CZECH TECHNICAL UNIVERSITY IN PRAGUE

Faculty of Nuclear Sciences and Physical Engineering

Department of Physics

DIPLOMA THESIS

Analysis of financial time series

Martin Prokš

Supervisor: Ing. Petr Jizba, Ph.D.

Prague, 2017

NASCANOVAT PODEPSANE ZADANI A
ULOZIT DO DVOU EPS SOUBORU
ZADANI1.EPS a ZADANI2.EPS

NASCANOVAT PODEPSANE ZADANI A
ULOZIT DO DVOU EPS SOUBORU
ZADANI1.EPS a ZADANI2.EPS

Acknowledgement

I would like to thank my supervisor, Ing. Petr Jizba, Ph.D., for his support and bottomless patience. I am also grateful to Ing. Václav Kůs, Ph.D. for his interesting lectures and inspiring conversations. Last but certainly not least, I wish to thank Dr. Hamid Shefaat and Dr. Tatjana Carle for providing intraday data and challenging discussion full of invaluable advice.

This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS16/239/OHK4/3T/14.

Prohlášení:

Prohlašuji, že jsem svou diplomovou práci vypracoval samostatně a použil jsem pouze podklady (literaturu, software, atd.) uvedené v přiloženém seznamu.

Nemám závažný důvod proti užití tohoto školního díla ve smyslu §60 Zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon).

V Praze dne ....

Title:
Analysis of financial time series

Author: Martin Prokš

Specialization: Mathematical physics

Sort of project: Diploma thesis

Supervisor: Ing. Petr Jizba, Ph.D.

Abstract: In the present thesis applications of information theory and entropy in general are discussed with emphasis on time series analysis. The two main concepts treated in details are Transfer entropy quantifying predictability in a time series and Superstatistics which well describes systems in local equilibrium. The former stands on rigorous footing of mathematical statistics, whereas the latter has its roots in Boltzmann-Gibbs entropy of statistical physics. Theoretical background for both ideas are reviewed, and their usefulness are demonstrated on financial time series. In the case of Superstatistics a broader model, namely, a transition between two Superstatistics on different time scales, is shown to be worthy of attention.

Key words: Shannon entropy, Rényi entropy, Information flow, Causality, Superstatistics

Název práce:
Analýza finančních časových řad

Autor: Martin Prokš

Zaměření: Matematická fyzika

Druh práce: Diplomová práce

Vedoucí práce: Ing. Petr Jizba, Ph.D.

Abstrakt: V této práci jsou diskutovány možné aplikace teorie informace a obecně entropie. Důraz je kladen na použití v analýze časových řad. Dva hlavní koncepty rozebrány podrobně v této práci jsou Transfer entropy, která měří předvídatelnost v časové řadě, a Superstatistika, která slouží jako dobrý popis pro systémy v lokální rovnováze. Transfer entropy stojí na rigorózních základech matematické statistiky, kdežto Superstatistika má kořeny v Botzmann-Gibbs entropii ze statistické fyziky. Teoretické základy pro oba pojmy jsou shrnuty a následně je jejich užitečnost demonstrována při aplikaci na finanční časové řady. Superstatistika je použita v širším pojetí. Přesněji je potvrzeno, že má smysl uvažovat obecnější model připouštějící přechod mezi dvěmi Superstatistikami na různých časových škálách.

Klíčová slova: Shannonova entropie, Rényiho entropie, Informační toky, Kauzalita, Superstatistika

Contents

List of Figures

List of Tables

Introduction

Although the origin of entropy comes from thermodynamics and statistical physics, the very same formula was later derived by Shannon when studying capacity of noisy channel. Thus relation between information theory and physics was suggested which gave mathematical justification for Maximum-entropy principle. Since then, there have been many advances in both information theory and statistical physics with, if any, not very clear relation between each other. Therefore, it is plausible to distinguish methods originating from rigorous information theory and methods coming from statistical physics with less firm foundations. In the present thesis we concentrate on applications to time series analysis and we pick up one representative from each of the two groups of methods. The one from information theory is Transfer entropy, and method originating in statistical physics, related to Tsallis entropy, is Superstatistics. These two methods are discussed in details and also applied on time series. A brief summary of each chapter follows.

In the first chapter Shannon entropy is introduced and its main properties are discussed. Our aim is to motivate abstract notation of information and also to give intuitive interpretation of the actual number which naturally leads to a glance at coding theory. A short note on a relation between Shannon information entropy and Boltzmann thermodynamic entropy is also found here.

The second chapter follows the steps taken by Rényi to generalize Shannon entropy into one-parametric class of entropies. The important issue here is arbitrariness of definition of generalized conditional entropy which plays a central role in generalizing Transfer entropy in the fourth chapter. As in the case of Shannon entropy the interpretation of Rényi entropy is also provided via coding theorem.

The first applications are encountered in the third chapter. The foremost intention is to suggest a replacement of auto-correlation function by mutual information which takes into account also nonlinear dependence. Next, Granger test for causality is reviewed, and its inappropriate use for nonlinear systems is pointed out. Main difficulties emerging from using information theoretical quantities are briefly presented as well.

The key concept of the thesis, Transfer entropy, is defined in the fourth chapter. Two versions of Transfer entropy are presented. The first original and mostly used Transfer entropy based on Shannon information entropy, and the second generalized version exploiting Rényi information entropy. Both of them are estimated from financial time series, and the final results may be found in the appendix B.

Finally, the last chapter sets theoretical background for the second central topic which is Superstatistics. After introducing Superstatistics, the new broader Superstatistical model which allows transition of Superstatistics at different times scales is discussed, and quantitative method for testing this new model is proposed. The result is that we can partially confirm conclusions made in [30].

Chapter 1

Information theory according to Shannon

Information theory was founded by Shannon in 1948, see [1], and it was originally intended to solve problem of reliable communication over an unreliable channel. Then it gradually spread to many fields. And now, after more than a half century, we can see broad applicability of information theory not only in communication theory, but even in physics, statistics or machine learning.

1.1 Shannon entropy

The main question of information theory is how we can measure information or uncertainty of random variable. First attempt to quantify information was performed by Hartley in 1928, see [2]. According to him, we need $\log_2 N$ units of information to describe (encode) particular element from some set consisting of N elements. The logarithmic measure provides additivity property, i.e. to select arbitrary element from two sets with N and M elements we need $\log_2 NM$ units of information. But this is just sum of necessary information to select element from the first set and then from the other.

Shannon extended this idea for sets with given probability distribution, i.e. provided we have additional knowledge about the set, and thereby proposed to measure the information by entropy.

Definition 1.1. *Let X be a discrete random variable with distribution $p(x)$. Then we define **entropy** of X as*

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

where $\mathcal{X}$ denotes set of all possible outcomes of X. For events x with probability $p(x) = 0$ we define summand by $\lim_{p\to 0} p\log p = 0$.

From definition it can be readily seen that entropy can be rewritten as expected value

$$H(X) = E\left[\log \frac{1}{p(X)}\right].$$

The expression $-\log p(x)$ is sometimes called measure of **surprise** of event x. It measures the uncertainty of the event before experiment or equivalently information that may be yielded by observing the event. The surprise is large for very unusual events due to logarithm around zero. On the other hand, observing of almost certain event does not surprise us so much (it gives us little information) since it is in some sense anticipated. Hence we can say that entropy is an expected value of measure of surprise.

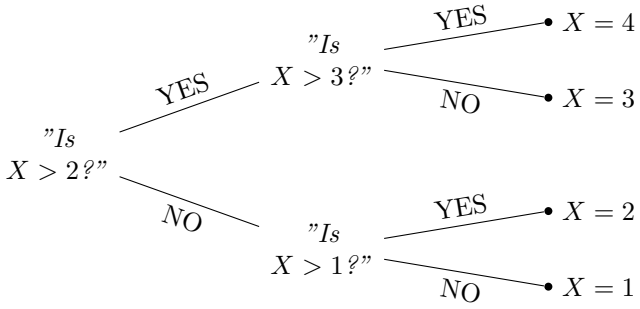
1.2 Coding theory and Huffman code

According to Shannon, the entropy is the average number of bits needed to optimally encode random variable X with its probability distribution $p(x)$. It means that entropy is average number of yes/no questions which bring us from absolute randomness to complete knowledge of random variable X, i.e. its occurred value. Proceeding with the following example we demonstrate this interpretation and also basic ideas of coding theory.

Let X be a random variable defined as

$$X = \begin{cases} 1 & \text{with probability } 1/2 \\ 2 & \text{with probability } 1/4 \\ 3 & \text{with probability } 1/8 \\ 4 & \text{with probability } 1/8 \end{cases}$$

Then there is at least $\log_2 4 = 2$ yes/no questions that completely determine actual value of random variable X. We can follow this diagram to determine the value of X.



In this case the number of questions does not depend on the actual value of X, and hence the averaged number of questions is $E[Q] = 2$ which would correspond to uniform distribution of X.

We can also simply ask: *"Is $X = 1$?"*, *"Is $X = 2$?"* and so on. This approach will require three questions in order that we can determine the arbitrary value of random variable X, but the actual number of questions depends on value of X. The averaged number of questions will be (notice we on purpose started asking from the most probable value that will turn up as a crucial aspect)

$$E[Q] = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{8} + 3 \cdot \frac{1}{8} = \frac{7}{4}. \tag{1.1}$$

Thus, we reduced the average number of questions by involving the additional information in the form of known probability distribution.

In order to encode the random variable X, we transfer this questionnaire into binary code with leading zeros followed by one on the i-th place representing yes for the i-th question. For example, the value 3 for X is encoded by 001. Generally every sequence of yes/no questions can be encoded in binary code, therefore finding the least average number of questions is equivalent to finding the shortest average binary code.

1.2.1 Huffman code

The main question in coding theory is how much we can shorten the code. The code should be instantaneous i.e. no code contains prefix of some other code. This requirement ensure instantaneous decoding i.e. we can decode every bit immediately without waiting for transmission of the whole code. Such a code is already uniquely decipherable. The existence of such a code is guaranteed by Kraft's inequality, see [3].

Theorem 1.1. *(Kraft's inequality)*
Let $\{x_1, \dots, x_N\}$ be possible outcomes that are encoded by sequences of characters from alphabet $\{0, \dots, D-1\}$. Then there is an instantaneous code with the lengths of sequences $\{l_1, \dots, l_N\}$ iff

$$\sum_{i=1}^{N} D^{-l_i} \leq 1. \tag{1.2}$$

Shannon solved the problem of redundancy when he proved the most significant theorem in coding theory.

Theorem 1.2. *(Shannon's noiseless coding theorem 1948)*
Let the lengths of codes $\{l_1, \dots, l_N\}$ satisfies inequality 1.2. Then the averaged length of code is bounded from below

$$E[L] \geq H(X).$$

Unfortunately the theorem claims nothing about construction of the code. It only states theoretical boundary for averaged length under which we cannot get. Lately Huffman published the construction of **optimal code** i.e. code that minimize average length

$$E[L] = \sum_{i=1}^{N} p_i l_i.$$

For fixed source, an analogy of random variable X, we readily find the optimum lengths l_i^* by minimizing $E[L]$ as a function of l_i subject to the Kraft inequality constraint. By Lagrange multipliers we derive

$$l_i^* = -\log_D p_i,$$

hence, the minimum average length is

$$E[L^*] = \sum_{i=1}^{N} p_i(-\log_D p_i) = H_D(X),$$

and from Shannon theorem it follows that $E[L^*]$ is minimum, and thus lengths $\{l_1^*, \ldots, l_N^*\}$ corresponds to optimal code. Since lengths must be integers, we generally achieve minimum lengths only for **D-adic probability** distributions i.e. for $\forall i\ \exists n \in \mathbb{N}$ such that $p_i = D^{-n}$

We may take $l_i = \lceil \log_D \frac{1}{p_i} \rceil$ these lengths satisfies Kraft inequality too because of the property of Ceiling function $x \leq \lceil x \rceil \leq x + 1$, and this choice of code lengths is called **Shannon-Fano code**. The averaged length then satisfies well known inequality

$$H_D(X) \leq E[L] < H_D(X) + 1. \tag{1.3}$$

In order to get closer to the boundary, we do not code only individual symbols but all sequences of say M symbols. Due to independence of symbols we get the entropy of sequence $H(X_1, \ldots, X_M) = MH(X)$ (we will state the properties of entropy later). The inequality 1.3 holds even for composed sequences thus, we have

$$H(X) \leq \frac{L}{M} < H(X) + \frac{1}{M},$$

and $\frac{L}{M}$ represents average length per symbol. We see that by coding longer sequences we can get arbitrarily close to the theoretical boundary even for non D-adic probability distribution.

Let us now discuss the construction of Huffman code. Huffman assumed instantaneous code and then derived optimal code by reasoning about properties of such code.

1. The length of more probable message must not be greater than length of less probable one. Hence after rearrangement of messages the following condition holds

$$p_1 \geq p_2 \geq \ldots \geq p_N,$$

$$l_1 \geq l_2 \geq \ldots \geq l_N.$$

2. Due to definition of instantaneous code, namely the prefix restriction, the two longest codewords must have the same length

$$l_{N-1} = l_N.$$

3. At least two and not more than D of the codewords with length l_N must differ only in the last bit/digit.

4. Each possible sequence of $l_N - 1$ digits must be used either as a codeword or must have one of its prefixes used as a codeword.

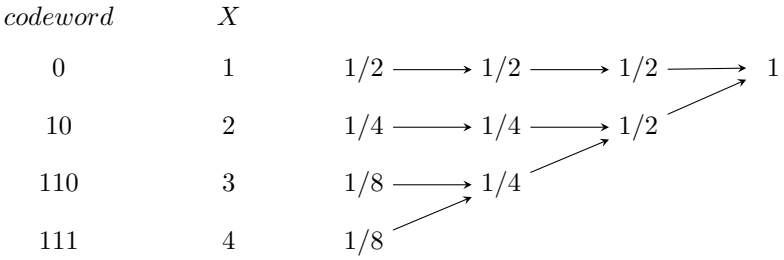
From these properties we can simply construct the optimal code. In what follows we assume $D = 2$ i.e. binary code. The construction is:

1. Assign to the two less probable messages 0 and 1. It will be their last digit in the codeword.

2. Combine these two messages into one with probability equal to sum of their probabilities.

3. Repeat all procedure with new set consisting of $N - 1$ messages until you have only one message.

We see that the codeword is created from the end to the beginning. It is worth illustrating the procedure by example. Recall the random variable X from the beginning of this section and encode it by Huffman optimal code.



The entropy of random variable X is

$$H(X) = \frac{1}{2} \cdot \log_2 2 + \frac{1}{4} \cdot \log_2 4 + \frac{1}{8} \cdot \log_2 8 + \frac{1}{8} \cdot \log_2 8 = \frac{7}{4},$$

and according to Shannon there is no code with average length less than $H(X)$. Let us see what expected length of Huffman code is

$$E[L] = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{8} + 3 \cdot \frac{1}{8} = \frac{7}{4}.$$

This result do not surprise us because we already know that Huffman code is optimal, and since probability distribution of X is 2-adic we reach the lowest boundary.

At the conclusion of coding procedure, let us note the connection between code and questions which bring us to complete knowledge about some system (random variable). We have already seen that every questionnaire can be rewritten into codeword and vice versa so that $E[Q] = E[L]$. And therefore since 1.1 we can say that we have by chance guessed the most effective questions. If we compare the way how we encoded the questions and Huffman code, we see that after swapping ones by zeros the codes exactly match.

Hence, we can ask whether looking for the most probable value in each step is generally the most effective way of determining the random variable. The answer is no. The proper questions may be obtained from Huffman code by determining the successive digits in the code (starting from the most significant bit i.e. from left to right). Thus after the first question: "Is X in a set A?" we must know what the first digit in the code is, and this is the requirement for appropriate choice of the set A.

1.3 Entropy interpretation in statistical physics

This section shows brief evolution of the word "entropy" as it emerges in two cognate fields of physics, namely thermodynamics and statistical physics. At the end, we will intimate connection with definition in information theory.

The term entropy was firstly introduced in thermodynamics by Clausius as a state function of thermodynamical system. More precisely only a differential of entropy was defined

$$\mathrm{d}S = \frac{\mathrm{d}Q}{T}. \tag{1.4}$$

The definition was motivated by the fact that heat received by a system during any reversible process depends on a path in a state space i.e. dQ is not a total differential of some state function. Luckily, for reversible processes there is always an integrating factor $\frac{1}{T}$ that changes Pfaff's form $\mathrm{d}Q$ into exact

differential and corresponding state function S is then called entropy. From relation 1.4 we also see that increase in entropy is associated with adding heat to the system. Since only differential was defined, the actual value of S depends on initial value S_0 independent of temperature and external parameters. However, it is proved that this initial value have to be function of number of particles otherwise Gibbs paradox arises.

Second law of thermodynamics states that for reversible (quasi-static) adiabatic process the entropy is conserved. However, for fast (irreversible also called dissipative) processes the entropy increases, and the difference $\mathrm{d}S > 0$ can be regarded as a measure of irreversibility or, in other words, the loss of information that is necessary for tracing the process back.

For isolated system entropy increases for non-static processes. Consider system at equilibrium state and by sudden change of external parameters shift it to new state (1) which is a non-equilibrium state due to the abrupt change. Then, provided the system is further isolated, it will aim to new equilibrium state (2) corresponding to new set of parameters. In this state the entropy takes maximum value, and the difference $\delta S = S_{(2)} - S_{(1)} > 0$ may represent distance of state (1) from equilibrium state (2).

Other interpretation of entropy comes from statistical physics, where it is considered to be a measure of the extent to which a system is disordered. And the value of entropy is logarithm of number of allowable configurations or micro-states of the system satisfying given constraint (observed macroscopic state), such as specific energy level. The Boltzmann equation expresses this interpretation

$$S = k \ln \Gamma. \tag{1.5}$$

Here Γ is number of micro-states also called thermodynamical probability of state, and thus maxim-entropy principle in the form of equation 1.5 says that system in equilibrium is in the most probable state [1].

In other words, every physical system is incompletely defined. We only know some macroscopic quantities and cannot specify the position and velocity of each molecule in the system. This lack of information is entropy i.e. entropy may be thought of as an amount of information about the system that is needed for description of microscopic structure.

Using equation 1.5 we may give more quantitative meaning to distance from equilibrium mentioned above. Imagine again system in equilibrium with maximum entropy $S_{eq} = k \ln \Gamma_{eq}$. If the system is driven out of equilibrium into new state by fluctuations, then this new state is bound to have lower entropy

[1] $\ln x$ is a monotone function so that maximizing S is equivalent to maximizing Γ.

and from 1.5 also less available micro-states $\Gamma < \Gamma_{eq}$. If we define probability of such fluctuation by $p = \frac{\Gamma}{\Gamma_{eq}}$, then we see that

$$p = \exp \frac{S - S_{eq}}{k},$$

i.e. the probability is exponentially damped by decrease in entropy.

Let us note that there were no clues that entropy defined by Shannon and that from statistical physics should be somehow related. Its the work of Jaynes who connected the information view of entropy with that from statistical physics or thermodynamics.

1.4 Basic properties of entropy

Firstly, someone may notice that we have not specified the base of logarithm. It is a common habit not to write the base as it is almost always assumed to be 2 in which case the entropy is measured in unit *bits*, which was adduced by J. W. Tukey. Nonetheless, we can sometimes encounter with natural logarithm which corresponds to unit called *nat*. For special purposes one is allowed to use arbitrary units (base of the logarithm). Fortunately, there is a simple rule for converting entropy between different basis D and D'. The rule reads

$$H_{D'}(X) = \log_{D'} D H_D(X).$$

Secondly, entropy of random variable X is independent of its possible values. It is only a function of probability distribution of X. Therefore entropy is often denoted by $H(p_1, \ldots, p_n)$, where $p_1, \ldots, p_n$ is the distribution of X, and the random variable is omitted. We can note that entropy is symmetric. It is intuitive requirement that measure of information should not depend on order of probabilities.

Entropy of random variable X is bounded. From definition it is evident that entropy is always positive since it is a sum of only positive values. On the other side, one can prove that entropy is also bounded from above, it is always less or equal than logarithm of number of possible outcomes of X.

Theorem 1.3. *Let X be discrete random variable and $|\mathcal{X}|$ denotes number of possible outcomes. Then*

$$0 \leq H(X) \leq \log |\mathcal{X}|$$

One may ask when these inequalities become equalities. The following theorem gives us the answer.

Theorem 1.4. *Let X be discrete random variable and $|\mathcal{X}|$ denotes number of possible outcomes. Then*

$$H(X) = 0 \quad \Leftrightarrow \quad \exists x \in \mathcal{X} \; p(x) = 1, \; \textit{and}$$

$$H(X) = \log|\mathcal{X}| \quad \Leftrightarrow \quad p(x) = \frac{1}{|\mathcal{X}|} \ \forall x \in \mathcal{X}.$$

The theorem claims that the entropy is equal to zero if and only if random variable X is deterministic constant i.e. X is distributed by Dirac distribution ($p(i) = 1$ for some i) also called degenerate random variable. The other equality is valid if and only if the distribution of X is uniform. It means that there is no outcome which we can somehow anticipate thus the system is completely unpredictable. Any non-uniform distribution may be understood as additional information, and therefore leads to decrease of entropy (or uncertainty).

According to theorem 1.4, we can imagine random variable X, that may represent some system, with entropy $H_2(X)$ (2 denotes units i.e. bits) as a system with $2^{H(X)}$ equally probable outcomes.

We have not justified the option for functional form of surprise $h(p) = -\log p(x)$. Clearly it satisfies two intuitive conditions required for measure of information, namely:

$$h(p) \text{ is nonnegative for } \forall p \in (0,1\rangle \tag{1.6}$$

$$\begin{gathered} h(p) \text{ is additive for independent events i.e.} \\ h(pq) = h(p) + h(q), \ p,q \in (0,1\rangle \end{gathered} \tag{1.7}$$

Someone may ask whether there is another function satisfying these two conditions (axioms). The answer is in the following theorem, see [4].

