21] B. Mandelbrot, R. Hudson, "The Misbehavior of Markets: A Fractal View of Financial Turbulence.", Basic Books (2006)

[22] R. N. Mantegna, H. E. Stanley, "Introduction to Econophysics: Correlations and Complexity in Finance.", Cambridge University Press (2007)

[23] R. Marschinski, H. Kantz, "Analysing the Information Flow Between Financial Time Series", Eur. Phys. J. B 30, 275-281

[24] H. Kantz, T. Schreiber "Nonlinear time series analysis.", Cambridge University Press (1997)

[25] M. Paluš, "From nonlinearity to causality: statistical testing and inference of physical mechanisms underlying complex dynamics.", Contemporary Physics 48(6):307-348 (2007) https://doi.org/10.1063/1.5019944

[26] M. Paluš and col., "Causality, dynamical systems and the arrow of time. ", Chaos Journal Vol.8 Issue 7. (2018) https://doi.org/10.1063/1.5019944

[27] M. Paluš, K. Hlavá˘cková-Schindler, M. Vejmelka, J. Bhattacharya, "Causality detection based on information-theoretic approaches in time series analysis.", Physics Reports 441 (2007) 1 – 46

[28] W. Paul, J. Baschnagel, "Stochastic Processes, From Physics to Finance.", Springer (1999)

[29] E. E. Peters, "Fractal Market Analysis: Applying Chaos Theory to Investment and Economics.", Wiley (1994)

[30] M. Prokopenko, J.T. Lizier, Don C. Price, "Causality detection based on information-theoretic approaches in time series analysis.", open access pdf (2013)

[31] M. Prokopenko, J.T. Lizier, D.C. Price, "On thermodynamic interpretation of transfer entropy." Entropy, 15(2), 524-543. https://doi.org/10.3390/e15020524 (2013)

[32] A. Rényi, "On measures of entropy and information.", Proc. Fourth Berkeley Symp. on Math. Statist. and Prob., Vol. 1 (Univ. of Calif. Press, 1961), 547-561

[33] A. Rényi, "Probability theory.", Akadémiai Kiadó, Budapest (1976)

[34] A. Rényi, "On algebra of distributions.", Selected Papers of A. Rényi, Vol. 1 Akadémiai Kiadó, Budapest (1976)

[35] A. Rényi, "On a new axiomatic theory of probability.", Selected Papers of A. Rényi, Vol. 1 Akadémiai Kiadó, Budapest (1976)

[36] A. Rényi, "Some fundamental questions of information theory (1960).", Selected Papers of A. Rényi, Vol. 2 Akadémiai Kiadó, Budapest (1976)

[37] A. Rényi, "A Diary on Information Theory", Wiley (1987)

[38] A. Rényi, "Letters on Probability.", Wayne State Univ Press, (1972)

[39] T. Schreiber, "Measuring Information Transfer.", Phys. Rev. Lett. 85 (2000) 461

[40] M. Schroeder, "Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise.", Dover Publications (2009)

[41] S. Thurner, R. Hanel, P. Klimek, "Introduction to the theory of complex systems.", Oxford University Press (2018)

[42] M. Vejmelka, M. Paluš, "Inferring the directionality of coupling with conditional mutual information.", Physical Review E77, (2008)

[43] T. Wegener, "Simulation and Analysis of the Lorenz System, Nonlinear Dynamics and Chaos.", open access pdf (2013)

[44] N. Wiener, "The theory of prediction.", McGraw-Hill, New York, (1956)

[45] figure provided by Ing. Hynek Lavicka, Ph.D.