Theorem 1.5. *The only function satisfying conditions 1.6 and 1.7 is*

$$h(p) = -c\log p, \qquad c \geq 0.$$

Here c corresponds only to different units used for measure of information (uncertainty). It is common to assume in addition to 1.6 and 1.7 also normalization

$$h\left(\frac{1}{2}\right) = 1$$

which sets the units to bits and $h(p) = -\log_2 p$.

1.5 Joint entropy

Similarly to the definition of entropy for one random variable we can define joint entropy for n random variables.

Definition 1.2. *Let $X_1, \ldots, X_n$ be n discrete random variables with joint distribution $p(x_1, \ldots, x_n)$. Then we define* ***joint entropy*** *of $X_1, \ldots, X_n$ as*

$$H(X_1, \ldots, X_n) = -\sum_{(x_1,\ldots,x_n)\in\mathcal{X}_1\times\ldots\times\mathcal{X}_n} p(x_1, \ldots, x_n)\log p(x_1, \ldots, x_n).$$

The relation between joint entropy and entropy of individual random variables states the following theorem which finds great applicability in data compression.

Theorem 1.6. *Let $X_1, \ldots, X_n$ be n discrete random variables with joint entropy $H(X_1, \ldots, X_n)$. Then*

$$H(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

and the equality holds iff the random variables $X_1, \ldots, X_n$ are mutually independent.

1.6 Conditional entropy

Let X and Y be random variables. Then for all y from possible outcomes of Y $p(X|Y = y)$ is a probability distribution of X. Therefore we can define entropy of X given $Y = y$

$$H(X|Y = y) = -\sum_{x \in \mathcal{X}} p(x|Y = y) \log p(x|Y = y).$$

Then the conditional entropy is defined as averaged entropy of random variable X under the assumption that the value of Y is known.

Definition 1.3. *Let X and Y be discrete random variables. Then the **conditional entropy** is defined as*

$$H(X|Y) = \sum_{y \in \mathcal{Y}} p(y) H(X|Y = y). \tag{1.8}$$

After inserting the definition of $H(X|Y = y)$ into expression 1.8 we get

$$H(X|Y) = -\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x, y) \log p(x|y).$$

Conditional entropy may bring a little bit of insight into difference between yielded information and uncertainty. Imagine we have event A that occurs with probability p, and after observing another event B, the probability of A changes to q. Thus, before happening B we get $\log_2 1/p$ bits of information from A, and provided B happened it changes to $\log_2 1/q$, and we can say that difference $\log_2 1/p - \log_2 1/q$ represents information gain.

We already know that joint entropy of two independent random variables is sum of individual entropies, and for dependent variables there is an inequality. With the help of conditional entropy we are able to write so-called **chain rule**.

Theorem 1.7. *(Chain rule)*
Let X and Y be discrete random variables then

$$H(X,Y) = H(X) + H(Y|X). \tag{1.9}$$

The relation 1.9 is valid also for conditional joint entropy, i.e.

$$H(X,Y|Z) = H(X|Z) + H(Y|Z,X).$$

Later we will use generalization for more than two random variables.

Theorem 1.8. *Let $X_1, \ldots, X_n$ be discrete random variables then*

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1),$$

where a notation $H(X_1|X_0, \ldots, X_1) = H(X_1)$ is used.

1.7 Relative entropy and mutual information

In what follows, we will define relative entropy also called Kullback divergence which is considered as a distance between probability distributions, even though neither triangle inequality nor symmetry property holds.

Definition 1.4. *Let $p(x)$ and $q(x)$ be two probability distributions and $q(x) \neq 0$ for $\forall x$, then* ***relative entropy*** *is defined as*

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}.$$

In analogy to conditional entropy, we may define also conditional Kullback divergence since

$$D(p(X|Y=y)||q(X|Y=y) = \sum_{x \in \mathcal{X}} p(X=x|Y=y) \log \frac{p(X=x|Y=y)}{q(X=x|Y=y)}$$

is well defined Kullback divergence for all $y \in \mathcal{Y}$, and therefore it is plausible to define **conditional Kullback divergence** as an average over all possible values of Y

$$D(p(X|Y)||q(X|Y)) = \sum_{y \in \mathcal{Y}} p(y) D(p(X|Y=y)||q(X|Y=y) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x,y) \log \frac{p(x|y)}{q(x|y)}. \tag{1.10}$$

It is worth mentioning that relative entropy is only a special case of general *f-divergence*.

Definition 1.5. *Let p and q be a discrete probability distributions with the same support S and f be a convex function defined for $t > 0$ and satisfies $f(1) = 0$ then* ***f-divergence*** *is defined as*

$$D_f(p||q) = \sum_{x \in S} q(x) f\left(\frac{p(x)}{q(x)}\right).$$

Hence we see that relative entropy emerges for $f(t) = t \log t$. General f-divergences are important in statistics where they are used as a different measures of distinction between probability distributions and are convenient minimizing functionals for testing quality of estimators of some unknown probability distribution.

In coding theory operational definition of relative entropy can be given as follows. $D(p||q)$ represents average number of unnecessarily bits used in encoding of random variable X if we use bad distribution $q(x)$ instead of underlaying probability distribution $p(x)$. That is instead of inequality 1.3 we have

$$H(X) + D(p||q) \leq E[L] < H(X) + D(p||q) + 1 \tag{1.11}$$

where expectation is taken with respect to $p(x)$, and the subscript denoting base of logarithm was dropped.

Relative entropy is used for defining mutual information of two random variables as a distance from total independence.

Definition 1.6. *Let X and Y be two discrete random variable with probability distributions $p(x)$, $p(y)$ respectively. Then* ***mutual information*** *is defined as*

$$I(X;Y) = D(p(x,y)||p(x)p(y)).$$

The mutual information represents amount of information about random variable X included in Y. Symmetry of mutual information is clear from definition and can be paraphrased as information about X in Y is equal to information about Y in X. It is useful to think of mutual information as an intersection of entropy (information) $H(X)$ and $H(Y)$ as is depicted in the figure 1.1.

Thus mutual information is convenient measure of dependence of random variables (or time series as we will see later). In fact, mutual information specifies how many bits in average we could predict about X from Y and vice versa. Due to symmetry it is not applicable to detect information flow between two time series because that should be directional.

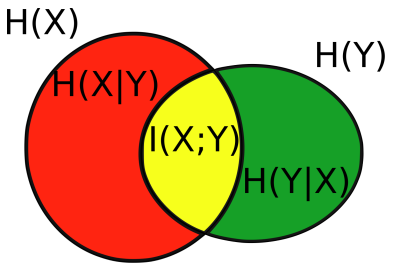


Figure 1.1: Relations between entropies

After some treatment we get relation between mutual entropy and entropy of random variable

$$H(X) = I(X;Y) + H(X|Y). \quad (1.12)$$

It flows from this equation that mutual information is the reduction in uncertainty after observing Y. Other relations between mutual information, conditional entropy and joint entropy may be figured out from figure 1.1.

The following expression is clear form 1.12 (since $H(X|X) = 0$) and intuitively reasonable as well

$$I(X;X) = H(X).$$

Generalization of 1.12 for more random variables is straightforward

$$I(X_1, \ldots, X_n; Y) = H(X_1, \ldots, X_n) - H(X_1, \ldots, X_n|Y).$$

From equation 1.12 we express the mutual information, and by conditioning both sides we get so-called **conditional mutual information**

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z). \quad (1.13)$$

This quantity is the reduction in the uncertainty of X due to knowledge of Y when Z is given, i.e., amount of information about X contained only in Y excluding possible intersection of $I(X;Y)$ and $I(X;Z)$ that may be thought of as a redundancy in variables X and Y given Z.

With the help of conditional mutual information we may obtain chain rule for mutual information analogous to one for joint entropy [2]

$$I(X_1, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, \ldots, X_1). \quad (1.14)$$

For more variable we cannot depict the situation in a Venn diagram because it would be indecipherable. But it is still possible to imagine that this relation just says:

> "Common information about n random variables in Y is a disjoint union of individual information about X_i in Y."

[2] The same notation as in theorem 1.8 applies here.

Another useful quantity which captures predictability property of mutual information and may be regarded as a **redundancy** can be defined as

$$R(X_1;\ldots;X_m) = \sum_{i=1}^{m} H(X_i) - H(X_1,\ldots,X_m)$$

which represents number of saved bits when group of m events are encoded with one codeword instead of encoding events separately. Clearly redundancy is 0 when all events are independent.

1.7.1 Jensen's inequality

Many important inequalities follow from Jensen's inequality which is valid for convex functions. Let us recall the definition.

Definition 1.7. *Let f be a real-valued function defined on $\langle a,b\rangle$. Then f is called convex if for $\forall x_1,x_2 \in (a,b)$ and $0 \leq \lambda \leq 1$*

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2).$$

f is called strictly convex if equality holds only if $\lambda = 0$ or $\lambda = 1$.

It is good to note that f is convex iff $-f$ is concave (definition of concave function differs only in opposite inequality). We will use this remark for logarithm, that is concave, in order to derive useful inequalities with the help of Jensen's inequality.

Theorem 1.9. *(Jensen's inequality)*
Let f be a convex function and X a random variable. Then

$$E[f(X)] \geq f(E[X])$$

and if f is strictly convex then the equality implies that random variable X is degenerate (i.e. $X = c$ with probability 1).

The following theorem is the key point for many important inequalities in information theory.

Theorem 1.10. *Let $p(x)$ and $q(x)$ be probability distributions and $q(x) \neq 0$ for $\forall x \in \mathcal{X}$ then*

$$D(p||q) \geq 0$$

with equality iff $p(x) = q(x)\ \forall x \in \mathcal{X}$.

Corollary 1.1. *For any two random variables*

$$I(X;Y) \geq 0$$

with equality iff X and Y are independent.

With this corollary it is easily seen from 1.12 that

$$H(X) \geq H(X|Y). \tag{1.15}$$

This means that knowing another random variable Y cannot increase uncertainty of X. But it is valid only on average, in special cases $H(X|Y = y)$ may be greater than $H(X)$.

1.8 Entropy rate

Entropy rate is defined for a stochastic processes to measure increase of joint entropy $H(X_1, \ldots, X_n)$ with respect to n.

Definition 1.8. *Let* $\mathbf{X} = \{X_n\}$ *be stochastic process. Then* ***entropy rate*** *of stochastic process* $\mathbf{X}$ *is*

$$H(\mathbf{X}) = \lim_{n \to +\infty} \frac{1}{n} H(X_1, \ldots, X_n),$$

provided the limit exists.

Let us calculate entropy rate for some stochastic processes:

1. Let X be a random variable with m equally distributed outcomes and consider stationary stochastic process $X_n = X \quad \forall n$. Then the sequence $(X_1, \ldots, X_n)$ has m^n equally probable results. Thus

$$H(\mathbf{X}) = \lim_{n \to +\infty} \frac{1}{n} H(X_1, \ldots, X_n) = \lim_{n \to +\infty} \frac{1}{n} \log m^n = \log m.$$

2. Consider sequence $(X_1, \ldots, X_n)$ of $i.i.d$ random variables. Then the entropy rate is

$$H(\mathbf{X}) = \lim_{n \to +\infty} \frac{1}{n} H(X_1, \ldots, X_n) = \lim_{n \to +\infty} \frac{1}{n} n H(X_1) = H(X_1).$$

These two examples are very simple, and the second one is just generalization of the first one. The resulting entropy rate can be guessed immediately without any calculation if we consider that entropy rate of stationary process characterizes measure of dependence in the process. Therefore, for every stationary process we have

$$H(\mathbf{X}) \leq H(X_1).$$

Next we define conditional entropy rate of stochastic process that is very helpful quantity in forecasting of future evolution of stochastic process (time series) because it tells us the uncertainty about the next step given all history.

Definition 1.9. *Let* $\mathbf{X} = \{X_n\}$ *be stochastic process. Then* ***conditional entropy rate*** *of stochastic process* $\mathbf{X}$ *is*

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

provided the limit exists.

The entropy rate represents entropy per symbol (or step in time series) whereas conditional entropy rate is conditional entropy of the last symbol given the past. These two quantities are generally distinct and even not necessarily exist. But for stationary processes the following theorem holds.

Theorem 1.11. *Let* $\mathbf{X} = \{X_n\}$ *be stationary stochastic process. Then* $H(\mathbf{X})$ *and* $H'(\mathbf{X})$ *exist and*

$$H(\mathbf{X}) = H'(\mathbf{X}).$$

We will not show prove of this theorem, but we should mention the interesting properties of stationary process that the prove takes advantage of.

Theorem 1.12. *Let* $\mathbf{X} = \{X_n\}$ *be stationary stochastic process. Then*

$$H(X_{n+1} | X_n, \ldots, X_1) \leq H(X_n | X_{n-1}, \ldots, X_1).$$

This means that for stationary processes the uncertainty of the next step given the past never increases.

1.9 Differential entropy

We shortly mention generalization of Shannon entropy for continuous random variables that is called differential entropy.

Definition 1.10. *Let* X *be random variable with probability density function* $f(x)$ *with support* S *then* ***differential entropy*** *is defined as*

$$h(X) = -\int_S f(x) \log f(x) \mathrm{d}x,$$

if the integral exist.

Likewise for discrete case the differential entropy is function only of probability density. But not every properties of discrete entropy are necessarily valid for continuous one. For instance, consider uniform distribution on interval $\langle 0, a \rangle$. Then we easily calculate $h(U) = \log a$, and for $a < 1$ we have negative entropy. Note also that differential entropy cannot be defined for δ_{x_0} distribution because $\log \delta_{x_0}$ is not well defined.

Entropy of Normal distribution Let us compute entropy of Normal distribution. We use convenient units (nats) and utilize known expression for Gauss integral

$$\int_{-\infty}^{+\infty} x^{2n} \exp\left(-\alpha x^2\right) \mathrm{d}x = \sqrt{\frac{\pi}{\alpha}} \frac{(2n-1)!!}{(2\alpha)^n}.$$

Then after little manipulation we get

$$H(X) = \frac{1}{2} \ln\left(2\pi\sigma^2\right) + \frac{1}{2}.$$

Unfortunately, this is not very useful because for small variance we get negative value of entropy.

Remark on Normal distribution We should mention that Normal distribution is the least biased distribution given mean and variance, in other words, it maximizes entropy constrained to fixed average value and standard deviation, i.e. it does not involve any other information and corresponds to maximum ignorance about the system. Similarly for n random variables and given covariance matrix entropy is maximized for n dimensional Gaussian distribution.

Exponential distribution Exponential distribution takes over the maximality property for positive random variables. The maximum entropy is

$$H(X) = \log(e\lambda).$$

The joint, conditional and other entropies are defined similarly to discrete ones, i.e. the sum is just replaced by integral.

Calculating entropy For calculating entropy of continuous random variable other approach can be used. We divide range of the random variable into N boxes of size ϵ and compute probabilities of these boxes

$$p_j = \int_{B_j} \rho(x) \mathrm{d}x.$$

Thus we get discrete random variables with N possible outcomes and may calculate its entropy. The entropy diverges with finer partitioning ($\epsilon \mapsto 0$), see [3] as it represents amount of information needed for specifying the state of the system with an accuracy ϵ. Consider easy example of uniform distribution on $\langle 0, 1\rangle$. Entropy of such random variable would mean average number of bits necessary to encode, in other words determine, arbitrary number from interval $\langle 0, 1\rangle$ and that is infinite (imagine any irrational number). Since in real world, i.e in real experiment, we are not able to distinguish all small details and any

measuring device can provide values only in some finite interval $[x - \epsilon, x + \epsilon]$, this infinity does not have to scary us.

Chapter 2

Rényi entropy

In this chapter we will generalize Shannon information measure according to Rényi, see [5]. We will follow intuitive way of Rényi to introduce new information measure, and furthermore, explore quantities related to information measure like relative information, conditional entropy or mutual information from other point of view in order to generalize them.

2.1 Information quantities

2.1.1 Entropy

To motivate Rényi entropy, we should have a look at the way how Shannon extended work of Hartley. Let $E = \bigcup_{k=1}^{n} E_k$ and E_k contains N_k elements. Then information necessary to characterize one of $N = \sum_{k=1}^{n} N_k$ equiprobable elements is $\log_2 N$. If we would like to know only the set in which the particular elements is, we can proceed as follows: Choosing arbitrary element can be done by first selecting E_k and then particular element from E_k. Since these two steps are independent, the additivity property of Hartley information measure claims

$$\log_2 N = H_k + \log_2 N_k,$$

where H_k represents information needed to specify set E_k. From this equation H_k can be readily obtained, and then it is reasonable to define H, information needed to specify the set which particular element belongs to, as a weighted sum of H_k and introduce probabilities $p_k = \frac{N_k}{N}$. Aforementioned procedure leads to already known Shannon's formula.

From above generalization of Hartley information measure we can see that Shannon information measure is based on two postulates (first of them was introduced by Hartley):

- Additivity - information gained from observing two independent events is the sum of the two partial ones

- Linear averaging - information gained from experiment that has n possible outcomes A_k with probabilities p_k $k = 1, \ldots, n$ is equal to linear average

$$\sum_{k=1}^{n} p_k H(A_k),$$

  where $H(A_k)$ denotes information gained from experiment when event A_k occurs.

Rényi was aware that there is no reason for restricting to linear average used by Shannon and considered Kolmogorov–Nagumo generalized mean

$$E_f[X] = f^{-1}(\sum p_i f(x_i)), \tag{2.1}$$

where f is continuous and strictly monotone (i.e. invertible). It represents the most general mean compatible with Kolmogorov axioms of probability theory. We may also encounter with name quasi-linear mean.

Hence, every continuous and strictly monotone function may define various measures of information i.e. various entropies. However, additivity postulate puts some constraints on possible functions f, namely it restricts f to only two options, linear $f(x) = cx$ and one-parametrized family of exponential functions chosen for later purposes in the form $f(x) = c(2^{(1-\alpha)x} - 1),\ \alpha \neq 1$, proof may be found in [6]. The linear function leads to already known Shannon entropy, and exponential function gives Renyi entropy

$$H_\alpha(\mathcal{P}) = \frac{1}{1-\alpha} \log_2 \sum_{i=1}^{N} p_i^\alpha. \tag{2.2}$$

Though, the left hand of 2.2 is defined for all $\alpha \neq 1$ we should disable non-positive values since for $\alpha < 0$ 2.2 becomes very sensitive to small probabilities. In other words, adding new event with zero probability makes $H_\alpha(\mathcal{P})$ infinite, and it is undesirable property because information measure should be function only of probability distribution and that remains unchanged after adding event with zero probability. For the same reason, $\alpha = 0$ is excluded since we get again value independent on probability distribution $H_0(\mathcal{P}) = \log_2 N$. Since the limit of 2.2 for α tends to 1 is well defined and equals to Shannon entropy[1], we may conclude that 2.2 defines eligible measure of information for $\alpha > 0$.

[1] It can be readily proven by L'Hopital rule

It is easily seen that H_α is non-negative and equals to 0 if and only if the probability distribution is degenerate, consider that $p_i^\alpha \leq p_i$ for $\forall i$ and $\alpha > 1$ where $p_i \in \langle 0, 1\rangle$ with equality iff $p_i = 0$ or $p_i = 1$ and opposite inequality for $\alpha < 1$. Using method of Lagrange multipliers, we may find that $H_\alpha \leq \log_2 N$, and thus we have the same boundary conditions as we have already seen for Shannon entropy

$$0 \leq H_\alpha \leq \log |\mathcal{X}| \qquad \alpha > 0.$$

By differentiation of 2.2 with respect to α and as a consequence of Jensen's inequality 1.9 applied to convex function $-\log x$ and random variable $X = p_k^{1-\alpha}$ with probability distribution $\frac{p_k^\alpha}{\sum p_k^\alpha}$, we find that Rényi entropy is decreasing function of α. This means that Shannon entropy may be regarded as a lower and upper boundary for Rényi's information measure of order $\alpha < 1$ and $\alpha > 1$ respectively.

2.1.2 Relative entropy revised

In preceding chapter we defined relative entropy without any motivation. Now we should have a look at this quantity in more detail to generalize it for Rényi's α information measure.

Relative entropy is connected with the idea of gain of information as can be seen from following example. Consider an experiment with $A_1, \ldots, A_n$ possible outcomes which occur with probabilities $p_1 = P(A_1), \ldots, p_n = P(A_n)$. Now we observe event B, and the probabilities change to $q_1 = P(A_1|B), \ldots, q_n = P(A_n|B)$. It is legitimate to ask how much information about the experiment we gained from observing event B. To answer this question, we first imagine only one outcome, say A_1. Before observing event B the outcome would give us $\log_2 1/p_1$ bits of information or equivalently the uncertainty of the outcome is $\log_2 1/p_1$. After occurring B, the uncertainty and possible information received from observing event A_1 change to $\log_2 1/q_1$. Hence we need $\log_2 1/p_1 - \log_2 1/q_1$ bits of information less than before and this decrease in uncertainty is equal to gain of information about A_1 observing B.

If we take into account all outcomes, we get n partial gains of information, and it is reasonable to assign the average of these gains to overall gain of information about experiment after observing event B. Notice that gain of information may be considered also as minus increase of uncertainty, and this brings us two possibilities to calculate overall gain of information. Either we take average of partial gains $\log_2 \frac{q_k}{p_k}$ or average increases of uncertainty $\log_2 \frac{p_k}{q_k}$ and the result multiply by (-1).

In Shannon's case of linear averaging both approaches leads to the same already known relative entropy. However, for generalized average, i.e. $E[X] = \frac{1}{1-\alpha} \log_2 \left(\sum p(x) 2^{(1-\alpha)x}\right)$, we get different results, see [5]. The first method

leads to undesirable properties of information gain, and hence the other method is used. Note that averaging is done with respect to $q \equiv \{q_k\}$ since we assume that B occurred, and hence q_k is the appropriate weight of event A_k.

Definition 2.1. *Let p and q be probability distributions on the same discrete probability space. Then gain of information of order α when p is replaced with q is*

$$D_\alpha(q||p) = \frac{1}{\alpha - 1} \log_2 \left(\sum_{k=1}^{n} \frac{q_k^\alpha}{p_k^{\alpha-1}} \right). \tag{2.3}$$

The properties of ordinary relative entropy are conserved and are again rooted in Jensen's inequality. We state one more property of D_α valid for all $\alpha > 0$

$$D_\alpha(q||u) = H_\alpha(u) - H_\alpha(q).$$

This relates gain of information with decrease of uncertainty after replacing the most ignorant distribution, i.e. uniform one u, with arbitrary distribution q. The prove is just inserting uniform distribution of size n to the definition.

2.1.3 Conditional entropy

Here we follow the same idea as in the first chapter only linear averaging is replaced by generalized mean. Having two random variables X and Y the remained uncertainty about X or information still gained from observing X after knowing that $Y = y_k$ is

$$H_\alpha(X|Y = y_k) = \frac{1}{1 - \alpha} \log_2 \left(\sum_{h=1}^{n} p_{h|k}^\alpha \right).$$

Then generalized averaging gives us conditional information of order α.

Definition 2.2. *Let X and Y be two discrete random variables with distribution p and q then conditional information of order α is defined as*

$$H_\alpha(X|Y) = \frac{1}{1 - \alpha} \log_2 \left(\sum_{h,k} \frac{r_{hk}^\alpha}{q_k^{\alpha-1}} \right),$$

where r_{hk} denotes joint probability distribution.

The inequality valid for Shannon conditional entropy is easily broaden to Rényi conditional entropy so we have

$$0 \leq H_\alpha(X|Y) \leq H_\alpha(X) \tag{2.4}$$

with the same conditions for equality as in Shannon's case, i.e. $H_\alpha(X|Y) = 0$ iff there is such a function g that $X = g(Y)$ and $H_\alpha(X|Y) = H_\alpha(X)$ iff X and Y are independent, see [5].

We remark another definition of conditional information that is based on the additive property of Shannon entropy for dependent variables, equation 1.9. We can postulate this equation also for Rényi entropy and define conditional entropy as

$$\tilde{H}_\alpha(X|Y) = H_\alpha(X,Y) - H_\alpha(Y) = \frac{1}{1-\alpha}\log_2\left(\frac{\sum\limits_{k=1}^{n} q_k^\alpha \left(\sum\limits_{h=1}^{m} p_{h|k}^\alpha\right)}{\sum\limits_{k=1}^{n} q_k^\alpha}\right), \quad (2.5)$$

where $H_\alpha(X,Y)$ is Rényi entropy of joint probability distribution.

Escort distribution

Imagine arbitrary probability distribution p then we can construct another probability distribution ρ called **escort distribution**

$$\rho_{qk} = \frac{p_k^q}{\sum\limits_{k=1}^{n} p_k^q}.$$

This new probability distribution has interesting property that it emphasizes probable events and suppresses rare ones for $q > 1$. The greater q, the more pronounced is the accentuation of probable events, i.e by choosing large q we restrict our interest on the center of probability distribution. On the other hand, $0 < q < 1$ highlights rare events and covers up most likely ones. As an example consider original distribution $N(0,1)$ then escort distribution corresponds to $N(0, \frac{1}{q})$ and depending on value of q bell curve get either narrower or more flatten.

Due to monotony of exponential function inequalities among probabilities remain unchanged and for q close to zero escort distribution tends to uniform distribution. This feature can be violated by allowing negative values of q which actually changes tails to peaks in probability distribution and vice versa.

Since escort distribution deforms original distribution, it is used in statistical physics for "zooming" in different regions of probability distribution. We shall note that escort distribution of escort distribution is also escort distribution with parameter $q = q_1 q_2$, i.e. escort distribution may be consider as a one-parametric group of transformations on probability distributions. Thus another "zooming" does not give us any new information.

We should also mention relation of Rényi entropy of escort distribution and entropy of original distribution

$$H_{1/q}(\rho_q) = H_q(p).$$

With the help of escort distribution we can rewrite equation 2.5 to

$$\tilde{H}_\alpha(X|Y) = \frac{1}{1-\alpha}\log_2\left(\sum_{k=1}^{n}\rho_{\alpha k}2^{(1-\alpha)H_\alpha(X|Y=y_k)}\right), \tag{2.6}$$

which means that to fulfill condition 1.9 we have to average with respect to escort distribution instead of original distribution.

It can be shown, see [6], that $\tilde{H}_\alpha(X|Y) = 0$ iff outcome of Y uniquely determines X and for independent random variables $\tilde{H}_\alpha(X|Y) = H_\alpha(X)$, but the opposite implication does not generally hold. $\tilde{H}_\alpha(X|Y) = H_\alpha(X)$ must hold for all $\alpha > 1$ or $0 < \alpha < 1$ to imply independence of X and Y, see [7].

2.1.4 Mutual information of order α

There are more ways how to define mutual information of order α. All of them are motivated by some relation valid for Shannon mutual information. The ambiguity is caused by the fact that all relations valid for Shannon mutual information cannot be simultaneously satisfied by any definition of Rényi mutual information. Hence, in application we should pick up such a definition that best fits our requirements.

The Shannon mutual information was defined in first chapter as a gain of information after replacing total independence by the joint distribution. Analogically, we could use equation 2.3 and define generalized mutual information in the same way. Unfortunately, this definition would violate desirable property of mutual information, namely

$$I_\alpha(X;Y) \leq H_\alpha(X), \tag{2.7}$$

which states that the information on X yielded by Y must not exceed uncertainty of X.

Mutual information may also be defined by the property of Shannon mutual information

$$I(X;Y) = H(X) - H(X|Y). \tag{2.8}$$

This would give us generalized mutual information in the form

$$I_\alpha(X;Y) = \frac{1}{1-\alpha}\log_2\left(\frac{\sum_{h=1}^{m}p_h^\alpha}{\sum_{h=1}^{m}\sum_{k=1}^{n}\frac{r_{hk}^\alpha}{q_k^{\alpha-1}}}\right). \tag{2.9}$$

However, Rényi preferred in his paper [5] another way of defining mutual information. He noticed that Shannon mutual information can be written as an average of information gain

$$I(X;Y) = \sum_{k=1}^{n} q_k D(P(X|Y = y_k)||P(X)).$$

Using equation 2.3 and generalized mean instead of linear averaging results in

$$I_\alpha(X;Y) = \frac{1}{1-\alpha} \log_2 \left(\sum_{k=1}^{n} \frac{q_k}{\sum\limits_{h=1}^{m} \frac{p_{h|k}^{\alpha}}{p_h^{\alpha-1}}} \right) \tag{2.10}$$

which satisfies 2.7 with the same conditions for equality as in Shannon's case. The drawback is that neither 2.10 nor 2.9 is symmetric, i.e. information on X gained from observing Y is generally distinct from information on Y from X.

It should be noted that 2.10 and 2.9 are different. First of them represents decrease of uncertainty while the second one is average information gain on X from observing Y.

In chapter 4 we will use different definition of mutual information that is based on property of Shannon mutual information

$$I(X;Y) = H(X) + H(Y) - H(X,Y).$$

By inserting Rényi entropies we arrive to the formula

$$I_\alpha(X;Y) = \frac{1}{1-\alpha} \log_2 \frac{\sum\limits_{h,k} (p_h q_k)^\alpha}{\sum\limits_{k,h} r_{hk}^\alpha}. \tag{2.11}$$

This quantity is symmetric and might have been obtained also by using equation 2.8 and the second definition of conditional entropy, equation 2.5.

Rényi rejected this definition because for Rényi information measure inequality

$$H_\alpha(X) + H_\alpha(Y) \geq H_\alpha(X,Y)$$

does not always hold. Hence, 2.11 can be negative, and according to Rényi it is inappropriate to have negative mutual information. However, it was examined in [7] that mutual information defined in 2.11 is negative if marginal events of X obtain higher probability at the cost of decrease of probability of central part of the distribution after observing Y. Such a feature can be handy in various applications, for example in finance as we will see in chapter 4.

2.2 Operational definition

Renyi entropy is information measure as well as Shannon entropy. Now we should address some possible ways how to interpret its actual value. This gives us basic view to particular problems in applications.

We already know that Shannon entropy emerged from coding theory where it represents the shortest average length of optimal code

$$H_1(p) \leq L(p) = \sum_{i=1}^{N} l_i p_i,$$

and the optimal lengths of individual symbols are related with their probabilities as

$$l_i^* = -\log_2 p_i.$$

That means highly improbable symbols corresponds to very long codewords in order to save short lengths for frequently transmitted symbols. Such a behavior is convenient for linear cost function occurring in transmitting where the bits are send one by one, hence, sending n bits takes n-times longer than sending one bit. Nevertheless, in some situations it is reasonable to use convenient cost function, for example, in storing data when exponential cost function may be used for "pricing" allocated free space. Thus, we are not interested in the shortest code but the cheapest one.

Campbell dealt with the problem of exponential weighting in his paper [8]. He proposed to minimize

$$C = \sum_{i=1}^{N} p_i D^{tl_i}$$

with respect to lengths l_i where t is some parameter related to the cost and D is number of symbols used for encoding messages. However, further analysis suggested to minimize logarithm of C

$$L(t) = \frac{1}{t} \log_D \left(\sum_{i=1}^{N} p_i D^{tl_i} \right) \tag{2.12}$$

so that an elegant connection with generalized mean 2.1 would emerge, i.e. 2.12 corresponds to Kolmogorov-Nagumo generalized mean with $f(x) = D^{tx}$.

The following theorem is the analogy of well-known Shannon noiseless channel theorem.

Theorem 2.1. *Let $l_1, \ldots, l_N$ satisfy* ***Kraft inequality***

$$\sum_{i=1}^{N} D^{-l_i} \leq 1,$$

then averaged length of optimal code with exponential cost is bounded from below

$$H_\alpha \leq L(t), \tag{2.13}$$

where $\alpha = 1/(t+1)$.

According to this theorem we must lengthen the code for highly probable symbols in order to be able to shorten improbable ones which would be otherwise strongly penalized by exponential cost function.

We have equality in 2.13 if

$$l_i = -\log_D \rho_{\alpha i}$$

where ρ_α is escort distribution, but this is actually the same result that we obtained for linear averaging except we replaced original distribution with escort distribution. For $t > 0$ we have $\alpha < 1$, and escort distribution properly enhances rare probabilities and suppresses likely ones so that we can use Shannon formula for optimal lengths. On the other hand, $-1 < t < 0$ corresponds to $\alpha > 1$ and probable events receives even shorter code. This may be helpful in the case when finite buffer is used for transmitting and we are interested in maximizing probability of sending message in one snapshot.

Aforementioned connection with classical coding procedure is also advantage in applicability of new coding theorem because we do not have to invent some new coding method which approaches the optimal lengths. We can just use Huffman code with escort distribution.

2.3 Axiomatization of information measure

Renyi compares information with energy because there was considered many different kinds of energy, and it took many years to discover that all of them are just one 'thing', the same discovery may come even for information, but in order to define different information measures, it is convenient to postulate some basic requirements that suitable information measure should fulfill. Fadeev proposed the following set of postulates:

1. Information measure is function only of probability distribution and has to be symmetric - $H(p_1, \ldots, p_n) = H(p_{\pi(1)}, \ldots, p_{\pi(n)})$ for any permutation π.

2. $H(p, 1-p)$ is a continuous function for $p \in \langle 0, 1\rangle$.

3. *normalization* - $H(\frac{1}{2}, \frac{1}{2}) = 1$.

4. $H(p_1, \ldots, p_n) = H(p_1 + p_2, p_3, \ldots, p_n) + (p_1 + p_2)H(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2})$.

The last axiom states that overall information needed for identification of particular message is independent on grouping of messages. That means that we can combine, say, two messages with probabilities p_1 and p_2 into one message, thus information needed for selecting one of these $n-1$ messages corresponds to the first term on the right side. When this new message occurs we examine which of the original two messages was actually sent, information necessary to this identification is the second term on the right side. The axiom demands that information needed for this procedure is equal to information needed for directly selecting particular message.

It can be shown, see [9], that these axioms holds if and only if Shannon information measure is used.

Theorem 2.2. *Let $p_1, \ldots, p_n$ be a probability distribution and H be an arbitrary function fulfilling postulates 1 to 4 above, then*

$$H(p_1, \ldots, p_n) = -\sum_{j=1}^{n} p_j \log_2 p_j.$$

The fourth axiom is somewhat too restrictive and precludes information measure of order α (Rényi's entropy). Therefore, Rényi weakened the fourth axiom by assuming only additivity of entropy for independent experiments and introduced new set of axioms that characterizes both Shannon and Rényi information measure. These new axioms are formulated for generalized probability distributions, i.e. including incomplete distributions for which $\sum p_i \leq 1$.

1. H is a symmetric function of the elements of generalized distribution .

2. $H(\{p\})$ is a continuous function of p for $p \in (0, 1\rangle$.

3. *normalization* - $H(\{\frac{1}{2}\}) = 1$.

4. *additivity* -

$$\begin{aligned} &H(\{p_1q_1, \ldots, p_nq_1, p_1q_2, \ldots, p_nq_2, \ldots, p_1q_m, \ldots, p_nq_m\}) = \\ &H(\{p_1, \ldots, p_n\}) + H(\{q_1, \ldots, q_m\}). \end{aligned} \tag{2.14}$$

5. *averaging* - There exists a strictly monotone and continuous function $g(x)$ such that for two generalized probability distributions $\{p_i\}$ and $\{q_k\}$ denote $W(\{p_i\}) = \sum p_i$, $W(\{q_k\}) = \sum q_k$ and if $W(\{p_i\}) + W(\{q_k\}) \leq 1$, then

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

Characterization of Shannon and Rényi entropy is then given by the following theorem.

Theorem 2.3. *Let $H(\{p_i\})$ be defined for all generalized probability distributions and satisfies axioms 1 to 4 and axiom 5 with $g_\alpha(x) = 2^{(\alpha-1)x},\ \alpha > 0,\ \alpha \neq 1$ and $g(x) = ax + b,\ a \leq 0$, then*

$$H(\{p_i\}) = \frac{1}{1-\alpha} \log_2 \left[\frac{\sum p_i^\alpha}{\sum p_i} \right] \text{ and}$$

$$H(\{p_i\}) = \frac{-\sum p_i \log_2 p_i}{\sum p_i} \text{ respectively.}$$

We mention one more set of axioms characterizing both Shannon and Rényi entropy that define also an conditional entropy in the form 2.6 which will be used in chapter 4.

1. Let X be a discrete random variable with probability distribution $\{p_i\}$, then $H(X)$ is a function only of $\{p_i\}$ and is continuous with respect to all its arguments.

2. For a given integer n $H(X)$ takes its maximum for $\{p_i = 1/n,\ i = 1, \ldots, n\}$ with the normalization $H(X) = 1$ for distribution $\{1/2, 1/2\}$.

3. For a given $\alpha \in \mathbb{R}$ and two random variables X, Y $H(X,Y) = H(X) + H(Y|X)$ with

$$H(X|Y) = g^{-1}\left(\sum_i \rho_{\alpha i} g\Big(H(Y|X = x_i) \Big) \right),$$

   where $\rho_{\alpha i}$ is escort distribution of probability distribution of X.

4. g is invertible and positive in $\langle 0, +\infty)$.

5. Let X be a random variable and $\{p_1, \ldots, p_n\}$ its distribution, and if X' has probability distribution $\{p_1, \ldots, p_n, 0\}$ then $H(X) = H(X')$. That is, adding an event of probability zero we do not gain any new information.

These axioms are generalization of Khinchin's axioms [10] in order to include Rényi entropy. It can be shown that the only possible functions in axiom 3 are either linear or exponential function which corresponds to Shannon and Rényi entropy, see [6].

Chapter 3

Correlation, memory and causality

In this chapter basic measures of correlation and memory are reminded, and their contemporary replacements from information theory are introduced. After correlation, term causality is discussed, and at the end of this chapter primary problems of information-theoretical approach are pointed out.

3.1 Characterizing and measuring correlation

The most known characterization of correlation is by **Pearson correlation coefficient**

$$R = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{E[(X-\mu_X)^2]E[(Y-\mu_Y)^2]}},$$

usually called just correlation coefficient which may lead to confusion. The reason for possible confusion is due to the fact that generally correlation should refer to any kind of relation, but Pearson coefficient measure just linear dependence.

Auto\Cross-correlation function measures exactly this coefficient between different values in stochastic process lagged in time or anything the index set refers to. The definition of auto-covariant function is

$$cov(t_1,t_2) = E[(X_{t_1} - \mu(t_1))(X_{t_2} - \mu(t_2))], \tag{3.1}$$

and auto-correlation function is obtained by dividing 3.1 by $\sigma_{t_1}\sigma_{t_2}$. In cross-covariant function the second time would refer to seres Y_t.

Such a characterization of dependence in stochastic process is sufficient for Gaussian processes for which all kind of dependence is included in auto-correlation function. This is due to the fact that Gaussian random variable is fully

specified by given mean and variance, and, in analogy, Gaussian process (its distribution) is completely defined by mean function $\mu(t)$ and auto-covariant function $cov(t, s)$ or instead of auto-covariant function, we may specify variance function $\gamma(t)$ and auto-correlation function $cor(t, s)$, see [11]. Since Gaussian processes showed fruitful applicability in practice mainly as a result of widely used ARIMA models, which are generally defined without reference to Gaussian noise, but in concrete application other noises used to be very rare, the auto-correlation function was long considered as a perfect measure of dependence in time series.

Nowadays it is generally accepted that neither linear ARIMA nor Gaussian process are satisfactory models for many phenomena one can encounter in biology, physics, finance or many other fields. Therefore, it is desired to generalize the Pearson correlation coefficient, and then auto\cross-correlation function, in an appropriate way so that it could incorporate all higher order dependence and thus would be a plausible measure of correlation.

In information theory context very promising candidate for such a measure is mutual information which takes advantage of information theoretical approach which means that it can capture all kinds of dependence because $I(X, Y) = 0$ iff X and Y are independent, see 1.1.

Figures 3.1 and 3.2 demonstrate this benefit of information approach. Figure 3.2 shows lagged mutual information of log-returns of London stock exchange index AIM100, and we see that unlike auto-correlation it detects nontrivial correlation even on one-hour time lag. The figure 3.2 nicely shows that original hypothesis by Bachelier that successive returns of stock prices are perfectly independent is wrong for empirical high frequency data and confirms Mandelbort's claim about history in returns or log-returns.

For another comparison of mutual information with conventional correlation coefficients, namely Pearson, Spearman and Kendall, see [12]. In that paper three examples were considered:

a) $Y = \epsilon$, $\epsilon \sim N(0, 1)$ and $X \sim Uniform([-3, 3])$

b) $Y = X + \epsilon$ and $X \sim Uniform([-3, 3])$

c) $Y = X^2 + \epsilon$ and $X \sim Uniform([-3, 3])$

In the paper is shown by simulating 10000 data points that all four quantities can detect no correlation in case a) and strong correlation in case b). However, in case c), i.e. non-linear correlation, all conventional coefficients fail in detection of correlation except mutual information.

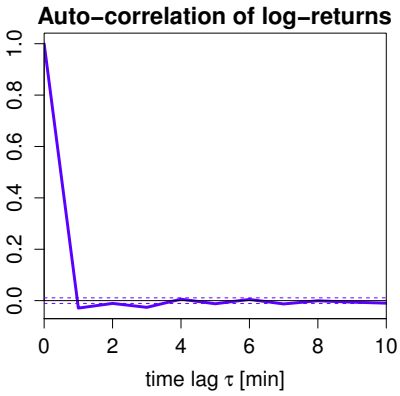


Figure 3.1: ACF of log-returns $r(t)$

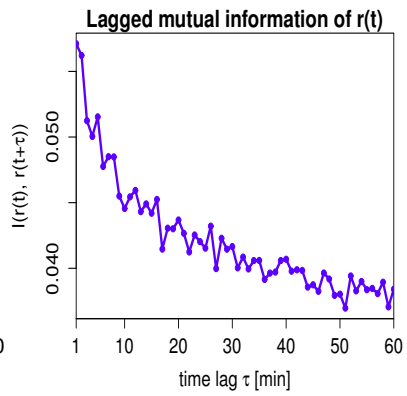


Figure 3.2: Lagged mutual information

3.1.1 Correlation vs Regression

These terms are very closely related but prone to confusion. Correlation means that there is some relation between two events or random variable, nevertheless, do not contain any information about a particular relation. This is the object of regression, i.e, to find appropriate functional relation, e.g. linear ($Y = F(X) = aX + b$), between X and Y which minimizes suitable measure of error in regression. The most widely used error measure is sum of squares of errors also called residuals $r_i = Y_i - F(X_i)$. Method using such a minimizing function is called method of **least squares**, and its popularity is due to the mathematical well behaved square function which may be differentiated, and thus easily minimized in contrast to absolute value error.

Note that term linear regression does not refer only to linear function F but to all functions which lead to linear optimizing equations, i.e. equations for parameters of regression. Linear regression is very favorable in practice since system of linear equations is well understood, and there are many methods how to solve them.

Pearson correlation coefficient is closely related to linear regression $Y = aX + b$ which may be seen from QQ-plot of realizations $\{x_i\}$ and $\{y_i\}$. According to sign of R we may better predict value of Y provided we know realization of X by simple rule, that when $x_i > E[X]$ then y_i is more likely to be also above the mean $E[Y]$, and a strength of this relation is proportional to $|R|$. That is R contains some extra information in contrast to mutual information. Since provided we know only mutual information, we have no clue how to improve prediction of y_i based on known value x_i. All we know is that there is in principle some possibility to better predict Y.

3.2 Memory in time series

Memory in a time series is classified according to auto-correlation function

Definition 3.1. *Stochastic process* $\{X_t\}$ *is said to have* ***short memory*** *if*

$$\sum_{\tau=-\infty}^{+\infty} |cor(t,\tau)| < +\infty \qquad for\ all\ t,$$

and if the series diverges then the process is said to have ***long memory****.*

Generalization were proposed in [13] which would take into account also non-linear dependency. The generalization is defined with help of mutual information as follows.

Definition 3.2. *Stochastic process* $\{X_t\}$ *is said to have* ***short memory in information*** *if*

$$\sum_{\tau=-\infty}^{+\infty} I(X_t, X_{t+\tau}) < +\infty \qquad for\ all\ t,$$

and if the series diverges then the process is said to have ***long memory in information****.*

Another kind of stochastic process related with memory is mean reverting process.

Definition 3.3. *Stochastic process* $\{X_t\}$ *is called* ***mean reverting process*** *if*

$$\lim_{\tau\to+\infty} cor(t,\tau) = 0, \qquad for\ all\ t$$

Trajectories of such process have tendency to fluctuate around the mean function, i.e. function $X_t - E[X_t]$ changes sign infinitely often. The most known example which does not possess mean-reverting property is Brownian motion. Even infinitely lagged values are still correlated. Its trajectories are "persistent", i.e. if any trajectory happens to be above mean then it is likely that it stays above the mean for ever.

Definition of mean reverting process is in [13] generalized to mean reverting process in information analogously as has been done for short\long memory process in information.

3.3 Granger causality

At first, let us highlight the difference between correlation and causality. Correlation is necessary condition for causality but by no means implies causality. So there may be correlated random variables which has absolutely no cause-effect relation.

There is no general definition of causality. The promising definition in case of deterministic causality may be given by two conditions:

1. if event A occurs then event B must occur
2. if event B occurs then event A must have occurred

Drawback of this deterministic causality is lack of their applicability since in real world strict determinism is very rare if not impossible. Thus adjustment is to be made to cast this defining properties into realm of probability.

Definition 3.4. *An event A is a cause to the event B in a probabilistic sense if two conditions are satisfied:*

1. *event A precedes event B*
2. *$P(A) \neq 0$ and $P(B|A) > P(B)$*

The imperfection of this definition resides in practical difficulties how to precisely state events A and B.

Wiener attempted to define causality in two signals as follows, see [14].

> 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, inspired by Wiener, slightly improved the above definition by requiring two conditions to be satisfied for cause effect relation.

1. the cause occurs before the effect
2. the cause contains information about the effect that is unique, and is in no other variable

Requirement concerning uniqueness of the information eliminates possible effect of the third hidden driving variable. On the other hand, it makes the causality generally uncheckable since in practice we will never be able to distinguish all hidden variables and exclude their effect.

The way how to overcome this seemingly hopeless causality detection is by checking both aims of causality. For instance, in case of stochastic process, when Granger's definition reduces to Wiener's one complemented by uniqueness of information, we should check whether

a) $\{X_t\}$ causes $\{Y_t\}$

b) $\{Y_t\}$ causes $\{X_t\}$

and only if just one test is positive we may say that one process causes the other.

3.3.1 Granger test

The last and the most practical point in testing causality between stochastic processes is question how to incorporate the abstract information in definition into forecasting. Granger proposed simple linear model

$$Y_t = a_0 + \sum_{k=1}^{L} b_{1k} y_{t-k} + \sum_{k=1}^{L} b_{2k} X_{t-k} + \xi_t. \tag{3.2}$$

So it is assumed that information from $\{X_t\}$ induces only linear change in process $\{Y_t\}$. Granger test is then based on testing null hypothesis of no causality, i.e. $b_{2k} = 0 \;\; k \in \{1, \ldots, L\}$ which leads to Granger-Sargent statistics

$$GS = \frac{(R_2 - R_1)/L}{R_1/(N - 2L)},$$

where R_1 is residual sum of squares under assumption of model 3.2, R_2 is residual sum of squares under assumption of null hypothesis and N is number of observed data points. The GS statistics has Fisher–Snedecor distribution with L and $N - 2L$ degrees of freedom.

An advantage of Granger test is that it is rigorous statistical test, hence, it gives us significant results. Furthermore, it is computationally quite undemanding. However, the test has little meaning when a priory chosen model 3.2 cannot be properly justified.

The problem with parametric Granger test streams from the fact that Granger test actually tests Wiener version of causality in sense of better predictability and not Granger version in sense of information. In chapter 4 we will see that information-theoretical quantity called transfer entropy may indeed measure this abstract information and hence may serve as a replacement for Granger causality test.

Similar note as in case of regression and correlation may be added to causality which simply says that there is in principle some possibility to use data from cause series to make better prediction about the effect series than without them. However, exact form of this dependence is another more difficult and more practically useful question which cannot be addressed by information-theoretical approach. Thus in situations when linear model is fairly plausible Granger test is invaluable.

3.4 Problems with information-theoretical approach

To motivate essential disadvantage of using information-theoretical quantities in time series analysis, we briefly bring up problem of inference of statistical

quantities from observed time series.

From time series auto-covariance function is estimated by

$$c_k = \frac{1}{N-|k|} \sum_{i=1}^{N-|k|} (X_i - \bar{X})(X_{i+|k|} - \bar{X}), \tag{3.3}$$

where $\bar{X}$ is sample mean. This estimator is unbiased provided we know the mean μ, but when mean is estimated by $\bar{X}$ then 3.3 becomes biased. We may also encounter estimator with prefactor $\frac{1}{N}$ instead of $\frac{1}{N-|k|}$ which has larger bias than 3.3, but on the other hand it has a smaller Mean square error, see [11]. Short discussion about bias and MSE is mentioned in Appendix A. Notice that 3.3 does not involve time thus at least weak stationarity is assumed.

Unfortunately, in contrast to auto\cross-correlation function which may be easily estimated by 3.3 whose properties are well-known and rigorously proven the estimation of mutual information from experimental data is more difficult. A compendium of available methods for estimating mutual information may be found in [14].

Let us briefly mention here method that was used in calculating mutual information of London stock index returns in figure 3.2. The method originated in theory of dynamical systems and was invented by Fraser and Swinney [15] who were interested in finding optimal time delay into reconstruction theorem, the cornerstone of dynamical systems, for more details about analysis of dynamical systems see [16].

They used simple plug-in estimator which means that only appropriate estimation of joint probability of two random variables is needed. The sophisticated technique of constructing proper partition for estimating joint probability makes it very robust, though difficult for implementation.

It is worth mentioning that in paper [15] random variables X and Y are considered continuous, however, mutual information is calculated for their discretization, and then

$$\lim_{n\to+\infty} i_n(X,Y) \tag{3.4}$$

is claimed to be a right value for mutual information of two continuous random variables. Where $i_n(X,Y)$ is a plug-in estimator of mutual information of discretized random variables in the n-th recursive step. Such a procedure would not give a correct value for differential entropy as was noted in section 1.9, but for mutual information the limit 3.4 indeed converges to $I(X,Y)$.

The iterative procedure for discretization is as follows:

1. Split support of X into two parts so that each part contains the same number of data points. Do the same for Y which gives the first partition G_1 of (X, Y) plane.

2. Create G_{n+1} from G_n by splitting each interval in discretization of X into two parts so that each box contains the same number of data points. Do the same for Y which as a result splits each element in G_n into four parts, see figure 3.3.

This approach gives for all n marginal distribution of X and Y equal to 2^{-n} independently on particular box which leads to significant simplification.

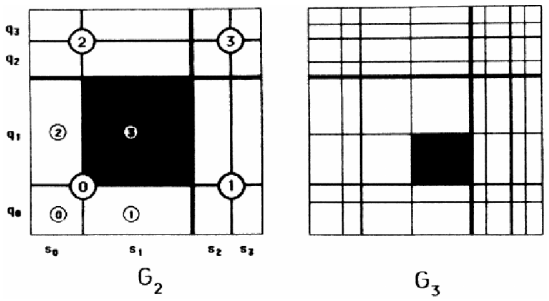


Figure 3.3: Successive steps in partitioning support of X and Y (adapted from [15])

The described algorithm of making finer partition seems to proceed in all parts of (X, Y) plane, but the key task is performed in recursive calculation of $I(X, Y)$ which stops whenever the element currently processing is flat, i.e. it does not have any finer structure, and it would not contribute anymore to total mutual information, see details in the paper [15].

The only remaining question is how to decide whether there is any further structure. Authors have chosen simple χ^2 test for testing flat multinomial distribution of two next iteration of finer partition on 20% confidence level. The recursion stops also when few points remains in the element, but the concrete number of points to stop is not stated.

Final remark about estimating mutual information concerns amount of data typically used. In dynamical systems when studying known system of equations, e.g. Lorentz attractor, no problem with data arises because we may simulate as much data as necessary, see [15] where 1 000 000 data points were generated. Therefore errors in estimating information-theoretical quantities are usually neglected. However, in finance analyzing real data we are more restricted by available "clean" data, even at high frequency acquisition.

Chapter 4

Transfer entropy

In this chapter we will see how the basic concepts of information theory can be exploited in time series analysis. In particular, we will measure information flow between two time series in order to detect any causality between them.

Transfer entropy was firstly introduced by Schriber [17] who applied it to biological data, and now we see applicability in many distinct fields. In this work we aim to financial time series likewise in [18] and [19]. Transfer entropy is very useful tool for cross-correlation and causality analysis of two time series. The huge advantage of transfer entropy is an independence on model used for modeling time series, i.e. model-free. Thus transfer entropy has broader applicability than Granger's method that assumes linear regressive model. In addition, transfer entropy is able to quantify information flow and not only reveal existence of causality.

Transfer entropy takes into account also higher order correlations, i.e. any kinds of dependency, and thus may show that two series are intertwined even if cross-correlation analysis points out no correlation.

4.1 Shannonian transfer entropy

We introduce transfer entropy in the similar way as was done in [18] and in its extended version [20] with slight modification in the form of time dependency.

Having discrete stochastic process $\mathbf{X} = \{X_t\}$ (time series), we define block entropy of order m and at time t as

$$H_{\mathbf{X}}(t, m) = -\sum p(x_t, x_{t-1}, \ldots, x_{t-m+1}) \log_2 p(x_t, x_{t-1}, \ldots, x_{t-m+1}) =,$$

where sum is over all possible m-tuples $(x_t, x_{t-1}, \ldots, x_{t-m+1})$ which we denote $x_t^{(m)}$ for the sake of brevity. We see that block entropy of stochastic process is

just joint entropy of m successive random variables $X_t, \ldots, X_{t-m+1} \equiv X_t^{(m)}$, i.e.

$$H_{\mathbf{X}}(t, m) = H(X_t^{(m)})$$

Block entropy represents, depending on point of view, either averaged uncertainty of next m values at time $t - m$ provided we have no extra knowledge about the process or information capacity about the process stored in m successive observation as a function of time. The difference

$$h_{\mathbf{X}}(t, m) = H_{\mathbf{X}}(t + 1, m + 1) - H_{\mathbf{X}}(t, m) = H(X_{t+1}, X_t^{(m)}) - H(X_t^{(m)})$$

is very important for predicting because it represents conditional entropy, see chain rule in 1.8 in chapter 1, at time t of the next step provided we know all m preceding values of the process. According to basic properties of conditional entropy 1.15 stated in chapter 1 we have inequality

$$0 \leq h_{\mathbf{X}}(t, m) = H(X_{t+1}|X_t^{(m)}) \leq H(X_{t+1}),$$

where $H(X_{t+1})$ is uncertainty of the next step without any extra information, for instance history of the process. The limit $\lim_{t\mapsto\infty} h_{\mathbf{X}}(t-1, t-1)$ is already known conditional entropy rate, definition 1.9.

For quantitative characterization it is convenient to introduce **relative explanation** that indicates percentage of predictability i.e. how much percent of information about the next step is stored in m preceding values

$$RE_{\mathbf{X}}(t, m) = 1 - \frac{h_{\mathbf{X}}(t, m)}{H(X_{t+1})} = \frac{I(X_{t+1}; X_t^{(m)})}{H(X_{t+1})}. \tag{4.1}$$

The relative explanation, especially its dependence on m, may also be used for characterizing stochastic processes. Imagine $RE_{\mathbf{X}}(t, m)$ remains zero independently on m for $\forall t$ this situation indicates totally random process since observing history of the process does not give us any new information about the next step. Such process is sometimes called *White noise*. Similarly, increase of $RE_{\mathbf{X}}(t, m)$ with respect to m until some fixed value M in which $RE_{\mathbf{X}}(t, m)$ levels off at value less than 1 suggests Markov process of order M. The third special case that can be detected by relative explanation is periodic process for which $RE_{\mathbf{X}}(t, m)$ reaches 1 for some M and $\forall t$, this value then corresponds to period of the process.

With conditional entropy in hand we can easily extend it for two stochastic processes $\mathbf{X}$, $\mathbf{Y}$ and get **transfer entropy** in the form

$$T_{\mathbf{Y}\mapsto\mathbf{X}}^{(m,l)}(t) = h_{\mathbf{X}}(t, m) - h_{\mathbf{XY}}(t, m, l), \tag{4.2}$$

where the conditional entropy for two processes is

$$h_{\mathbf{XY}}(t, m, l) = H(X_{t+1}, X_t^{(m)}, Y_t^{(l)}) - H(X_t^{(m)}, Y_t^{(l)}) = H(X_{t+1}|X_t^{(m)}, Y_t^{(l)}) \tag{4.3}$$

where $X_t^{(m)}$ and $Y_t^{(l)}$ substitutes history in X and Y respectively, i.e. $X_t^{(m)} \equiv X_t, \ldots, X_{t-m+1}$ and similarly $Y_t^{(l)} \equiv Y_t, \ldots, Y_{t-l+1}$. Thus,

$$T^{(m,l)}_{\mathbf{Y}\mapsto\mathbf{X}}(t) = H(X_{t+1}|X_t^{(m)}) - H(X_{t+1}|X_t^{(m)}, Y_t^{(l)}). \tag{4.4}$$

From equation 4.4 we see that transfer entropy is always nonnegative since any extra knowledge about random variable never increase uncertainty, see inequality 1.15, and transfer entropy vanish if and only if the next step in $\mathbf{X}$ process is independent on history of $\mathbf{Y}$ up to $t-l+1$, i.e. independent of block of random variables $Y_t^{(l)}$.

For numerical evaluation of transfer entropy, we need to plug definition of conditional entropy into 4.4 so we get

$$T^{(m,l)}_{\mathbf{Y}\mapsto\mathbf{X}}(t) = -\sum_{x_{t+1},x_t^{(m)}} p(x_{t+1}, x_t^{(m)}) \log_2 p(x_{t+1}|x_t^{(m)}) + \sum_{x_{t+1},x_t^{(m)},y_t^{(l)}} p(x_{t+1}, x_t^{(m)}, y_t^{(l)}) \log_2 p(x_{t+1}|x_t^{(m)}, y_t^{(l)})$$

and using property of joint distribution $p(x_{t+1}, x_t^{(m)}) = \sum_{y_t^{(l)}} p(x_{t+1}, x_t^{(m)}, y_t^{(l)})$ we arrive to final explicit formula for transfer entropy

$$T^{(m,l)}_{\mathbf{Y}\mapsto\mathbf{X}}(t) = \sum p(x_{t+1}, x_t^{(m)}, y_t^{(l)}) \log_2 \frac{p(x_{t+1}|x_t^{(m)}, y_t^{(l)})}{p(x_{t+1}|x_t^{(m)})}, \tag{4.5}$$

where sum is taken over all possible outcomes of $(x_{t+1}, x_t^{(m)}, y_t^{(l)})$. In fact, in the form 4.5 Schreiber [17] originally defined Transfer entropy as a conditional Kullback divergence, i.e. deviation of $p(x_{t+1}|x_t^{(m)})$ from the generalized Markov property $p(x_{t+1}|x_t^{(m)}, y_t^{(l)})$ averaged over all possible realizations of $(x_t^{(m)}, y_t^{(l)})$.

It is convenient to state even in words what transfer entropy means.

$T^{(m,l)}_{\mathbf{Y}\mapsto\mathbf{X}}(t)$ = Uncertainty about the next step in $\mathbf{X}$ knowing the history of $\mathbf{X}$
− Uncertainty about the next step in $\mathbf{X}$ knowing the history of $\mathbf{X}$ and $\mathbf{Y}$

By using definition of conditional mutual information 1.13 [1] we may rewrite 4.4 in the form of flow of information from process $\mathbf{Y}$ to $\mathbf{X}$

$$\begin{aligned} T^{(m,l)}_{\mathbf{Y}\mapsto\mathbf{X}} &= I(X_{t+1}; Y_t^{(l)}|X_t^{(m)}) \\ &= I(X_{t+1}; X_t^{(m)}, Y_t^{(l)}) - I(X_{t+1}; X_t^{(m)}), \end{aligned} \tag{4.6}$$

where in the second equality we used chain rule 1.14 and symmetry of both mutual information and conditional mutual information.

[1] Note that conditioning is independent on order $H(X|Y,Z) = H(X|Z,Y)$.

The causality or directionality of transfer entropy is provided by non-symmetry property of conditional mutual information [2]. So we can measure flow from $\mathbf{Y}$ to $\mathbf{X}$ and vice versa, and according to sign of difference between these two flows, we may conclude which of them is superior and which of them is subordinate in sense of information production which actually means cause/effect detection.

We see that transfer entropy depends on two parameters (m and l). These parameters should correspond to order of Markov process, i.e. $\mathbf{X}$ and $\mathbf{Y}$ should be Markov process with order m and l respectively. The advantage of Markov process is that we can calculate the genuine transfer entropy that is only burdened with statistical error while for non-Markov process we should take all history in both series to obtain true value of transfer entropy, but that is in practice impossible. Consequently, by taking only limited history in target series we may erroneously regard information from the rest of the history of target series as incoming from source series, see equation 4.6. Thus generally speaking, low m overestimates transfer entropy while low l, i.e. short history in source series, underestimates transfer entropy. To avoid spurious information flow from target series, it is common to set $l = 1$ and m as large as possible.

Interpretation of actual number To get a better understanding of the actual value of transfer entropy, it is convenient to examine ratio of transfer entropy and conditional entropy, which Marchinski called **relative explanation added**

$$REA(m, l, t) = \frac{T^{(m,l)}_{\mathbf{Y}\mapsto\mathbf{X}}(t)}{h_{\mathbf{X}}(t, m)}. \tag{4.7}$$

This quantity tell us how many percent of information about the next step in process $\mathbf{X}$ can be gained from history of $\mathbf{Y}$ provided we already know history of $\mathbf{X}$.

Stationarity assumption As for now we have seen that there is no problem with generalization of transfer entropy to time dependent quantity. The reason for stationarity assumption, mentioned in almost all papers dealing with transfer entropy, arises in practical application since we have to somehow obtain the probability distribution in equation 4.5. This is done by observing one long realization of process and next computing the relative frequencies, hence, the processes in consideration should be strictly speaking also ergodic.

4.2 Rényian transfer entropy

Generalization of Shannonian transfer entropy for Renyi entropy may be done according to information representation of transfer entropy, i.e., equations 4.4

[2] $I(X;Y|Z) = I(Y;X|Z)$ but $I(X;Y|Z) \neq I(X;Z|Y)$

and 4.6, see [7],

$$\begin{aligned} T^{(m,l)}_{q;\mathbf{Y}\mapsto\mathbf{X}}(t) &= H_q(X_{t+1}|X_t^{(m)}) - H_q(X_{t+1}|Y_t^{(l)}, X_t^{(m)}) \\ &= I_q(X_{t+1}; X_t^{(m)}, Y_t^{(l)}) - I_q(X_{t+1}; X_t^{(m)}). \end{aligned} \tag{4.8}$$

Using definition of conditional entropy in equation 2.5 and mutual information in form 2.11, we can rewrite aforementioned equation to

$$\begin{aligned} T^{(m,l)}_{q;\mathbf{Y}\mapsto\mathbf{X}}(t) &= \frac{1}{1-q}\log_2 \frac{\sum \rho_q(x_t^{(m)}) p^q(x_{t+1}|x_t^{(m)})}{\sum \rho_q(x_t^{(m)}, y_t^{(l)}) p^q(x_{t+1}|x_t^{(m)}, y_t^{(l)})} \\ &= \frac{1}{1-q}\log_2 \frac{\sum \rho_q(x_t^{(m)}) p^q(y_t^{(l)}|x_t^{(m)})}{\sum \rho_q(x_{t+1}, x_t^{(m)}) p^q(y_t^{(l)}|x_{t+1}, x_t^{(m)})}. \end{aligned} \tag{4.9}$$

As we know from chapter 2, more definition of mutual information and conditional entropy exists, hence, different generalization of Shannonian transfer entropy may be received. Our choice is motivated by attractive properties of Rényian transfer entropy defined by 4.9 for financial time series.

Namely, it can be interpreted as a rating factor which quantifies a gain/loss in the risk concerning the behavior of the next step in X after we take into account the historical values of a time series Y. The positive value means decrease of risk, and negative value occurs when the knowledge of history in series Y broadens the tail part of distribution of the next step in X more than does only knowledge of history in X. This perception flows from already mentioned properties of mutual information defined in 2.11.

4.3 Simulated data

In this section we mention basic features of plug-in estimator of transfer entropy used in later analysis. To test the estimator, we simulated simple linear coupling

$$X(t) = r(t) + \epsilon Y(t-1), \tag{4.10}$$

$$Y(t) = s(t), \tag{4.11}$$

where $r(t)$ and $s(t)$ are two uncorrelated white noise processes, i.e. its distribution is $N(0,1)$.

4.3.1 Shannonian flow

Firstly, we derive analytical solution for Shannonian transfer entropy using $S = 3$ bins in coarse graining of both continuous time series X and Y. The decision for only three bins is due to lack of real data used in later analysis and the fact that for more bins one needs huge amount of data for reasonable results. On the other hand, three bins is the minimum that can incorporate non-linear dependency.

We use the same notation as above, i.e., $x_t^{(m)} = (x_t, \dots, x_{t-m+1})$ and due to stationarity we get $x_t^{(m)} \stackrel{d}{=} (x_m, \dots, x_1)$, and from now on we will omit redundant time subscript. In our analysis we use $l = 1$ as it is common practice when limited amount of data are available, see [18], hence we can write y_0 instead of $y^{(l)}$. Then transfer entropy 4.5 may be written as

$$T_{\mathbf{Y}\mapsto\mathbf{X}}^{(m,1)} = \sum_{x_{m+1}} \sum_{x^{(m)}} \sum_{y_0} p(x_{m+1}, x^{(m)}, y_0) \log_2 \frac{p(x_{m+1}, x^{(m)}, y_0) p(x^{(m)})}{p(x^{(m)}, y_0) p(x_{m+1}, x^{(m)})}. \quad (4.12)$$

From equation 4.10 and 4.11 we see that successive values of X process are mutually independent and identically distributed, this is clearly valid also for Y and its distribution is $N(0, 1+\epsilon^2)$ and $N(0,1)$ respectively. Due to independence of $x_t^{(m)}$ on y_0 we may rewrite joint probabilities in equation 4.12

$$p(x_{m+1}, x^{(m)}, y_0) = p(x_{m+1}, y_0) p(x^{(m)}), \quad (4.13)$$

$$p(x^{(m)}, y_0) = p(y_0) p(x^{(m)}), \quad (4.14)$$

$$p(x_{m+1}, x^{(m)}) = p(x_{m+1}) p(x^{(m)}). \quad (4.15)$$

After inserting equations 4.13, 4.14 and 4.15 into 4.12 and simplification of the fraction, we arrive at

$$T_{\mathbf{Y}\mapsto\mathbf{X}}^{(m,1)} = \sum_{x_{m+1}} \sum_{x^{(m)}} \sum_{y_0} p(x_{m+1}, y_0) p(x^{(m)}) \log_2 \frac{p(x_{m+1}, y_0)}{p(y_0) p(x_{m+1})}. \quad (4.16)$$

Next, we can sum over all m-tuples $x^{(m)}$, and owing to identical distribution of both X and Y, we can use just x and y instead of x_{m+1} and y_0. Finally, we get transfer entropy in the form

$$T_{\mathbf{Y}\mapsto\mathbf{X}}^{(m,1)} = \sum_{x,y=1}^{S} p(x,y) \log_2 \frac{p(x,y)}{p(y)p(x)}. \quad (4.17)$$

We see that in this special case transfer entropy does not depend on parameter m, and all variations are caused only by chosen partitioning. In what follows we derive transfer entropy for equiprobable bins and discretization obtained by standard deviation.

We will need joint probability density function to calculate probabilities occurring in equation 4.17. The density may be written in the form

$$\rho(x,y) = \frac{1}{2\pi} \exp\left\{ -\frac{(x-\epsilon y)^2 - y^2}{2} \right\}, \quad (4.18)$$

which follows from definitional equations 4.10 and 4.11 [3]. Then, necessary probabilities are obtained by integrating over appropriate limits depending on chosen partition.

[3] Note that x and y represents $X(t+1)$ and $Y(t)$ respectively.

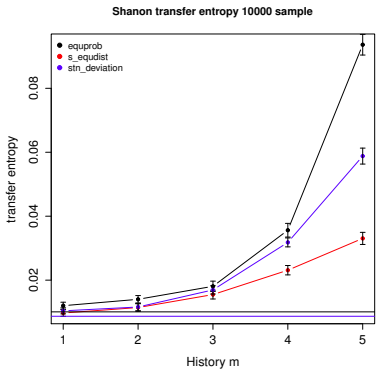


Figure 4.1: Transfer entropy

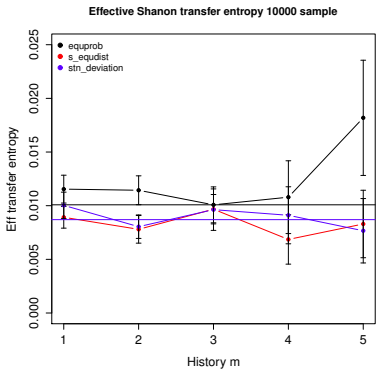


Figure 4.2: Effective transfer entropy

Equiprobable partition For equiprobable partition, i.e. $p(x) = p(y) = \frac{1}{S}$ for all bins, we get after simple manipulation

$$2\log_2 S + \sum_{x,y=1}^{S} p(x,y)\log_2 p(x,y). \tag{4.19}$$

Standard deviation partition Partitioning to three bins distinguishing between high drop, high rise and slight change, where high drop is considered decrease of more than one standard deviation and similarly the high rise, results in

$$\sum_{x,y=1}^{3} p(x,y)\log_2 p(x,y) - \sum_{x=1}^{3} p(x)\log_2 p(x) - \sum_{y=1}^{3} p(y)\log_2 p(y). \tag{4.20}$$

Numerical evaluation of equations 4.19 and 4.20 gives theoretical values 0.010 bits and 0.009 bits respectively.

Now we want to examine convergence of our estimator for these two partitions and estimate its standard errors. For this purpose, we generated time series X and Y of length 10000 data points and calculated transfer entropy from X to Y as a function of history length m for both partitioning. The results are depicted in figure 4.1 along with straight lines denoting precise theoretical values and errorbars obtained by bootstrap method when we set bootstrap sample length only 20 because of huge computation time demand, for a short introduction to Bootstrap method see Appendix A.

From figure 4.1 we can see that transfer entropy increases with m, but in our example whatever partitioning we use it should remain constant for all m. This spurious increase is caused by finite sample effect and is much more emphasized for larger alphabet, i.e. more number of bins. In [18] the same example was

History m	equprob	equdist	stn_deviation
1	0.0027	0.0016	0.0023
2	0.0032	0.0029	0.0042
3	0.0060	0.0054	0.0046

Table 4.1: Standard errors

studied as a function of sample length, and it was shown that transfer entropy approaches its theoretical value very slowly. Therefore, Marchinsky introduced **Effective transfer entropy**

$$T^{Eff}_{Y\to X} = T_{Y\to X} - T_{Y_{sh}\to X}, \tag{4.21}$$

where Y_{sh} means that original time series Y was shuffled, and hence all possible correlation between X and Y vanished. Thus, no information flow should be detected. However, numerical calculation shows increase with m of transfer entropy from Y_{sh} to X similar to one observed in case of transfer entropy from Y to X. Marchinsky then assigned $T_{Y_{sh}\to X}$ to finite sample effect and suggested to use Effective transfer entropy 4.21 instead of 4.5.

Estimation of Effective entropy is depicted in figure 4.2. Though it is clear that Effective entropy is much closer to theoretical values for both partitions than transfer entropy estimation (notice different scale on y axis) it still considerably fluctuates for different values of m even for relatively large sample size used $N = 10000$, see also [18] where comparison between transfer entropy and Effective transfer entropy was done for sample size up to 60000.

Due to statistical fluctuation we would like to pick up such a partition that is the most robust with respect to finite sample effects. For this reason, we estimated transfer entropy and Effective transfer entropy even for one other partition which have drawback that its theoretical value cannot be calculated, and thus its consistency is not justified by simulation. Nevertheless, after experience with relative consistency for equiprobable and standard-deviation coarse graining we assume that our transfer entropy estimator should be consistent for any partitioning.

The extra partition mentioned above is equidistant one, i.e. it divides range of time series to S equidistant bins (in our example $S = 3$). In figure 4.2 can be seen that this partition is rather stable with respect to m and therefore should be used in later application. In order to more advocate the choice of equidistant partition, we performed the same calculation as above for $N = 2500$ which is the minimum length of time series that we analyzed. This calculation showed that equidistant partition had the lowest standard error see table 4.1 for small $m \in \{1, 2\}$.

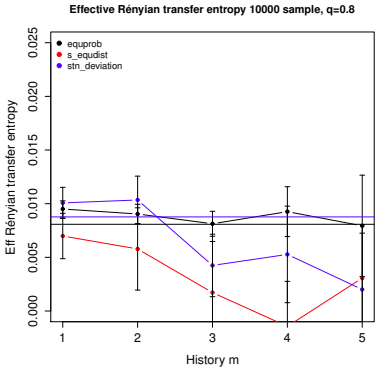


Figure 4.3: Rényian effective transfer entropy, $q = 0.8$

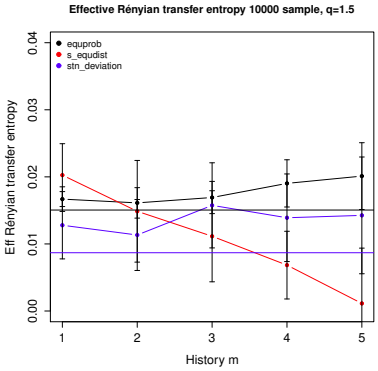


Figure 4.4: Rényian effective transfer entropy, $q = 1.5$

4.3.2 Rényian flow

The same linear coupling was analyzed even with help of Rényian transfer entropy. Here, instead of struggling with explicit formula 4.9 that is suitable for unknown systems, we profit from symbolic representation of transfer entropy, equation 4.8. Using the same argumentation about independence as used in Shannonian case we get

$$\begin{aligned} T^{(m,l)}_{q;\mathbf{Y}\mapsto\mathbf{X}}(t) &= H_q(X_{t+1}) + H_q(Y_t) - H_q(X_{t+1}, Y_t) \\ &= \frac{1}{1-q}\left(\log_2 \sum_y p^q(y) + \log_2 \sum_x p^q(x) - \log_2 \sum_{x,y} p^q(x,y)\right). \end{aligned} \tag{4.22}$$

We can use already obtained probabilities and get results for equprobable bins and standard deviation partitioning for two different values of parameter $q = 1.5$ and $q = 0.8$. Consequently the same simulation was performed, but now equidistant partitioning does not look as the best option, and it seems that equprobable partitioning is the most suitable one, see figures 4.3 and 4.4.

In order to compare Shannonian and Rényian transfer entropy, it is convenient to calculate them with the same partitioning, and from preceding simulation example we see that it is impossible to determine one universal partitioning that would fit to all cases. Hence, more careful analysis is necessary to get plausible results especially with real data as we will see in the next section.

Symbolic representation of time series The problem how to make a proper discretion of some system is dealt in mathematical branch called **Symbolic dynamics**. Generally, symbolic dynamics deals with problem how to assign symbols to continuous variable, i.e. discretization, in such a way that

new symbolic variable would contain as much information about original one as possible. In the case of time series, we get new series of symbols, and we then examine this discretized version and want to infer some statistical properties of original one. In fact, time series are usually already discretized in time so we can say that every time series analysis use some kind of symbolic dynamics approach even though it may not be apparent. Unfortunately, no general rule exist, and thus in practice we have to find "quasi-optimal" discretization with help of trial and error.

4.4 Real markets analysis

We have obtained minute data of 11 biggest stock exchanges in period from 1st July 2012 to 1st October 2012. Before any numerical analysis we have to appropriately prepare the data. That is done by excluding any no-trading periods (holidays, nighttime) in both series. After this we obtain different time series where time axis become so-called *trading time*. The drawback is that separated points in original time series may become close neighbors in new time series. Nevertheless small number of such points precludes statistical significant errors.

Due to different time zones and trading hours of particular stock exchanges, it is impossible to exploit minute data to measure information flow between Asia and the other continents. Unfortunately, later analysis showed non-stationarity which spoiled also possibility to measure interrelatedness between Europe and USA. Hence, we measure information flow only within continents.

We would like to analyze information flow in whole period, however, first look at data reveals non stationarity of time series, see figure B.10 in appendix B depicting variances calculated in individual blocks along with its error bars and with variance taken from the whole series of London index AIM100. Similar behavior was observed for the other indices. Note that we transformed series of minute closure prices s_n to log-returns

$$X_i = \log s_i - \log s_{i-1}$$

before analysis, and this new series still preserves non-stationarity. Since our basic analysis assumes stationarity of time series, we had to select only part of data where all indices in particular continent have at least approximately the same mean and variance within their blocks.

In the case of Europe we analyzed three indices, namely AIM100 of London stock exchange, DAX and EURO STOXX 50 which is composed of 50 largest stocks in Eurozone and should represents summary for all Europe. Due to stationary issue we had to select only data in time period from $23rd$ of August to $7th$ of September in which all three idices may be considered stationary. Only

two weeks may seem rather short, but acquisition at high frequency assures sufficient amount of data $\simeq 5000$.

In America the biggest stock indices were selected DJI - Dow Jones Industrial Average, NYA - New York Stock Exchange and CCMP - NASDAQ Composite Index. These indexes appeared to be approximately stationary at the end of our examined period, and we could pick up larger data set composed of $\simeq 6000$ data points from 7th to 29th of August.

Many big stock exchanges are situated in the east coast of China and in Japan. Five indices were available, namely, Shenzhen, Korea, Hong Kong, Shanghai and Tokyo. Unfortunately, Asia stock exchanges close around lunch time for one and half hour, and moreover, there is different time zone in China and Japan and thus after filtration little data have left. It was impossible to find common time period in which all fife Asian idices had been simultaneously stationary. Hence, we had to restrict ourselves to only Shanghai Stock Exchange Composite Index, Korea Stock Exchange KOSPI 200 Index and HSI - The Hang Seng Index Hong Kong in time period from 13th of August to 6th of September. This time period gave us over 3000 data points.

4.4.1 Choice of parameters

Due to non-stationarity we have quite small amount of "clean" data therefore actual values of transfer entropy are subject to huge statistical errors. Thus it is difficult to compare flows from different stock exchanges. The errors are more enhanced for larger m and l, and we have to trade off between statistical errors and bias caused by underestimation of history parameter m.

Effective transfer entropy has higher errors (twice the error of transfer entropy) because it is sum of two transfer entropies. Moreover, for small m the correction of transfer entropy is not very significant, and for this reason we decided to use only transfer entropy and parameters $m = l = 1$ as had been done in [19] where similar amount of data had been analyzed.

This approach leads to slight overestimation of actual values but allows significant comparing between both directions and various indexes that is crucial in our case since we analyze flow only inside continents, and these systems are rather close to equilibrium, thus, small flows appear. Hence, we should interpret calculated results more in qualitative sense, for instance detecting major (leading) stock exchange, than quantitative description like a precise number of bits that flows from one series to another.

After careful examination of errors for three mentioned partitioning for all indexes we decided to use equiprobable partitioning that was used also in [18].

4.4.2 Numerical results

Numerical results are presented in appendix B. q parameter of transfer entropy selects the part of distribution in which we are interested, so we see that more information is exchanged in central part of distribution $q = 1.5$ than tail part $q = 0.8$. We can also see that the highest information flow is in Asia followed by America, and in Europe the analyzed stock exchanges are only slightly coupled. (Note that different scales are used in heat maps.)

In order to determine major stock exchange in each continent, we calculate **net flow**

$$T_{net} = T_{\mathbf{Y}\mapsto\mathbf{X}} - T_{\mathbf{X}\mapsto\mathbf{Y}}.$$

According to sign of this quantity, we find out which of these two series produces more information which is a typical feature of leading markets. For conclusion see figures B.11, B.12 and B.13 in appendix B where only significant differences are depicted, i.e. $|T_{net}| > error_{\mathbf{X}\mapsto\mathbf{Y}} + error_{\mathbf{Y}\mapsto\mathbf{X}}$.

Note that net flow characterizes the causality since if there is non-zero flow in each directions it may correspond to the existence of common driving process, but that would imply that the information flow should be the same in both directions. Therefore, if the flow is asymmetric, i.e. $|T_{net}|$, then one series may be the cause of the other.

4.4.3 Time dependent information flow

For DAX and SX5E where the biggest amount of data are available, we calculate information flow as a function of time. The whole series is divided into 12 blocks corresponding roughly to one weak. In each block data are considered stationary, and transfer entropy is calculated.

We can see from figure 4.6 that information flow from SX5E to DAX indeed changes during examined period. Nevertheless, we cannot say that it changes in the form depicted in the figure because errors of our estimator overlaps for successive weeks. All we can say is that there is significantly higher influence at the beginning of examined period than around 5th week which is period of lower interrelatedness and is followed by another time period of stronger correlation around 8th week. At the end of examined period we see restoration of previous lower connected state.

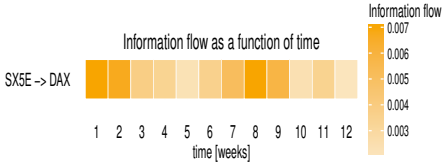


Figure 4.5: Heat map of Shannonnian information flow from Europian index SX5E to DAX as a function of time.

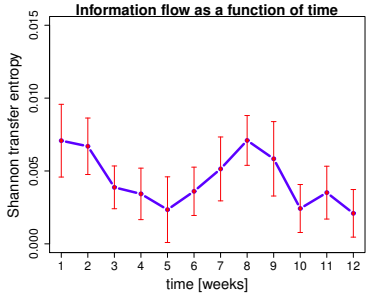


Figure 4.6: Shannonnian information flow from Europian index SX5E to DAX as a function of time.

Chapter 5

Superstatistics

In this chapter we introduce Superstatistics which is a concept devised by Beck for systems with fluctuating intensive parameter, e.g. temperature, which are therefore in non-equilibrium state. The idea first appeared in the paper [21], and the term Superstatistics was coined later in the successive paper [22].

The assumption is that the system is in non-equilibrium steady state and is composed of many cells which are locally in equilibrium but with different values of intensive parameter, e.g. temperature. This intensive parameter in each cell changes on a long time scale T much larger than relaxation time of the cell.

5.1 Illustrative example

In the original paper [21] Beck's aim was to provide a reason for fruitful applications of Tsallis statistics, which describes non-extensive systems, on a microscopic level, i.e. to give a dynamical explanation for Tsallis statistic. He succeeded in his goal for systems with fluctuating intensive parameter (in [21] for statistical validation energy dissipation rate was used as an intensive parameter).

The reasoning is as follows. Consider Brownian particle whose velocity u is a solution of Langevin equation

$$m\frac{\mathrm{d}u}{\mathrm{d}t}(t) = -\gamma u(t) + L(t), \tag{5.1}$$

where $\gamma > 0$ is a friction constant, m is a mass of the Brownian particle and $L(t)$ is White noise, i.e. Gaussian process with correlation function $E[L(t)L(t')] = \sigma^2\delta(t-t')$ and $E[L(t)] = 0$. σ is a strength of the random force and from microscopical point of view is the property of tiny particles composing the surroundings, e. g. liquid, of Brownian particle.

Solution of the equation 5.1 can be found in [23], and the result is

$$u(t) = u_0 \exp\left(-\frac{\gamma t}{m}\right) + \frac{1}{m}\int_0^t L(t') \exp\left(-\frac{\gamma(t-t')}{m}\right) \mathrm{d}t' \tag{5.2}$$

where the integral of White noise need to be interpreted as an Ito integral which is motivated by formal definition of White noise as a derivative of Brownian motion[1], see [24].

$$Y(t) = \frac{1}{m}\int_0^t L(t') \exp\left(-\frac{\gamma(t-t')}{m}\right) \mathrm{d}t' = \frac{\sigma}{m}\int_0^t \exp\left(-\frac{\gamma(t-t')}{m}\right) \mathrm{d}B(t')$$

It is known, see e.g. [24], that Ito integral of non-random function $f(t,t')$ is a Gaussian stochastic process with zero expectation value and covariance function given by

$$Cov(Y(t), Y(t+h)) = \int_0^t f(t,t') f(t+h,t') \mathrm{d}t' \tag{5.3}$$

which in our case for

$$f(t,t') = \frac{\sigma}{m} \exp\left(-\frac{\gamma(t-t')}{m}\right)$$

gives after short calculation

$$Cov(Y(t), Y(t+h)) = \frac{\sigma^2}{2m\gamma}\left(1 - \exp\left(-\frac{2\gamma t}{m}\right)\right) e^{-\frac{\gamma}{m}h} \tag{5.4}$$

Thus, probability distribution of $u(t)$ is $N(\mu, D^2)$, where

$$\mu = u_0 \exp\left(-\frac{\gamma t}{m}\right),$$

$$D^2 = \frac{\sigma^2}{2m\gamma}\left(1 - \exp\left(-\frac{2\gamma t}{m}\right)\right)$$

We see that on the time scale larger than $\tau = \frac{m}{\gamma}$, dependent on parameters m and γ, we may consider $e^{-\frac{2\gamma t}{m}} \approx 0$ and the process $u(t)$ may be regarded as strictly stationary because for Gaussian processes weak stationarity[2] implies strict stationarity. This stationary solution corresponds to equilibrium state, and τ can be identified with relaxation time needed by the system to reach equilibrium.

[1] Vaguely $L(t') = \frac{\mathrm{d}B(t')}{\mathrm{d}t}$, therefore $L(t')\mathrm{d}t' = \mathrm{d}B(t')$.

[2] Recall that weak stationarity means that a mean function $\mu(t) = E[u(t)]$ is independent on time, and covariance function is a function only of time lag $cov(t,t') = cov(t-t')$.

With the help of Equipartition theorem

$$\frac{1}{2}mE[u^2] = \frac{1}{2}k\Theta$$

we may relate variance of the velocity of the Brownian particle to temperature Θ of the heat bath with which the system is in equilibrium[3] and consequently introduce inverse temperature

$$\beta = \frac{1}{k\Theta} = \frac{2\gamma}{\sigma^2}.$$

The probability distribution for u is then

$$p(u|\beta) = \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{1}{2}\beta m u^2\right), \tag{5.5}$$

which we write as a conditional probability distribution on purpose since the idea of Superstatistics is to elevate the parameter β to a random variable. Henceforth we set $m = 1$ in order to get the same result as Beck.

As the next step we imagine either lots of Brownian particles in separated environments with inverse temperature distribution $f(\beta)$ from which data about their velocities are collecting for time period T. Or equivalently one Brownian particle in a changing environment when the inverse temperature β takes different value, taken as a realization of a random variable with probability distribution $f(\beta)$, in each non-overlapping time window of length T. In this case the data are collecting over much longer period of time. The overall effect is that we obtain data from mixture of Gaussian distributions with different variance.

It is worth noting that we assume the data acquisition to occur in discrete time as is often the case in practice. For instance, we may perform n measurements of each particle's velocity during time T which means the time elapsed between consecutive measurements is T/n. This time must be sufficiently longer than relaxation time in order to ensure that the particle has reached equilibrium when the second method for data acquisition is used. In either cases the result is that we end up with N times n values for velocities, where N is number of examined Brownian particles or number of time periods T during which one particular Brownian particle was observed.

Distribution of velocities acquired in such a way is given by continuous mixing of 5.5

$$p(u) = \int p(u|\beta)f(\beta)\mathrm{d}\beta. \tag{5.6}$$

[3]Note that equilibrium is reached after $t > \tau$ when average velocity is consider to be zero, therefore $E[u^2] = Var[u]$.

Beck pointed out in [21] that a particular choice[4] for probability distribution $f(\beta)$ of intensive parameter

$$f(\beta) = \frac{1}{\Gamma(\frac{n}{2})} \left(\frac{n}{2\beta_0}\right)^{\frac{n}{2}} \beta^{\frac{n}{2}-1} \exp\left(-\frac{n\beta}{2\beta_0}\right), \qquad \beta > 0 \tag{5.7}$$

leads to Tsallis statistics. It is easily seen since integration[5] in 5.6 gives

$$p(u) = \frac{\Gamma\left(\frac{n}{2}+\frac{1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(\frac{\beta_0}{\pi n}\right)^{\frac{1}{2}} \frac{1}{\left(1+\frac{\beta_0}{n}u^2\right)^{\frac{n+1}{2}}}, \tag{5.8}$$

and after identification

$$q = \frac{n+3}{n+1}, \qquad \tilde{\beta} = \frac{\beta_0(n+1)}{n}, \tag{5.9}$$

we end up with Tsallis distribution 5.11 derived in [25] also called *q-Gaussian distribution*.

Short clarification is in order at this point. Tsallis entropy would lead to distribution of energy, but when we identify temperature according to Equipartition theorem in local equilibrium we actually consider the Brownian particle as a free particle. Hence the energy which would lead to distribution 5.8 from maximum entropy principle is $\frac{1}{2}mu^2$.

5.2 Generalized Boltzmann factor

This rather surprising result that a statistical mixing of systems with varying Gamma-distributed intensive parameter leads to probability distribution which may also be obtained by using maximum entropy principle with respect to Tsallis non-extensive entropy and canonical ensemble, i.e. averaged energy constraint, motivated Beck to a bold proposal for generalizing standard Boltzmann-Gibbs statistics. For such generalization Beck suggested name ***Superstatistics***, which emphases the fact that it is statistics of statistics as will be seen later, and it includes both Tsallis and Boltzmann-Gibbs as a special cases.

The above mentioned example with particle in changing environment is not, of course, the only driving force which motivated Beck to introduce new statistics. The major compulsion comes from experiments. There have been never ending discussions about foundations of Boltzmann-Gibbs entropy and its universal applicability, see e.g. [26], and indeed it has been shown that some systems, in particular non-extensive systems, are not well described by Boltzmann

[4] Gamma probability distribution but referred to as χ^2 distribution in Beck's papers.

[5] The only step in integration is to recall definition of Γ function $\Gamma(n) = \int_0^{+\infty} x^{n-1}e^{-x}\mathrm{d}x$.

distribution $p_i \propto e^{-\beta E_i}$. Therefore Tsallis in [25] defined his non-extensive entropy

$$S_q = k\frac{1-\sum_i p_i^q}{q-1} \tag{5.10}$$

which after extremalization procedure subjected to constraint (the same result is obtained if the standard average is used with $q' = 2 - q$, see [27] for more details)

$$\sum_{i=1}^{W} p_i^q E_i = U$$

gives power-law probability distribution on states

$$p_i \propto \frac{1}{\left(1+\tilde{\beta}(q-1)E_i\right)^{\frac{1}{q-1}}}. \tag{5.11}$$

Although this one-parametrized generalization, which for $q \to 1$ recovers Boltzmann distribution [6], correctly characterizes many systems which violate Boltzmann distribution, there are systems for which even Tsallis distribution seems inadequate as highlighted in [22]. Therefore another generalization is needed.

Generalized Boltzmann factor or **effective Boltzmann factor** is defined in [22] as

$$B(E) = \int_0^{+\infty} f(\beta)e^{-\beta E}\mathrm{d}\beta, \tag{5.12}$$

where now β can be any intensive parameter not necessarily temperature. The definition is motivated by the example with Brownian particle discussed earlier where Hamiltonian was that of free particle $\frac{1}{2}u^2$ ($m = 1$) and $f(\beta)$ Gamma distribution. Here comes the name Superstatistics as explained in [22] since $B(E)$ is obtained as a statistics of a statistics (averaging Boltzmann statistics $e^{-\beta E}$ with respect to $f(\beta)$ - statistics of β).

One may think that any distribution $f(\beta)$ would be possible. However, some reasonable constrains need to be put on conceivable distributions in order to get physically relevant statistics, i.e. distribution on phase space.

- The new statistics must be normalizable, i.e. $\int_0^{+\infty} \rho(E)B(E)\mathrm{d}E < +\infty$, where $\rho(E)$ is density of states.
- The standard Boltzmann factor should be recovered when β becomes constant, i.e. no fluctuation of intensive parameter.

[6] $\lim_{q\to 1}\left(1+\tilde{\beta}(q-1)E_i\right)^{\frac{1}{1-q}} = \exp\left(\lim_{q\to 1}\frac{\ln(1+\tilde{\beta}(q-1)E_i)}{1-q}\right) \overset{\mathrm{LH}}{=} \exp\left(\lim_{q\to 1}\frac{-\tilde{\beta}E_i}{1+\tilde{\beta}(q-1)E_i}\right) = e^{-\tilde{\beta}E_i}$

5.3 Simple examples of Superstatistics

When β is to relate to any physical intensive parameter it has to be positive, therefore $f(\beta)$ needs to be zero for negative β. The list of the most common Superstatistics and their corresponding generalized Boltzmann factor follows.

1. *uniform distribution* - Uniform on interval $[a, a+b]$

$$f(\beta) = \frac{1}{b},$$

$$\beta_0 = E[\beta] = a + \frac{b}{2}, \qquad \sigma^2 = Var[\beta] = \frac{b^2}{12},$$

$$B(E) = \frac{1}{bE}\left(e^{-(\beta_0 - 0.5b)E} - e^{-(\beta_0 + 0.5b)E}\right).$$

2. *2-level distribution* - two possible values a and $a+b$ with equal probability

$$f(\beta) = 0.5\delta(a) + 0.5\delta(a+b),$$

$$\beta_0 = E[\beta] = a + \frac{b}{2}, \qquad \sigma^2 = Var[\beta] = \frac{b^2}{4},$$

$$B(E) = \frac{e^{-\beta_0 E}}{2}\left(e^{0.5bE} + e^{-0.5bE}\right).$$

3. *Gamma distribution*[7] - $Gamma(a, b)$, $a = \frac{n}{2\beta_0}, b = \frac{n}{2}$

$$f(\beta) = \frac{1}{\Gamma(\frac{n}{2})}\left(\frac{n}{2\beta_0}\right)^{\frac{n}{2}} \beta^{\frac{n}{2}-1} \exp\left(-\frac{n\beta}{2\beta_0}\right),$$

$$\beta_0 = E[\beta], \qquad \sigma^2 = Var[\beta] = \frac{2\beta_0^2}{n},$$

$$B(E) = \left(1 + \frac{2\beta_0}{n}E\right)^{-\frac{n}{2}}.$$

4. *Log-normal distribution* -

$$f(\beta) = \frac{1}{\beta\sqrt{2\pi s^2}} \exp\left(-\frac{(\log\beta - \log\mu)^2}{2s^2}\right),$$

$$\beta_0 = E[\beta] = \mu e^{0.5s^2}, \qquad \sigma^2 = Var[\beta] = \mu^2 e^{s^2}(e^{s^2} - 1),$$

$B(E)$ does not have analytical expression.

[7] Beck often refers to χ^2 distribution which is strictly speaking only special case for $a = \frac{1}{2}$, $b = \frac{n}{2}$. Nevertheless, we will here, for better orientation in Beck's papers, regard χ^2 distribution as a synonym for Gamma distribution.

5. *Inverse-χ^2 distribution*[8] -

$$f(\beta) = \frac{1}{\Gamma(\frac{n}{2})} \left(\frac{b_0 n}{2}\right)^{\frac{n}{2}} \beta^{-\frac{n}{2}-1} \exp\left(-\frac{nb_0}{2\beta}\right),$$

$$\beta_0 = E[\beta] = \frac{nb_0}{n-2}, \qquad \sigma^2 = Var[\beta] = \frac{n^2 b_0^2}{(n-2)^2(\frac{n}{2}-2)}, \qquad n > 4$$

$B(E)$ does not have analytical expression.

Interesting remark is that for low energy E and small fluctuation of intensive parameter β all Superstatistics with finite variance $\sigma^2 = Var[\beta]$ are indistinguishable in a sense that they have the same first-order correction to ordinary Boltzmann factor.

This universal property may be easily proven by moment expansion which is just Taylor expansion about the mean.

$$E[g(X)] = E[g(\mu_X) + g'(\mu_X)(X - \mu_X) + \frac{1}{2}g''(\mu)(X - \mu_X)^2 + O(X^3)],$$

therefore

$$E[g(X)] \approx g(\mu_X) + \frac{1}{2}g''(\mu_X)Var[X]. \tag{5.13}$$

Since $B(E)$ is expectation value of $e^{-\beta E}$, we apply 5.13 for $g(x) = e^{-x}$ and random variable $X = E\beta$.

$$B(E) \approx e^{-\beta_0 E} + \frac{1}{2}e^{-\beta_0 E}\sigma^2 E^2 = e^{-\beta_0 E}(1 + \frac{1}{2}\sigma^2 E^2) \tag{5.14}$$

valid for small fluctuations of β and low energy.

Due to this universal property maximum entropy principle was suggested for Superstatistics in [22]. Since extremalisation of Tsallis entropy leads to Gamma Superstatistics, we may by Taylor expansion in 5.11 identify entropic parameter q.

Note that Boltzmann factor is given without normalization, thus $B(E)$ is just Tsallis distribution without normalization constant (partition sum Z_q)

$$B(E) = \frac{1}{\left(1 + \tilde{\beta}(q-1)E\right)^{\frac{1}{q-1}}} = \exp\left(-\frac{1}{q-1}\ln\left(1 + \tilde{\beta}(q-1)E\right)\right),$$

and using expansion of logarithm[9]

$$B(E) \approx e^{-\tilde{\beta}E}\exp\left(\frac{1}{2}\tilde{\beta}^2(q-1)E^2\right) \approx e^{-\tilde{\beta}E}\left(1 + \frac{1}{2}\tilde{\beta}^2(q-1)E^2\right). \tag{5.15}$$

[8] The same note that was mentioned in connection with χ^2 and Gamma distribution applies here, i.e. strictly speaking this is called Inverse-Gamma distribution.

[9] $\log(1+x) = x - \frac{x^2}{2} + O(x^3), x \to 0$

Next $\beta_0 = \tilde{\beta}$ and after comparing with 5.14 we get

$$(q-1)\beta_0^2 = \sigma^2$$

and equivalently

$$q = \frac{E[\beta^2]}{E[\beta]^2}. \tag{5.16}$$

Maximum entropy principle for general Superstatistics $f(\beta)$ is then simply extremalization of Tsallis entropy with parameter q given by 5.16 and should yield approximately good predictions for low energies. Note that from 5.16 follows $q \geq 1$, i.e. entropic parameters $q < 1$ cannot be easily realized by Superstatistic approach.

5.3.1 Universality classes

After extensive application of Superstatistical models to various physical and nonphysical systems, three Superstatistics appeared to be highly successful. These are χ^2 Superstatistics, Log-normal Suprestatistics and Inverse χ^2 Superstatistics. Their respective outstanding agreement with experimental measurements are departure delay on British rail network data, velocity difference of a test particle in Lagrangian turbulence and cancer survival data, see [28], [29] and references therein for more details.

Inspiration and partial justifications for these universal Superstatistics may come from microscopic point of view as discussed in [28]. Let us consider three different ways how fluctuation in β may arise.

1. *Gamma distribution*
   Imagine large number of small independent random variables ξ_j. If properly normalized their sum is standard normal random variable

   $$X_i = \frac{\sum_j^J \xi_j - d_J}{c_J} \sim N(0,1).$$

   Since β needs to be positive, simple model may be $\beta \propto X_i^2$, and if there are n such independent random variables all contributing to fluctuation of β, we may write

   $$\beta = \frac{\beta_0}{n}\sum_{i=1}^{n} X_i^2 = \frac{1}{n}\sum_{i=1}^{n}(\sqrt{\beta_0}X_i)^2,$$

   i.e. β is equal to average contribution of degrees of freedom and proportionality constant β_0 may be interpreted as a variance of one degree of freedom $Y_i = \sqrt{\beta_0}X_i$.

Probability density function of β is easily found when we recall that sum of n squares of $N(0,1)$ random variables is χ^2 random variable with n degrees of freedom which is just the special case of Gamma distribution $Gamma(\frac{1}{2}, \frac{n}{2})$

$$f(y) = \frac{1}{\Gamma(\frac{n}{2})} \left(\frac{1}{2}\right)^{\frac{n}{2}} y^{\frac{n}{2}-1} e^{-\frac{x}{2}},$$

and after rescaling [10]

$$f(\beta) = \frac{1}{\Gamma(\frac{n}{2})} \left(\frac{n}{2\beta_0}\right)^{\frac{n}{2}} \beta^{\frac{n}{2}-1} \exp\left(-\frac{n\beta}{2\beta_0}\right),$$

we end up with $Gamma(\frac{n}{2\beta_0}, \frac{n}{2})$ distribution.

2. *Inverse χ^2 distribution*
Similarly as in case the of Gamma distribution the inverse Gamma distribution may emerge when

$$\beta = \beta_0 \frac{1}{\frac{1}{n}\sum_{i=1}^{n} X_i^2} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \left(\frac{X_i}{\sqrt{\beta_0}}\right)^2},$$

where now $\frac{1}{\beta_0}$ is variance of one degree of freedom.

We get probability density function using the definition of Inverse χ^2 distribution, i.e. if $Z \sim \chi^2$ then $Y = \frac{1}{Z} \sim Inv\text{-}\chi^2$ with density

$$f(y) = \frac{1}{\Gamma(\frac{n}{2})} \left(\frac{1}{2}\right)^{\frac{n}{2}} \left(\frac{1}{y}\right)^{\frac{n}{2}+1} e^{\frac{1}{2y}},$$

and rescaling gives

$$f(\beta) = \frac{1}{\Gamma(\frac{n}{2})} \left(\frac{\beta_0 n}{2}\right)^{\frac{n}{2}} \beta^{-\frac{n}{2}-1} e^{-\frac{n\beta_0}{2\beta}}$$

which is Inverse-Gamma distribution.

3. *Log-normal distribution*
Consider again small independent random variables ξ_j which are now assumed to be positive. If these variables contribute as a multiplicative precess $X_i = \prod_j^J \xi_j$, then for large J $\frac{\ln X_i - d_J}{c_J}$ is standard normal random variable, and by definition X_i is Log-normal random variable. β is now assumed in the form

$$\beta = \prod_{i=1}^{n} X_i.$$

[10] Note $Y \sim f_Y(y)$ then $Z = aY \sim f_Z(z) = f_Y(\frac{z}{a})\frac{1}{a}$.

Company		Sector	Stock exchange
AA	Alcola Inc.	basic materials	NYSE
KO	The Coca-Cola Company	consumer goods	NYSE
BAC	Bank of America Corporation	financial	NYSE
JNJ	Johnson & Johnson	healthcare	NYSE
GE	General Electric Company	industrial goods	NYSE
WMT	Wal-Mart Stories Inc.	services	NYSE
INTC	Intel Corporation	technology	NASDAQ

Table 5.1: List of companies used in analysis

Hence $\ln \beta$ is a sum of normal random variables and thus also normal random variable. Therefore β is Log-normal random variable with probability density [11]

$$f(\beta) = \frac{1}{\beta\sqrt{2\pi s^2}} \exp\left(-\frac{(\log\beta - \log\mu)^2}{2s^2}\right).$$

The above mentioned reasoning not only gives possible idea about microscopic dynamics responsible for fluctuation of intensive parameter but also makes fitting Superstatistical models much easier. Since by restricting to three possible functional forms of $f(\beta)$, we arrive in fact to parametric statistics which is much more tractable than non-parametric one. We will later take advantage of these universality classes.

5.4 Transition between Superstatistics

Superstatistics is definitely a great idea which has justifiable motivation, and furthermore, it has been successfully fitted on empirical data. Therefore, there is little doubt about usefulness of Superstatistics, however, new broader model was recently introduced by father of Superstatistics in [30]. It is claimed that transition of Superstatistics is possible when we look at the time series at different time scales. It is questionable if this new model is valid. Although this new idea was also tested in the same paper, the crude method used is not very convincing. Thus new method is needed.

In this section we use the same dataset as in [30], i.e. stock prices of seven companies from different sectors recorded on the minute-tick basis during period from the $2nd$ Jan, 1998, to $22nd$ May, 2013, see table 5.1, and propose finer quantitative method for testing transition of Superstatistics.

[11] Simply derived using $Y \sim f_Y(y)$ then $Z = F(Y) \sim f_Z(z) = f_Y(F^{-1}(z))|(F^{-1})'|(z)$ for F regular map.

5.4.1 Data preprocessing

When analyzing financial time series of stock prices $\{S_i\}$, the typical quantity worth of attention is series of so-called log-returns

$$r_i = \log \frac{S_{i+1}}{S_i}.$$

Hence, instead of $\{S_i\}$ we analyze $\{r_i\}$. r_i is an increment of logarithm of stock price during sampling time τ, in our case $\tau = 1$ minute, which corresponds to time scale at which we observe the system (stock price). It is important that the time scale is fixed, in other words the time series must be sampled equidistantly in time.

Deeper data examination shows various errors in data, which occurred probably during acquisition since some values are missing, and hence data are not spaced equidistantly in time. These errors were eliminated by selecting only increments occurred during one minute which also effectively removes problem with overnight jumps.

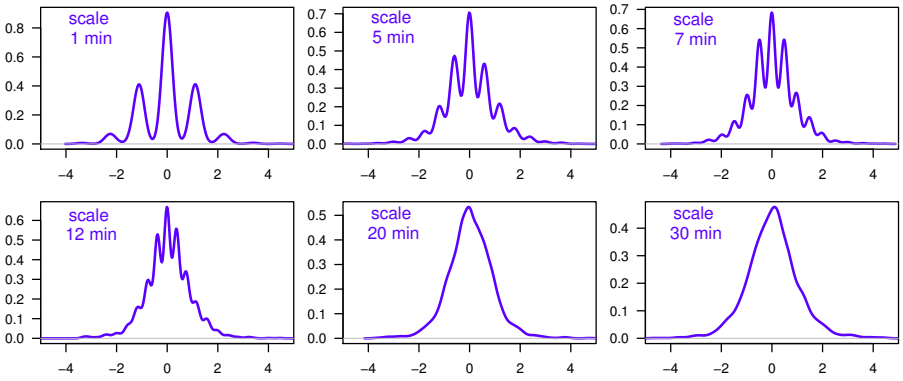


Figure 5.1: Artificial structure in log-returns time series, in particular company KO.

Another problem was occurrence of quite artificial structure likely due to very fine recording of data when liquidity of individual stocks, even of large companies used in analysis, is in question. The artificial structure is shown in figure 5.1 for company KO. The figure shows kernel density estimate of probability density for the first 2000 log-returns at various time scales. Note that 2000 data points are clearly enough in order that the discrete levels could be considered significant. This artificial structure does occur also at different time positions in the time series, i.e. not necessarily at the beginning, and all studied time series possess the same flaw.

It is worth noting that discovery of this structure happened rather incidentally since the probability density estimate of log-returns for the whole series

does not show any discreteness in log-returns. The explanation may be that the discrete levels are slightly shifted after some time period. From a physical point of view, we may say that for minute scale system has not yet reached equilibrium in the cell, i.e. assumption $\tau > \tau_r$, where τ_r is relaxation time of a cell, does not hold.

One feasible solutions is to aggregate data into higher time scale in which single stock becomes liquid and log-returns continuous up to recording precision. By successive aggregation we find out that as a convenient sampling interval may be considered 20 minutes as can be seen in figure 5.1 since artificial structure vanishes. This 20−minutes sampling is adequate also for time series of the other companies. So we consider new time series constructed as [12]

$$r_j^{(20)} = \sum_{i=1}^{20} r_{i+(j-1)20}^{(min)} \qquad j \in \{1, \ldots, \lfloor \frac{N}{20} \rfloor\}, \tag{5.17}$$

where N is the length of the original series sampled every minute. In what follows we for brevity suppress the superscript indicating scale of the time series and use only 20-min scale data unless stated otherwise. In paper [30] such adjustments were not done, and it may be the reason for slightly different conclusions.

5.4.2 Optimal window width

When we have reliable data then the next step is to determine time interval in which intensive parameter stays constant. We proceed with the same method as in [30], that is estimating kurtosis of log-returns for various length T of window, and then select optimal T_{op} for which average kurtosis $\bar{\kappa}_T$ over all windows is the closest to one of normal distribution, $\kappa = 3$. Biased so-called *moment estimator* for kurtosis was used

$$\hat{\kappa} = \frac{m_4}{m_2^2} = \frac{\frac{1}{T}\sum_{i=1}^{T}(r_i - \bar{r})^4}{\left(\frac{1}{T}\sum_{i=1}^{T}(r_i - \bar{r})^2\right)^2}, \tag{5.18}$$

where $\bar{r}$ is sample mean of log-returns in given window. It is standard estimator, and moreover, according to [31] has the lowest mean square error for normal sample.

Slight modification was made in favor of further analysis. Since we are eventually interested in distribution of variances, we prefer more data points for variance, and the accuracy of estimated variance in each block is secondary because the errors cancel each other out in density estimation. Therefore, we

[12] Note that log-returns r are additive when changing to higher time scale.

introduce a small threshold ϵ, and consider the optimal T as the lowest one for which average kurtosis $\bar{\kappa}_T$ over $n = \lfloor \frac{N}{T} \rfloor$ blocks satisfies

$$|\bar{\kappa}_T - 3| < \epsilon, \tag{5.19}$$

where the threshold is chosen $\epsilon = 0.1$. This ensures the longest possible series of sample variances without significant departure form $\kappa = 3$, see figure 5.2.

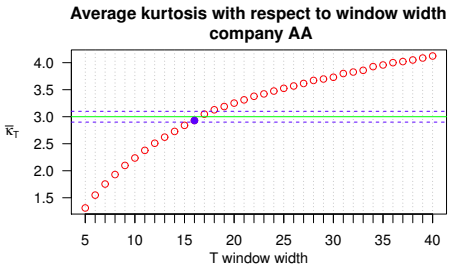


Figure 5.2: Finding optimal block width, $\epsilon = 0.1$.

For subsequent comparison of variance distribution on different time scales, it is convenient to normalize log-returns to zero mean and unit variance.

$$u_i = \frac{r_i - \bar{r}}{s}, \tag{5.20}$$

where $\bar{r}$ and s^2 is sample mean and sample variance of the whole series, respectively.

Having optimal window width and normalized log-returns, we estimate variance in each block with unbiased estimator

$$s_j^2 = \frac{1}{T_{op} - 1} \sum_{i=1}^{T_{op}} (u_{i+(j-1)T_{op}} - \mu_j)^2 \qquad j \in \{1, \ldots, \lfloor \frac{N}{T_{op}} \rfloor\}, \tag{5.21}$$

where N is the length of the new series sampled every 20 min, and μ_j denotes sample mean in a particular block

$$\mu_j = \frac{1}{T_{op}} \sum_{i=1}^{T_{op}} u_{i+(j-1)T_{op}}. \tag{5.22}$$

By this procedure we obtain $\lfloor \frac{N}{T_{op}} \rfloor$ values for variance or temperature in physical jargon, however, in Superstatistics we need β which is inverse temperature

$$\beta_j = \frac{1}{s_j^2}. \tag{5.23}$$

In the figure 5.3 the procedure for obtaining values for inverse temperature is depicted.

5.4.3 Selecting proper Superstatistics

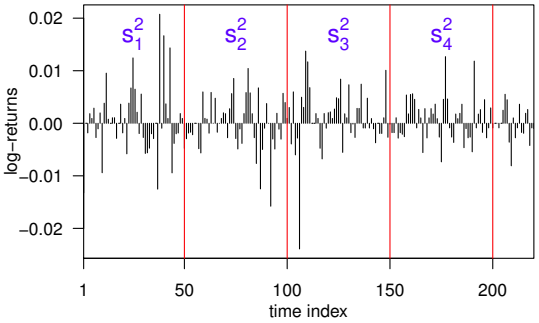


Figure 5.3: Illustrative figure for temperature estimation procedure

Next step would be to find probability distribution of β. This task is in general very difficult, nevertheless, in Superstatistic there are universality classes, and therefore we may greatly simplify our problem by restricting ourself to three two-parametric families of probability distributions. These are Gamma distribution, Log-normal distribution and Inverse Gamma distribution.

In case of parametric fitting we need to choose method for parameters estimation. There are many of them, e.g. minimum distance method, moment method or maximum likelihood method. We use maximum likelihood method so that it will not interfere with distance measures used for subsequent goodness of fit.

Once we have optimal parameters, we can ask which of the three considered distributions is the best fit for inverse temperatures β. Goodness of fit is usually measured by various distances between fully specified distribution function, i.e. all its parameters have to be given, and empirical distribution function. We use three distances, namely, Kolmogorov-Smirnov distance

$$D_n = \sup_x |F_n(x) - F(x)|, \tag{5.24}$$

Cramér-von Mises distance

$$C_n = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 \mathrm{d}F(x) \tag{5.25}$$

and Anderson–Darling distance

$$A_n = n \int_{-\infty}^{+\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x))} \mathrm{d}F(x), \tag{5.26}$$

where $F(x)$ is fully specified distribution function and $F_n(x)$ is empirical distribution function

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(u_i \leq x). \tag{5.27}$$

	Log-norm	**Gamma**	**Inv-Gamma**
Kolmogorov-Smirnov	0.032	0.050	0.117
Cramer-von Mises	1.26	3.04	22.06
Anderson-Darling	8.57	16.18	121.88

Table 5.2: Values for various distance measures for different probability distributions, company AA, scale 20 min.

Corresponding statistical tests for goodness of fit make use of these distances in order to test the hypothesis that data comes from given distribution $F(x)$. Therefore, one may feel the temptation to exploit one of the goodness of fit test. It is, however, important to note that due to high dependence in β, equivalently in volatility, as seen in the figure 5.4 of autocorrelation function, we cannot use standard procedure and test goodness of fit by these tests. Not mentioning complications which arise when parameters are estimated from data, as is our case, and proper p-value of the tests would have to be calculated by simulations.

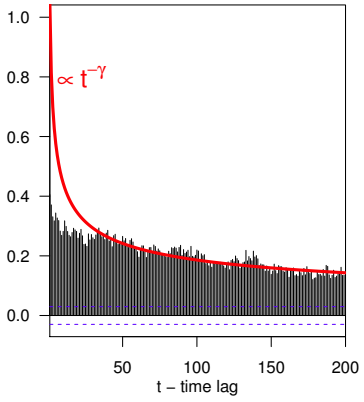


Figure 5.4: ACF of inverse temperature of AA company at scale 20 min, $\gamma = 0.377$.

Nevertheless, even for dependent data we consider distances as a convenient way to distinguish between the three probability distributions. Thus, we calculate distances for each company at a particular time scale and claim the distribution with the lowest distance as an optimal Superstatistics of β.

However, each distance has specific properties, e.g. Cramér-von Mises distance, eq. 5.25, measures quadratic errors between distribution function and empirical distribution function without any weighting factor so all discrepancies are equal while in the case of Anderson-Darling distance, eq. 5.26, squared errors are weighted by $\frac{1}{F(x)(1-F(x)}$ which gives higher importance to tail discrepancies. Therefore, there should be no wonder that different distances may point out different results. Yet in spite of distance characteristics later analysis shows that conclusions more or less coincide regardless of distance measures employed.

In the table 5.2 values for considered distance measures are shown for the three universality-classes probability distributions and dataset corresponding to company Alcola Inc. Values corresponds to time scale 20 minutes which is the smallest reliable scale. We can conclude from the table 5.2 that for small time scales the best Superstatistic for β from the three distributions taken into account is Log-normal distribution. This is in agreement with the claim

	Log-normal	Gamma	Inv-Gamma
Kolmogorov-Smirnov	0.077	0.045	0.133
Cramer-von Mises	0.153	0.032	0.770
Anderson-Darling	1.151	0.255	4.590

Table 5.3: Values for various distance measures for different probability distributions, company AA, scale 390 min, i.e. 1 trading day.

made in [30] where the statement was supported only by visual inspection of histogram and fitted distribution.

5.4.4 Addressing the transition

The main point of the article [30] was to show that it is possible to observe different Superstatistics at different time scales, i.e. transition of Superstatistics may occur. The idea was demonstrated by fitting probability distribution of β at two outlying time scales, namely, minute scale and daily scale. Tables 5.2 and 5.3 advocate the idea in a quantitative way. We see that transition indeed occurs, at least for company AA, from Log-normal distribution at small time scale to Gamma distribution at daily scale.

To get a better picture of the transition, it is worth calculating statistical distances for more than two time scales and see how the distances behave with respect to time scale. We calculated distances for scales ranging from 20 minutes to 1000 minutes $\sim$ 3 days (for higher time scales more data would be necessary). Results for short time scales are depicted in the figure 5.5 from which we see that only for very short time scales Log-normal regime prevails.

Unfortunately, using this crude method to find transition point is not very reliable because when individual distance measures approach each other their characteristics may need to be taken into account, and simple comparison of actual values cannot be considered as a definite criteria. Nevertheless, in the figure 5.6 which shows distance measures behavior on higher time scales we see that around 400 minutes, i.e. around one trading day, we can claim change of Superstatistics as quite significant, if Kolmogorov distance is disregarded as being the least relevant due to its instability in the presence of outliers.

The discussion above confirms observation made in the article [30] for company AA. However, according to [30] transition of Superstatistics applies also to the other companies. That it is not true can be seen in the figure 5.9 and 5.8 where the same distance measures are shown for company BAC. Log-normal regime appears to continue even on long time scales, and no transition of Superstatistics is observed. This naturally does not exclude existence of transition on much longer time scales but clearly reveals drawback of visual examination of fitted histogram. In truth, authors of [30] admit that it is by no means easy to

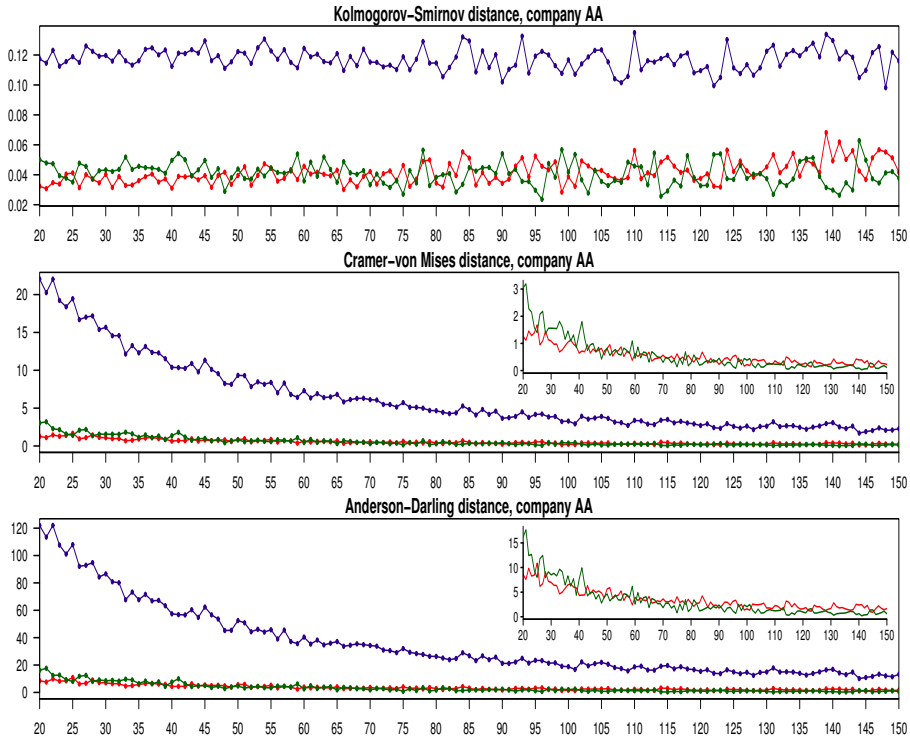


Figure 5.5: Three statistical distance measures for considered probability distributions are shown as a function of time scale in interval from 20 minutes to 2.5 hours (150 min), dataset Alcola Inc. Blue: Inv-Gamma distribution, Green: Gamma distribution, Red: Log-normal distribution.

distinguish between different Superstatistics, and this is also slightly indicated in figures 5.6 and 5.8 where we need to "zoom" in order to distinguish between different Superstatistics. It is also due to the fact that for higher time scales we have less data points for inverse temperature β, and therefore statistical distances has less statistical power to distinguish between two probability distribution. Note that for larger time scales even Inverse Gamma Superstatistics becomes plausible, but it is only because of lack of data.

We analyzed all seven companies and can conclude that for four of them, namely AA, INTC, KO and WMT, the transition between Superstatistics can be confirmed by using distance measures. On the other hand, companies BAC, GE and JNJ do not exhibit transition of Superstatistics at least at time scales less then one trading day.

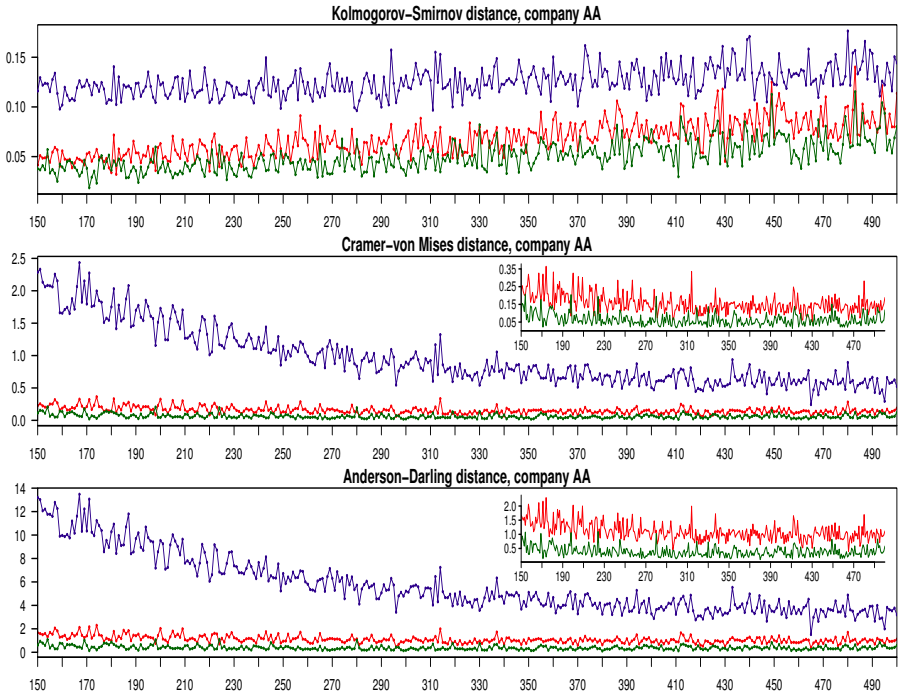


Figure 5.6: Three statistical distance measures for different probability distributions are shown as a function of time scale in interval from 2.5 hours (150 min) to 500 minutes, dataset Alcola Inc. Blue: Inv-Gamma distribution, Green: Gamma distribution, Red: Log-normal distribution.

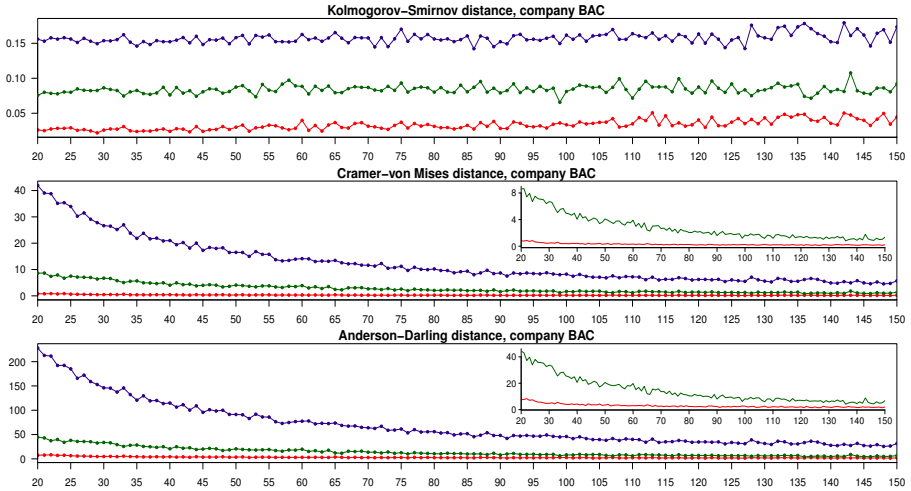


Figure 5.7: Three statistical distance measures for different probability distributions are shown as a function of time scale in interval from 20 minutes to 2.5 hours (150 min), dataset Bank of America Corporation. Blue: Inv-Gamma distribution, Green: Gamma distribution, Red: Log-normal distribution.

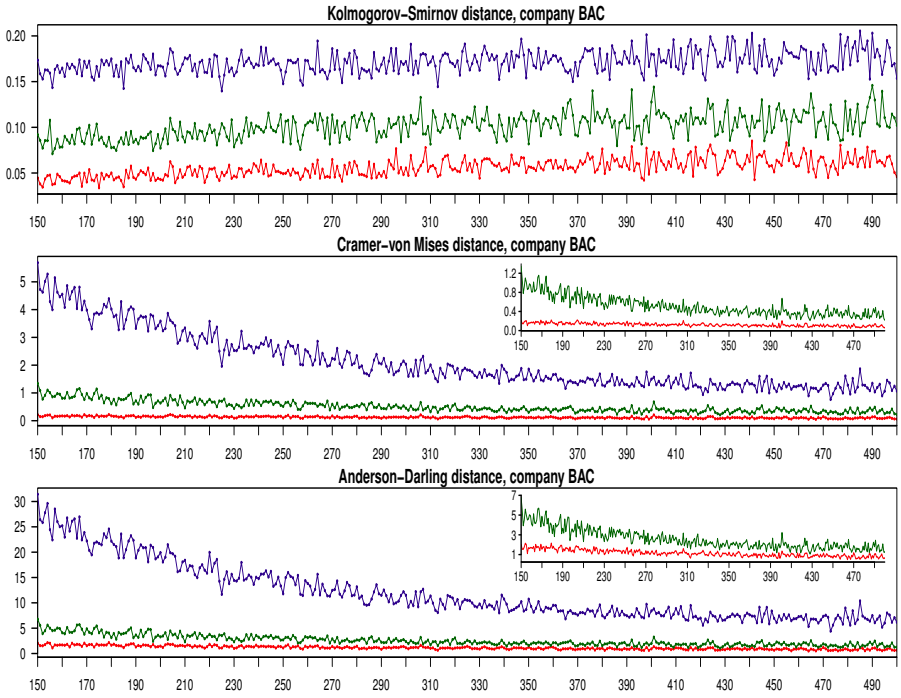


Figure 5.8: Three statistical distance measures for different probability distributions are shown as a function of time scale in interval from 2.5 hours (150 min) to 500 minutes, dataset Bank of America Corporation. Blue: Inv-Gamma distribution, Green: Gamma distribution, Red: Log-normal distribution.

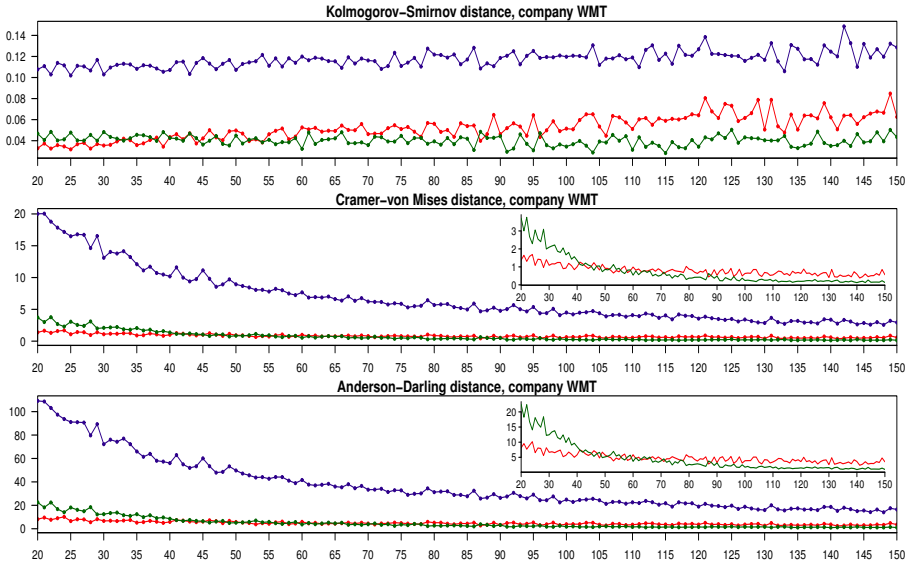


Figure 5.9: Very illustrative example of transition.

Appendix A

Estimators, errors and Bootstrap

A.1 Estimators

From measured data $(x_1, \ldots, x_N)$ we want to infer some feature A of the whole population, for example population mean μ or variance σ^2. The proper estimator $\hat{A}_N(x_1, \ldots, x_N)$ of some parameter A should have the following properties.

- *consistency* - $\lim_{N \mapsto \infty} \hat{A}_N = A \qquad$ *in probability*
- *unbiasedness* - $Bias(\hat{A}) = E[\hat{A}_N] - A = 0 \quad for\ all\ N$
- *efficiency* - Any other estimator of A fulfilling the conditions above must have higher variance.

For example, biased estimator of variance is

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2. \tag{A.1}$$

This is so called **plug-in estimator** since we plug in empirical distribution function into expression for variance instead of proper unknown one, this is a common practice in estimating parameters. However, this estimator is biased and the unbiased one is

$$s^2 = \frac{N}{N-1} \hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2, \tag{A.2}$$

where $\bar{x}$ is sample mean and so-called Bessel's correction was used. Nevertheless, since for large data sample N prefactor $\frac{N}{N-1} \approx 1$ the relation A.2 shows that A.1 is **asymptotically unbiased**.

Since every estimator is a random variable we would like to know variance of our estimator. The square root of the variance (standard deviation) is then called **standard error** and is used for error bars in plotting.

It is worth taking a look at the simples case, i.e. population mean. Its variance, provided data are independent, is $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{N}$. We see that it depends on unknown population variance σ^2, therefore we are forced estimate both population mean and its standard error. In order to correctly estimate standard error of $\bar{x}$ we need to find suitable estimator $\hat{\sigma}$ of population standard deviation σ. Such an estimator is, see [32],

$$\hat{\sigma} = K_N s = \sqrt{\frac{N-1}{2}} \frac{\Gamma\left(\frac{N-1}{2}\right)}{\Gamma\left(\frac{N}{2}\right)} s. \tag{A.3}$$

Nevertheless, asymptotic behavior shows that for $N > 10$ it is reasonable to use $K_N = 1$, as it is very common in practice, i.e. $\hat{\sigma} = s$ and thus $\hat{\sigma}_{\bar{x}} = s/\sqrt{N}$ which is just square root of A.2. The expression A.3 demonstrates the important feature of estimators:

> Unbiased estimator of a function of some parameter may not be, in general, obtained by simply plugging unbiased estimator of the parameter into the given function.

The standard error of variance estimator is

$$\sigma_{s^2} = \sigma^2 \sqrt{\frac{2}{N-1}}, \tag{A.4}$$

which is easily obtained from A.2 when we notice that $\frac{(N-1)s^2}{\sigma^2}$ has chi-squared distribution with $N-1$ degrees which has variance $2(N-1)$.

The same trick leads to standard error of estimator A.1

$$sd(\hat{\sigma^2}) = \frac{\sqrt{2(N-1)}\sigma^2}{N}. \tag{A.5}$$

Comparing A.4 and A.5 we see that biased version has lower standard error. In order to judge which estimator is better we introduce another quantity so-called **Mean square error (MSE)**

$$MSE(\theta) = E[(\hat{\theta} - \theta)^2].$$

MSE is convenient because it may be written in the following form [1]

$$MSE(\theta) = Var[\hat{\theta}] + Bias(\hat{\theta})^2$$

[1] Just add and subtract $E[\hat{\theta}]$

which properly incorporates both standard error and bias of the estimator. Since $\text{Bias}(\hat{\sigma^2}) = -\frac{\sigma^2}{N}$ we can calculate

$$MSE(\hat{\sigma^2}) = \frac{\sigma^4(2N-1)}{N^2}$$

which is smaller than $MSE(s^2)$. Thus in sense of Mean square error estimator A.1 seems to be better choice.

Aforementioned discussion demonstrates general problem in selecting suitable estimator and shows that biased estimators are also important, i.e. unbiasedness is not the only adequate criterion.

A.2 Bootstrap

These two examples are specific because there is analytically derived standard error of the estimator. In practice we need to estimate other, much more difficult, parameters, i.e. function of hidden probability distribution $\theta = t(F)$. For this purpose, we suggest some estimator $\hat{\theta}$ of θ and then we would like to know its standard error. In many cases we are not able to derive exact probability distribution of estimator $\hat{\theta}$ and in these situations bootstrap method may be helpful. Bootstrap has been using since 1979 when computers power became capable of processing huge amount of data in reasonable time, see [33].

The main idea behind bootstrap is very simple but demands immense computational effort that is why it emerged quite recently. Let sample values $\mathbf{x} = (x_1, \ldots, x_n)$ are given from experiment. Then for this values we calculate estimate $\hat{\theta}(\mathbf{x})$ of our desired parameter and in order to find out standard error of the estimator we resamle the data to get so-called *bootstrap sample* $\mathbf{x}^*$, i.e. we draw n values with replacement from original dataset $\mathbf{x}$. Thus some values may repeat in the new sample and other ones may be missing.

As a next step, we evaluate the estimator for this new sample to get new estimate $\hat{\theta}(\mathbf{x}^*)$. We repeat the same procedure many times until we get sufficient number of values $\{\hat{\theta}(\mathbf{x}_1^*), \ldots, \hat{\theta}^*(\mathbf{x}_m^*)\}$ for statistical inference. The bootstrap estimate of standard error is just standard deviation calculated from $\{\hat{\theta}(\mathbf{x}_1^*), \ldots, \hat{\theta}^*(\mathbf{x}_m^*)\}$. The length m of bootstrap sample is usually taken in range $25 - 200$, see [33].

A.2.1 Problem with bootstrap

Unfortunately, some problems emerge when we try to apply bootstrap method in time series analysis since bootstrap method assumes that data in original sample are i.i.d. The identically-distributed restriction may be satisfied for stationary time series but the independence is general problem in most time

series, in fact dependence of successive values is the reason for introducing time series.

The simplest solution is differencing time series and hope that new time series of differences is already independent, as it is the case, e.g. for random walk. The differencing of a time series is based on some a priori known structure or model of the system. Hence, provided we have faithful model of data we may bootstrap only the extracted random already independent noise or residuals and then reconstruct new resampled time series. However, in many cases the model is unknown and we are left with nonparametric bootstrap.

We analyze financial data, particularly stock indexes which are, more precisely its logarithm, according to obsolete theory motivated by Bachelier regarded as a random walk. Therefore, someone would expect there is no problem since we know the model, nevertheless, empirical analysis shows that this theory is not satisfactory and even the differences are dependent. The dependence may not be obvious since auto-correlation function suggests uncorrelated data but as we have already seen auto-correlation cannot detect nonlinear correlation and more precise analysis with mutual information points out non-negligible dependence. Thus it is clear that simple bootstrap sample loses the correlation structure and hence cannot faithfully represent original data.

For dependent data with unknown structure we have to use improved bootstrap method called **moving blocks bootstrap**. Instead of resampling bare data we resample all blocks of given length l. The procedure is as follows:

1. From original data we construct $n - l + 1$ overlapping blocks.

$$B_1 = (X_1, \ldots, X_l),\ B_2 = (X_2, \ldots, X_{l+1}), \ldots,\ B_{n-l+1} = (X_{n-l+1}, \ldots, X_n)$$

2. Then, provided l divides n, we generate $b = n/l$ random numbers uniformly distributed on $n - l + 1$ and accordingly select blocks from which we consequently compose new series.

When l does not divide n we "circle" original data, i.e. at the end we take also blocks $B_{n-l+2} = (X_{n-l+2}, \ldots, X_n, X_1)$ and so on. This procedure is recommended even in case when l divides n because otherwise data points at the beginning and at the end of the series would occur less frequently in bootstrapped time series.

Note that the longer the block the more dependency of original time series is preserved in bootstrapped one. On the other hand, the longer blocks means less variety which leads to underestimation of standard error calculated from bootstrap sample. Hence a trade off must be done in choosing right block length l.

Appendix B

Figures

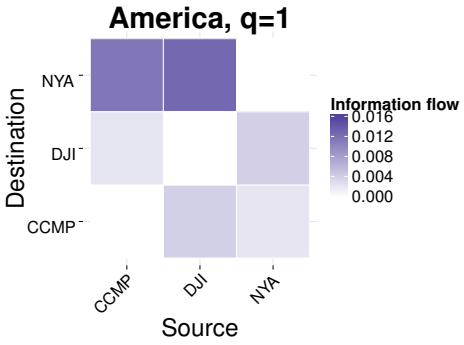


Figure B.1: America, heat map of Shanonian transfer entropy

	NYA	CCMP	DJI
NYA	0 0	0.0117 ± 0.0027	0.0128 ± 0.0022
CCMP	0.0022 ± 0.0010	0 0	0.0040 ± 0.0012
DJI	0.0041 ± 0.0011	0.0022 ± 0.0013	0 0

Table B.1: America, Shanonian transfer entropy

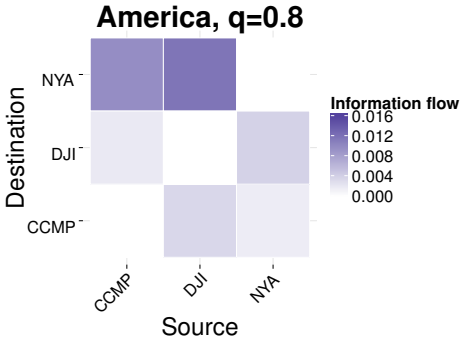


Figure B.2: America, heat map of Rényian transfer entropy q=0.8

	NYA	CCMP	DJI
NYA	0 0	0.0097 ± 0.0016	0.0118 ± 0.0018
CCMP	0.0016 ± 0.0008	0 0	0.0034 ± 0.0012
DJI	0.0038 ± 0.0009	0.0018 ± 0.0011	0 0

Table B.2: America, Rényian transfer entropy q=0.8

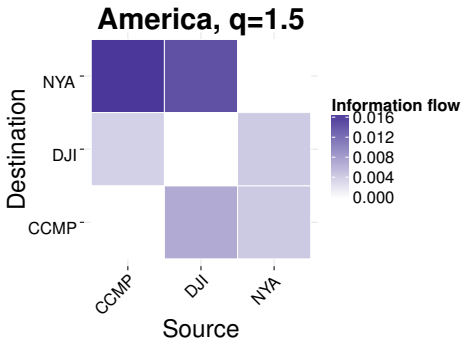


Figure B.3: America, heat map of Rényian transfer entropy q=1.5

	NYA	CCMP	DJI
NYA	0 0	0.0170 ± 0.0027	0.0148 ± 0.0018
CCMP	0.0046 ± 0.0014	0 0	0.0071 ± 0.0020
DJI	0.0045 ± 0.0016	0.0040 ± 0.0015	0 0

Table B.3: America, Rényian transfer entropy q=1.5

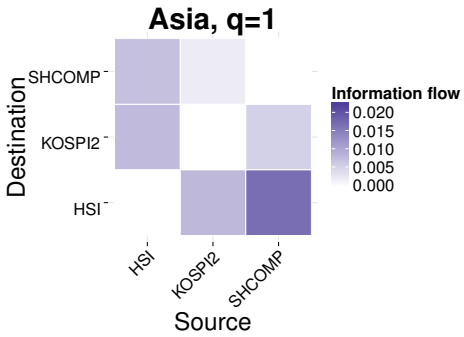


Figure B.4: Asia, heat map of Shanonian transfer entropy

	KOSPI2	HSI	SHCOMP
KOSPI2	0 0	0.0078 ± 0.0027	0.0053 ± 0.0015
HSI	0.0080 ± 0.0022	0 0	0.0166 ± 0.0049
SHCOMP	0.0021 ± 0.0015	0.0072 ± 0.0024	0 0

Table B.4: Asia, Shanonian transfer entropy

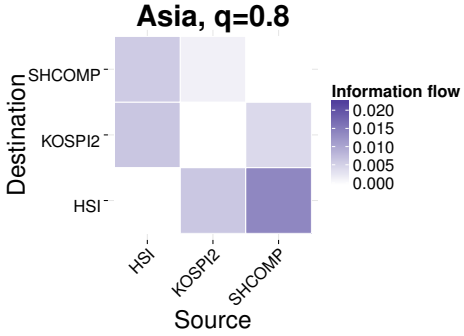


Figure B.5: Asia, heat map of Rényian transfer entropy q=0.8

	KOSPI2	HSI	SHCOMP
KOSPI2	0 0	0.0066 ± 0.0019	0.0045 ± 0.0021
HSI	0.0064 ± 0.0018	0 0	0.0138 ± 0.0032
SHCOMP	0.0018 ± 0.0012	0.0059 ± 0.0018	0 0

Table B.5: Asia, Rényian transfer entropy q=0.8

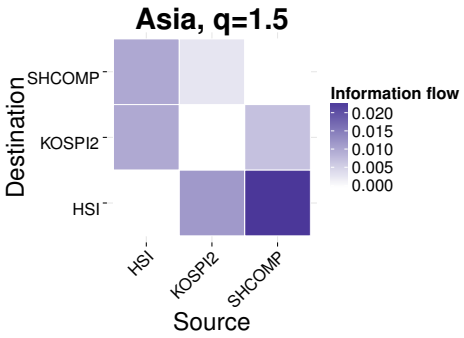


Figure B.6: Asia, heat map of Rényian transfer entropy q=1.5

	KOSPI2	HSI	SHCOMP
KOSPI2	0 0	0.0098 ± 0.0035	0.0071 ± 0.0031
HSI	0.0117 ± 0.0038	0 0	0.0230 ± 0.0047
SHCOMP	0.0030 ± 0.0022	0.0099 ± 0.0028	0 0

Table B.6: Asia, Rényian transfer entropy q=1.5

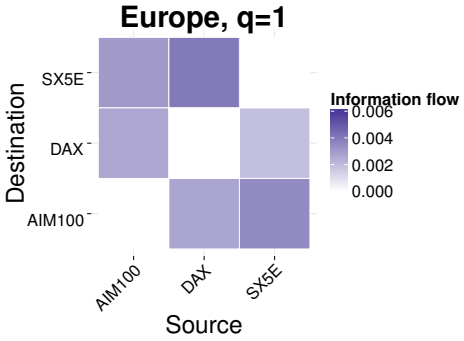


Figure B.7: Europe, heat map of Shanonian transfer entropy

	SX5E	AIM100	DAX
SX5E	0 0	0.0031 ± 0.0008	0.0041 ± 0.0012
AIM100	0.0036 ± 0.0012	0 0	0.0028 0.0010
DAX	0.0019 ± 0.0006	0.0027 ± 0.0009	0 0

Table B.7: Europe, Shanonian transfer entropy

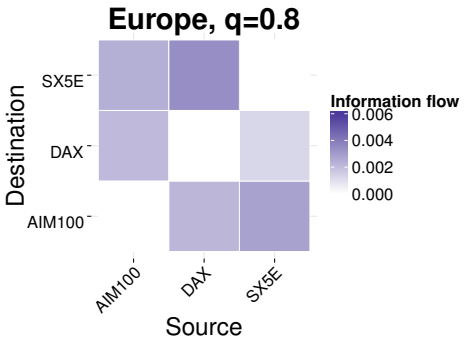


Figure B.8: Europe, heat map of Rényian transfer entropy q=0.8

	SX5E	AIM100	DAX
SX5E	0 0	0.0024 ± 0.0008	0.0035 ± 0.0011
AIM100	0.0028 ± 0.0009	0 0	0.0022 ± 0.0006
DAX	0.0013 ± 0.0010	0.0021 ± 0.0009	0 0

Table B.8: Europe, Rényian transfer entropy q=0.8

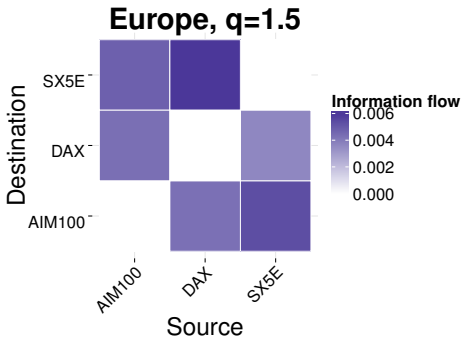


Figure B.9: Europe, heat map of Rényian transfer entropy q=1.5

	SX5E	AIM100	DAX
SX5E	0 0	0.0049 ± 0.0012	0.0061 ± 0.0017
AIM100	0.0054 ± 0.0016	0 0	0.0044 ± 0.0013
DAX	0.0037 ± 0.0011	0.0044 ± 0.0011	0 0

Table B.9: Europe, Rényian transfer entropy q=1.5

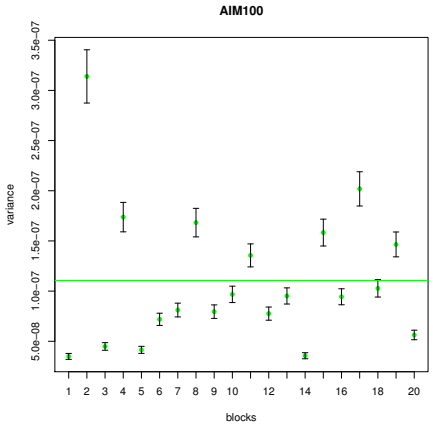


Figure B.10: The whole time series of London stock index was divided into 20 non-overlapping blocks and in each block variance of log-returns was calculated. For weak stationary time series all values should stay within a standard deviation around overall variance (the green line). Depicted behavior suggests non-stationarity.

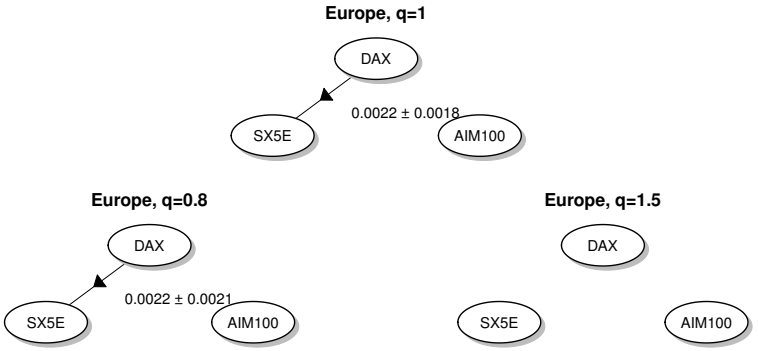


Figure B.11: Detected causality relations in European stock indices.

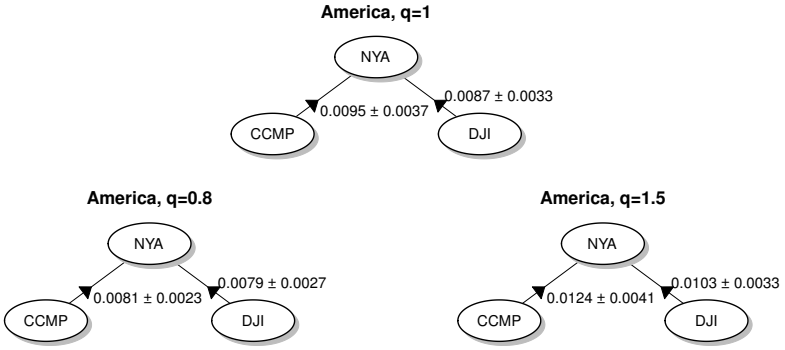


Figure B.12: Detected causality relations in American stock indices.

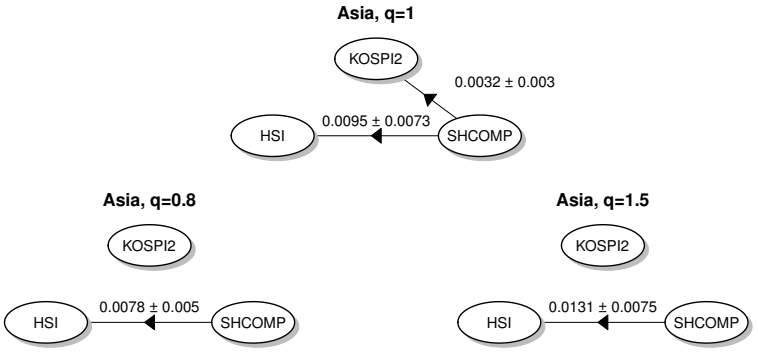


Figure B.13: Detected causality relations in Asian stock indices.

Bibliography

[1] C.E. Shannon, The Mathematical Theory of Communication, University of Illinois Press, New York, 1949.

[2] R. V. Hartley, Transmission of information. Bell System Technical Journal, 7 (1928) 535-563

[3] Thomas M. Cover, Joy A. Thomas. Elements of Information Theory, second edition. Wiley 2006.

[4] J. Aczél, Z. Daróczy, Measures of Information and their Characterizations, Academic Press, New Yourk, 1975.

[5] A. Rényi, Selected Papers of Alfred Rényi, vol. 2, Akademia Kiado, Budapest, 1976.

[6] P. Jizba, T. Arimitsu, Ann. Phys. (NY) 312 (2004) 17.

[7] P. Jizba, H. Kleinert, Mohammad Shefaat, Renyi's information transfer between financial time series, Physica A 2012, 391, 2971–2989.

[8] L.L. Campbell, Inf. Control 8 (1965) 423.

[9] A. Rényi, Probability Theory, North-Holland, Amsterdam, 1970.

[10] A. I. Khinchin, Mathematical Foundations of Information Theory, Dover Publications, Inc., New York, 1957.

[11] M. B. Priestley, Spectral analysis and time series, Volume 1, Academic press, 1981.

[12] C.J. Cellucci, A.M. Albano and P.E. Rapp, Phys. Rev. E 71, 66208 (2005)

[13] F.M. Aparicio, A. Escribano, Information-theoretic analysis of serial dependence and cointegration, Stud. Nonlinear Dynam. Econometr. 3 (1998) 119–140.

[14] K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, J. Bhattacharya, Causality detection based on information-theoretic approaches in time series analysis, Phys. Reports, 441(1):1, 2007

[15] A. Fraser, H. Swinney, Independent coordinates for strange attractors from mutual information, Phys. Rev. A 33 (1986) 1134–1140.

[16] H. Kantz, T. Schreiber, Nonlinear Time Series Analysis, Cambridge university press, 2003.

[17] T. Schreiber, Phys. Rev. Lett. 85 (2000) 461.

[18] R. Marschinski, H. Kantz, Eur. Phys. J. B 30 (2002) 275.

[19] O. Kwon, J.-S. Yang, Europhys. Lett. 82 (2008) 68003

[20] R. Marschinski, L. Matassini, Financial Markets as a Complex System: A Short Time Scale Perspective, Research Notes in Economics & Statistics, Deutsche Bank Research.

[21] C. Beck, Dynamical foundatiouns of nonextensive statistical mechanics, Phys. Rev. Lett. 87, 180601 (2001).

[22] C. Beck, E.G.D. Cohen, Superstatistics, Physica A 322, 267 (2003).

[23] J. Korbel, Applications of Multifractals in Financial Markets (Diploma thesis), FNSPE, 2012.

[24] F. C. Klebaner, Introduction to Stochastic Calculus with Applications, second edition, Imperial College Press, 2005.

[25] C. Tsallis, Possible Generalization of Boltzmann-Gibbs Statistics, Journal of Statistical Physics, Vol. 52, 479 (1988).

[26] E. G. D. Cohen, Statistics and dynamics, Physica A 305, 19 (2002).

[27] C. Tsallis, R. S. Mendes, A. R. Plastino, The role of constraints within generalized nonextensive statistics, Physica A 261, 534 (1998).

[28] C. Beck, Recent developments in superstatistics, Brazilian Journal of Physics, vol. 39, no. 2A, 2009.

[29] C. Beck, Generalized statistical mechanics for superstatistical systems, Phil. Trans. R. Soc. A (2011) 369, 453–465.

[30] C. Beck, D. Xu, Transition from lognormal to χ^2-superstatistics for financial time series, Physica A 453, 173 (2016).

[31] Joanes, D. N., C. A. Gill, Comparing Measures of Sample Skewness and Kurtosis, Journal of the Royal Statistical Society, Series D (The Statistician) 47, no. 1 (1998): 183-89.

[32] E.L. Lehmann, G. Casella, Theory of Point Estimation, second edition, Springer, 1998.

[33] B. Efron, R. J. Tibshirani, An introduction to the Bootstrap, Chapman & Hall, Ind., 1993.